Differential Equations

and

Linear Algebra

A Course for
Science and Engineering

Volume I: Chapters 1-7
Volume II: Chapters 8-12

by
Grant B. Gustafson

ISBN **Volume I**: 9798705491124
ISBN **Volume II**: 9798711123651

Contents

Foreword

Organization

Each chapter of the text is organized into sections that represent one or two classroom lectures of 50 minutes each. Outside work for these divisions requires one to six hours, depending upon the depth of study.

Each section within a chapter consists of three distinct **parts**. The divisions represent the **lecture**, **examples** and **technical details**. Generally, proofs of theorems or long justifications of formulas are delayed until after the examples. The lectures contain only the briefest examples, figures and illustrations.

A key to a successful course is a weekly session dedicated to review, drill, answers, solutions, exposition and exam preparation. While group meetings are important, individual effort is required to flesh out the details and to learn the subject in depth. The textbook design supports targeted self-study through its examples, exercises and odd exercise solutions.

There is a defense for this style of presentation, matched equally by a long list of criticisms. The *defense* is that this style represents how material is presented in classroom lectures, and how the topics are studied in the private life of a student. It is unexpected to read everything in a textbook and the style addresses the issue of what to skip and what to read in detail. The *criticisms* include a departure from standard textbooks, which intermix theory and examples and proofs. Page flipping criticism applies to the printed textbook. The PDF textbook has embedded links.

Prerequisites

Beginning sections of chapters require college algebra, basic high school geometry, coordinate geometry, elementary trigonometry, differential calculus and integral calculus. Several variable calculus and linear algebra are assumed for certain advanced topics. Instructors are the best judges of what to include and what to skip, concerning advanced topics in this textbook.

Survey course

A complete **survey course** in differential equations for engineering and science can be constructed from the lectures and examples, by skipping the technical details supplied in the text. A deeper introduction to the subject is obtained by reading the details. Such survey courses will necessarily contact more chapters and trade the depth of a conventional course for a faster pace, easier topics, and more variety.

Conventional Course

Differential equations courses at the undergraduate level will present some or all of the technical details in class, as part of the lecture. Deeper study with technical details is warranted for specialties like physics and electrical engineering. Hybrid courses that combine the conventional course and the engineering course can be realized.

To the Student

Expertise in background topics is expected only after review and continued use in the course, especially by writing solutions to exercises.

Instructors are advised that an exercise list and subsequent evaluation of the work is essential for successful classroom use of the text.

The text has nearly 3,600 exercises, supported by textbook examples and odd-numbered solutions. Solutions are located in the PDF textbook + solution manual.

To learn the subject, not only is it required to *solve exercises*, but to *write exercises*, which is not different from writing in a foreign language.

Writing exercises requires two or more drafts and a final *presentation*. Engineering paper and lineless duplicator paper encourage final reports with adequate white space between equations. Pencil and eraser save time. Pens and word processors waste time.

Contributions to legibility, organization and presentation of hand-written exercises were made at The University of Utah, by numerous creative engineering, computer science, physics, biology and mathematics students, over the years 1990-2019. Their ideas produced the *suggestions* below in Table 1, which were applied to the text examples, illustrations and exercise solutions.

Table 1. Suggestions for Hand-Written Exercises 1990-2019

1. A report is hand-written by pencil on lineless paper or engineering paper. It starts with a problem statement followed perhaps by a final answer summary. Supporting material appears at the end, like a tax return.
2. Mathematical notation is on the left, text on the right, often a 60% to 40% ratio. One equal sign per line, equations justified left or aligned on equal signs. Vertical white space separates equation displays.
3. Text is left-justified in a column on the right. It contains explanations, references by keyword or page number, statements and definitions, references to delayed details like long calculations, graphics, answer checks.
4. Every report has an answer check. It is usual to see *back of book* as the only detail. Proofs have no answer check.
5. No suggestion is a rule: invent and develop your own style.

Work, School and Family

The textbook and the solution manual were designed for students who study in isolation, their university schedule driven by their jobs and family. In spite of forced isolation from the classroom, working students with families seek help from others through telephone calls, online search, internet messaging, email, office visits to the university, study groups, supplemental and online instruction.

Chapter 1

Fundamentals

Contents

Introduced here are notation, definitions and background results suitable for use in differential equations.

Prerequisites include college algebra, coordinate geometry, differential calculus and integral calculus. The examples and exercises include a review of some calculus topics, especially derivatives, integrals, numerical integration, hand and computer graphing. A significant part of the review is algebraic manipulation of logarithms, exponentials, sines and cosines.

The chapter starts with differential equations applications that require only a background from pre-calculus: exponential and logarithmic functions. No differential equations background is assumed or used. Differential equations are defined and insight is given into the notion of *answer* for differential equations in science and engineering applications.

Basic topics: direction fields, phase line diagrams and bifurcation diagrams, which require only a calculus background. Applications of these topics appear later in the text, after more solution methods have been introduced.

Advanced topics: existence-uniqueness theory and implicit functions. Included are practical computer algebra system methods to assist with finding solutions, verifying equations, modeling, and related topics.

1.1 Exponential Modeling

Three model differential equations are studied through a variety of specific applications. All applications use the calculus exponential function $y(t) = y_0 e^{kt}$.

Three Examples

These applications are studied:

Growth–Decay Models
Newton Cooling
Verhulst Logistic Model

It is possible to solve a variety of differential equations without reading this book or any other differential equations text. Given in the table below are three exponential models and their known solutions, all of which will be derived from principles of elementary differential calculus.

Growth-Decay $\qquad \dfrac{dA}{dt} = kA(t),\ A(0) = A_0$

$$\boxed{A(t) = A_0 e^{kt}}$$

Newton Cooling $\qquad \dfrac{du}{dt} = -h(u(t) - u_1),\ u(0) = u_0$

$$\boxed{u(t) = u_1 + (u_0 - u_1)e^{-ht}}$$

Verhulst Logistic $\qquad \dfrac{dP}{dt} = (a - bP(t))P(t),\ P(0) = P_0$

$$\boxed{P(t) = \frac{aP_0}{bP_0 + (a - bP_0)e^{-at}}}$$

These models and their solution formulas form a foundation of intuition for all of differential equation theory. Considerable use will be made of the models and their solution formulas.

The *physical meanings* of the constants k, A_0, h, u_1, u_0, a, b, P_o and the variable names $A(t)$, $u(t)$, $P(t)$ are given below, as each example is discussed.

Background

Mathematical background used in exponential modeling is limited to algebra and basic calculus. The following facts are assembled for use in applications. Prime notation is used: $' = \frac{d}{dt}$ and sometimes $' = \frac{d}{dx}$.

$\ln e^x = x,\ e^{\ln y} = y$	In words, the exponential and the logarithm are inverses. The domains are $-\infty < x < \infty$, $0 < y < \infty$.
$e^0 = 1,\ \ln(1) = 0$	Special values, usually memorized.
$e^{a+b} = e^a e^b$	In words, the exponential of a sum of terms is the product of the exponentials of the terms.
$(e^a)^b = e^{ab}$	Negatives are allowed, e.g., $(e^a)^{-1} = e^{-a}$.
$\left(e^{u(t)}\right)' = u'(t)e^{u(t)}$	The *chain rule* of calculus implies this formula from the identity $\frac{d}{dx}e^x = e^x$ where $x = u(t)$ and $\frac{dx}{dt} = u'(t)$.
$\ln AB = \ln A + \ln B$	In words, the logarithm of a product of factors is the sum of the logarithms of the factors.
$B\ln(A) = \ln\left(A^B\right)$	Negatives are allowed, e.g., $-\ln A = \ln(1/A)$ when $B = -1$.
$(\ln\lvert u(t)\rvert)' = \dfrac{u'(t)}{u(t)}$	The identity $\frac{d}{dx}\ln(x) = 1/x$ implies this general version by the *chain rule* applied with $x = u(t)$, $\frac{dx}{dt} = u'(t)$.

Applied topics using exponentials inevitably lead to equations involving logarithms. Conversion of exponential equations to logarithmic equations, and the reverse, happens to be an important subtopic of differential equations. The examples and exercises contain typical calculations.

Growth-Decay Model

Growth and decay models in science are based upon the exponential equation

$$(1) \qquad y = y_0 e^{kx}, \quad y_0 \text{ and } k \text{ constant.}$$

The exponential e^{kx} increases if $k > 0$ and decreases if $k < 0$. A model based upon the exponential is called a **growth model** if $k > 0$ and a **decay model** if $k < 0$. Examples of growth models include population growth and compound interest. Examples of decay models include radioactive decay, radiocarbon dating and drug elimination. Typical growth and decay curves appear in Figure 1.

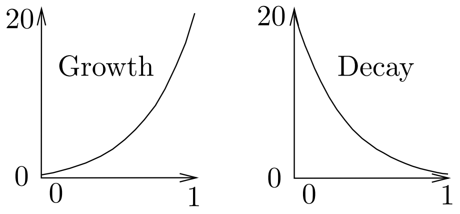


Figure 1. Growth and decay curves.

Definition 1.1 (Growth-Decay Equation)
The differential equation

$$\frac{dy}{dx} = ky \tag{2}$$

is called a **growth-decay** differential equation.

A solution of (2) is $y(x) = y_0 e^{kx}$; see the verification on page 10. It is possible to show directly that the differential equation has no other solutions, hence the terminology *the solution* $y = y_0 e^{kx}$ is appropriate; see the verification on page 11. The solution $y = y_0 e^{kx}$ in (1) satisfies the **growth-decay initial value problem**

$$\frac{dy}{dx} = ky, \quad y(0) = y_0. \tag{3}$$

The **initial condition** $y(0) = y_0$ means $y = y_0$ at $x = 0$. It can be written as $y(x)|_{x=0} = y_0$.

How to Solve a Growth-Decay Equation

Numerous applications to first order differential equations are based upon equations that have the general form $\frac{dy}{dx} = ky$. Whenever this form is encountered, immediately the solution is known: $y(x) = y_0 e^{kx}$. The symbol y_0 is a constant known as the **initial state**, because $e^{kx} = 1$ at $x = 0$ implies $y(x)$ equals y_0 at $x = 0$.

Newton Cooling Model

If a fluid is held at constant temperature, then the cooling of a body immersed in the fluid is subject to **Newton's cooling law**:

> The rate of temperature change of the body is proportional to the difference between the body's temperature and the fluid's constant temperature.

Translation to mathematical notation gives the differential equation

$$\frac{du}{dt} = -h(u(t) - u_1) \tag{4}$$

where $u(t)$ is the temperature of the body, u_1 is the constant ambient temperature of the fluid and $h > 0$ is a constant of proportionality.

A typical instance is the cooling of a cup of hot chocolate in a room. Here, u_1 is the wall thermometer reading and $u(t)$ is the reading of a dial thermometer immersed in the chocolate drink.

Theorem 1.1 (Solution of Newton's Cooling Equation)
The change of variable $y(t) = u(t) - u_1$ translates the cooling equation $du/dt = -h(u-u_1)$ into the growth-decay equation $\frac{d}{dt}y(t) = -hy(t)$. Therefore, the cooling solution is given in terms of $u_0 = u(0)$ by the equation

$$u(t) = u_1 + (u_0 - u_1)e^{-ht}. \tag{5}$$

The result is proved on page 11. It shows that a cooling model is just a translated growth-decay model. The solution formula (5) can be expressed in words as follows:

> The dial thermometer reading of the hot chocolate equals the wall thermometer reading plus an exponential decay term.

Cooling problems have *curious extra conditions*, usually involving physical measurements, for example the three equations

$$u(0) = 100, \quad u(1) = 90 \quad \text{and} \quad u(\infty) = 22.$$

The extra conditions implicitly determine the actual values of the three undetermined parameters h, u_1, u_0. The logic is as follows. Equation (5) is a relation among 5 variables. Substitution of values for t and u eliminates 2 of the 5 variables and gives an equation for u_1, u_0, h. The system of three equations in three unknowns can be solved for the actual values of u_1, u_0, h.

Stirring Effects

Exactly how to maintain a constant ambient temperature is not addressed by the model. One method is to stir the liquid, as in Figure 2, but the mechanical energy of the stirrer will inevitably appear as heat in the liquid. In the simplest case, stirring effects add a fixed constant temperature S_0 to the model. For slow stirring, $S_0 = 0$ is assumed, which is the above model.

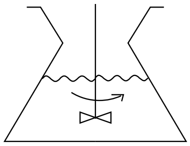


Figure 2. Flask Cooling with Stirring.

Population Modeling

The **human population** of the world reached seven billion in 2011. The estimate for year 2021 is more than 7.7 billion.[1]

World Population Estimate
2020
7,794,798,739
Source: U.S. Census Bureau

The term **population** refers to humans. In literature, it may also refer to bacteria, insects, rodents, rabbits, wolves, trees, yeast and similar living things that have birth rates and death rates.

Malthusian Population Model

A constant birth rate or a constant death rate is unusual in a population, but these ideal cases have been studied. The biological reproduction law is called **Malthus' law**:

> The population flux is proportional to the population itself.

This biological law can be written in calculus terms as

$$\frac{dP}{dt} = kP(t)$$

where $P(t)$ is the population count at time t. The reasoning is that **population flux** is the expected change in population size for a unit change in t, or in the limit, dP/dt. A careful derivation of such calculus laws from English language appears in Appendix A.1.

The theory of growth-decay differential equations implies that population studies based upon Malthus's law employ the exponential algebraic model

$$P(t) = P_0 e^{k(t-t_0)}.$$

The number k is the difference of the birth and death rates, or **combined birth-death rate**, t_0 is the **initial time** and P_0 is the **initial population size** at time $t = t_0$.

[1] Reference: `https://www.worldometers.info/population/`

Verhulst Logistic Model

The population model $P' = kP$ was studied around 1840 by the Belgian demographer and mathematician Pierre-Francois Verhulst (1804–1849) in the special case when k depends on the population size $P(t)$. Under Verhulst's assumptions, $k = a - bP$ for positive constants a and b, so that $k > 0$ (growth) for populations P smaller than a/b and $k < 0$ (decay) when the population P exceeds a/b. The result is called the **logistic equation**:

$$P' = (a - bP)P. \tag{6}$$

Verhulst established the limit formula

$$\lim_{t\to\infty} P(t) = a/b, \tag{7}$$

which has the interpretation that initial populations $P(0)$, regardless of size, will after a long time stabilize to size approximately a/b. The constant a/b is called the **carrying capacity** of the population.

Limit formula (7) for $a > 0$ follows from solution formula (8) below.

Theorem 1.2 (Verhulst Logistic Solution)
The change of variable $y(t) = P(t)/(a - bP(t))$ transforms the logistic equation $P'(t) = (a - bP(t))P(t)$ into the growth-decay equation $y'(t) = ay(t)$. Then the logistic equation solution is given by

$$P(t) = \frac{aP(0)}{bP(0) + (a - bP(0))e^{-at}}. \tag{8}$$

The derivation appears on page 11. The impact of the result is that a logistic model transforms to a growth-decay model via a fractional change of variable. The **Verhulst logistic model** reduces to the **Malthus model** when $b = 0$. Then solution formula (8) reduces to the solution $y = y_0 e^{at}$ of growth-decay equation $y' = ay$, where $y = P, y_0 = P(0)$. Solution formula (8) remains valid regardless of the signs of a and b, provided the quotient is defined. Case $a = b = 0$ means $P'(t) = 0$ and constant population $P(t) = P_0$.

Examples

Example 1.1 (Growth-Decay)
Solve the initial value problem

$$\frac{dy}{dx} = 2y, \quad y(0) = 4.$$

Solution: This is a growth-decay equation $y' = ky$, $y(0) = y_0$ with $k = 2$, $y_0 = 4$. One way to decide on the constant k is to compute y'/y from the given differential equation. Then $y'/y = 2$ implies $k = 2$. Therefore, the solution is $y = y_0 e^{kx} = 4e^{2x}$. No method is required to *solve* the equation $y' = 2y$, because of the theory on page 3.

Example 1.2 (Newton Cooling)

Solve the initial value problem

$$\frac{du}{dt} = -3(u(t) - 72), \quad u(0) = 190.$$

Solution: This is a Newton cooling equation $u' = -h(u - u_1)$, $u(0) = u_0$ with $h = 3$, $u_1 = 72$, $u_0 = 190$. Therefore, the solution is $u(t) = u_1 + (u_0 - u_1)e^{-ht} = 72 + 118e^{-3t}$. No method is required to *solve* the equation $u' = -3(u - 72)$, because of the theorem on page 5.

To eliminate memorization, use the substitution $y = u - u_1$ to transform the problem to the growth-decay model $y' = -hy$ with solution $y = y_0 e^{-ht}$. Then back-substitute $y = u - u_1$ to solve for $u(t)$.

In this particular case, let $y = u - 72$ to get $y' = -3y$, then $y = y_0 e^{-3t}$ and finally $u - 72 = y_0 e^{-3t}$. Value y_0 equals $y(0)$. It is determined by the condition $y(t) + 72 = u(t) = 190$ at $t = 0$ (supplied as $u(0) = 190$) to give $y_0 = 118$ and then the final answer is $u(t) = 72 + 118e^{-3t}$.

Example 1.3 (Verhulst Logistic)

Solve the initial value problem

$$\frac{dP}{dt} = (1 - 2P)P, \quad P(0) = 500.$$

Solution: This is a Verhulst logistic equation $P' = (a - bP)P$, $P(0) = P_0$ with $a = 1$, $b = 2$, $P_0 = 500$. Therefore, the solution is

$$P(t) = \frac{500}{1000 - 999e^{-t}}.$$

No method is required to *solve* the equation $P' = (1 - 2P)P$, because of the formula supplied by Theorem 1.2.

Because of Verhulst solution formula complexity, there is no practical shortcut to obtain the solution. The easiest route is to use the solution formula in Theorem 1.2.

Example 1.4 (Standing Room Only)

Justify the estimate 2600 for the year in which each human has only one square foot of land to stand upon. Assume the Malthus model $P(t) = 3.34e^{0.02(t-1965)}$, with t in years and P in billions.

Solution: The mean radius of the earth is 3965 miles or $20,935,200$ feet. The surface area formula $4\pi r^2$ gives $5,507,622$ billion square feet. About 20% of this is land, or $1,101,524$ billion square feet.

The estimate 2600 is obtained by solving for t years in the equation

$$3.34e^{0.02(t-1965)} = 1101524.$$

The college algebra details:

$$e^{0.02(t-1965)} = \frac{1101524}{3.34}$$

Isolate the exponential on the left. Solving for t.

$\ln e^{0.02(t-1965)} = \ln 329797.6$	Simplify the right side and take the logarithm of both sides.
$0.02(t-1965) = 12.706234$	On the right, compute the logarithm. Use $\ln e^u = u$ on the left.
$t = 1965 + \dfrac{12.706234}{0.02}$	Solve for t.
$= 2600.3.$	About the year 2600.

Example 1.5 (Rodent Growth)

A population of two rodents in January reproduces to population sizes 20 and 110 in June and October, respectively. Determine a Malthusian law for the population and test it against the data.

Solution: However artificial this example might seem, it is almost a real experiment; see Braun [Braun1986], Chapter 1, and the reference to rodent *Microtus Arvallis Pall.*

The law proposed is $P = 2e^{2t/5}$, which is 40% growth, $k = 2/5$. For a 40% rate, $P(6) \approx 2e^{12/5} = 22.046353$ and $P(10) \approx 2e^{2(10)/5} = 109.1963$. The agreement with the data is reasonable. It remains to explain how this "40% law" was invented.

The Malthusian model $P(t) = P_0 e^{kt}$, with t in months, fits the three data items $P(0) = 2$, $P(6) = 20$ and $P(10) = 110$ provided $P_0 = 2$, $2e^{6k} = 20$ and $2e^{10k} = 110$. The exponential equations are solved for $k = \ln(10)/6$ and $k = \ln(55)/10$, resulting in the two growth constants $k = 0.38376418$ and $k = 0.40073332$. The average growth rate is 39.2%, or about 40%.

Example 1.6 (Flask Cooling)

A flask of water is heated to 95C and then allowed to cool in ambient room temperature 21C. The water cools to 80C in three minutes. Verify the estimate of 48 minutes to reach 23C.

Solution: Basic modeling by Newton's law of cooling gives the temperature as $u(t) = u_1 + (u_0 - u_1)e^{-kt}$ where u_1, u_0 and k are parameters. Three conditions are given in the English statement of the problem.

$u(\infty) = 21$	The ambient air temperature is 21C.
$u(0) = 95$	The flask is heated at $t = 0$ to 95C.
$u(3) = 80$	The flask cools to 80C in three minutes.

In the details below, it will be shown that the parameter values are $u_1 = 21$, $u_0 - u_1 = 74$, $k = 0.075509216$. Then $u(t) = 21 + 74e^{-0.075509216t}$, t in minutes.

To find u_1:

$21 = u(\infty)$	Given ambient temperature condition.
$= \lim_{t\to\infty} u(t)$	Definition of $u(\infty)$.
$= \lim_{t\to\infty} u_1 + (u_0 - u_1)e^{-kt}$	Definition of $u(t)$.
$= u_1$	The exponential has limit zero.

To calculate $u_0 - u_1 = 74$ from $u(0) = 95$:

$95 = u(0)$	Given initial temperature condition.
$\quad = u_1 + (u_0 - u_1)e^{-k(0)}$	Definition of $u(t)$ at $t = 0$.
$\quad = 21 + u_0 - u_1$	Use $e^0 = 1$.

Therefore, $u_0 - u_1 = 95 - 21 = 74$.

Computation of k starts with the equation $u(3) = 80$, which reduces to $21+74e^{-3k} = 80$. This exponential equation is solved for k as follows:

$e^{-3k} = \dfrac{80-21}{74}$	Isolate the exponential factor on the left side of the equation.
$\ln e^{-3k} = \ln \dfrac{80-21}{74}$	Take the logarithm of both sides.
$-3k = \ln(59/74)$	Simplify the fraction. Apply $\ln e^u = u$ on the left.
$k = \dfrac{1}{3}\ln(74/59)$	Divide by -3, then on the right use $-\ln x = \ln(1/x)$.

The estimate $u(48) \approx 23$ will be verified. The time t at which $u(t) = 23$ is found by solving the equation $21 + 74e^{-kt} = 23$ for t. A checkpoint is $-kt = \ln(2/74)$, from which t is isolated on the left. After substitution of $k = 0.075509216$, the value is $t = 47.82089$.

Example 1.7 (Baking a Roast)

A beef roast at room temperature 70F is put into a 350F oven. A meat thermometer reads 100F after four minutes. Verify that the roast is done (340F) in 120 minutes.

Solution: The roast is done when the thermometer reads 340F or higher. If $u(t)$ is the meat thermometer reading after t minutes, then it must be verified that $u(120) \geq 340$.

Even though the roast is heating instead of cooling, the beef roast temperature $u(t)$ after t minutes is given by the Newton cooling equation $u(t) = u_1 + (u_0 - u_1)e^{-kt}$, where u_1, u_0 and k are parameters. Three conditions appear in the statement of the problem:

$u(\infty) = 350$	The ambient oven temperature is 350F.
$u(0) = 70$	The beef is 70F at $t = 0$.
$u(4) = 100$	The roast heats to 100F in four minutes.

As in the *flask cooling example*, page 9, the first two relations above lead to $u_1 = 350$ and $u_0 - u_1 = -280$. The last relation determines k from the equation $350 - 280e^{-4k} = 100$. Solving by the methods of the flask cooling example gives $k = \frac{1}{4}\ln(280/250) \approx 0.028332171$. Then $u(120) = 350 - 280e^{-120k} \approx 340.65418$.

Details and Proofs

Growth-Decay Equation Existence Proof. It will be verified that $y = y_0e^{kx}$ is a solution of $y' = ky$. It suffices to expand the left side (LHS) and right side (RHS) of the differential equation and compare them for equality.

$\text{LHS} = \dfrac{dy}{dx}$	The left side of $\dfrac{dy}{dx} = ky$ is dy/dx.

$$
\begin{aligned}
&= \frac{d}{dx}\left(y_0e^{kx}\right) && \text{Substitute } y = y_0e^{kx}.\\
&= y_0ke^{kx} && \text{Apply the rule } (e^u)' = u'e^u.\\
\text{RHS} &= ky && \text{The right side of } \frac{dy}{dx} = ky \text{ is } ky.\\
&= k(y_0e^{kx}) && \text{Substitute } y = y_0e^{kx}.
\end{aligned}
$$

Therefore, LHS = RHS. ■

Growth-Decay Equation Uniqueness Proof. It will be shown that $y = y_0e^{kx}$ is the only solution of $y' = ky$, $y(0) = y_0$. The idea is to reduce the question to the application of a result from calculus. This is done by a clever change of variables, which has been traced back to Kümmer (1834).[2]

Assume that y is a given solution of $y' = ky$, $y(0) = y_0$. It has to be shown that $y = y_0e^{kx}$.

Define $v = y(x)e^{-kx}$. This defines a change of variable from y into v. Then

$$
\begin{aligned}
v' &= (e^{-kx}y)' && \text{Compute } v' \text{ from } v = e^{-kx}y.\\
&= -ke^{-kx}y + e^{-kx}y' && \text{Apply the product rule } (uy)' = u'y + uy'.\\
&= -ke^{-kx}y + e^{-kx}(ky) && \text{Use the differential equation } y' = ky.\\
&= 0. && \text{The terms cancel.}
\end{aligned}
$$

In summary, $v' = 0$ for all x. The calculus result to be applied is:

> The only function $v(x)$ that satisfies $v'(x) = 0$ on an interval is $v(x) =$ constant.

The conclusion is $v(x) = v_0$ for some constant v_0. Then $v = e^{-kx}y$ gives $y = v_0e^{kx}$. Setting $x = 0$ implies $v_0 = y_0$ and finally $y = y_0e^{kx}$. ■

Newton Cooling Solution Verification (Theorem 1.1). The substitution $A(t) = u(t) - u_1$ will be applied to find an equivalent growth-decay equation:

$$
\begin{aligned}
\frac{dA}{dt} &= \frac{d}{dt}\left(u(t) - u_1\right) && \text{Definition of } A = u - u_1.\\
&= u'(t) - 0 && \text{Derivative rules applied.}\\
&= -h(u(t) - u_1) && \text{Cooling differential equation applied.}\\
&= -hA(t) && \text{Definition of } A.
\end{aligned}
$$

The conclusion is that $A'(t) = -hA(t)$. Then $A(t) = A_0e^{-ht}$, from the theory of growth-decay equations. The substitution gives $u(t) - u_1 = A_0e^{-ht}$, which is equivalent to equation (5), provided $A_0 = u_0 - u_1$. ■

Logistic Solution Verification (Theorem 1.2). Given $a > 0$, $b > 0$ and the logistic equation $P' = (a - bP)P$, the plan is to derive the solution formula

$$P(t) = \frac{aP(0)e^{at}}{bP(0)e^{at} + a - bP(0)}.$$

[2] The German mathematician E. E. Kümmer, in his paper in 1834, republished in 1887 in *J. für die reine und angewandte Math.*, considered changes of variable $y = wv$, where w is a given function of x and v is the new variable that replaces y.

Assume $P(t)$ satisfies the logistic equation. Suppose it has been shown (see below) that the variable $u = P/(a - bP)$ satisfies $u' = au$. By the exponential theory, $u = u_0 e^{at}$, hence

$$\begin{aligned}
P &= \frac{au}{1+bu} && \text{Solve } u = P/(a-bP) \text{ for } P \text{ in terms of } u.\\
&= \frac{au_0e^{at}}{1+bu_0e^{at}} && \text{Substitute } u = u_0e^{at}.\\
&= \frac{ae^{at}}{1/u_0 + be^{at}} && \text{Divide by } u_0.\\
&= \frac{ae^{at}}{(a-bP(0))/P(0) + be^{at}} && \text{Use } u_0 = u(0) \text{ and } u = P/(a-bP).\\
&= \frac{aP(0)e^{at}}{bP(0)e^{at} + a - bP(0)}. && \text{Formula verified.}
\end{aligned}$$

The derivation using the substitution $u = P/(a - bP)$ requires only differential calculus. The substitution was found by afterthought, already knowing the solution; historically, integration methods have been applied.

The change of variables $(t, P) \to (t, u)$ given by the equation $u = P(a - bP)$ is used to justify the relation $u' = au$ as follows.

$$\begin{aligned}
u' &= \left(\frac{P}{a-bP}\right)' && \text{It will be shown that } u' = au.\\
&= \frac{P'(a-bP) - P(-bP')}{(a-bP)^2} && \text{Quotient rule applied.}\\
&= \frac{aP'}{(a-bP)^2} && \text{Simplify the numerator.}\\
&= \frac{a(a-bP)P}{(a-bP)^2} && \text{Substitute } P' = (a-bP)P.\\
&= au && \text{Substitute } u = P/(a-bP).
\end{aligned}$$

This completes the motivation for the formula. To verify that it works in the differential equation is a separate issue, which is settled in the exercises.

Exercises 1.1

Growth-Decay Model

Solve the given initial value problem using the growth-decay formula; see page 3 and Example 1.1 page 7.

1. $y' = -3y$, $y(0) = 20$

2. $y' = 3y$, $y(0) = 1$

3. $3A' = A$, $A(0) = 1$

4. $4A' + A = 0$, $A(0) = 3$

5. $3P' - P = 0$, $P(0) = 10$

6. $4P' + 3P = 0$, $P(0) = 11$

7. $I' = 0.005I$, $I(t_0) = I_0$

8. $I' = -0.015I$, $I(t_0) = I_0$

9. $y' = \alpha y$, $y(t_0) = 1$

10. $y' = -\alpha y$, $y(t_0) = y_0$

Growth-decay Theory

11. Graph without a computer $y = 10(2^x)$ on $-3 \le x \le 3$.

12. Graph without a computer $y = 10(2^{-x})$ on $-3 \le x \le 3$.

13. Find the doubling time for the growth model $P = 100e^{0.015t}$.

14. Find the doubling time for the growth model $P = 1000e^{0.0195t}$.

15. Find the elapsed time for the decay model $A = 1000e^{-0.11237t}$ until $|A(t)| < 0.00001$.

16. Find the elapsed time for the decay model $A = 5000e^{-0.01247t}$ until $|A(t)| < 0.00005$.

Newton Cooling Recipe

Solve the given cooling model. Follow Example 1.2 on page 8.

17. $u' = -10(u - 4)$, $u(0) = 5$

18. $y' = -5(y - 2)$, $y(0) = 10$

19. $u' = 1 + u$, $u(0) = 100$

20. $y' = -1 - 2y$, $y(0) = 4$

21. $u' = -10 + 4u$, $u(0) = 10$

22. $y' = 10 + 3y$, $y(0) = 1$

23. $2u' + 3 = 6u$, $u(0) = 8$

24. $4y' + y = 10$, $y(0) = 5$

25. $u' + 3(u + 1) = 0$, $u(0) = -2$

26. $u' + 5(u + 2) = 0$, $u(0) = -1$

27. $\alpha' = -2(\alpha - 3)$, $\alpha(0) = 10$

28. $\alpha' = -3(\alpha - 4)$, $\alpha(0) = 12$

Newton Cooling Model

The cooling model $u(t) = u_0 + A_0e^{-ht}$ is applied; see page 4. Methods parallel those in the flask cooling example, page 9, and the baking example, page 10.

29. (Ingot Cooling) A metal ingot cools in the air at temperature 20C from 130C to 75C in one hour. Predict the cooling time to 23C.

30. (Rod Cooling) A plastic rod cools in a large vat of 12-degree Celsius water from 75C to 20C in 4 minutes. Predict the cooling time to 15C.

31. (Murder Mystery) A body discovered at 1:00 in the afternoon, March 1, 1929, had temperature 80F. Assume outdoor temperature 50F from 9am. Over the next hour the body's temperature dropped to 76F. Estimate the date and time of the murder.

32. (Time of Death) A dead body found in a 40F river had body temperature 70F. The coroner requested that the body be left in the river for 45 minutes, whereupon the body's temperature was 63F. Estimate the time of death, relative to the discovery of the body.

Verhulst Model

Solve the given Verhulst logistic equation using formula (8). Follow Example 1.3 on page 8.

33. $P' = P(2 - P)$, $P(0) = 1$

34. $P' = P(4 - P)$, $P(0) = 5$

35. $y' = y(y - 1)$, $y(0) = 2$

36. $y' = y(y - 2)$, $y(0) = 1$

37. $A' = A - 2A^2$, $A(0) = 3$

38. $A' = 2A - 5A^2$, $A(0) = 1$

39. $F' = 2F(3 - F)$, $F(0) = 2$

40. $F' = 3F(2 - F)$, $F(0) = 1$

Inverse Modeling

Given the model, find the differential equation and initial condition.

41. $A = A_0e^{4t}$

42. $A = A_0e^{-3t}$

43. $P = 1000e^{-0.115t}$

44. $P = 2000e^{-7t/5}$

45. $u = 1 + e^{-3t}$

46. $u = 10 - 2e^{-2t}$

47. $P = \dfrac{10}{10 - 8e^{-2t}}$

48. $P = \dfrac{5}{15 - 14e^{-t}}$

49. $P = \dfrac{1}{5 - 4e^{-t}}$

50. $P = \dfrac{2}{4 - 3e^{-t}}$

Populations

Use Malthusian population theory page 6 and Malthusian model $P(t) = P_0e^{kt}$. Methods appear in Examples 1.4 and 1.5 page 8.

51. **(World Population)** The world population of $5,500,000,000$ people was increasing at a rate of $250,000$ people per day in June of 1993. Predict the date when the population reaches 10 billion.

52. **(World Population)** Suppose the world population at time $t = 0$ is 5.5 billion and increases at rate $250,000$ people per day. How many years before that was the population one billion?

53. **(Population Doubling)** A population of rabbits increases by 10% per year. In how many years does the population double?

54. **(Population Tripling)** A population of bacteria increases by 15% per day. In how many days does the population triple?

55. **(Population Growth)** Trout in a river are increasing by 15% in 5 years. To what population size does 500 trout grow in 15 years?

56. **(Population Growth)** A region of 400 acres contains 1000 forest mushrooms per acre. The population is decreasing by 150 mushrooms per acre every 2 years. Find the population size for the 400-acre region in 15 years.

Verhulst Equation

Write out the solution to the given differential equation and report the carrying capacity $M = \lim_{t\to\infty} P(t)$.

57. $P' = (1 - P)P$

58. $P' = (2 - P)P$

59. $P' = 0.1(3 - 2P)P$

60. $P' = 0.1(4 - 3P)P$

61. $P' = 0.1(3 + 2P)P$

62. $P' = 0.1(4 + 3P)P$

63. $P' = 0.2(5 - 4P)P$

64. $P' = 0.2(6 - 5P)P$

65. $P' = 11P - 17P^2$

66. $P' = 51P - 13P^2$

Logistic Equation

The following exercises use the Verhulst logistic equation $P' = (a - bP)P$, page 6. Some methods appear on page 11.

67. **(Protozoa)** Experiments on the protozoa *Paramecium* determined growth rate $a = 2.309$ and carrying capacity $a/b = 375$ using initial population $P(0) = 5$. Establish the formula $P(t) = \dfrac{375}{1 + 74e^{-2.309t}}$.

68. **(World Population)** Demographers projected the world population in the year 2000 as 6.5 billion, which was corrected by census to 6.1 billion. Use $P(1965) = 3.358 \times 10^9$, $a = 0.029$ and carrying capacity $a/b = 1.0760668 \times 10^{10}$ to compute the logistic equation projection for year 2000.

69. **(Harvesting)** A fish population satisfying $P' = (a - bP)P$ is subjected to harvesting, the new model being $P' = (a - bP)P - H$. Assume $a = 0.04$, $a/b = 5000$ and $H = 10$. Using algebra, rewrite it as $P' = a(\alpha - P)(P - \beta)$ in terms of the roots α, β of $ay - by^2 - H = 0$. Apply the change of variables $u = P - \beta$ to solve it.

70. (Extinction) Let an endangered species satisfy $P' = bP^2 - aP$ for $a > 0$, $b > 0$. The term bP^2 represents births due to chance encounters of males and females, while the term aP represents deaths. Use the change of variable $u = P/(bP - a)$ to solve it. Show from the answer that initial population sizes $P(0)$ below a/b become extinct.

71. (Logistic Answer Check) Let $P = au/(1 + bu)$, $u = u_0 e^{at}$, $u_0 = P_0/(a - bP_0)$. Verify that $P(t)$ is a solution the differential equation $P' = (a - bP)P$ and $P(0) = P_0$.

72. (Logistic Equation) Let k, α, β be positive constants, $\alpha < \beta$. Solve $w' = k(\alpha - w)(\beta - w)$, $w(0) = w_0$ by the substitution $u = (\alpha - w)/(\beta - w)$, showing that $w = (\alpha - \beta u)/(1 - u)$, $u = u_0 e^{(\alpha - \beta)kt}$, $u_0 = (\alpha - w_0)/(\beta - w_0)$. This equation is a special case of the harvesting equation $P' = (a - bP)P + H$.

Growth-Decay Uniqueness Proof

73. State precisely and give a calculus text reference for *Rolle's Theorem*, which says that a function vanishing at $x = a$ and $x = b$ must have slope zero at some point in $a < x < b$.

74. Apply Rolle's Theorem to prove that a differentiable function $v(x)$ with $v'(x) = 0$ on $a < x < b$ must be constant.

1.2 Exponential Application Library

The model differential equation $y' = ky$, and its variants via a change of variables, appears in various applications to biology, chemistry, finance, science and engineering. All the applications below use the exponential model $y = y_0 e^{kt}$.

Light Intensity
Electric Circuits
Drug Dosage
Radioactive Decay
Chemical Reactions
Drug Elimination
Continuous Interest
Radiocarbon Dating

Light Intensity

Physics defines the **lumen unit** to be the light flux through a solid unit angle from a point source of 1/621 watts of yellow light.[3] The lumen is designed for measuring **brightness**, as perceived by the human eye. The **intensity** $E = \frac{F}{A}$ is the flux F per unit area A, with units Lux or Foot-candles (use $A = 1\text{m}^2$ or $A = 1\text{ft}^2$, respectively). At a radial distance r from a point source, in which case $A = 4\pi r^2$, the intensity is given by the **inverse square law**

$$E = \frac{F}{4\pi r^2}.$$

An **exposure meter**, which measures incident or reflected light intensity, consists of a body, a photocell and a readout in units of Lux or Foot-candles. Light falling on the photocell has energy, which is transferred by the photocell into electrical current and ultimately converted to the readout scale.

In classical physics experiments, a jeweler's bench is illuminated by a source of 8000 lumens. The experiment verifies the inverse square law, by reading an exposure meter at 1/2, 1 and 3/2 meters distance from the source.

As a variant on this experiment, consider a beaker of jeweler's cleaning fluid which is placed over the exposure meter photocell; see Figure 3. Successive meter readings with beaker depths of 0, 5, 10, 15 centimeters show that fluid **absorption** significantly affects the meter readings. Photons[4] striking the fluid convert into heat, which accounts for the rapid loss of intensity at depth in the fluid.

[3] Precisely, the wavelength of the light is 550-nm. The unit is equivalent to one **candela**, one of the seven basic SI units, which is the luminous intensity of one sixtieth of a square centimeter of pure platinum held at 1770C.

[4] A photon is the quantum of electromagnetic radiation, of energy $h\nu$, where ν is the radiation frequency and h is Planck's constant.

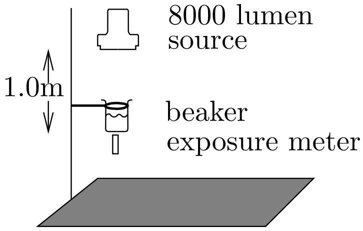


Figure 3. Jeweler's bench experiment. The exposure meter measures light intensity at the beaker's base.

Empirical evidence from experiments suggests that light intensity $I(x)$ at a depth x in the fluid *changes at a rate proportional to itself*, that is,

$$\frac{dI}{dx} = -kI. \tag{9}$$

If I_0 is the surface intensity at zero depth ($x = 0$) and $I(x)$ is the intensity at depth x meters, then the theory of growth-decay equations applied to equation (9) gives the solution

$$I(x) = I_0 e^{-kx}. \tag{10}$$

Equation (10) says that the intensity $I(x)$ at depth x is a percentage of the surface intensity $I(0) = I_0$, the percentage decreasing with depth x.

Electric Circuits

Classical physics analyzes the RC-circuit in Figure 4 and the LR-circuit in Figure 5. The physics background will be reviewed.

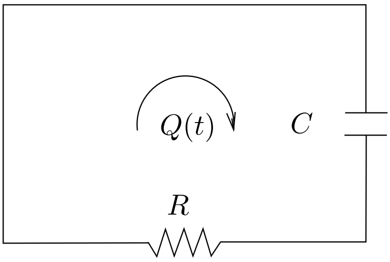


Figure 4. An RC-Circuit, no emf.

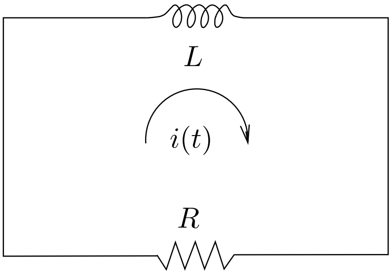


Figure 5. An LR-Circuit, no emf.

First, the **charge** $Q(t)$ in coulombs and the **current** $I(t)$ in amperes are related by the rate formula $I(t) = Q'(t)$. We use prime notation $' = \frac{d}{dt}$. Secondly, there are some empirical laws that are used. There is **Kirchhoff's voltage law**:

The algebraic sum of the voltage drops around a closed loop is zero.

Kirchhoff's **node law** is not used here, because only one loop appears in the examples.

There are the **voltage drop formulas** for an inductor of L henrys, a resistor of R ohms and a capacitor of C farads:

Faraday's law $\qquad V_L = LI'$

Ohm's law	$V_R = RI$
Coulomb's law	$V_C = Q/C$

In Figure 4, Kirchhoff's law implies $V_R + V_C = 0$. The voltage drop formulas show that the charge $Q(t)$ satisfies $RQ'(t) + (1/C)Q(t) = 0$. Let $Q(0) = Q_0$. Growth-decay theory, page 3, gives $Q(t) = Q_0 e^{-t/(RC)}$.

In Figure 5, Kirchhoff's law implies that $V_L + V_R = 0$. By the voltage drop formulas, $LI'(t) + RI(t) = 0$. Let $I(0) = I_0$. Growth-decay theory gives $I(t) = I_0 e^{-Rt/L}$.

In summary:

RC-**Circuit**	$Q = Q_0 e^{-t/(RC)}$, $RQ' + (1/C)Q = 0$, $Q(0) = Q_0$
LR-**Circuit**	$I = I_0 e^{-Rt/L}$, $LI' + RI = 0$, $I(0) = I_0$.

The ideas outlined here are illustrated in Examples 1.9 and 1.10, page 22.

Interest

The notion of **simple interest** is based upon the financial formula

$$A = (1 + r)^t A_0$$

where A_0 is the initial amount, A is the final amount, t is the number of years and r is the **annual interest rate** or **rate per annum** (5% means $r = 5/100$). The **compound interest** formula is

$$A = \left(1 + \frac{r}{n}\right)^{nt} A_0$$

where n is the number of times to compound interest per annum. Use $n = 4$ for **quarterly interest** and $n = 360$ for **daily interest**.

The topic of **continuous interest** rests on the limit formula

$$\lim_{n\to\infty} \left(1 + \frac{r}{n}\right)^{nt} = e^{rt}. \tag{11}$$

Replacement of simple interest by the exponential limit leads to the **continuous interest formula**

$$A = A_0 e^{rt}$$

which by the growth-decay theory arises from the initial value problem

$$\begin{cases} A'(t) = rA(t), \\ A(0) = A_0. \end{cases}$$

Shown on page 27 are the details for taking the limit as $n \to \infty$ in the compound interest formula. In analogy with population theory, the following statement can be made about continuous interest.

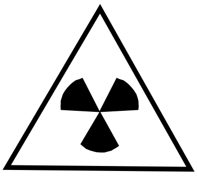
> The amount accumulated by continuous interest increases at a rate proportional to itself.

Applied often in interest calculations is the **geometric sum formula**:

$$1 + r + \cdots + r^n = \frac{r^{n+1} - 1}{r - 1}.$$

Cross-multiplication of identity () by $r - 1$ gives a useful factorization, which for $n = 2$ is the college algebra identity $(1 + r + r^2)(r - 1) = r^3 - 1$.

Radioactive Decay



A constant fraction of the atoms present in a radioactive isotope will spontaneously decay into another isotope of the identical element or else into atoms of another element. Empirical evidence gives the following decay law:

> A radioactive isotope decays at a rate proportional to the amount present.

In analogy with population models the differential equation for radioactive decay is

$$\frac{dA}{dt} = -kA(t),$$

where $k > 0$ is a physical constant called the **decay constant**, $A(t)$ is the number of atoms of radioactive isotope and t is measured in years.

Radiocarbon Dating

The decay constant $k \approx 0.0001245$ is known for carbon-14 (^{14}C). The model applies to measure the date that an organism died, assuming it metabolized atmospheric carbon-14.

The idea of radiocarbon dating is due to Willard S. Libby[5] in the late 1940s. The basis of the chemistry is that radioactive carbon-14, which has two more electrons than stable carbon-12, gives up an electron to become stable nitrogen-14. Replenishment of carbon-14 by cosmic rays keeps atmospheric carbon-14 at a nearly constant ratio with ordinary carbon-12 (this was Libby's assumption). After death, the radioactive decay of carbon-14 depletes the isotope in the organism. The percentage of depletion from atmospheric levels of carbon-14 gives a measurement that dates the organism.

[5] Libby received the Nobel Prize for Chemistry in 1960.

Definition 1.2 (Half-Life)
The **half-life** of a radioactive isotope is the time T required for half of the isotope to decay. In functional notation, it means $A(T) = A(0)/2$, where $A(t) = A(0)e^{kt}$ is the amount of isotope at time t.

For carbon-14, the half-life is 5568 years plus or minus 30 years, according to Libby (some texts and references give 5730 years). The decay constant $k \approx 0.0001245$ for carbon-14 arises by solving for $k = \ln(2)/5568$ in the equation $A(5568) = \frac{1}{2}A(0)$. Experts believe that carbon-14 dating methods tend to underestimate the age of a fossil.

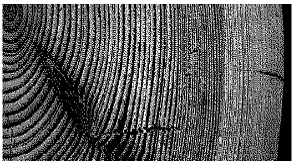
Uranium-238 undergoes decay via alpha and beta radiation into various nuclides, the half-lives of which are shown in Table 1. The table illustrates the range of possible half-lives for a radioactive substance.

Table 1. Uranium-238 Nuclides by Alpha or Beta Radiation.

Nuclide	Half-Life
uranium-238	4,500,000,000 years
thorium-234	24.5 days
protactinium-234	1.14 minutes
uranium-234	233,000 years
thorium-230	83,000 years
radium-236	1,590 years
radon-222	3.825 days
polonium-218	3.05 minutes
lead-214	26.8 minutes
bismuth-214	19.7 minutes
polonium-214	0.00015 seconds
lead-210	22 years
bismuth-210	5 days
polonium-210	140 days
lead-206	stable

Tree Rings

Libby's work was based upon calculations from sequoia tree rings. Later investigations of 4000-year old trees showed that carbon ratios have been nonconstant over past centuries.



Libby's method is advertised to be useful for material 200 years to $40,000$ years old. Older material has been dated using the ratio of disintegration byproducts of potassium-40, specifically argon-40 to calcium-40.

An excellent reference for dating methods, plus applications and historical notes on the subject, is Chapter 1 of Braun [Braun1986].

Chemical Reactions

If the molecules of a substance decompose into smaller molecules, then an empirical law of **first-order reactions** says that the decomposition rate is proportional to the amount of substance present. In mathematical notation, this means

$$\frac{dA}{dt} = -hA(t)$$

where $A(t)$ is the amount of the substance present at time t and h is a physical constant called the **reaction constant**.

The **law of mass action** is used in chemical kinetics to describe **second-order reactions**. The law describes the amount $X(t)$ of chemical C produced by the combination of two chemicals A and B. The empirical law says that the rate of change of X is proportional to the product of the amounts left of chemicals A and B. which is the rate equation

$$X' = k(\alpha - X)(\beta - X), \quad X(0) = X_0. \tag{12}$$

Symbols k, α and β are physical constants, $\alpha < \beta$; see Zill-Cullen [Zill-C], Chapter 2. The substitution $u = (\alpha - X)/(\beta - X)$ is known to transform (12) into $u' = k(\alpha - \beta)u$. See page 11 for the technique. More details are in the exercises. The solution of mass–action model (12):

$$X(t) = \frac{\alpha - \beta u(t)}{1 - u(t)}, \quad u(t) = u_0 e^{(\alpha-\beta)kt}, \quad u_0 = \frac{\alpha - X_0}{\beta - X_0}. \tag{13}$$

Drug Elimination

Some drugs are eliminated from the bloodstream by an animal's body in a predictable fashion. The amount $D(t)$ in the bloodstream declines at a rate proportional to the amount already present. Modeling drug elimination exactly parallels radioactive decay, in that the translated mathematical model is

$$\frac{dD}{dt} = -hD(t),$$

where $h > 0$ is a physical constant, called the **elimination constant** of the drug.

Oral drugs must move through the digestive system and into the gut before reaching the bloodstream. The model $D'(t) = -hD(t)$ applies only after the drug has reached a stable concentration in the bloodstream and the body begins to eliminate the drug.

Examples

Example 1.8 (Light Intensity in a Lake)

Light intensity in a lake is decreased by 75% at depth one meter. At what depth is the intensity decreased by 95%?

Solution: The answer is 2.16 meters (7 feet, $1\frac{1}{16}$ inches). This depth will be justified by applying the light intensity model $I(x) = I_0 e^{-kx}$, where I_0 is the surface light intensity.

At one meter the intensity is $I(1) = I_0 e^{-k}$, but also it is given as $0.25 I_0$. The equation $e^{-k} = 0.25$ results, to determine $k = \ln 4 \approx 1.3862944$. To find the depth x when the intensity has decreased by 95%, solve $I(x) = 0.05 I_0$ for x. The value I_0 cancels from this equation, leaving $e^{-kx} = 1/20$. The usual logarithm methods give $x \approx 2.2$ meters, as follows:

$$\begin{aligned}
\ln e^{-kx} &= \ln(1/20) && \text{Take the logarithm across } e^{-kx} = 1/20.\\
-kx &= -\ln(20) && \text{Use } \ln e^u = u \text{ and } -\ln u = \ln(1/u).\\
x &= \frac{\ln(20)}{k} && \text{Divide by } -k.\\
&= \frac{\ln(20)}{\ln(4)} && \text{Use } k = \ln(4).\\
&\approx 2.16 \text{ meters.} && \text{Only 5\% of the surface intensity remains at 2.16 meters.}
\end{aligned}$$

Example 1.9 (Circuit: RC)

Solve the RC-circuit equation $RQ' + (1/C)Q = 0$ when $R = 2$, $C = 10^{-2}$ and the voltage drop across the capacitor at $t = 0$ is 1.5 volts.

Solution: The charge is $Q = 0.015e^{-50t}$.

To justify this equation, start with the voltage drop formula $V_C = Q/C$, page 17. Then $1.5 = Q(0)/C$ implies $Q(0) = 0.015$. The differential equation is $Q' + 50Q = 0$. The solution from page 3 is $Q = Q(0)e^{-50t}$. Then the equation for the charge in coulombs is $Q(t) = 0.015e^{-50t}$.

Example 1.10 (Circuit: LR)

Solve the LR-circuit equation $LI' + RI = 0$ when $R = 2$, $L = 0.1$ and the resistor voltage drop at $t = 0$ is 1.0 volts.

Solution: The solution is $I = 0.5e^{-20t}$. To justify this equation, start with the voltage drop formula $V_R = RI$, page 17. Then $1.0 = RI(0)$ implies $I(0) = 0.5$. The differential equation is $I' + 20I = 0$; page 3 gives the solution $I = I(0)e^{-20t}$.

Example 1.11 (Compound Interest: Auto Loan)

Compute the fixed monthly payment for a 5-year auto loan of $\$18,000$ at 9% per annum, using (a) daily interest and (b) continuous interest.

Solution: The payments are (a) \$373.9361 and (b) \$373.9360, which differ by hundredths of a cent; details below.

Let $A_0 = 18000$ be the initial amount. It will be assumed that the first payment is due after 30 days and monthly thereafter. To simplify the calculation, a **day** is defined to be 1/360th of a year, regardless of the number of days in that year, and payments are applied every 30 days. Late fees apply if the payment is not received within the **grace period**, but it will be assumed here that all payments are made on time.

Part (a). The **daily interest rate** is 0.09/360 applied for 1800 days (5 years). Between payments P, daily interest is applied to the balance $A(t)$ owed after t periods. The balance grows between payments and then decreases on the day of the payment. The problem is to *find* P *so that* $A(1800) = 0$.

Payment P is subtracted every 30 days, which changes the loan balance $B(n)$ after n days. Define $R = 0.09/360$ (9% daily interest), $B(0) = 18000$, $Z = (1 + R)^{30} = 1.007527251$. Then

$B(30) = B(0)(1 + R)^{30} - P$	Balance after 1 month.
$B(60) = B(30)(1 + R)^{30} - P$	Balance after 2 months.
$\quad = B(0)Z^2 - PZ - P$	Expand using $Z = (1+R)^{30}$ and $B(30) = B(0)Z - P$.
$B(30k) = B(0)Z^k - P\left(1 + \cdots + Z^{k-1}\right)$	For $k = 1, 2, 3, \ldots$.
$\quad = B(0)Z^k - P\dfrac{Z^k - 1}{Z - 1}$	Geometric sum formula page 19 with ratio r replaced by Z.
$0 = B(0)Z^{60} - P\dfrac{Z^{60} - 1}{Z - 1}$	Use $B(1800) = 0$, which corresponds to $k = 60$.
$P = B(0)(Z - 1)\dfrac{Z^{60}}{Z^{60} - 1}$	Solve for P.
$P = 373.9361355$	By `maple`, given $B(0) = 18000$ and $Z = 1.007527251$.

Part (b). The details are the same except for the method of applying daily interest. The daily interest rate remains $R = 0.09/360$. Equation (11) will be used in the form $\left(1 + \dfrac{r}{n}\right)^{nt} \approx e^{rt}$ as $n \to \infty$. Let $n = 360$. Define r and t by the equations $nt = 30$ and $\frac{r}{n} = R$. Replace Z in Part (a): $Z = (1 + R)^{30} \approx e^{rt}$. Then $Z = e^{nRt} = e^{30R} = 1.007528195$ (pause here to confirm). The details:

$B(30) = B(0)Z - P$	Balance after 1 month.
$B(60) = B(30)Z - P$	Balance after 2 months.
$\quad = B(0)Z^2 - PZ - P$	Expand using $Z = (1+R)^{30}$ and $B(30) = B(0)Z - P$.
$B(30k) = B(0)Z^k - P\left(1 + \cdots + Z^{k-1}\right)$	For $k = 1, 2, 3, \ldots$.
$\quad = B(0)Z^k - P\dfrac{Z^k - 1}{Z - 1}$	Geometric sum formula page 19 with ratio r replaced by Z.
$0 = B(0)Z^{60} - P\dfrac{Z^{60} - 1}{Z - 1}$	Use $B(1800) = 0$, which corresponds to $k = 60$.
$P = B(0)(Z - 1)\dfrac{Z^{60}}{Z^{60} - 1}$	Solve for P.
$P = 373.9460360$	By `maple`, given $B(0) = 18000$ and $Z = 1.007528195$.

Example 1.12 (Effective Annual Yield)

A bank advertises an effective annual yield of 5.73% for a certificate of deposit with continuous interest rate 5.5% per annum. Justify the rate.

Solution: The **effective annual yield** is the simple annual interest rate which gives the same account balance after one year. The issue is whether one year means 365 days or 360 days, since banks do business on a 360-day cycle.

Suppose first that one year means 365 days. The model used for a saving account is $A(t) = A_0 e^{rt}$ where $r = 0.055$ is the interest rate per annum. For one year, $A(1) = A_0 e^r$. Then $e^r = 1.0565406$, that is, the account has increased in one year by 5.65%. The *effective annual yield* is 0.0565 or 5.65%.

Suppose next that one year means 360 days. Then the bank pays 5.65% for only 360 days to produce a balance of $A_1 = A_0 e^r$. The extra 5 days make 5/360 years, therefore the bank records a balance of $A_1 e^{5r/360}$ which is $A_0 e^{365r/360}$. The rate for 365 days is then 5.73%, by the calculation

$$\frac{365}{360} 0.0565406 = 0.057325886.$$

Example 1.13 (Retirement Funds)

An engineering firm offers a starting salary of 40 thousand per year, which is expected to increase 3% per year. Retirement contributions are 11% of salary, deposited monthly, growing at 6% continuous interest per annum. The company advertises a million dollars in retirement funds after 40 years. Justify the claim.

Solution: Answer: $1,108,233.90$ in the retirement account after 40 years.

After 39 years of 3% yearly salary increases the initial salary of $\$40,000$ increases to $40000(1.03)^{39} = \$126,681$. In year $n \geq 1$, the 11% retirement contribution is computed from monthly salary $\frac{40000}{12}(1.03)^{n-1}$. The retirement account can be viewed as a 6% continuous interest savings account with monthly deposit. The amount deposited changes each month, which complicates the computation.

Continuous interest rates are $r = 0.06$ (annual) and $s = 0.06/12$ (monthly). Define $R = 1.03$ and $P_0 = 40000/12$. Define monthly salary $P_1 = 40000/12$ for year 1. For year $n \geq 1$ define monthly salary $P_n = P_0 \, R^{n-1}$, because paychecks increase by 3% each year. Define A_n be the amount in the retirement account at the start of year n. The retirement account has zero balance $A_1 = 0$ at the start of employment. Define the monthly retirement contribution in year n to be $R_n = 0.11 \, P_n$.

During the first year, the retirement account gets 12 deposits of R_1 dollars. Monthly continuous interest at s% is applied and re-deposited into the account. The account balance is $A_1 e^s + R_1 e^s$ at the end of month 1, $(A_1 e^s + R_1 e^s)e^s + R_1 e^s$ at the end of month 2, and so on. Then:

$$A_2 = A_1 \, e^{12s} + R_1 \left(e^s + \cdots + e^{12s}\right)$$

Continuous interest at monthly rate $s = 0.06/12$ on the retirement account balance for months 1–12.

$$= A_1 \, e^{12s} + R_1 \frac{e^{12s} - 1}{1 - e^{-s}}$$

Geometric sum with common ratio e^s. The denominator is $e^{-s}(e^s - 1)$.

$$= 4546.026266.$$

Retirement balance at the start of year 2.

$$A_{n+1} = A_n e^{12s} + R_n \frac{e^{12s} - 1}{1 - e^{-s}}$$

General recursion to be proved by induction. The details are omitted.

$$A_{n+1} = \frac{e^{12s}-1}{1-e^{-s}} \sum_{k=1}^{n} R(k)(e^{12s})^{n-k} \qquad \text{Solved recursion. Details below.}$$

The advertised retirement fund after 40 years should be the amount A_{41}, which is obtained by setting $n = 40$ in the last equality: $A_{41} = 1,108,233.904$.

A solved recursion is not required if computer programming is used in a loop to evaluate A_{n+1}.

```
# Maple
s:=0.06/12;P:=n->(40000/12)*(1.03)^(n-1);R:=n->0.11*P(n);
X:=0;for j from 1 to 40 do
X:=X*exp(12*s)+R(j)*(exp(12*s)-1)/(1-exp(-s));end do;
```

Recursion Details.

The recursion is $A_{n+1} = A_nW + R_nZ$ where $W = e^{12s}$ and $Z = \dfrac{e^{12s}-1}{1-e^{-s}}$. The steps used to solve the recursion:

$$\begin{aligned}
A_2 &= A_1W + R_1Z \\
A_3 &= A_2W + R_2Z \\
&= (A_1W + R_1Z)W + R_2Z \\
&= A_1W^2 + Z(R_1W + R_2) \\
&= A_1W^2 + Z\textstyle\sum_{k=1}^{2} R_kW^{2-k} \\
A_4 &= A_3W + R_3Z \\
&= (A_1W^2 + Z(R_1W + R_2))W + R_3Z \\
&= A_1W^3 + Z(R_1W^2 + R_2W + R_3) \\
&= A_1W^3 + Z\textstyle\sum_{k=1}^{3} R_kW^{3-k}
\end{aligned}$$

Induction details are omitted.

Example 1.14 (Half-life of Radium)

A radium sample loses 1/2 percent due to disintegration in 12 years. Verify the half-life of the sample is about $1,660$ years.

Solution: The decay model $A(t) = A_0e^{-kt}$ applies. The given information $A(12) = 0.995A(0)$ reduces to the exponential equation $e^{-12k} = 0.995$. Solve for k with logarithms: $k = \ln(1000/995)/12$. The half-life T satisfies $A(T) = \frac{1}{2}A(0)$, which reduces to $e^{-kT} = 1/2$. Since k is known, the value T can be found as $T = \ln(2)/k \approx 1659.3909$ years.

Example 1.15 (Radium Disintegration)

The disintegration reaction

$$_{88}R^{226} \longrightarrow {}_{88}R^{224}$$

of radium-226 into radon has a half-life of 1700 years. Compute the decay constant k in the decay model $A' = -kA$.

Solution: The half-life equation is $A(1700) = \frac{1}{2}A(0)$. Since $A(t) = A_0e^{-kt}$, the equation reduces to $e^{-1700k} = 1/2$. The latter is solved for k by logarithm methods (see page 8), giving $k = \ln(2)/1700 = 0.00040773364$.

Example 1.16 (Radiocarbon Dating)

The ratio of carbon-14 to carbon-12 in a dinosaur fossil is 6.34 percent of the current atmospheric ratio. Verify the dinosaur's death was about 22,160 years ago.

Solution: The method due to Willard Libby will be applied, using his assumption that the ratio of carbon-14 to carbon-12 in living animals is equal to the atmospheric ratio. Then carbon-14 depletion in the fossil satisfies the decay law $A(t) = A_0e^{-kt}$ for some parameter values k and A_0.

Assume the half-life of carbon-14 is 5568 years. Then $A(5568) = \frac{1}{2}A(0)$ (see page 20). This equation reduces to $A_0e^{-5568k} = \frac{1}{2}A_0e^0$ or $k = \ln(2)/5568$. In short, k is known but A_0 is unknown. It is not necessary to determine A_0 in order to do the verification.

At the time t_0 in the past when the organism died, the amount A_1 of carbon-14 began to decay, reaching the value $6.34A_1/100$ at time $t = 0$ (the present). Therefore, $A_0 = 0.0634A_1$ and $A(t_0) = A_1$. Taking this last equation as the starting point, the final calculation proceeds as follows.

$$\begin{aligned}
A_1 &= A(t_0) && \text{The amount of carbon-14 at death is } A_1,\ -t_0 \text{ years ago.}\\
&= A_0e^{-kt_0} && \text{Apply the decay model } A = A_0e^{-kt} \text{ at } t = t_0.\\
&= 0.0634A_1e^{-kt_0} && \text{Use } A_0 = 6.34A_1/100.
\end{aligned}$$

The value A_1 cancels to give the new relation $1 = 0.0634e^{-kt_0}$. The value $k = \ln(2)/5568$ gives an exponential equation to solve for t_0:

$$\begin{aligned}
e^{kt_0} &= 0.0634 && \text{Multiply by } e^{kt_0} \text{ to isolate the exponential.}\\
\ln e^{kt_0} &= \ln(0.0634) && \text{Take the logarithm of both sides.}\\
t_0 &= \frac{1}{k}\ln(0.0634) && \text{Apply } \ln e^u = u \text{ and divide by } k.\\
&= \frac{5568}{\ln 2}\ln(0.0634) && \text{Substitute } k = \ln(2)/5568.\\
&= -22157.151 \text{ years.} && \text{By calculator. The fossil's age is the negative.}
\end{aligned}$$

Example 1.17 (Percentage of an Isotope)

A radioactive isotope disintegrates by 5% in ten years. By what percentage does it disintegrate in one hundred years?

Solution: The answer is not 50%, as is widely reported by lay persons. The correct answer is 40.13%. It remains to justify this non-intuitive answer.

The model for decay is $A(t) = A_0e^{-kt}$. The decay constant k is known because of the information ... *disintegrates by 5% in ten years*. Translation to equations produces $A(10) = 0.95A_0$, which reduces to $e^{-10k} = 0.95$. Solving with logarithms gives $k = 0.1\ln(100/95) \approx 0.0051293294$.

After one hundred years, the isotope present is $A(100)$, and the percentage is $100\frac{A(100)}{A(0)}$. The common factor A_0 cancels to give the percentage $100e^{-100k} \approx 59.87$. The reduction is 40.13%.

To reconcile the lay person's answer, observe that the amounts present after one, two and three years are $0.95A_0$, $(0.95)^2A_0$, $(0.95)^3A_0$. The lay person should have guessed 100 times $1 - (0.95)^{10}$, which is 40.126306. The common error is to simply multiply 5%

by the ten periods of ten years each. By this erroneous reasoning, the isotope would be depleted in two hundred years, whereas the decay model says that about 36% of the isotope remains!

Example 1.18 (Chemical Reaction)

The manufacture of t-butyl alcohol from t-butyl chloride is made by the chemical reaction

$$(CH_3)_3CCL + NaOH \longrightarrow (CH_3)_3COH + NaCL.$$

Model the production of t-butyl alcohol, when N% of the chloride remains after t_0 minutes.

Solution: It will be justified that the model for alcohol production is $A(t) = C_0(1-e^{-kt})$ where $k = \ln(100/N)/t_0$, C_0 is the initial amount of chloride and t is in minutes.

According to the theory of first-order reactions, the model for chloride depletion is $C(t) = C_0e^{-kt}$ where C_0 is the initial amount of chloride and k is the reaction constant. The alcohol production is $A(t) = C_0 - C(t)$ or $A(t) = C_0(1 - e^{-kt})$. The reaction constant k is found from the initial data $C(t_0) = \frac{N}{100}C_0$, which results in the exponential equation $e^{-kt_0} = N/100$. Solving the exponential equation gives $k = \ln(100/N)/t_0$.

Example 1.19 (Drug Dosage)

A veterinarian applies general anesthesia to animals by injection of a drug into the bloodstream. Predict the drug dosage to anesthetize a 25-pound animal for thirty minutes, given:

1. The drug requires an injection of 20 milligrams per pound of body weight in order to work.
2. The drug eliminates from the bloodstream at a rate proportional to the amount present, with a half-life of 5 hours.

Solution: The answer is about 536 milligrams of the drug. This amount will be justified using exponential modeling.

The drug model is $D(t) = D_0e^{-ht}$, where D_0 is the initial dosage and h is the elimination constant. The half-life information $D(5) = \frac{1}{2}D_0$ determines $h = \ln(2)/5$. Depletion of the drug in the bloodstream means the drug levels are always decreasing, so it is enough to require that the level at 30 minutes exceeds 20 times the body weight in pounds, that is, $D(1/2) > (20)(25)$. The critical value of the initial dosage D_0 then occurs when $D(1/2) = 500$ or $D_0 = 500e^{h/2} = 500e^{0.1\ln(2)}$, which by calculator is approximately 535.88673 milligrams.

Drugs like sodium pentobarbital behave somewhat like this example, although injection in a single dose is unusual. An intravenous drip can sustain the blood levels of the drug, keeping the level closer to the target 500 milligrams.

Details and Proofs

Verification of Continuous Interest by Limiting. Derived here is the continuous interest formula by limiting as $n \to \infty$ in the compound interest formula.

$$\left(1+\frac{r}{n}\right)^{nt} = B^{nt}$$

In the exponential rule $B^x = e^{x\ln B}$, the base is $B = 1 + r/n$.

$$= e^{nt\ln B}$$

Use $B^x = e^{x\ln B}$ with $x = nt$.

$$= e^{\dfrac{r\ln(1+u)}{u}t}$$

Substitute $u = r/n$. Then $u \to 0$ as $n \to \infty$.

$$\approx e^{rt}$$

Because $\ln(1+u)/u \approx 1$ as $u \to 0$, by L'Hospital's rule.

Exercises 1.2

Light Intensity

The following exercises apply the theory of light intensity on page 16, using the model $I(t) = I_0 e^{-kx}$ with x in meters. Methods parallel Example 1.8 on page 21.

1. The light intensity is $I(x) = I_0 e^{-1.4x}$ in a certain swimming pool. At what depth x does the light intensity fall off by 50%?

2. The light intensity in a swimming pool falls off by 50% at a depth of 2.5 meters. Find the depletion constant k in the exponential model.

3. Plastic film is used to cover window glass, which reduces the interior light intensity by 10%. By what percentage is the intensity reduced, if two layers are used?

4. Double-thickness colored window glass is supposed to reduce the interior light intensity by 20%. What is the reduction for single-thickness colored glass?

RC-Electric Circuits

In the exercises below, solve for $Q(t)$ when $Q_0 = 10$ and graph $Q(t)$ on $0 \le t \le 5$.

5. $R = 1$, $C = 0.01$.

6. $R = 0.05$, $C = 0.001$.

7. $R = 0.05$, $C = 0.01$.

8. $R = 5$, $C = 0.1$.

9. $R = 2$, $C = 0.01$.

10. $R = 4$, $C = 0.15$.

11. $R = 4$, $C = 0.02$.

12. $R = 50$, $C = 0.001$.

LR-Electric Circuits

In the exercises below, solve for $I(t)$ when $I_0 = 5$ and graph $I(t)$ on $0 \le t \le 5$.

13. $L = 1$, $R = 0.5$.

14. $L = 0.1$, $R = 0.5$.

15. $L = 0.1$, $R = 0.05$.

16. $L = 0.01$, $R = 0.05$.

17. $L = 0.2$, $R = 0.01$.

18. $L = 0.03$, $R = 0.01$.

19. $L = 0.05$, $R = 0.005$.

20. $L = 0.04$, $R = 0.005$.

Interest and Continuous Interest

Financial formulas which appear on page 18 are applied below, following the ideas in Examples 1.11, 1.12 and 1.13, pages 22 and 24.

21. **(Total Interest)** Compute the total daily interest and also the total continuous interest for a 10-year loan of $5,000$ dollars at 5% per annum.

22. **(Total Interest)** Compute the total daily interest and also the total continuous interest for a 15-year loan of $7,000$ dollars at $5\frac{1}{4}$% per annum.

23. **(Monthly Payment)** Find the monthly payment for a 3-year loan of $8,000$ dollars at 7% per annum compounded continuously.

24. **(Monthly Payment)** Find the monthly payment for a 4-year loan of $7,000$ dollars at $6\frac{1}{3}\%$ per annum compounded continuously.

25. **(Effective Yield)** Determine the effective annual yield for a certificate of deposit at $7\frac{1}{4}\%$ interest per annum, compounded continuously.

26. **(Effective Yield)** Determine the effective annual yield for a certificate of deposit at $5\frac{3}{4}\%$ interest per annum, compounded continuously.

27. **(Retirement Funds)** Assume a starting salary of $35,000$ dollars per year, which is expected to increase 3% per year. Retirement contributions are $10\frac{1}{2}\%$ of salary, deposited monthly, growing at $5\frac{1}{2}\%$ continuous interest per annum. Find the retirement amount after 30 years.

28. **(Retirement Funds)** Assume a starting salary of $45,000$ dollars per year, which is expected to increase 3% per year. Retirement contributions are $9\frac{1}{2}\%$ of salary, deposited monthly, growing at $6\frac{1}{4}\%$ continuous interest per annum. Find the retirement amount after 30 years.

29. **(Actual Cost)** A van is purchased for $18,000$ dollars with no money down. Monthly payments are spread over 8 years at $12\frac{1}{2}\%$ interest per annum, compounded continuously. What is the actual cost of the van?

30. **(Actual Cost)** Furniture is purchased for $15,000$ dollars with no money down. Monthly payments are spread over 5 years at $11\frac{1}{8}\%$ interest per annum, compounded continuously. What is the actual cost of the furniture?

Radioactive Decay

Assume the decay model $A' = -kA$ from page 19. Below, $A(T) = 0.5A(0)$ defines the *half-life* T. Methods parallel Examples 1.14– 1.17 on pages 25– 26.

31. **(Half-Life)** Determine the half-life of a radium sample which decays by 5.5% in 13 years.

32. **(Half-Life)** Determine the half-life of a radium sample which decays by 4.5% in 10 years.

33. **(Half-Life)** Assume a radioactive isotope has half-life 1800 years. Determine the percentage decayed after 150 years.

34. **(Half-Life)** Assume a radioactive isotope has half-life 1650 years. Determine the percentage decayed after 99 years.

35. **(Disintegration Constant)** Determine the constant k in the model $A' = -kA$ for radioactive material that disintegrates by 5.5% in 13 years.

36. **(Disintegration Constant)** Determine the constant k in the model $A' = -kA$ for radioactive material that disintegrates by 4.5% in 10 years.

37. **(Radiocarbon Dating)** A fossil found near the town of Dinosaur, Utah contains carbon-14 at a ratio of 6.21% to the atmospheric value. Determine its approximate age according to Libby's method.

38. **(Radiocarbon Dating)** A fossil found in Colorado contains carbon-14 at a ratio of 5.73% to the atmospheric value. Determine its approximate age according to Libby's method.

39. **(Radiocarbon Dating)** In 1950, the Lascaux Cave in France contained charcoal with 14.52% of the carbon-14 present in living wood samples nearby. Estimate by Libby's method the age of the charcoal sample.

40. **(Radiocarbon Dating)** At an excavation in 1960, charcoal from building material had 61% of the carbon-14 present in living wood nearby. Estimate the age of the building.

41. **(Percentage of an Isotope)** A radioactive isotope disintegrates by 5% in 12 years. By what percentage is it reduced in 99 years?

42. **(Percentage of an Isotope)** A radioactive isotope disintegrates by 6.5% in 1, 000 years. By what percentage is it reduced in 5, 000 years?

Chemical Reactions

Assume below the model $A' = kA$ for a first-order reaction. See page 21 and Example 1.18, page 27.

43. **(First-Order $A + B \longrightarrow C$)** A chemical reaction produces $X(t)$ grams of product C from 50 grams of chemical A and 32 grams of catalyst B. The reaction uses 1 gram of A to 4 grams of B. Variable t is in minutes. Justify for some constant K the model $\frac{dX}{dt} = K\left(50 - \frac{1}{5}X\right)\left(32 - \frac{4}{5}X\right)$ and calculate $\lim_{t\to\infty} X(t) = 40$.

44. **(First-Order $A + B \longrightarrow C$)** A first order reaction produces product C from chemical A and catalyst B. Model the production of C using a grams of A and b grams of B, assuming initial amounts M of A and N of B, $M < N$.

45. **(Law of Mass-Action)** Consider a second-order chemical reaction $X(t)$ with $k = 0.14$, $\alpha = 1$, $\beta = 1.75$, $X(0) = 0$. Find an explicit formula for $X(t)$ and graph it on $t = 0$ to $t = 2$.

46. **(Law of Mass-Action)** Consider a second-order chemical reaction $X(t)$ with $k = 0.015$, $\alpha = 1$, $\beta = 1.35$, $X(0) = 0$. Find an explicit formula for $X(t)$ and graph it on $t = 0$ to $t = 10$.

47. **(Mass-Action Derivation)** Let k, α, β be positive constants, $\alpha < \beta$. Solve $X' = k(\alpha - X)(\beta - X)$, $X(0) = X_0$ by the substitution $u = (\alpha - X)/(\beta - X)$, showing that $X = (\alpha - \beta u)/(1 - u)$, $u = u_0 e^{(\alpha-\beta)kt}$, $u_0 = (\alpha - X_0)/(\beta - X_0)$.

48. **(Mass-Action Derivation)** Let k, α, β be positive constants, $\alpha < \beta$. Define $X = (\alpha - \beta u)/(1 - u)$, where $u = u_0 e^{(\alpha-\beta)kt}$ and $u_0 = (\alpha - X_0)/(\beta - X_0)$. Verify by calculus computation that (1) $X' = k(\alpha - X)(\beta - X)$ and (2) $X(0) = X_0$.

Drug Dosage

Employ the drug dosage model $D(t) = D_0 e^{-ht}$ given on page 21. Apply the techniques of Example 1.19, page 27.

49. **(Injection Dosage)** Bloodstream injection of a drug into an animal requires a minimum of 20 milligrams per pound of body weight. Predict the dosage for a 12-pound animal which will maintain a drug level 3% higher than the minimum for two hours. Assume half-life 3 hours.

50. **(Injection Dosage)** Bloodstream injection of an antihistamine into an animal requires a minimum of 4 milligrams per pound of body weight. Predict the dosage for a 40-pound animal which will maintain an antihistamine level 5% higher than the minimum for twelve hours. Assume half-life 3 hours.

51. **(Oral Dosage)** An oral drug with half-life 2 hours is fully absorbed into the bloodstream in 45 minutes, blood level 63% of the dose. Assume 500 milligrams in the first dose is fully absorbed at $t = 0$. A second dose is taken 1 hour later to maintain a blood level of at least 180 milligrams for 2.5 hours. Explain why 1 hour might be reasonable.

52. **(Oral Dosage)** An oral drug with half-life 2 hours is fully absorbed into the bloodstream in 45 minutes, blood level 63% of the dose. Determine three (small) dosage amounts, and their administration time, which keep the blood level above 180 milligrams but below 280 milligrams over three hours.

1.3 Differential Equations of First Order

The nature of a solution is studied through possible representations as explicit or implicit equations, numeric tables and graphical visualization.

First Order Differential Equation

The equation

$$y'(x) = f(x, y(x)) \tag{1}$$

is called a **first order differential equation**. The function $f(x, y)$ is defined in a region D of the xy-plane. In most physical applications f is continuous in D or else it has simple discontinuities, such as those caused by switches.

Cited below are some striking examples of first order differential equations in science and engineering.

$$\frac{dy}{dx} = F(x)$$

The **fundamental theorem of calculus**, Appendix A, implies that $y(x) = \int_{x_0}^{x} F(t)dt$ satisfies differential equation $y' = F(x)$.

$$\frac{du}{dt} = -k(u - u_1)$$

Cooling of a body with temperature $u(t)$ in a medium of temperature u_1 obeys Newton's law of cooling. Symbol k is the **cooling constant**.

$$\frac{dQ}{dt} = k(T^4 - T_0^4)$$

Stefan's radiation law models the heat lost by a body of temperature T in a medium of temperature T_0 due to thermal radiation.

$$\frac{dy}{dt} = -h\sqrt{|y(t)|}$$

Tank draining obeys **Torricelli's law**, where h is a constant and y is the fluid depth in the tank at time t.

$$\frac{dP}{dt} = kP$$

Population dynamics may assume Malthus's reproduction law: *the population changes at a rate proportional to the present population P.*

$$\frac{dv}{dt} = F/m$$

Free fall velocity $v(t)$ of a mass m accelerating due to constant gravitational force F obeys Newton's second law $F = ma$, where a is the acceleration.

$$\frac{dy}{dx} = k(a^2 - x^2)$$

Boat trajectory for a river crossing, with the fastest current in the center, can be modelled by the distance x from the center and the distance $y(x)$ downstream.

Symbolic Formula for $y(x)$ is Unlikely

A quadratic equation $ax^2 + bx + c = 0$ has *numerical answer* $x = -b/(2a) \pm \frac{\sqrt{b^2 - 4ac}}{2a}$. Differential equations have answers that are graphs, represented by *functions* $y(x)$. Sadly, it is generally impossible to write down a symbolic formula for the **answer** $y(x)$ to a given differential equation $\frac{dy}{dx} = f(x, y(x))$.

Applied Models

Science and engineering modelers are not much interested in *solving a differential equation.* They use differential equations to express or define a variable via a mathematical model. Initially, during modeling stages, theoretical existence suffices for the variable's resultant function. After proper modeling, analytical and numerical methods might be applied to actually find the function. In summary:

> Differential equations are used in application modeling to **define** or **express** a variable/function of the physical parameters.

Tables, Formulas and Graphs as Answers

An answer to a differential equation problem is given in various forms, suited to the intended application. The most common forms are **tables**, **equations** and **graphs**. Answers are related to the notion of a **solution**, which is a precise mathematical term, defined below.

Definition 1.3 (Solution)
Let $f(x, y)$ be defined for $a < x < b$ and $c < y < d$. A **solution** $y(x)$ to the differential equation $\frac{dy}{dx} = f(x, y)$ on the interval (a, b) is a function $y(x)$ defined for $a < x < b$ such that

> (1) The left side $y'(x)$ of the differential equation and the right side $f(x, y(x))$ are defined for each $a < x < b$.
>
> (2) Substitution of $y(x)$ in each side gives symbolically equal expressions for each value of x in the domain $a < x < b$.

Often solution formulas contain physical constants represented as symbols, like R and L in an RL-circuit equation. In such cases the definition is modified to say *each side gives symbolically equal expressions for all symbols.*

Extensions. The definition can be restated for half-open intervals, closed intervals and intervals in which one or both endpoints are infinite. If $f(x, y)$ contains discontinuous switches, then the definition of solution is relaxed, possibly excluding points of discontinuity.

Impulse Modeling. The definition does not apply as stated to the case when $f(x, y)$ contains impulses (hammer hits or instantaneous injection of energy). Laplace Theory provides an accessible introduction.

Definition 1.4 (Equilibrium Solution)
A constant solution $y(x) = k$ to the differential equation $y' = f(x, y)$ is called an **equilibrium solution**.

Equivalent terms. Literature may use **rest solution** and/or **steady state solution**. The meaning: $y(x)$ equals a number k for all values of x. Function y satisfies $y' = 0$: the motion is at rest. Steady-state behavior means *after a long time*, then k is the constant limit of a time-varying solution $y(x)$ as $x \to \infty$ (time=x). Symbol t is often used instead of x for models with time domain, in which case the differential equation becomes $y'(t) = f(t, y(t))$ and $' = d/dt$.

To illustrate the notion of equilibrium solution, consider $y' = y(1 - y)$. This equation has two equilibrium solutions $y = 0$ and $y = 1$. They are found by formal substitution of $y = k$ into $y' = y(1 - y)$ and then solving for k in the formal equation $0 = k(1 - k)$.

The equation $y' = x(1 - y)$ has equilibrium solution $y = 1$. The equation $x = 0$ is not an equilibrium solution: it is a **red herring**, often reported in error. The formal equation $0 = x(1 - k)$ is solved for k with symbol x allowed to assume all possible values. Then $x \neq 0$ forces $k = 1$. The expected report: equilibrium solution $y = 1$.

Definition 1.5 (Initial Value Problem)
The **initial value problem** for a first order equation $y' = f(x, y)$ on $a < x < b$ is the problem of finding a solution $y(x)$ on $a < x < b$ which in addition satisfies an **initial condition** of the form $y = y_0$ at $x = x_0$.

Notation. An initial condition may be given in compact notation $y(x_0) = y_0$. Substitution notation can be used as in integration theory, e.g., $\int_0^1 x\,dx = (x^2/2)\big|_{x=0}^{x=1}$. For instance, if $y = x + 10$ is the expected solution, then $y(0) = 10$ is the same as $(x + 10)\big|_{x=0} = 10$. In general, the notation is $y(x)\big|_{x=x_0} = y_0$.

To make sense of the initial condition, $f(x, y)$ must have (x_0, y_0) in its domain of definition, that is, $a < x_0 < b$ and $c < y_0 < d$. Similar statements apply to more general domains.

Uniqueness

In typical applications, just one solution is isolated by the initial condition. Having just one solution is not obvious on physical grounds; see Example 1.20. **Non-uniqueness** allows modeling an answer like $y = 1 + x^3$ through an initial value problem, while a numerical procedure computes a different answer like $y = 1$. **Uniqueness** forces the modeler and the solver to find the same answer. The jobs of scientists and engineers include keeping computers from producing nonsense numbers and incorrect graphs. It is possible for bad modeling, which allows non-uniqueness, to cause bad results to come off the computer. In summary:

> Numerical answers and computer graphs obtained from the differential equation $y' = f(x, y)$ are **nonsense** unless the model has a unique solution.

Explicit and Implicit Equations

Equations that represent answers to first order differential equations are either **implicit** or **explicit**. An equation with y isolated on the left side and right side independent of y is called **explicit**. Otherwise, the equation is called **implicit**. Some examples:

$y = \sin x + e^{-x}$	Equations treated in differential calculus are **explicit** equations.
$y = f(x)$	Equations given in abstract functional notation are **explicit** equations.
$y = 1 + \pi$	Constant equations are **explicit** equations.
$2y = 1$	An **implicit** equation (y not isolated left). Can be converted to explicit equation $y = 1/2$.
$x + y = 1$	As written, y is not isolated on the left, so it is an **implicit** equation. It can be converted to the **explicit** equation $y = 1 - x$.
$x^2 + y^2 = 1$	The equation of a circle is an **implicit** equation.
$f(x, y) = c$	Abstract level curve equations are assumed to be in **implicit** form. To convert to **explicit** form, solve for y in terms of x.
$x + y^2 = 1$	As written, y is not isolated on the left, so it is an **implicit** equation. It converts to two **explicit** equations $y = \sqrt{1-x}$ and $y = -\sqrt{1-x}$.

Definition 1.6 (Explicit Equation)
An xy-equation is **explicit** if exactly y appears on the left, followed by an equal sign, followed by an expression independent of y. In functional notation, the equation must have the form $y = f(x)$.

Any equation that is not explicit is called **implicit**.

Illustrations. Equations $2y = x$, $-y = 1 + x$ and $xy = 1$ are implicit, but they can be converted by algebra into the explicit equations $y = x/2$, $y = -1 - x$, $y = 1/x$. Any explicit equation can be re-written in infinitely many ways as a implicit equation.

Numeric Tables

A **numeric table** is a list of x, y values. Tables are finite lists. Typical numeric tables appear in Examples 1.22 and 1.23 on page 36.

A numeric table for solution $y(x)$ of differential equation $y' = f(x, y)$ can be generated by a **numerical method**. Normally, the x-values are equally spaced on some interval. A specific numerical method is applied to find each of the y-values. The most elementary numerical methods are Euler's method, Heun's method and the Runge-Kutta method.

A numeric table in current scientific literature may assume that x or y is a *vector variable*. The effect is to allow numeric tables with multiple columns.

Graphics

Graphs of solutions to differential equations $y' = f(x, y)$ can be generated by hand from numeric data. The most popular method for hand-graphing is the **connect-the-dots method**. This method constructs a graph as straight-line connections of the data points. An illustration is Example 1.24, page 37.

Curve library methods and computer methods for graphing equations and numerical data sets are considered elsewhere; see Appendix A.2. The methods apply especially if the curve is given by an equation, either explicit or implicit.

Examples

Example 1.20 (IVP with Two Solutions)

Display an answer check for the initial value problem (IVP) on interval $x \geq 0$, showing that it has two solutions: (1) $y(x) = x^2/4$ and (2) constant solution $y(x) = 0$.

$$y' = \sqrt{|y|}, \quad y(0) = 0,$$

Solution: The example is important, because modern computer algebra systems allow numeric methods to be blindly applied to examples like this one. No error messages are emitted by such computer programs. Failures, routinely blamed on computers, can be the result of unexpected modeling intricacies.

The example is curiously close to the tank-draining equation $y' = -h\sqrt{y}$ based upon *Torricelli's law*, page 31. Arguments that an equation *physically has a unique solution* are unheard by computer programs: the programs are not smarter than the humans who employ them.

The tank draining problem has no unique solution for $y(0) = 0$, because solutions must be defined on $-H < x < H$, not on a half-interval like $x \geq 0$. Condition $y(0) = 0$ means the tank is empty at time $x = 0$. An empty tank could have occurred any time *before* time $x = 0$. There is no way from data $y(0) = 0$ to determine when the tank emptied, so there are infinitely many events that could lead to $y(0) = 0$, all of which are solutions to the differential equation model.

The verification involves two steps:

(a) The differential equation $y' = \sqrt{|y|}$ has y as a solution.

(b) The initial condition $y(0) = 0$ ($y = 0$ at $x = 0$) is satisfied.

Answer Check for Solution $y(x) = x^2/4$.
In both steps (a) and (b), the verification amounts to expanding the left hand side (LHS) and right hand side (RHS) of the equalities, then a check is made for equality of the LHS and RHS, for all symbols. The details for $y = x^2/4$ are as follows.

LHS $= y'$	The left side of $y' = \sqrt{\vert y\vert}$ is y'.
$= (x^2/4)'$	The solution being tested is $y = x^2/4$.
$= x/2$,	and
RHS $= \sqrt{\vert y\vert}$	The right side of $y' = \sqrt{\vert y\vert}$.
$= \sqrt{\vert x^2/4\vert}$	Because $y = x^2/4$.
$= x/2$	Because $x \ge 0$.

Therefore, LHS = RHS, and step (a) is finished.

Answer Check for Initial Condition $y(0) = 0$.
To complete step (b), proceed similarly:

LHS $= y(0)$	Initial condition $y(0) = 0$ left side.
$= (x^2/4)\big\vert_{x=0}$	The solution being tested is $y = x^2/4$.
$= 0$	
$=$ RHS	The right side of $y(0) = 0$.

Answer Check for Solution $y(x) = 0$.
The details for the constant solution $y(x) = 0$ are similar. As a mental exercise, repeat the steps above with $x^2/4$ replaced by 0, to verify steps (a) and (b) for the constant solution $y(x) = 0$.

Example 1.21 (Implicit and Explicit Equations)

Classify $1 + e^y = x^2$ as implicit or explicit. If implicit, then find an explicit representation for y in terms of x.

Solution: The equation is classified as *implicit*, because y is not isolated on the left side. Conversion to explicit form uses college algebra, as follows.

$1 + e^y = x^2$	Given equation. Solving for y.
$e^y = x^2 - 1$	Isolate y-terms on the left.
$\ln e^y = \ln \vert x^2 - 1\vert$	Take the logarithm of both sides.
$y = \ln \vert x^2 - 1\vert$	Simplify the left side. Identity $\ln e^u = u$ applied.

Example 1.22 (Verify a Numerical Table)

Verify Table 2 using the explicit equation $y = 1 - x + 2x^2$.

Table 2. Numerical data for an explicit equation.

x	0.0	0.1	0.2	0.3	0.4
y	1.0	0.92	0.88	0.88	0.92

Solution: Each column is verified in an identical way. For example, column 2 is checked by substitution of $x = 0.1$ into $y = 1 - x + 2x^2$ to obtain $y = 1 - 0.1 + 2(0.1)^2 = 0.92$.

Example 1.23 (Verify an Approximation Table)

Verify Table 3 using the approximation formula $y(x+0.1) \approx y(x) + 0.1(x + y(x))$.

Table 3. Numerical data for an approximation formula.

x	y
0.0	1.0
0.1	1.1
0.2	1.22
0.3	1.362
0.4	1.5282

Solution: The formula is applied as a **recursion**, which is a set of formulas which generate from a *given table pair* x, y the *next table pair* X, Y via the relations

$$X = x + 0.1, \quad Y = y + 0.1(x + y).$$

Important in the mathematical translation is the elimination of the approximation symbol ($\approx$) and the use of equal signs ($=$) in the final relations.

Each row is verified in an identical way. For example, row 3 is checked by substitution of data from the previous row. Items $x = 0.1$ and $y = 1.1$ from row 2 are substituted into $X = x+0.1$ and $Y = y+0.1(x+y)$ to obtain $X = 0.2$ and $Y = 1.22$. The approximations 0.2, 1.22 are then copied to row 3 of the table.

Example 1.24 (Hand Graphing of Numeric Data)

Graph on engineering paper the piecewise-defined function $y(x)$ using six data points from $x = 0$ to $x = 1/2$ in steps of 0.1.

$$y(x) = \begin{cases} 1.1x + 1.10 & 0.0 \le x \le 0.1, \\ -1.6x + 1.37 & 0.1 < x \le 0.2, \\ 1.5x + 0.75 & 0.2 < x \le 0.3, \\ 0.1x + 0.90 & 0.3 < x \le 0.4, \\ -0.2x + 2.10 & 0.4 < x \le 0.5. \end{cases}$$

Solution: The xy-data points for $y(x)$ are

$$\begin{array}{ccc} (0.0, 1.10), & (0.1, 1.21), & (0.2, 1.05), \\ (0.3, 1.20), & (0.4, 1.30), & (0.5, 1.10). \end{array}$$

Engineering paper divisions are set for this example at 0.1 horizontal and 0.1 vertical. The origin will be $x = 0.0$, $y = 1.0$. The first step is to plot the points as *dots*. The second step connects the dots with straight lines, just as in children's *connect-the-dot* puzzles. The graphic appears in Figure 6. The figure is correct between data points because $y(x)$ is piecewise linear. Generally, the connect-the-dot method makes errors between data points.

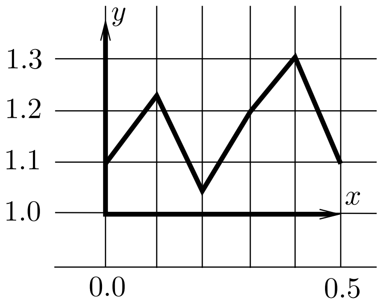


Figure 6. Engineering paper graphic of numeric data.

Exercises 1.3

Solution Verification

Given the differential equation, initial condition and proposed solution y, verify that y is a solution. Don't try to *solve* the equation!

1. $\dfrac{dy}{dx} = y$, $y(0) = 2$, $y = 2e^x$

2. $y' = 2y$, $y(0) = 1$, $y = e^{2x}$

3. $y' = y^2$, $y(0) = 1$, $y = (1-x)^{-1}$

4. $\dfrac{dy}{dx} = y^3$, $y(0) = 1$,
$y = (1-2x)^{-1/2}$

5. $D^2y(x) = y(x)$, $y(0) = 2$,
$Dy(0) = 2$, $y = 2e^x$

6. $D^2y(x) = -y(x)$, $y(0) = 0$,
$Dy(0) = 1$, $y = \sin x$

7. $y' = \sec^2 x$, $y(0) = 0$, $y = \tan x$

8. $y' = -\csc^2 x$, $y(\pi/2) = 0$,
$y = \cot x$

9. $y' = e^{-x}$, $y(0) = -1$, $y = -e^{-x}$

10. $y' = 1/x$, $y(1) = 1$, $y = \ln x$

Explicit and Implicit Solutions

Identify the given solution as *implicit* or *explicit*. If *implicit*, then solve for y in terms of x by college algebra methods.

11. $y = x + \sin x$

12. $y = x + \sin x$

13. $2y + x^2 + x + 1 = 0$

14. $x - 2y + \sin x + \cos x = 0$

15. $y = e^\pi$

16. $e^y = \pi$

17. $e^{2y} = \ln(1+x)$

18. $\ln|1 + y^2| = e^x$

19. $\tan y = 1 + x$

20. $\sin y = (x-1)^2$

Tables and Explicit Equations

For the given explicit equation, make a table of values $x = 0$ to $x = 1$ in steps of 0.2.

21. $y = x^2 - 2x$

22. $y = x^2 - 3x + 1$

23. $y = \sin \pi x$

24. $y = \cos \pi x$

25. $y = e^{2x}$

26. $y = e^{-x}$

27. $y = \ln(1+x)$

28. $y = x\ln(1+x)$

Tables and Approximate Equations

Make a table of values $x = 0$ to $x = 1$ in steps of 0.2 for the given approximate equation. Identify precisely the *recursion* formulas applied to obtain the next table pair from the previous table pair.

29. $y(x+0.2) \approx y(x)+0.2(1-y(x))$, $y(0) = 1$

30. $y(x+0.2) \approx y(x)+0.2(1+y(x))$, $y(0) = 1$

31. $y(x+0.2) \approx y(x)+0.2(x-y(x))$, $y(0) = 0$

32. $y(x+0.2) \approx y(x)+0.2(2x+y(x))$, $y(0) = 0$

33. $y(x+0.2) \approx y(x)+0.2(\sin x + xy(x))$, $y(0) = 2$

34. $y(x+0.2) \approx y(x)+0.2(\sin x - x^2y(x))$, $y(0) = 2$

35. $y(x+0.2) \approx y(x)+0.2(e^x - 7y(x))$, $y(0) = -1$

36. $y(x+0.2) \approx y(x)+0.2(e^{-x} - 5y(x))$, $y(0) = -1$

37. $y(x+0.2) \approx y(x)+0.1(e^{-2x} - 3y(x))$, $y(0) = 2$

38. $y(x+0.2) \approx y(x)+0.2(\sin 2x - 2y(x))$, $y(0) = 2$

Hand Graphing

Make a graphic by hand on engineering paper, using 6 data points. Cite the divisions assigned horizontally and vertically. Label the axes and the center of coordinates. Supply one sample hand computation per graph. Employ a computer program or calculator to obtain the data points.

39. $y = 5x^3$, $x = 0$ to $x = 1$.

40. $y = 3x$, $x = 0$ to $x = 1$.

41. $y = 2x^5$, $x = 0$ to $x = 1$.

42. $y = 3x^7$, $x = 0$ to $x = 1/2$.

43. $y = 2x^4$, $x = 0$ to $x = 1$.

44. $y = 3x^6$, $x = 0$ to $x = 1$.

45. $y = \sin x$, $x = 0$ to $x = \pi/4$.

46. $y = \cos x$, $x = 0$ to $x = \pi/4$.

47. $y = \dfrac{x+1}{x+2}$, $x = 0$ to $x = 1$.

48. $y = \dfrac{x-1}{x+1}$, $x = 0$ to $x = 1$.

49. $y = \ln(1+x)$, $x = 0$ to $x = 1$.

50. $y = \ln(1+2x)$, $x = 0$ to $x = 1$.

1.4 Direction Fields

The **method of direction fields** is a graphical method for displaying the general shape and behavior of solutions to $y' = f(x,y)$. It persists as a fundamental topic because it *does not require solving the differential equation* $y' = f(x,y)$. The uniform grid method and the isocline method are introduced, for computer and hand construction of direction fields.

Euler's Visualization

L. Euler (1707–1783) discovered a way to draw a graphic showing the behavior of all solutions to a given differential equation, *without solving the equation.* The graphic is built from a grid of points arranged on a graph window. Paired with each grid point is a line segment centered on the grid point. The line segments are non-overlapping. Euler's idea is to replace the differential equation model $y' = f(x,y)$ by a graphical model.

Definition 1.7 (Direction Field for $y' = f(x,y)$)
A graph window plus pairs of grid points and non-overlapping line segments is called a **direction field**, provided the line segment at grid point (x_0, y_0) coincides with the tangent line to the solution $y(x)$ of the initial value problem $\begin{cases} y' = f(x,y), \\ y(x_0) = y_0. \end{cases}$ The line segment at grid point (x_0, y_0) is forced to have slope $m = f(x_0, y_0)$.

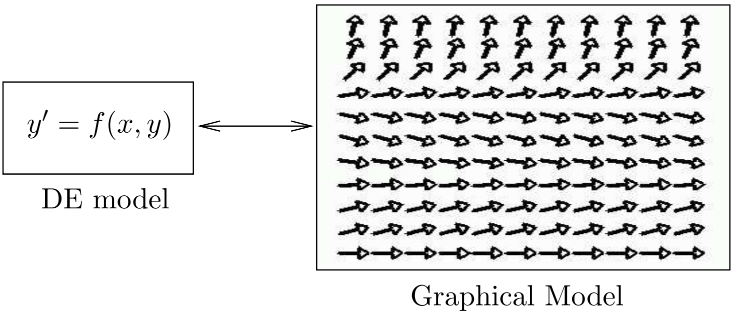


Figure 7. Model Replacement.

A differential equation model $y' = f(x,y)$ is replaced by a direction field model. The graphic can be enriched with a few edge-to-edge solution curves.

Important: We don't have to know a formula for $y(x)$, because $y'(x_0)$ can be computed from its equivalent formula $y'(x_0) = f(x_0, y_0)$, a number that depends only on the grid point (x_0, y_0) and the function $f(x,y)$.

Solution Curves and Direction Fields

Euler's visualization idea begins with the direction field, drawn for some graph window, with pairs of grid points and line segments dense enough to cover most of the white space in the graph window. The *theory* used in Euler's idea consists of a short list of facts:

1. Solutions of $y' = f(x, y)$ don't cross.
2. A tangent to an edge-to-edge solution $y(x)$ nearly matches tangents to nearby direction field segments.
3. Direction field segments are solutions of $y' = f(x, y)$, to pixel resolution.

Details 1: If solutions $y_1(x)$, $y_2(x)$ cross at $x = x_0$, then let $y_0 = y_1(x_0) = y_2(x_0)$ and consider the initial value problem

$$\begin{cases} y' = f(x, y), \\ y(x_0) = y_0. \end{cases}$$

We assume solutions to all such initial value problems are *unique*. This implies $y_1(x) = y_2(x)$ for $|x - x_0|$ small. Hence crossings are impossible. The analysis implies that two solutions which touch must coalesce.

Direction field segments represent solution curves, so they must be constructed not to touch each other. Edge-to-edge solution curves cannot cross a direction field segment, but they may coincide with a direction field segment, to pixel resolution.

Details 2: Tangent vectors for $\vec{\mathbf{r}} = x\vec{\imath} + y(x)\vec{\jmath}$ are drawn from $\vec{\mathbf{r}}\,' = \vec{\imath} + y'(x)\vec{\jmath} = \vec{\imath} + f(x, y)\vec{\jmath}$. Continuity of f implies that the vector $\vec{\mathbf{r}}\,'$ is to pixel resolution identical for all (x, y) sufficiently close to a grid point (x_0, y_0). This is why an edge-to-edge solution passes grid points with tangent vector *nearly* matching the tangent vector of nearby segments.

Details 3: Each segment is a tangent line $y = y_0 + m(x - x_0)$, constructed with slope $m = y'(x_0)$. It approximates the curve $y(x)$ local to the contact point (x_0, y_0). Graphically, a short tangent line coalesces with the solution curve near the contact point, to pixel resolution.

The tangent line approximation is called **Euler's approximation**. Correct pronunciation is **Oiler**. To make the audience giggle, pronounce it **Yuler**.

Rules for Drawing Threaded Solutions

A direction field graphic replaces all the information supplied by the equation $y' = f(x, y)$. The equation is tossed aside and not used.

Visualization of all solutions involves drawing a small number of edge-to-edge solutions $y(x)$ onto the direction field graph window. We will use just two abbreviated rules (see **1** and **2** in Table 4).

Table 4. Two Rules for Drawing Edge-to-Edge Solutions

Abbreviated Threading Rules

1. Solutions don't cross.
2. Nearby tangents nearly match.

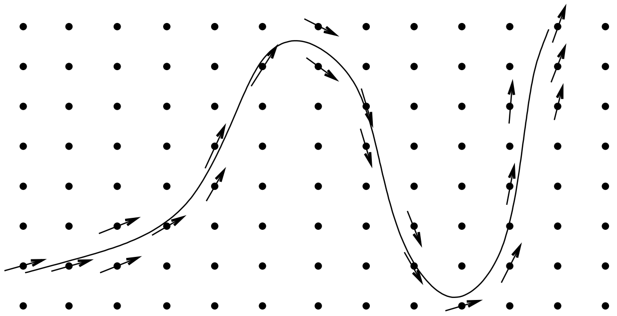


Figure 8. Threading a Solution Edge-to-Edge.

Shown in Figure 8 is a threaded solution curve for $y' = f(x, y)$ plus nearby grid points and relevant line segments (arrows). The solution threads its way through the direction field, matching tangents at nearby grid points. Arrows that touch a threaded curve must coalesce with the curve (solutions don't cross).

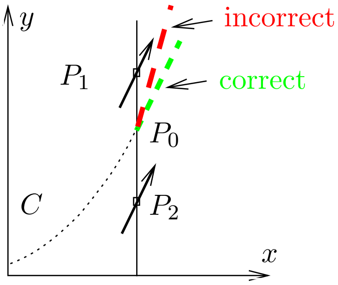


Figure 9. Threading Rules.
Solution curve C threads from the left edge and meets a line at point P_0. The line contains two nearby grid points P_1, P_2. The tangent at P_0 must nearly match direction field arrows at P_1, P_2.

Tangent Matching Explained.
The slopes of the tangents in Figure 9 are given by $y' = f(x, y)$. For points $P_1 = (x_1, y_1)$, $P_0 = (x_0, y_0)$ and $P_2 = (x_2, y_2)$, the slopes are $f(x_1, y_1)$, $f(x_0, y_0)$, $f(x_2, y_2)$. If the points P_0, P_1, P_2 are close, then continuity of f implies all three slopes are *nearly equal.*

How to Construct a Direction Field Graphic

Window Invent the graph x-range and y-range. $\boxed{1}$

Grid Plot a uniform grid of N grid points within the graph window. Invent N to populate the graphical white space, $N \approx 50$ for hand work. $\boxed{2}$

Field	Draw at each grid point (x^*, y^*) a short tangent vector $\epsilon\vec{T}$, where $\vec{T} = \vec{\imath} + f(x^*, y^*)\vec{\jmath}$. $\boxed{3}$
Threaded Solutions	Draw additional edge-to-edge threaded solutions into the remaining white space of the graphic. $\boxed{4}$

Construction Notes.

$\boxed{1}$ The window should include all significant equilibrium solutions, that is, solutions y = constant of $y' = f(x, y)$, which plot as horizontal lines. Physically interesting initial conditions $(x_0, y(x_0))$ should be added.

$\boxed{2}$ The isocline method might also be used to select grid points. For details on both methods, see the next subsection.

$\boxed{3}$ The arrow shaft is a **replacement curve** for the solution of $y' = f(x, y)$ through grid point (x^*, y^*) on a small x-interval, called a **lineal element**.

$\boxed{4}$ Threading is educated guesswork, discussed above, in Figures 8 and 9. If possible, choose (x_0, y_0) on the left window edge, then thread the solution until it exits the window top, bottom or right.

Direction fields are used *infra* in phase portraits of two-dimensional systems of differential equations.

Two Methods for Selecting Grid Points

There are two standard methods for selecting grid points, called the **uniform grid method** and the **isocline grid method**. The methods may be combined in some applications.

Uniform Grid

Two positive increment parameters n and m are supplied along with a graph window $a \le x \le b$, $c \le y \le d$. Hand work usually starts with $n = m = 11$; computer software starts with $n = m = 21$.

The nm grid points are defined for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ by the equations $x_i = a + (b - a)(i - 1)/(n - 1)$, $y_j = c + (d - c)(j - 1)/(m - 1)$.

Isocline Grid

A graph window $a \le x \le b$, $c \le y \le d$ is given plus a list of invented slopes $M_1, \ldots, M_p$ for the lineal elements.

To define the grid points, select the number $n > 0$ of grid points to be drawn on each isocline. Construct n equally-spaced horizontal lines (or vertical lines). Define grid points as intersections of the lines with all the implicit curves $f(x, y) = M_\ell$, $\ell = 1, \ldots, p$.

Along the implicit curve $f(x, y) = M_\ell$, within the graph window, mark each grid point and draw a lineal element, each element of exactly the same slope M_ℓ, for $\ell = 1, \ldots, p$.

The two methods are applied in Examples 1.27 and 1.28, page 45. Illustrated for the isocline method are possibilities such as graph window clipping and fine-tuning of the slopes to allow the grid points to fill the window.

Grid points in the *isocline method* are intersections of equally-spaced lines with the implicit curves $f(x, y) = M_\ell$. Lineal elements sketched along this curve all have slope M_ℓ and therefore they can be drawn with reduced effort.

How to Make Lineal Elements

A lineal element is a line segment centered at a grid point. They should not touch, because they represent, to pixel resolution, non-crossing solution curves on a short x-interval. Choose H to be not greater than the minimum distance between pairs of grid points. Initially, one can guess the value of H, then adjust the value after seeing the result. Define

$$h = \frac{H}{2\sqrt{1 + |f(x_0, y_0)|^2}}.$$

Then a lineal element of length H is defined by the midpoint (x_0, y_0) and the two endpoints $(x_0 - h, y_0 - hM)$ and $(x_0 + h, y_0 + hM)$, where $M = f(x_0, y_0)$.

This choice insures lineal elements do not touch. It is possible to erase the line segment to the left or right of the grid point without losing much information. Arrow heads can be added to show the tangent direction.

Examples

Example 1.25 (Window and Grid)

Choose a graph window for the differential equation $y' = y^2(2 - y)(1 + y)$ which includes the equilibrium solutions. Draw a 5×5 uniform grid on the graph window and plot the equilibrium solutions. Do not draw the direction field nor threaded solutions.

Solution: Let $f(x, y) = y^2(2-y)(1+y)$. Then $y = k$ is a constant solution of $y' = f(x, y)$ exactly when $0 = k^2(2 - k)(1 + k)$. The values $k = -1, 0, 2$ give horizontal lines $y = -1$, $y = 0$, $y = 2$. These lines are called equilibrium solutions; they are constant solutions of the differential equation. Accordingly, a graph window containing the equilibria is $-3 \leq x \leq 3$, $-2 \leq y \leq 3$. The 25 grid points are obtained by the formulas $x_k = -3 + 6(k/4)$, $k = 0, \ldots, 4$ and $y_j = -2 + 5(j/4)$, $j = 0, \ldots, 4$. The plot is done by hand. A computer plot appears in Figure 10.

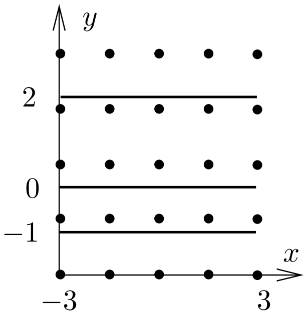


Figure 10. A graph window with uniform grid and equilibria.
Three equilibrium solutions $y = -1$, $y = 0$, $y = 2$ appear plus 25 grid points on the graph window $|x| \le 3$, $-2 \le y \le 3$.

Example 1.26 (Threading a Solution)

Starting at the black dots in the direction field graphic of Figure 11, thread three solution curves.

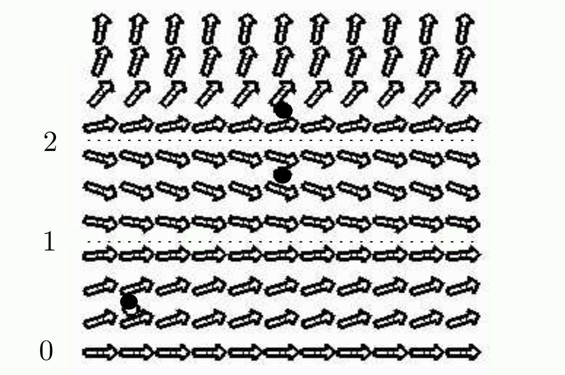


Figure 11. A direction field.
A field for the differential equation $y' = y(2-y)(1-y)$ is plotted on graph window $0 \le x \le 3$, $0 \le y \le 3$. The black dots are at $(0.25, 0.4)$, $(1.5, 2.25)$ and $(1.5, 1.65)$.

Solution: A plot appears in Figure 12.

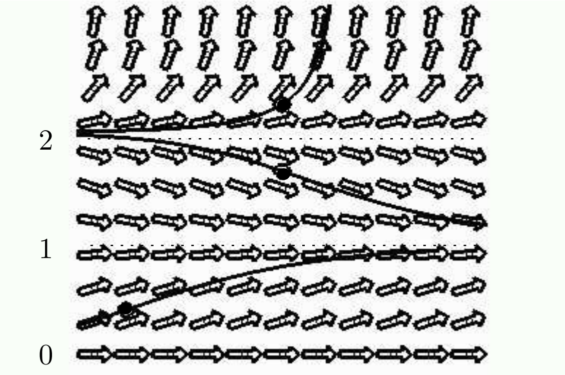


Figure 12. Threaded solutions.
The graph window is $0 \le x \le 3$, $0 \le y \le 3$. Threaded curves cannot cross equilibrium solutions $y = 0$, $y = 1$ and $y = 2$.

A threaded solution matches its tangents with nearby lineal elements of the direction field in Figure 11; see page 41 for an explanation. Each threaded curve represents a solution of the differential equation through the given dot on the entire interval $0 \le x \le 3$, whereas the lineal elements represent solutions through the grid point on a very short x-interval.

Example 1.27 (Uniform Grid Method)

Make a direction field of 11×11 points for $y' = x + y(1-y)$ on $-1 \le x \le 1$, $-2 \le y \le 2$.

Solution: Let $f(x,y) = x + y(1-y)$. The 121 grid points are the pairs (x,y) where $x = -1$ to 1 in increments of 0.2 and $y = -2$ to 2 in increments of 0.4. The minimum distance between grid points is $H = 0.2$.

We will generate the endpoints of the lineal element at $x_0 = -0.4$, $y_0 = 1.6$. It will be shown that the first endpoint is $(-0.34076096, 1.5194349)$. This point can be located from (x_0, y_0) by traveling distance $H/2$ at slope $M = -1.36$.

$M = f(x_0, y_0)$	The line segment slope for Euler's rule.
$= x_0 + y_0(1 - y_0)$	Apply $f(x,y) = x + y(1-y)$.
$= -1.36,$	Use the first point $x_0 = -0.4$, $y_0 = 1.6$.
$h = \dfrac{H}{2\sqrt{1+M^2}}$	Apply the formula $h = (H/2)/\sqrt{1 + f(x_0,y_0)^2}$.
$= 0.059239045,$	Use $H = 0.2$ and $f(x_0, y_0) = M = -1.36$.
$X = x_0 + h$	Compute the x-coordinate of the second point.
$= -0.34076096$	Use $x_0 = -0.4$ and $h = 0.059239045$.
$Y = y_0 + hf(x_0, y_0)$	Compute the y-coordinate of the second point.
$= 1.5194349$	Use values $y_0 = 1.6$, $f(x_0, y_0) = M = -1.36$, $h = 0.059239045$.

The second endpoint $(-0.459239045, 1.6805651)$, at distance $H/2$ from the grid point, in the opposite direction, can be found by minor changes to the above calculation. Automation of this process is necessary because 121 such calculations are required. Some basic `maple` code appears below which computes the 121 pairs of points for the direction field, then plots a replica of the field. The graphic appears in Figure 13. The code adapts to numerical laboratories like `matlab`, `octave` and `scilab`, which may or may not have a suitable direction field library, depending on the version.

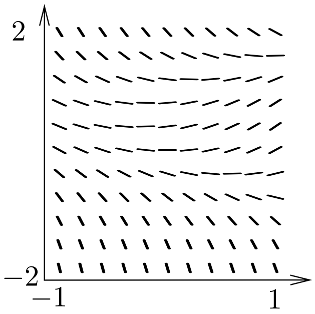


Figure 13. Direction field for the equation $y' = x + y(1 - y)$.
The uniform grid method is used on graph window $-1 \le x \le 1$, $-2 \le y \le 2$. There are 121 grid points.

```
a:=-1:b:=1:c:=-2:d:=2:n:=11:m:=11:
H:=(b-a)/(n-1):K:=(d-c)/(m-1):HH:=0.15:
f:=(x,y)->x+y*(1-y): X:=t->a+H*(t-1):Y:=t->c+K*(t-1):P:=[]:
for i from 1 to n do for j from 1 to m do
  x0:=X(i):y0:=Y(j):M:=evalf(f(x0,y0)):
  h:=evalf((HH/2)/sqrt(1+M^2)):
  Seg:=[[x0-h,y0-h*M],[x0+h,y0+h*M]]:
  if (P = []) then P:=Seg: next: fi: P:=P,Seg:
od:od:
opts:=scaling=constrained,color=black,thickness=3,axes=boxed;
plot([P],opts);
```

Versions of `maple` since V 5.1 have a `DEtools` package which simplifies the process of making a direction field. In `mathematica`, a similar command exists.

```
with(DEtools):  de:=diff(y(x),x)=x+y(x)*(1-y(x)):  # Maple
opts:=arrows=LINE,dirgrid=[11,11];
DEplot(de,y(x),x=-1..1,y=-2..2,opts);
```

```
<< Graphics\PlotField.m # Mathematica
PlotVectorField[1,x+y (1-y),x,-1,1,y,-2,2]
```

Resources for computer-assisted direction fields with interactive threaded solutions include `Maple, Mathematica, Matlab`. Each system has a steep learning curve, the time investment well worth the effort expended.

Example 1.28 (Isocline Method)

Make a direction field by hand using the isocline method for the differential equation $y' = x + y(1-y)$ on $-1 \le x \le 1$, $-1 \le y \le 2$.

Solution: Let $f(x,y) = x + y(1-y)$ and let M denote the slope of a replacement lineal element. The isoclines are defined by $f(x,y) = M$. It has the standard equation $(y-1/2)^2 = x - M + 1/4$, which is a parabola with center $(M - 1/4, 1/2)$ opening to the right. The algebra details:

$x + y(1-y) = M$	The equation $f(x,y) = M$ expanded.
$y^2 - y = x - M$	Multiply by -1 and move $-x$ to the right side.
$y^2 - y + \frac{1}{4} = x - M + \frac{1}{4}$	Apply *square completion*: add the square of half the coefficient of y to both sides.
$(y - \frac{1}{2})^2 = x - M + \frac{1}{4}$	Write the left side as a perfect square. It has the form of the standard curve library equation $Y^2 = X$. See Appendix A.4.

The basic requirement for slope M selection is that the set of grid points obtained below fills the white space of the graph window. Briefly, some portion of each parabola has to intersect the graph window. By experiment, the slopes M to be used in the isocline method will be selected as $M = 1/4 + (-3)$ to $M = 1/4 + (1)$ in increments of 0.2 to identify 21 isoclines:

Isocline Equation	**Slope** M
$(y-1/2)^2 = x - (-3)$	$0.25 + (-3)$
$(y-1/2)^2 = x - (-2.8)$	$0.25 + (-3) + 0.2$
$\vdots$	$\vdots$
$(y-1/2)^2 = x - (1)$	$0.25 + (-3) + 4.0$

To define the grid points, let $y = -1$ to 2 in increments of 0.3 to make 11 horizontal lines. The intersections of these 11 lines with the 21 parabolas define at least 100 grid points inside the graph window. It is possible to graph rapidly the 21 parabolas, because they are translates of the standard parabola $Y^2 = X$.

The replacement lineal elements on each parabola are sketched rapidly as follows. Using pencil and paper, graph accurately the first lineal element on the isocline curve, using associated slope M. Rotate the paper until the lineal element is vertical. Draw additional lineal elements at the remaining grid points of the isocline curve as vertical lines. Accuracy improves with the use of a drawing easel, T-square and triangle.

A computer graphic is shown in Figure 14 which closely resembles a hand-made graphic. Compare it to the uniform grid method graphic in Figure 13, page 46.

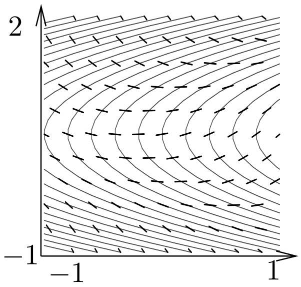


Figure 14. Direction field for the differential equation $y' = x + y(1 - y)$.
The isocline method is applied using graph window $-1 \le x \le 1$, $-1 \le y \le 2$. Parabolas are isoclines. Grid points are intersections of direction field lineal elements with isoclines. Lineal elements are computed along equally-spaced horizontal lines.

The `maple` code that produced Figure 14 is included below to show machine equivalents of a hand computation. The ordering of the code: the lineal elements are drawn in `Plot1`, then the isocline curves are drawn in `Plot2`. Then two graphics are superimposed. A key detail is solving $f(x, y_0) = M$ for $x = x_0$ to locate a grid point (x_0, y_0) and then construct the lineal element. Factor H is adjusted to keep lineal elements from touching.

```
with(plots):
getGridSegments:=proc(c,d,n,m,H,slopes,f)
local M,Y0,j,k,h,x0,y0,Seg,P;
Y0:=unapply(c+(d-c)*(t-1)/(m-1),t): P:=[]:
for j from 1 to n do
M:=slopes[j]:h:=evalf(H*0.5/sqrt(1+M^2)):
for k from 1 to m do # loop on m horiz lines
y0:=Y0(k): x0:=solve(f(x,y0)=M,x);# (x0,y0)=GridPoint
Seg:=[[x0-h,y0-h*M],[x0+h,y0+h*M]]:# lineal element
if P=[] then P:=Seg: next: fi:
P:=P,Seg:od:od: return P; end proc:

a:=-1:b:=1:c:=-1:d:=2:m:=11:
H:=0.1:f:=(x,y)->x+y*(1-y):
opts:=color=BLACK,thickness=4,axes=none,scaling=constrained:
Window:=x=a..b,y=c..d: n:=21;
slopes:=[seq(-3+4*(t-1)/(n-1),t=1..n)];# guesswork
P:=getGridSegments(c,d,n,m,H,slopes,f);
Plot1:=plot([P],Window,opts):
eqs:=[seq(f(x,y)=slopes[j],j=1..n)]:
Plot2:=implicitplot(eqs,Window):
display([Plot1,Plot2]);
```

Exercises 1.4

Window and Grid

Find the equilibrium solutions, then determine a graph window which includes them and construct a 5 × 5 uniform grid. Follow Example 1.25.

1. $y' = 2y$

2. $y' = 3y$

3. $y' = 2y + 2$

4. $y' = 3y - 2$

5. $y' = y(1 - y)$

6. $y' = 2y(3 - y)$

7. $y' = y(1 - y)(2 - y)$

8. $y' = 2y(1 - y)(1 + y)$

9. $y' = 2(y - 1)(y + 1)^2$

10. $y' = 2y^2(y-1)^2$

11. $y' = (x+1)(y+1)(y-1)y$

12. $y' = 2(x+1)y^2(y+1)(y-1)^2$

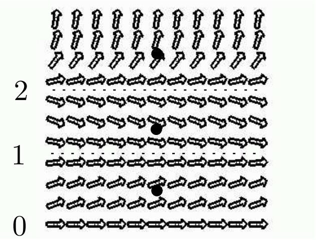
13. $y' = (x+2)y(y-3)(y+2)$

14. $y' = (x+1)y(y-2)(y+3)$

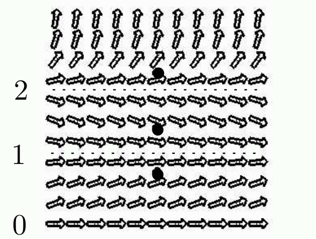
Threading Solutions

Each direction field below has window $0 \le x \le 3$, $0 \le y \le 3$. Start each threaded solution at a black dot and continue it left and right across the field. Dotted horizontal lines are equilibrium solutions. See Example 1.26.

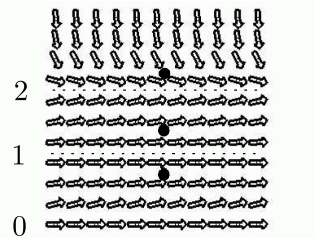
15.



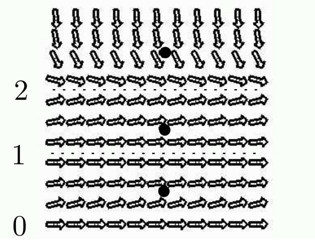
16.



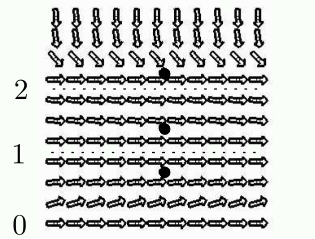
17.



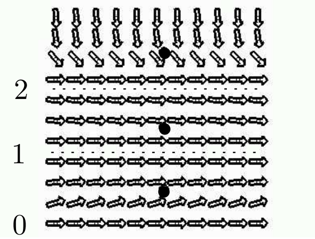
18.



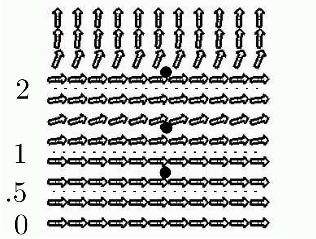
19.



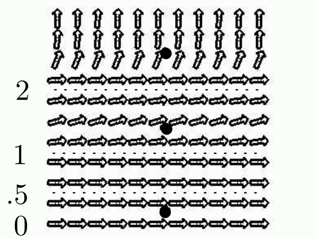
20.



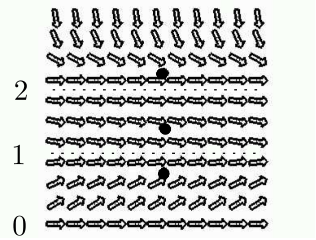
21.



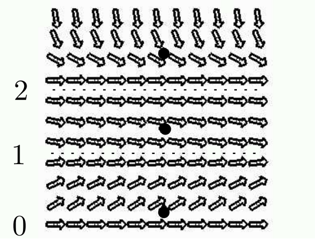
22.



23.



24.



Uniform Grid Method

Apply the uniform grid method as in Example 1.27, page 45 to make a direction field of 11×11 grid points for the given differential equation on $-1 \le x \le 1$, $-2 \le y \le 2$. If using a computer program, then use about 20×20 grid points.

25. $y' = 2y$

26. $y' = 3y$

27. $y' = 1 + y$

28. $y' = 2 + 3y$

29. $y' = x + y(2 - y)$

30. $y' = x + y(1 - 2y)$

31. $y' = 1 + y(2 - y)$

32. $y' = 1 + 2y(2 - y)$

33. $y' = x - y$

34. $y' = x + y$

35. $y' = y - \sin(x)$

36. $y' = y + \sin(x)$

Isocline Method

Apply the isocline method as in Example 1.28, page 47 to make a direction field of about 11×11 points for the given differential equation on $0 \le x \le 1$, $0 \le y \le 2$. Computer programs are used on these kinds of problems to find grid points as intersections of isoclines and horizontal lines. Graphics are expected to be done by hand. Extra isoclines can fill large white spaces.

37. $y' = x - y^2$

38. $y' = 2x - y^2$

39. $y' = 2y/(x + 1)$

40. $y' = -y^2/(x + 1)^2$

41. $y' = \sin(x - y)$

42. $y' = \cos(x - y)$

43. $y' = xy$

44. $y' = x^2y$

45. $y' = xy + 2x$

46. $y' = x^2y + 2x^2$

1.5 Phase Line Diagrams

Technical publications may use special diagrams to display **qualitative information** about the equilibrium points of the differential equation

$$y'(x) = f(y(x)). \tag{1}$$

The right side of this equation is independent of x, hence there are no external control terms that depend on x. Due to the lack of external controls, the equation is said to be **self-governing** or **autonomous**.

Definition 1.8 (Phase Line Diagram)
A **phase line diagram** for the autonomous equation $y' = f(y)$ is a line segment with labels **sink**, **source** or **node** (definitions below), one mark and label for each root y of $f(y) = 0$, i.e., each equilibrium; see Figure 15.

The labels **sink**, **source**, **node** are borrowed from the theory of fluids and they have the following **special definitions**:[6]

Sink $y = y_0$ The equilibrium $y = y_0$ *attracts* nearby solutions at $x = \infty$: for some $H > 0$, $|y(0) - y_0| < H$ implies $|y(x) - y_0|$ decreases to 0 as $x \to \infty$.

Source $y = y_1$ The equilibrium $y = y_1$ *repels* nearby solutions at $x = \infty$: for some $H > 0$, $|y(0) - y_1| < H$ implies that $|y(x) - y_1|$ increases as $x \to \infty$.

Node $y = y_2$ The equilibrium $y = y_2$ is neither a sink nor a source.

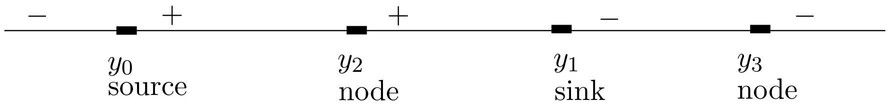


Figure 15. A phase line diagram along the y-axis.
A plus sign means $f(y) > 0$ for y between equilibria. A minus sign means $f(y) < 0$ for y between equilibria. A sign change minus to plus is a source y_0 , plus to minus is a sink y_1. No sign change, plus to plus or minus to minus is a node – y_2, y_3 are nodes.

Figure 15 shows that classifications **source, sink, node** (or **spout, funnel, node**) can be decided from the signs of $f(y)$ left and right of an equilibrium point.

Scalar function $f(y)$ must be one-signed on the y-interval between adjacent equilibrium points, because $f(y) = 0$ means y is an equilibrium point.

A phase line diagram summarizes the contents of a direction field and all equilibrium solutions. It is used to efficiently draw threaded curves across the graph

[6] It is for geometric intuition that the current text section requires monotonic behavior in the definition of a sink. In applied literature a sink is defined by $\lim_{x\to\infty} |y(x) - y_0| = 0$, an easy transition for most, although unnecessarily abstract. See page 55 for definitions of **attracting** and **repelling** equilibria.

window, producing a **phase portrait** for $y' = f(y)$. The drawing rules increase in number, however for the special equation $y' = f(y)$ neither grid points nor a direction field are used.

Drawing Phase Portraits

A phase line diagram is used to draw a **phase portrait** of threaded solutions and equilibrium solutions by using the three rules below, justified on page 54.

Three Drawing Rules for $y' = f(y)$

1. Equilibrium solutions are horizontal lines in the phase diagram.
2. Threaded solutions of $y' = f(y)$ don't cross. In particular, they don't cross equilibrium solutions.
3. A threaded non-equilibrium solution that starts at $x = 0$ at a point y_0 must be increasing if $f(y_0) > 0$, and decreasing if $f(y_0) < 0$.

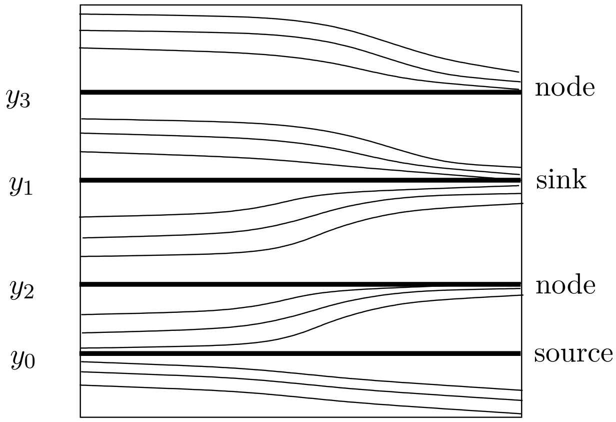


Figure 16. A phase portrait for an autonomous equation $y' = f(y)$.
The graphic is drawn directly from phase line diagram Figure 15, using rules **1**, **2**, **3**. While not a replica of an accurately constructed computer graphic, the general look of threaded solutions is sufficient for intuition. Labels **source**, **sink**, **node** are essential. Alternate labels: **spout**, **funnel**, **node**.

Table 5. Equilibria Classification by Signs of $f(y)$

Classification	Sign of $f(y)$ left	Sign of $f(y)$ right
Source [Spout]	MINUS	PLUS
Sink [Funnel]	PLUS	MINUS
Node	PLUS	PLUS
Node	MINUS	MINUS

Drain and Spout

In the theory of fluids, **source** means fluid is created and **sink** means fluid is lost. A memory device for these concepts is the kitchen water spout, which is the *source*, and the kitchen drain, which is the *sink*.

Figure 17. A source or a spout.
A water **spout** from a kitchen faucet or a spray-can is a **source**. Pencil traces in a figure represent flow lines in the fluid.

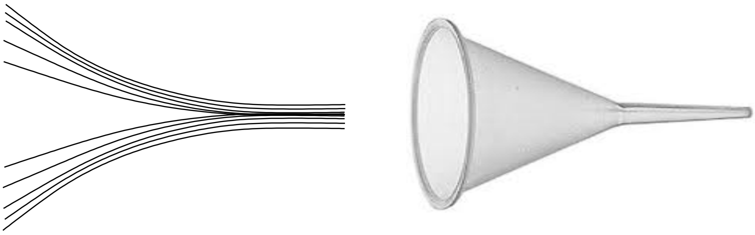


Figure 18. A sink or a funnel.
A **funnel** rotated 90 degrees has the shape of a **sink**. A drain in the kitchen sink has the same geometry. The lines drawn in a funnel figure can be visualized as traces of flow lines or dust particles in the fluid, going down the drain.

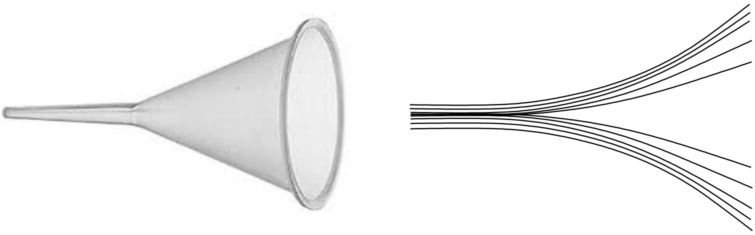


Figure 19. Video replay in reverse time.
A video of a funnel or sink played backwards looks like a source or spout.

Justification of the Three Drawing Rules:
Rule **1**: The curve $y =$ constant is a horizontal line.

Rule **2**: Two solutions $y_1(x)$, $y_2(x)$ that touch at $x = x_0$, $y = y_0$ must coalesce: both solutions satisfy $y' = f(t)$, $y(x_0) = y_0$, then Picard's theorem says $y_1(x) = y_2(x)$ for small $|x - x_0|$. The Picard-Lindelöf theorem hypotheses are met by examples herein and by the bulk of applied problems.

Rule **3**: let $y_1(x)$ be a solution with $y_1'(x) = f(y_1(x))$ either positive or negative at $x = 0$. If $y_1'(x_1) = 0$ for some $x_1 > 0$, then let $c = y_1(x_1)$ and define equilibrium solution $y_2(x) = c$. Then solution y_1 crosses the equilibrium solution y_2 at $x = x_1$, violating rule **2**.

Stability Test

The terms **stable equilibrium** and **unstable equilibrium** refer to the predictable plots of nearby solutions. The term **stable** means that solutions that start near the equilibrium will stay nearby as $x \to \infty$. The term **unstable** means *not stable.* Therefore, a sink is stable and a source is unstable.

Definition 1.9 (Stable Equilibrium)
An equilibrium y_0 of $y' = f(y)$ is **stable** provided for given $\epsilon > 0$ there exists some $H > 0$ such that $|y(0) - y_0| < H$ implies solution $y(x)$ exists for $x \ge 0$ and $|y(x) - y_0| < \epsilon$ for all $x \ge 0$.

The solution $y = y(0)e^{kx}$ of the equation $y' = ky$ exists for $x \ge 0$. Properties of exponentials justify that the equilibrium $y = 0$ is a sink for $k < 0$, a source for $k > 0$ and just stable for $k = 0$.

Definition 1.10 (Attracting and Repelling Equilibria)
An equilibrium $y = y_0$ is **attracting** provided $\lim_{x\to\infty} y(x) = y_0$ for all initial data $y(0)$ with $0 < |y(0) - y_0| < h$ and $h > 0$ sufficiently small. An equilibrium $y = y_0$ is **repelling** provided $\lim_{x\to-\infty} y(x) = y_0$ for all initial data $y(0)$ with $0 < |y(0) - y_0| < h$ and $h > 0$ sufficiently small.

The **stability test** below in Theorem 1.3 is motivated by the vector calculus results $\mathbf{Div}(\mathbf{P}) < 0$ for a sink and $\mathbf{Div}(\mathbf{P}) > 0$ for a source, where $\mathbf{P}$ is the velocity field of the fluid and **Div** is divergence. Justification is postponed to page 60.

Theorem 1.3 (Stability and Instability Conditions)
Let f and f' be continuous. The equation $y' = f(y)$ has a *sink* at $y = y_0$ provided $f(y_0) = 0$ and $f'(y_0) < 0$. An equilibrium $y = y_1$ is a *source* provided $f(y_1) = 0$ and $f'(y_1) > 0$. There is *no test* when f' is zero at an equilibrium. The no-test case can sometimes be decided by an additional test:

(a) Equation $y' = f(y)$ has a *sink* at $y = y_0$ provided $f(y)$ changes sign from positive to negative at $y = y_0$.

(b) Equation $y' = f(y)$ has a *source* at $y = y_0$ provided $f(y)$ changes sign from negative to positive at $y = y_0$.

Phase Line Diagram for the Logistic Equation

The model logistic equation $y' = (1 - y)y$ is used to produce the phase line diagram in Figure 20. The logistic equation is discussed on page 6, in connection with the Malthusian population equation $y' = ky$. The letters S and U are used for a stable sink and an unstable source, while N is used for a node. Details are in Example 1.30, page 58.

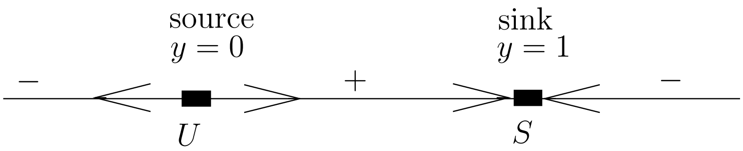


Figure 20. A phase line diagram for $y' = (1 - y)y$.
The equilibrium $y = 0$ is an unstable source (a spout) and equilibrium $y = 1$ is a stable sink (a funnel).

Arrowheads are used to display the **repelling** or **attracting** nature of the equilibrium.

Direction Field Plots for $y' = f(y)$

A direction field for an autonomous differential equation $y' = f(y)$ can be constructed in two steps.

Step 1. Draw grid points and line segments along the y-axis.

Step 2. Duplicate the y-axis direction field at even divisions along the x-axis.

Duplication is justified because $y' = f(y)$ does not depend on x, which means that the slope assigned to a line segment at grid points $(0, y_0)$ and (x_0, y_0) are identical.

The following facts are assembled for reference:

Fact 1. An equilibrium is a horizontal line. It is *stable* if all solutions starting near the line remain nearby as $x \to \infty$.

Fact 2. Solutions don't cross. In particular, any solution that starts above or below an equilibrium solution must remain above or below.

Fact 3. A solution curve of $y' = f(y)$ rigidly moved to the left or right will remain a solution, i.e., the translate $y(x - x_0)$ of a solution to $y' = f(y)$ is also a solution.

A phase line diagram is merely a summary of the solution behavior in a direction field. Conversely, an independently made phase line diagram can be used to enrich the detail in a direction field.

Fact 3 will create additional threaded solutions from an initial threaded solution by translation. Threaded solutions with turning points will have translations with turning points marching monotonically to the left, or to the right.

Bifurcation Diagrams

The phase line diagram has a close relative called a **bifurcation diagram**. The purpose of the diagram is to display qualitative information about equilibria, across all equations $y' = f(y)$, obtained by varying physical parameters appearing implicitly in f. In the simplest cases, each parameter change to $f(y)$ produces one phase line diagram and the two-dimensional stack of these phase line diagrams is the bifurcation diagram (see Figure 21).

Fish Harvesting

To understand the reason for such diagrams, consider a private lake with fish population $y(t)$. The population is harvested at rate k fish per year. A suitable sample logistic model is

$$\frac{dy}{dt} = y(4 - y) - k$$

where the constant harvesting rate k is allowed to change. Given some relevant values of k, a field biologist would produce corresponding phase line diagrams, then display them by vertical stacking to obtain a two-dimensional diagram like Figure 21.

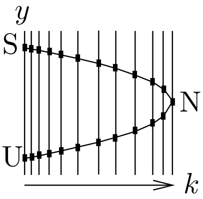


Figure 21. A bifurcation diagram.
Legend: U=Unstable, S=Stable, N=node.
The fish harvesting diagram consists of stacked phase-line diagrams.

In Figure 21, the vertical axis represents initial values $y(0)$ and the horizontal axis represents the harvesting rate k. Each phase line diagram has two equilibria, one stable and one unstable, except the rightmost diagram, which has exactly one equilibrium.

The bifurcation diagram shows how the number of equilibria and their classifications *sink*, *source* and *node* change with the harvesting rate.

Shortcut methods exist for drawing bifurcation diagrams and these methods have led to succinct diagrams that remove the phase line diagram detail. The basic idea is to eliminate the vertical lines in the plot, and replace the equilibria **dots** by a curve, essentially obtained by **connect-the-dots**. In current literature, Figure 21 is ofteb replaced by the more succinct Figure 22.

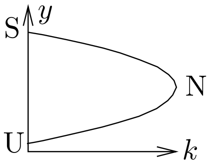


Figure 22. A succinct bifurcation diagram for fish harvesting.
The vertical axis y represents initial population and the horizontal axis k is the harvesting rate.
Legend: U=Unstable, S=Stable, N=node.

Stability and Bifurcation Points

Biologists call a fish population *stable* when the fish reproduce at a rate that keeps up with harvesting. Bifurcation diagrams show how to stock the lake and harvest it in order to have a stable fish population.

A point $N = (k_0, y_0)$ in a bifurcation diagram is called a **bifurcation point** provided small local changes to k result in a sudden change in qualitative behavior. In Figure 22, the sudden change in qualitative behavior is from one unstable equilibrium to two equilibria, one stable and one unstable. Some facts about Figure 22:

[**1**] The **carrying capacity** M for harvesting rate k is found from a point (k, M) on the upper curve. Symbol M is the largest population size for a **stable fish population**.

[**2**] The **minimum stocking size** m for harvesting rate k is found from a point (k, m) on the lower curve .

[**3**] **Extinction** results for harvesting rates $k > k_0$. Extinction means all solutions limit to zero at $t = \infty$.

[**4**] **Extinction** results for harvesting rates k and initial population y with (k, y) in the region below the lower curve.

Some combinations are obvious, e.g., a harvest of 2 thousand per year from an equilibrium population of about 4 thousand fish. Less obvious is a **sustainable harvest** of about 4 thousand fish with an equilibrium population of about 2 thousand fish, detected from the portion of the curve near bifurcation point N.

Examples

Example 1.29 (No Test in Sink–Source Theorem 1.3)

Find an example $y' = f(y)$ which has an unstable node at $y = 0$ and no other equilibria.

Solution: Let $f(y) = y^2$. The equation $y' = f(y)$ has an equilibrium at $y = 0$. In Theorem 1.3, there is a *no test* condition $f'(0) = 0$.

A computer algebra system can determine $y = 1/(1/y(0) - x)$:

```
dsolve(diff(y(x),x)=y(x)^2,y(x));   maple
ode2('diff(y,x) = y^2,y,x);         Maxima
```

Solutions with $y(0) < 0$ limit to the equilibrium solution $y = 0$, but positive solutions "blow up" before $x = \infty$ at $x = 1/y(0)$. The equilibrium $y = 0$ is an unstable node, that is, it is not a source nor a sink.

The same conclusions are obtained from basic calculus, without solving the differential equation. The reasoning: y' has the sign of y^2, then $y' \geq 0$ implies $y(x)$ increases. The equilibrium $y = 0$ behaves like a source when $y(0) > 0$. For $y(0) < 0$, again $y(x)$ increases, but in this case the equilibrium $y = 0$ behaves like a sink. Accordingly, $y = 0$ is not a source nor a sink, but a node.

Example 1.30 (Phase Line Diagram)

Verify the phase line diagram in Figure 23 for the logistic equation $y' = (1 - y)y$, using Theorem 1.3.

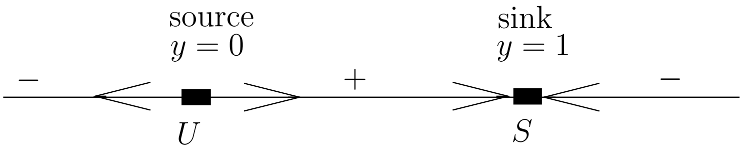


Figure 23. Phase line diagram for $y' = (1 - y)y$.

Solution: Let $f(y) = (1 - y)y$. To justify Figure 23, there are three steps:

1. Find the equilibria. Answer: $y = 0$ and $y = 1$.
2. Find the signs PLUS and MINUS.
3. Apply Theorem 1.3 to show $y = 0$ is a source and $y = 1$ is a sink.

The plan is to first compute the equilibrium points.

$(1 - y)y = 0$	Solving $f(y) = 0$ for equilibria.
$y = 0,\ y = 1$	Roots found.

The signs $\boxed{+}$ and $\boxed{-}$ appearing in Figure 20 are labels that mean f is positive or negative on the interval between adjacent equilibria.

A sign of plus or minus is determined by the sign of $f(x)$ for x between equilibria. To justify this statement, suppose both signs occur, $f(x_1) > 0$ and $f(x_2) < 0$. Then continuity of f implies $f(x) = 0$ for a point x between x_1, x_2, which is impossible on an interval free of roots.

The method to determine the signs, plus or minus, then reduces to evaluation of $f(x)$ for an invented sample x chosen between two equilibria, for instance:

$f(-1) = (y - y^2)\big|_{x=-1} = -2$ The sign is MINUS. Chosen was $x = -1$, which is in the interval $-\infty < x < 0$.

$f(0.5) = (y - y^2)\big|_{x=0.5} = 0.25$ The sign is PLUS. Chosen was $x = 0.5$, which is in the interval $0 < x < 1$.

$f(2) = (y - y^2)\big|_{x=2} = -2$ The sign is MINUS. Chosen was $x = 2$, which is in the interval $1 < x < \infty$.

We will apply Theorem 1.3. The plan is to find $f'(y)$ and then evaluate f' at each equilibrium. An alternative technique is to apply Theorem 1.3, part **(a)** or **(b)**, which is the method of choice in practise.

$f'(y) = (y - y^2)'$ Find f' from $f(y) = (1 - y)y$.

$= 1 - 2y$ Derivative $f'(y) = \frac{df}{dy}$ found.

$f'(0) = 1$ Positive means it is a *source* (*spout*), by Theorem 1.3.

$f'(1) = -1$ Negative means it is a *sink* (*funnel*), by Theorem 1.3.

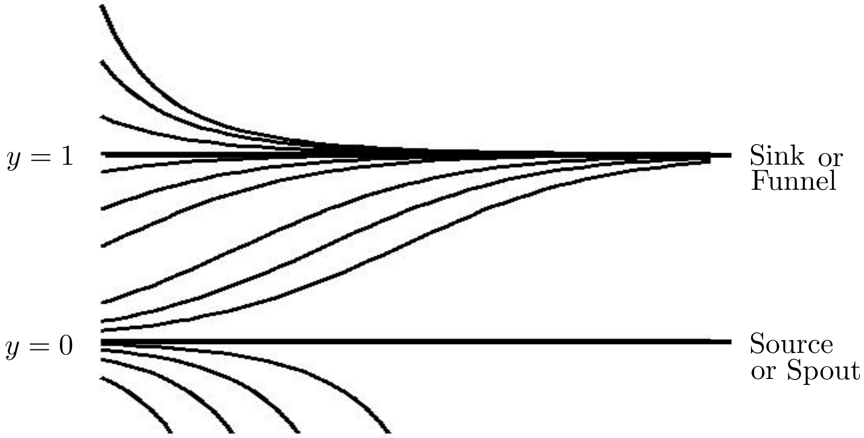


Figure 24. Phase portrait for $y' = (1 - y)y$.
Drawn from the phase line diagram of Example 1.30.

Example 1.31 (Phase Portrait)

Justify the phase portrait in Figure 24 for the logistic equation $y' = (1 - y)y$, using the phase line diagram constructed in Example 1.30.

Solution:

Drawing rules. The phase line diagram contains all essential information for drawing threaded curves. Threaded solutions have to be either horizontal (an equilibrium solution), increasing or decreasing. Optional is representation of turning points.

Translations. Because translates of solutions are also solutions and solutions are unique, then the drawing of an increasing or decreasing threaded curve determines the shape of all nearby threaded curves. There is **no option** for drawing nearby curves!

Explanation. The phase portrait is drawn by moving the phase line diagram to the y-axis of the graph window $0 \le x \le 6$, $-0.5 \le y \le 2$. The graph window was selected by first including the equilibrium solutions $y = 0$ and $y = 1$, then growing the window after an initial graph. Each equilibrium solution produces a horizontal line, i.e., lines $y = 0$ and $y = 1$. The signs copied to the y-axis from the phase line diagram tell us how to draw a threaded curve, either increasing (PLUS) or decreasing (MINUS).

Labels. It is customary to use labels **sink**, **source**, **node** or the alternates **spout**, **funnel**, **node**. Additional labels are **Stable** and **Unstable**. The only stable geometry is a **sink** (**funnel**).

Example 1.32 (Bifurcation Diagram)

Verify the fish harvesting bifurcation diagram in Figure 21.

Solution: Let $f(y) = y(4 - y) - k$, where k is a parameter that controls the harvesting rate per annum. A phase line diagram is made for each relevant value of k, by applying Theorem 1.3 to the equilibrium points. First, the *equilibria* are computed, that is, the roots of $f(y) = 0$:

$$\begin{aligned}
&y^2 - 4y + k = 0 && \text{Standard quadratic form of } f(y) = 0.\\
&y = \frac{4 \pm \sqrt{4^2 - 4k}}{2} && \text{Apply the quadratic formula.}\\
&\quad = 2 + \sqrt{4-k},\ \ 2 - \sqrt{4-k} && \text{Evaluate. Real roots exist only for } 4 - k \ge 0.
\end{aligned}$$

In preparation to apply Theorem 1.3, the derivative f' is calculated and then evaluated at the equilibria:

$$\begin{aligned}
&f'(y) = (4y - y^2 - k)' && \text{Computing } f' \text{ from } f(y) = (4-y)y - k.\\
&\qquad\ = 4 - 2y && \text{Derivative found.}\\
&f'(2 + \sqrt{4-k}) = -2\sqrt{4-k} && \text{Negative means a \textit{sink}, by Theorem 1.3.}\\
&f'(2 - \sqrt{4-k}) = 2\sqrt{4-k} && \text{Positive means a \textit{source}, by Theorem 1.3.}
\end{aligned}$$

A typical phase line diagram then looks like Figure 15, page 51. In the ky-plane, sources go through the curve $y = 2 - \sqrt{4-k}$ and sinks go through the curve $y = 2 + \sqrt{4-k}$. This justifies the bifurcation diagram in Figure 21, and also Figure 22, except for the common point of the two curves at $k = 4$, $y = 2$.

At this common point, the differential equation is $y' = -(y-2)^2$. This equation is studied in Example 1.29, page 58; a change of variable $Y = 2 - y$ shows that the equilibrium is a node.

Proofs and Details

Stability Test Proof: Let f and f' be continuous. It will be justified that the equation $y' = f(y)$ has a *stable* equilibrium at $y = y_0$, provided $f(y_0) = 0$ and $f'(y_0) < 0$. The *unstable* case is left for the exercises.

We show that f changes sign at $y = y_0$ from positive to negative, as follows, hence the hypotheses of **(a)** hold. Continuity of f' and the inequality $f'(y_0) < 0$ imply $f'(y) < 0$ on some small interval $|y - y_0| \le H$. Therefore, $f(y) > 0 = f(y_0)$ for $y < y_0$ and $f(y) < 0 = f(y_0)$ for $y > y_0$. This justifies that the hypotheses of **(a)** apply. We complete the proof using only these hypotheses.

Global existence. It has to be established that some constant $H > 0$ exists, such that $|y(0) - y_0| < H$ implies $y(x)$ exists for $x \ge 0$ and $\lim_{x\to\infty} y(x) = y_0$. To define $H > 0$, assume $f(y_0) = 0$ and the change of sign condition $f(y) > 0$ for $y_0 - H \le y < y_0$, $f(y) < 0$ for $y_0 < y \le y_0 + H$.

Assume that $y(x)$ exists as a solution to $y' = f(y)$ on $0 \le x \le h$. It will be established that $|y(0) - y_0| < H$ implies $y(x)$ is monotonic and satisfies $|y(x) - y_0| \le Hh$ for $0 \le x \le h$.

The constant solution y_0 cannot cross any other solution, therefore a solution with $y(0) > y_0$ satisfies $y(x) > y_0$ for all x. Similarly, $y(0) < y_0$ implies $y(x) < y_0$ for all x.

The equation $y' = f(y)$ dictates the sign of y', as long as $0 < |y(x) - y_0| \le H$. Then $y(x)$ is either decreasing $(y' < 0)$ or increasing $(y' > 0)$ towards y_0 on $0 \le x \le h$, hence $|y(x) - y_0| \le H$ holds as long as the monotonicity holds. Because the signs endure on $0 < x \le h$, then $|y(x) - y_0| \le H$ holds on $0 \le x \le h$.

Extension to $0 \le x < \infty$. Differential equations extension theory applied to $y' = f(y)$ says that a solution satisfying on its domain $|y(x)| \le |y_0| + H$ may be extended to $x \ge 0$. This dispenses with the technical difficulty of showing that the domain of $y(x)$ is $x \ge 0$. Unfortunately, details of proof for extension results require more mathematical background than is assumed for this text; see Birkhoff-Rota [BirkRota], which justifies the extension from the Picard theorem.

Limit at $x = \infty$. It remains to show that $\lim_{x\to\infty} y(x) = y_1$ and $y_1 = y_0$. The limit equality follows because y is monotonic. The proof concludes when $y_1 = y_0$ is established.

Already, $y = y_0$ is the only root of $f(y) = 0$ in $|y - y_0| \le H$. This follows from the change of sign condition in **(a)**. It suffices to show that $f(y_1) = 0$, because then $y_1 = y_0$ by uniqueness.

To verify $f(y_1) = 0$, apply the fundamental theorem of calculus with $y'(x)$ replaced by $f(y(x))$ to obtain the identity

$$y(n+1) - y(n) = \int_n^{n+1} f(y(x))dx.$$

The integral on the right limits as $n \to \infty$ to the constant $f(y_1)$, by the integral mean value theorem of calculus, because the integrand has limit $f(y_1)$ at $x = \infty$. On the left side, the difference $y(n+1) - y(n)$ limits to $y_1 - y_1 = 0$. Therefore, $0 = f(y_1)$.

The additional test stated in the theorem is the observation that internal to the proof we used only the change of sign of f at $y = y_0$, which was deduced from the sign of the derivative $f'(y_0)$. If $f'(y_0) = 0$, but the change of sign occurs, then the details of proof still apply. ■

Exercises 1.5

Stability-Instability Test

Find all equilibria for the given differential equation and then apply Theorem 1.3, page 55, to obtain a classification of each equilibrium as a **source**, **sink** or **node**. Do not draw a phase line diagram.

1. $P' = (2 - P)P$

2. $P' = (1 - P)(P - 1)$

3. $y' = y(2 - 3y)$

4. $y' = y(1 - 5y)$

5. $A' = A(A - 1)(A - 2)$

6. $A' = (A - 1)(A - 2)^2$

7. $w' = \dfrac{w(1 - w)}{1 + w^2}$

8. $w' = \dfrac{w(2 - w)}{1 + w^4}$

9. $v' = \dfrac{v(1 + v)}{4 + v^2}$

10. $v' = \dfrac{(1 - v)(1 + v)}{2 + v^2}$

Phase Line Diagram

Draw a phase line diagram, with detail similar to Figure 20.

11. $y' = y(2 - y)$

12. $y' = (y + 1)(1 - y)$

13. $y' = (y - 1)(y - 2)$

14. $y' = (y - 2)(y + 3)$

15. $y' = y(y - 2)(y - 1)$

16. $y' = y(2 - y)(y - 1)$

17. $y' = \dfrac{(y - 2)(y - 1)}{1 + y^2}$

18. $y' = \dfrac{(2 - y)(y - 1)}{1 + y^2}$

19. $y' = \dfrac{(y - 2)^2(y - 1)}{1 + y^2}$

20. $y' = \dfrac{(y - 2)(y - 1)^2}{1 + y^2}$

Phase Portrait

Draw a phase portrait of threaded curves, using the phase line diagram constructed in the previous ten exercises.

21. $y' = y(2 - y)$

22. $y' = (y + 1)(1 - y)$

23. $y' = (y - 1)(y - 2)$

24. $y' = (y - 2)(y + 3)$

25. $y' = y(y - 2)(y - 1)$

26. $y' = y(2 - y)(y - 1)$

27. $y' = \dfrac{(y - 2)(y - 1)}{1 + y^2}$

28. $y' = \dfrac{(2 - y)(y - 1)}{1 + y^2}$

29. $y' = \dfrac{(y - 2)^2(y - 1)}{1 + y^2}$

30. $y' = \dfrac{(y - 2)(y - 1)^2}{1 + y^2}$

Bifurcation Diagram

Draw a stack of phase line diagrams and construct from it a succinct bifurcation diagram with abscissa k and ordinate $y(0)$. Don't justify details at a bifurcation point.

31. $y' = (2 - y)y - k$

32. $y' = (3 - y)y - k$

33. $y' = (2 - y)(y - 1) - k$

34. $y' = (3 - y)(y - 2) - k$

35. $y' = y(0.5 - 0.001y) - k$

36. $y' = y(0.4 - 0.045y) - k$

Details and Proofs

Supply details for the following statements.

37. **(Stability Test)**
Verify **(b)** of Theorem 1.3, page 55, by altering the proof given in the text for **(a)**.

38. **(Stability Test)**
Verify **(b)** of Theorem 1.3, page 55, by means of the change of variable $x \to -x$.

39. (Autonomous Equations) Let $y' = f(y)$ have solution $y(x)$ on $a < x < b$. Then for any c, $a < c < b$, the function $z(x) = y(x+c)$ is a solution of $z' = f(z)$.

40. (Autonomous Equations) The method of isoclines can be applied to an autonomous equation $y' = f(y)$ by choosing equally spaced horizontal lines $y = c_i$, $i = 1, \ldots, k$. Along each horizontal line $y = c_i$ the slope is a constant $M_i = f(c_i)$, and this determines the set of invented slopes $\{M_i\}_{i=1}^{k}$ for the method of isoclines.

1.6 Computing and Existence

The initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{1}$$

is studied here from a computational viewpoint. Answered are some basic questions about practical and theoretical computation of solutions:

- Why can numerical methods fail in problem (1)?
- What hypotheses for (1) make it possible to use numerical methods?
- When does (1) have a symbolic solution, that is, a solution described by an xy-equation?

Three Key Examples

The range of unusual behavior of solutions to $y' = f(x, y)$, $y(x_0) = y_0$ can be illustrated by three examples.

(A) $y' = 3(y-1)^{2/3}$, $y(0) = 1$.
The right side $f(x, y)$ is continuous. It has two solutions $y = 1 + x^3$ and $y = 1$.

(B) $y' = \dfrac{2y}{x-1}$, $y(0) = 1$.
The right side $f(x, y)$ is discontinuous. It has infinitely many piecewise-defined solutions
$$y = \begin{cases} (x-1)^2 & x < 1, \\ c(x-1)^2 & x \ge 1. \end{cases}$$

(C) $y' = 1 + y^2$, $y(0) = 0$.
The right side $f(x, y)$ is differentiable. It has unique solution $y = \tan(x)$, but $y(x_0) = \infty$ at finite time $x_0 = \pi/2$.

Numerical method failure can be caused by multiple solutions to problem (1), e.g., examples (A) and (B), because a numerical method is going to compute just *one* answer; see Example 1.33, page 69. Multiple solutions are often signaled by discontinuity of either f or its partial derivative f_y. In (A), the right side $3(y-1)^{2/3}$ has an infinite partial at $y = 1$, while in (B), the right side $2y/(x-1)$ is infinite at $x = 1$.

Simple jump discontinuities, or **switches**, appear in modern applications of differential equations. Therefore, it is important to allow $f(x, y)$ to be discontinuous, in a limited way, but multiple solutions must be avoided, e.g., example (B). An important success story in electrical engineering is circuit theory with periodic and piecewise-defined inputs. See Example 1.34, page 70.

Discontinuities of f or f_y in problem (1) should raise questions about the applicability of numerical methods. Exactly why there is not a precise and foolproof test to predict failure of a numerical method remains to be explained.

Theoretical solutions exist for problem (1), if $f(x,y)$ is *continuous*. See Peano's theorem, page 68. This solution may blow up in a finite interval, e.g., $y = \tan(x)$ in example (C). See Example 1.35, page 71.

No symbolic closed-form solution formula exists as a result of the basic theory. In part, this dilemma is due to the possibility of multiple solutions, if f is only continuous, e.g., example (A). Picard's iteration provides assumptions to give a symbolic solution formula. However, Picard's formula is currently impractical for applied mathematics. Additional general assumptions do not seem to help. *There is in general no symbolic solution formula available for use in applied mathematics.*

Exactly one theoretical solution exists in problem (1), provided $f(x,y)$ and $f_y(x,y)$ are continuous; see the Picard–Lindelöf theorem, page 68. The situation with numerical methods improves dramatically: the most popular methods work on a computer.

Why Not "Put it on the computer?"

Typically, scientists and engineers rely upon computer algebra systems and numerical laboratories, e.g., `maple`, `mathematica` and `matlab`.

Computerization for differential equation models constantly improves, with the advent of computer algebra systems and ever-improving numerical methods. Indeed, neither an advanced degree in mathematics nor a wizard's hat is required to query these systems for a closed-form solution formula. Many cases are checked systematically in a few seconds.

Fail-safe mechanisms usually do not exist for applying modern software to the initial value problem

$$\frac{dy}{dx} = f(x, y(x)), \quad y(x_0) = y_0.$$

For instance, the initial value problem $y' = 3(y-1)^{2/3}$, $y(0) = 1$ entered into computer algebra system `maple` reports the *solution* $y = 1 + x^3$. But the obvious equilibrium solution $y = 1$ is unreported. The `maple` numeric solver silently accepts the same problem and solves to obtain the solution $y = 1$. To experience this, execute the `maple` code below.

```
de:=diff(y(x),x)=3*(y(x)-1)^(2/3):  ic:=y(0)=1:
dsolve({de,ic},y(x));                    # Symbolic sol
p:=dsolve({de,ic},y(x),numeric); p(1); # Numerical sol
```

There was a report improvement in `maple` version 10. In later versions, $y(x) = 1$ was reported for both. The inference for the maple user is that there is a unique solution, but the model has multiple solutions, making both reports incorrect. Computer algebra system `Maxima` has similar issues.

Numerical instability is typically not reported by computer software. To understand the difficulty, consider the differential equation

$$y' = y - 2e^{-x}, \quad y(0) = 1.$$

The symbolic solution is $y = e^{-x}$. Attempts to solve the equation numerically will inevitably compute the nearby solutions $y = ce^x + e^{-x}$, where c is small. As x grows, the numerical solution grows like e^x, and $|y| \to \infty$. For example, `maple` computes $y(30) \approx -72557$, but $e^{-30} \approx 0.94 \times 10^{-13}$. In reality, the solution $y = e^{-x}$ cannot be computed. The `maple` code:

```
de:=diff(y(x),x)=y(x)-2*exp(-x):  ic:=y(0)=1:
sol:=dsolve({de,ic},y(x),numeric):  sol(30);
```

Mathematical model formulation seems to be an essential skill which does not come in the colorfully decorated package from the software vendor. It is this creative skill that separates the practicing scientist from the person on the street who has enough money to buy a computer program.

Closed-Form Existence-Uniqueness Theory

The **closed-form existence-uniqueness theory** describes models

$$(2) \qquad y' = f(x, y), \quad y(x_0) = y_0$$

for which a closed-form solution is known, as an equation of some sort. The objective of the theory for first order differential equations is to obtain existence and uniqueness by exhibiting a solution formula. The mathematical literature which documents these models is too vast to catalog in a textbook. We discuss only the most popular models.

`Dsolve` Engine in `Maple`

The computer algebra system has an implementation for some specialized equations within the closed-form theory. Below are some of the equation types examined by `maple` for solving a differential equation using classification methods. Not everything tried by `maple` is listed, e.g., Lie symmetry methods, which are beyond the scope of this text.

Equation Type	Differential Equation
Quadrature	$y' = F(x)$
Linear	$y' + p(x)y = r(x)$
Separable	$y' = f(x)g(y)$
Abel	$y' = f_0(x) + f_1(x)y + f_2(x)y^2 + f_3(x)y^3$
Bernoulli	$y' + p(x)y = r(x)y^n$

Clairaut	$y = xy' + g(y')$
d'Alembert	$y = xf(y') + g(y')$
Chini	$y' = f(x)y^n - g(x)y + h(x)$
Homogeneous	$y' = f(y/x), \quad y' = y/x + g(x)f(y/x),$ $y' = (y/x)F(y/x^\alpha),$ $y' = F((a_1x + a_2y + a_3)/(a_4x + a_5y + a_6))$
Rational	$y' = P_1(x, y)/P_2(x, y)$
Ricatti	$y' = f(x)y^2 + g(x)y + h(x)$

Not every equation can be solved as written — restrictions are made on the parameters. Omitted from the above list are **power series methods** and differential equations with **piecewise-defined coefficients**. They are part of the closed-form theory using specialized representations of solutions.

Special Equation Preview

The program here is to catalog a short list of first order equations and their known solution formulas. The formulas establish existence of a solution to the given initial value problem. They preview what is possible; details and examples appear elsewhere in the text. The issue of uniqueness is often routinely settled, as a separate issue, by applying the Picard-Lindelöf theorem. See Theorem 1.5, page 68, *infra*.

First Order Linear $y' + p(x)y = r(x)$ $y(x_0) = y_0$	Let p and r be continuous on $a < x < b$. Choose any (x_0, y_0) with $a < x_0 < b$. Then $y = y_0 e^{-\int_{x_0}^x p(t)dt} + \int_{x_0}^x r(t)e^{\int_t^x p(s)ds}dt$.
First Order Separable $y' = F(x)G(y)$ $y(x_0) = y_0$	Let $F(x)$ and $G(y)$ be continuous on $a < x < b$, $c < y < d$. Assume $G(y) \neq 0$ on $c < y < d$. Choose any (x_0, y_0) with $a < x_0 < b$, $c < y_0 < d$. Then $W(Y) = \int_{y_0}^Y du/G(u)$ is invertible and $y(x) = W^{-1}\left(\int_{x_0}^x F(t)dt\right)$.
First Order Analytic $y' = a(x)y + b(x)$ $y(x_0) = y_0$	Assume a and b have power series expansions in $\lvert x - x_0\rvert < h$. Then the power series $y(x) = \sum_{n=0}^\infty y_n(x - x_0)^n$ is convergent in $\lvert x - x_0\rvert < h$ and the coefficients y_n are found from the recursion $n!y_n = \left(\frac{d}{dx}\right)^{n-1}(a(x)y(x) + b(x))\Big\rvert_{x=x_0}$.

General Existence-Uniqueness Theory

The **general existence-uniqueness theory** describes the features of a differential equation model which make it possible to compute both theoretical and numerical solutions.

Modelers who create differential equation models generally choose the differential equation based upon the intuition gained from the *closed-form theory* and the *general theory*. Sometimes, modelers are lucky enough to refine a model to some known equation with closed-form solution. Other times, they are glad for just a numerical solution to the problem. In any case, they want a model that is *tested* and *proven* in applications.

General Existence Theory in Applications

For scientists and engineers, the results can be recorded as the following statement:

> In applications it is usually enough to require $f(x, y)$ and $f_y(x, y)$ to be continuous. Then the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ is a well-tested model to which classical numerical methods apply.

The general theory results to be stated are due to **Peano** and to **Picard-Lindelöf**. The techniques of proof require advanced calculus, perhaps graduate real-variable theory as well.

Theorem 1.4 (Peano)
Let $f(x, y)$ be continuous in a domain D of the xy-plane and let (x_0, y_0) belong to the interior of D. Then there is a small $h > 0$ and a function $y(x)$ continuously differentiable on $|x - x_0| < h$ such that $(x, y(x))$ remains in D for $|x - x_0| < h$ and $y(x)$ is one solution (many more might exist) of the initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0.$$

Definition 1.11 (Picard Iteration)
Define the constant function $y_0(x) = y_0$ and then define by iteration

$$y_{n+1}(x) = y_0 + \int_{x_0}^{x} f(t, y_n(t))dt.$$

The sequence $y_0(x)$, $y_1(x)$, ... is called the **sequence of Picard iterates** for $y' = f(x, y)$, $y(x_0) = y_0$. See Example 1.36 for computational details.

Theorem 1.5 (Picard-Lindelöf)
Let $f(x, y)$ and $f_y(x, y)$ be continuous in a domain D of the xy-plane. Let (x_0, y_0) belong to the interior of D. Then there is a small $h > 0$ and a *unique* function $y(x)$ continuously differentiable on $|x - x_0| < h$ such that $(x, y(x))$ remains in D for $|x - x_0| < h$ and $y(x)$ solves

$$y' = f(x, y), \quad y(x_0) = y_0.$$

The equation

$$\lim_{n\to\infty} y_n(x) = y(x)$$

is satisfied for $|x - x_0| < h$ by the Picard iterates $\{y_n\}$.

The **Picard iteration** is the replacement for a closed-form solution formula in the general theory, because the Picard-Lindelöf theorem gives the uniformly convergent infinite series *solution*

$$y(x) = y_0 + \sum_{n=0}^{\infty} \left(y_{n+1}(x) - y_n(x)\right). \tag{3}$$

Well, yes, it is a solution formula, but from examples it is seen to be currently impractical. There is as yet no known practical solution formula for the general theory.

The condition that f_y be continuous can be relaxed slightly, to include such examples as $y' = |y|$. The replacement is the following.

Definition 1.12 (Lipschitz Condition)

Let $M > 0$ be a constant and f a function defined in a domain D of the xy-plane. A **Lipschitz condition** is the inequality $|f(x, y_1) - f(x, y_2)| \le M|y_1 - y_2|$, assumed to hold for all (x, y_1) and (x, y_2) in D. The most common way to satisfy this condition is to require the partial derivative $f_y(x, y)$ to be continuous (Exercises page 73). A key example is $f(x, y) = |y|$, which is non-differentiable at $y = 0$, but satisfies a Lipschitz condition with $M = 1$.

Theorem 1.6 (Extended Picard-Lindelöf)

Let $f(x, y)$ be continuous and satisfy a Lipschitz condition in a domain D of the xy-plane. Let (x_0, y_0) belong to the interior of D. Then there is a small $h > 0$ and a *unique* function $y(x)$ continuously differentiable on $|x - x_0| < h$ such that $(x, y(x))$ remains in D for $|x - x_0| < h$ and $y(x)$ solves

$$y' = f(x, y), \quad y(x_0) = y_0.$$

The equation

$$\lim_{n\to\infty} y_n(x) = y(x)$$

is satisfied for $|x - x_0| < h$ by the Picard iterates $\{y_n\}$.

Example 1.33 (Numerical Method Failure)

A project models $y = 1 + x^3$ as the solution of the problem $y' = 3(y-1)^{2/3}$, $y(0) = 1$. Computer work gives the solution as $y = 1$. Is it bad computer work or a bad model?

Solution: It is a bad model, explanation to follow.

Solution verification. One solution of the initial value problem is given by the equilibrium solution $y = 1$. Another is $y = 1 + x^3$. Both are verified by direct substitution, using methods similar to Example 1.20, page 35.

Bad computer work? Technically, the computer made no mistake. Production numerical methods use only $f(x, y) = 3(y - 1)^{2/3}$ and the initial condition $y(0) = 1$. They apply a fixed algorithm to find successive values of y. The algorithm is expected to be successful, that is, it will compute a list of data points which can be graphed by "connect-the-dots." The majority of numerical methods applied to this example will compute $y = 1$ for all data points.

Bad model? Yes, it is a bad model, because the model does not *define* the solution $y = 1 + x^3$. The lesson here is that knowing a solution to an equation does not guarantee a numerical laboratory will be able to compute it.

Detecting a bad model. The right side $f(x, y)$ of the differential equation is continuous, but not continuously differentiable, therefore Picard's theorem does not apply, although Peano's theorem says a solution exists. Peano's theorem allows multiple solutions, but Picard's theorem does not. Sometimes, the only signal for non-uniqueness is the failure of application of Picard's theorem.

Physically significant models can have multiple solutions. A key example is the tank draining equation of E. Torricelli (1608-1647). A simple instance is $y' = -2\sqrt{|y|}$, in which $y(x)$ is the water height in some cylindrical tank at time x. No water in the tank means the water height is $y = 0$. An initial condition $y(0) = 0$ does not determine a unique solution, because the tank could have drained at some time in the past. For instance, if the tank drained at time $x = -1$, then a piecewise defined solution is $y(x) = (x+1)^2$ for $-\infty < x < -1$ and $y(x) = 0$ for $x \ge -1$. Most numerical methods applied to $y = -2\sqrt{|y|}$, $y(0) = 0$, compute $y(x) = 0$ for all x, which illustrates the inability of computer software to detect multiple solution errors.

Example 1.34 (Switches)

The problem $y' = f(x, y)$, $y(0) = y_0$ with

$$f(x, y) = \begin{cases} 0 & 0 \le x \le 1, \\ 1 & 1 < x < \infty, \end{cases}$$

has a piecewise-defined discontinuous right side $f(x, y)$. Solve the initial value problem for $y(x)$.

Solution: The solution $y(x)$ is found by dissection of the problem according to the two intervals $0 \le x \le 1$ and $1 < x < \infty$ into the two differential equations $y' = 0$ and $y' = 1$. By the fundamental theorem of calculus, the answers are $y = y_0$ and $y = x + y_1$, where y_1 is to be determined. At the common point $x = 1$, the two solutions should agree (we ask y to be continuous at $x = 1$), therefore $y_0 = 1 + y_1$, giving the final solution

$$y(x) = \begin{cases} y_0 & \text{on } 0 \le x \le 1, \\ x - 1 + y_0 & \text{on } 1 < x < \infty. \end{cases}$$

The function $y(x)$ is continuous, but y' is discontinuous at $x = 1$. The differential equation $y' = f(x, y)$ and initial condition $y(0) = y_0$ are formally satisfied.

Is there another continuous solution? No, because the method applied here assumed that $y(x)$ worked in the differential equation, and if it did, then it had to agree with $y = y_0$ on $0 \le x < 1$ and $y = x - 1 + y_0$ on $1 < x < \infty$, by the fundamental theorem of calculus.

Technically adept readers will find a flaw in the solution presented here, because of the treatment of the point $x = 1$, where y' does not exist. The flaw vanishes if we agree to verify the differential equation except at finitely many x-values where y' is undefined.

Example 1.35 (Finite Blowup)

Verify that $dx/dt = 2 + 2x^2$, $x(0) = 0$ has the unique solution $x(t) = \tan(2t)$, which approaches infinity in finite time $t = \pi/4$.

Solution: The function $x(t) = \tan(2t)$ works in the initial condition $x(0) = 0$, because $\tan(0) = \sin(0)/\cos(0) = 0$. The well-known trigonometric identity $\sec^2(2t) = 1 + \tan^2(2t)$ and the differentiation identity $x'(t) = 2\sec^2(2t)$ shows that the differential equation $x' = 2 + 2x^2$ is satisfied. The equation $x(\pi/4) = \infty$ is verified from $\tan(2t) = \sin(2t)/\cos(2t)$. Uniqueness follows from Picard's theorem, because $f(t, x) = 2 + 2x^2$ and $f_x(t, x) = 4x$ are continuous everywhere.

Example 1.36 (Picard Iterates)

Compute the Picard iterates $y_0, \ldots, y_3$ for the initial value problem $y' = f(x, y)$, $y(0) = 1$, given $f(x, y) = x - y$.

Solution: The answers are

$$\begin{aligned}
y_0(x) &= 1,\\
y_1(x) &= 1 - x + x^2/2,\\
y_2(x) &= 1 - x + x^2 - x^3/6,\\
y_3(x) &= 1 - x + x^2 - x^3/3 + x^4/24.
\end{aligned}$$

The details for all computations are similar. A sample computation appears below.

$y_0 = 1$	This follows from $y(0) = 1$.
$y_1 = 1 + \int_0^x f(t, y_0(t))dt$	Use the formula with $n = 1$.
$= 1 + \int_0^x (t - 1)dt$	Substitute $y_0(x) = 1$, $f(x, y) = x - y$.
$= 1 - x + x^2/2$	Evaluate the integral.

The exact answer is $y = x - 1 + 2e^{-x}$. A Taylor series expansion of this function motivates why the Picard iterates converge to $y(x)$. See the exercises for details, page 73.

The `maple` code which does the various computations appears below. The code involving `dsolve` is used to compute the exact solution and the series solution.

```
y0:=x->1:f:=(x,y)->x-y:
y1:=x->1+int(f(t,y0(t)),t=0..x):
y2:=x->1+int(f(t,y1(t)),t=0..x):
y3:=x->1+int(f(t,y2(t)),t=0..x):
y0(x),y1(x),y2(x),y3(x);
de:=diff(y(x),x)=f(x,y(x)): ic:=y(0)=1:
dsolve({de,ic},y(x)); dsolve({de,ic},y(x),series);
```

Exercises 1.6

Multiple Solution Example

Define $f(x, y) = 3(y - 1)^{2/3}$. Consider $y' = f(x, y)$, $y(0) = 1$.

1. Do an answer check for $y(x) = 1$. Do a second answer check for $y(x) = 1 + x^3$.

2. Let $y(x) = 1$ on $0 \le x \le 1$ and $y(x) = 1 + (x-1)^3$ for $x \ge 1$. Do an answer check for $y(x)$.

3. Does $f_y(x, y)$ exist for all (x, y)?

4. Verify that Picard's theorem does not apply to $y' = f(x, y)$, $y(0) = 1$, due to discontinuity of f_y.

5. Verify that Picard's theorem applies to $y' = f(x, y)$, $y(0) = 2$.

6. Let $y(x) = 1 + (x + 1)^3$. Do an answer check for $y' = f(x, y)$, $y(0) = 2$. Does another solution exist?

Discontinuous Equation Example

Consider $y' = \dfrac{2y}{x-1}$, $y(0) = 1$. Define $y_1(x) = (x-1)^2$ and $y_2(x) = c(x-1)^2$. Define $y(x) = y_1(x)$ on $-\infty < x < 1$ and $y(x) = y_2(x)$ on $1 < x < \infty$. Define $y(1) = 0$.

7. Do an answer check for $y_1(x)$ on $-\infty < x < 1$. Do an answer check for $y_2(x)$ on $1 < x < \infty$. Skip condition $y(0) = 1$.

8. Justify one-sided limits $y(1+) = y(1-) = 0$. The functions y_1 and y_2 join continuously at $x = 1$ with common value zero and the formula for $y(x)$ gives one continuous formal solution for each value of c (∞-many solutions).

9. (a) For which values of c does $y_2'(1)$ exist? (b) For which values of c is $y_2(x)$ continuously differentiable?

10. Find all values of c such that $y_2(x)$ is a continuously differentiable function that satisfies the differential equation and the initial condition.

Finite Blowup Example

Consider $y' = 1 + y^2$, $y(0) = 0$. Let $y(x) = \tan x$.

11. Do an answer check for $y(x)$.

12. Find the partial derivative f_y for $f(x, y) = 1 + y^2$. Justify that f and f_y are everywhere continuous.

13. Justify that Picard's theorem applies, hence $y(x)$ is the only possible solution to the initial value problem.

14. Justify for $a = -\pi/2$ and $b = \pi/2$ that $y(a+) = -\infty$, $y(b-) = \infty$. Hence $y(x)$ blows up for finite values of x.

Numerical Instability Example

Let $f(x, y) = y - 2e^{-x}$.

15. Do an answer check for $y(x) = e^{-x}$ as a solution of the initial value problem $y' = f(x, y)$, $y(0) = 1$.

16. Do an answer check for $y(x) = ce^x + e^{-x}$ as a solution of $y' = f(x, y)$.

Multiple Solutions

Consider the initial value problem $y' = 5(y-2)^{4/5}$, $y(0) = 2$.

17. Do an answer check for $y(x) = 2$. Do a second answer check for $y(x) = 2 + x^5$.

18. Verify that the hypotheses of Picard's theorem fail to apply.

19. Find a formula which displays infinitely many solutions to $y' = f(x, y)$, $y(0) = 2$.

20. Verify that the hypotheses of Peano's theorem apply.

Discontinuous Equation

Consider $y' = \dfrac{y}{x-1}$, $y(0) = 1$. Define $y(x)$ piecewise by $y(x) = -(x-1)$ on $-\infty < x < 1$ and $y(x) = c(x-1)$ on $1 < x < \infty$. Leave $y(1)$ undefined.

21. Do an answer check for $y(x)$. The initial condition $y(0) = 1$ applies only to the domain $-\infty < x < 1$.

22. Justify one-sided limits $y(1+) = y(1-) = 0$. The piecewise definitions of $y(x)$ join continuously at $x = 1$ with common value zero and the formula for $y(x)$ gives one continuous formal solution for each value of c (∞-many solutions).

23. (a) For which values of c does $y'(1)$ exist? (b) For which values of c is $y(x)$ continuously differentiable?

24. Find all values of c such that $y(x)$ is a continuously differentiable function that satisfies the differential equation and the initial condition.

Picard Iteration

Find the Picard iterates y_0, y_1, y_2, y_3.

25. $y' = y + 1$, $y(0) = 2$

26. $y' = 2y + 1$, $y(0) = 0$

27. $y' = y^2$, $y(0) = 1$

28. $y' = y^2$, $y(0) = 2$

29. $y' = y^2 + 1$, $y(0) = 0$

30. $y' = 4y^2 + 4$, $y(0) = 0$

31. $y' = y + x$, $y(0) = 0$

32. $y' = y + 2x$, $y(0) = 0$

Picard Iteration and Taylor Series

Find the Taylor polynomial $P_n(x) = y(0) + y'(0)x + \cdots + y^{(n)}(0)x^n/n!$ and compare with the Picard iterates. Use a computer algebra system, if possible.

33. $y' = y$, $y(0) = 1$, $n = 4$,
$y(x) = e^x$

34. $y' = 2y$, $y(0) = 1$, $n = 4$,
$y(x) = e^{2x}$

35. $y' = x - y$, $y(0) = 1$, $n = 4$,
$y(x) = -1 + x + 2e^{-x}$

36. $y' = 2x - y$, $y(0) = 1$, $n = 4$,
$y(x) = -2 + 2x + 3e^{-x}$

Numerical Instability

Use a computer algebra system or numerical laboratory. Let $f(x, y) = y - 2e^{-x}$.

37. Solve $y' = f(x, y)$, $y(0) = 1$ numerically for $y(30)$.

38. Solve $y' = f(x, y)$, $y(0) = 1+0.0000001$ numerically for $y(30)$.

Closed–Form Existence

Solve these initial value problems using a computer algebra system.

39. $y' = y$, $y(0) = 1$

40. $y' = 2y$, $y(0) = 2$

41. $y' = 2y + 1$, $y(0) = 1$

42. $y' = 3y + 2$, $y(0) = 1$

43. $y' = y(y - 1)$, $y(0) = 2$

44. $y' = y(1 - y)$, $y(0) = 2$

45. $y' = (y - 1)(y - 2)$, $y(0) = 3$

46. $y' = (y - 2)(y - 3)$, $y(0) = 1$

47. $y' = -10(1 - y)$, $y(0) = 0$

48. $y' = -10(2 - 3y)$, $y(0) = 0$

Lipschitz Condition

Justify the following results.

49. The function $f(x, y) = x - 10(2 - 3y)$ satisfies a Lipschitz condition on the whole plane.

50. The function $f(x, y) = ax + by + c$ satisfies a Lipschitz condition on the whole plane.

51. The function $f(x, y) = xy(1 - y)$ satisfies a Lipschitz condition on $D = \{(x, y) : |x| \le 1, \quad |y| \le 1\}$.

52. The function $f(x, y) = x^2y(a - by)$ satisfies a Lipschitz condition on $D = \{(x, y) : x^2 + y^2 \le R^2\}$.

53. If f_y is continuous on D and the line segment from (x, y_1) to (x, y_2) is in D, then $f(x, y_1) - f(x, y_2) = \int_{y_1}^{y_2} f_y(x, u)du$.

54. If f and f_y are continuous on a disk D, then f is Lipschitz with $M = \max_D\{|f_y(x, u)|\}$.

Chapter 2

First Order Differential Equations

Contents

The subject of the chapter is the first order differential equation

$$y' = f(x, y).$$

The study includes closed-form solution formulas for special equations and some applications to science and engineering.

2.1 Quadrature Method

The **method of quadrature** refers to the technique of integrating both sides of an equation, hoping thereby to extract a solution formula.

The term **quadrature** originates in ancient geometry, where it means *finding area* of a plane figure, by constructing a square of equal area.[1] Numerical quadrature computes areas enclosed by plane curves from approximating rectangles, by

[1] See Katz, Victor J. (1998) *A History of Mathematics: An Introduction* (2nd edition) Addison Wesley Longman, ISBN 0321016181, and Wikipedia: http://en.wikipedia.org/wiki/Quadrature.

algorithms such as the rectangular rule and Simpson's rule. For symbolic problems, the task is overtaken by Newton's integral calculus. The naming convention follows computer algebra system `maple`.

Fundamental Theorem of Calculus

The foundation of the study of differential equations rests with Isaac Newton's discovery on instantaneous velocities. Details of the calculus background required appears in Appendix A.1, page 1005.

Theorem 2.1 (Fundamental Theorem of Calculus I)
Let G be continuous and let F be continuously differentiable on $[a,b]$. Then

(a) $F(b) - F(a) = \displaystyle\int_a^b \frac{dF}{dx}(x)dx,$

(b) $\displaystyle\frac{d}{dx}\int_a^x G(t)dt = G(x).$

Theorem 2.2 (Fundamental Theorem of Calculus II)
Let $G(x)$ be continuous and let $y(x)$ be continuously differentiable on $[a,b]$. Then for some constant c,

(a) $y(x) = \displaystyle\int \frac{dy}{dx}dx + c,$

(b) $\displaystyle\frac{d}{dx}\int G(x)dx = G(x).$

Part (a) of the fundamental theorem is used to find a candidate solution to a differential equation.

Part (b) of the fundamental theorem is used in differential equations to do an answer check.

The Method of Quadrature

The method is applied to differential equations $y' = f(x,y)$ in which f is independent of y. Then symbol y is absent from $f(x,y)$, which implies $f(x,y)$ is constant or else $f(x,y)$ depends only on the symbol x. The model differential equation then has the form $y' = F(x)$ where F is a given function of the single variable F.

(i) **To solve for $y(x)$ in $\dfrac{dy}{dx} = F(x)$, integrate on variable x across the equation, then use the Fundamental Theorem of Calculus.**

(ii) **Check the answer.**

Indefinite Integral Shortcut. Integrate across the equation with indefinite integrals, then collect all integration constants into symbol c.

Solution with Symbol c. Symbol c initially appears in the expression obtained for y. If no initial condition was given, then the answer for y is this expression, which contains the unresolved symbol c. Experts call this expression the **general solution**.

Solution with No symbol c. If an initial condition is given in the form $y = y_0$ at $x = x_0$ (same as $y(x_0) = y_0$), then symbol c can be resolved. For instance, if the answer is $y = 2(x - 1) + c$ and the initial condition is $y(-1) = 3$, then $y = 2(x - 1) + c$ with $x = -1, y = 3$ becomes $3 = 2(-1 - 1) + c$, and then $c = 7$. Experts call the xy–expression with c eliminated a **particular solution**.

Theorem 2.3 (Existence-Uniqueness for Quadrature Equations)
Let $F(x)$ be continuous on $a < x < b$. Assume $a < x_0 < b$ and $-\infty < y_0 < \infty$. Then the initial value problem

$$y' = F(x), \quad y(x_0) = y_0 \tag{1}$$

has on interval $a < x < b$ the unique solution

$$y(x) = y_0 + \int_{x_0}^{x} F(t)dt. \tag{2}$$

Details of proof appear on page 79.

Examples

Example 2.1 (Quadrature)
Solve $y' = 3e^x$, $y(0) = 0$.

Solution:

The final answer is $y = 3e^x - 3$. An answer check appears in the next example.

Details. The *shortcut* is applied.

$\frac{dy}{dx} = 3e^x$	Copy the differential equation.
$\int \frac{dy}{dx}dx = \int 3e^x dx$	Integrate across the equation on x.
$y(x) + c_1 = \int 3e^x dx$	Fundamental theorem of calculus, page 75.
$y(x) + c_1 = 3e^x + c_2$	Integral table.
$y(x) = 3e^x + c$	Where $c = c_2 - c_1$ is a constant.

The answer is $y = 3e^x + c$. The symbol c is to be resolved from the **initial condition** $y(0) = 0$, as follows.

$0 = y(0)$	Copy the initial condition (sides reversed).
$= (3e^x + c)\|_{x=0}$	Insert $y = 3e^x + c$, the proposed solution.
$= 3e^0 + c$	Substitute $x = 0$.

$= 3 + c$	Use $e^0 = 1$.
$c = -1$	Equation $0 = 3 + c$ solved for c.

Candidate solution. Back-substitute the symbol c value $c = -1$ into the answer $y = 3e^x + c$ to obtain the candidate solution $y = 3e^x + (-3)$. This answer can contain errors, in general, due to integration and arithmetic mistakes.

Example 2.2 (Answer Check)

Given $y' = 3e^x$, $y(0) = 0$ and candidate solution $y(x) = 3e^x - 3$, display an answer check.

Solution: There are **two panels** in this answer check: **Panel 1**: differential equation check, **Panel 2**: initial condition check.

Panel 1. We check the answer $y = 3e^x - 3$ for the differential equation $y' = 3e^x$.

The steps are:

LHS $= y'$	Left side of the differential equation.
$= (3e^x - 3)'$	Substitute the expression for y.
$= 3e^x - 0$	Sum rule, constant rule and $(e^u)' = u'e^u$.
= RHS	Solution verified.

Panel 2. Let's check the answer $y = 3e^x - 3$ against the initial condition $y(0) = 0$. Expected is an immediate mental check that $e^0 = 1$ implies the correctness of $y(0) = 0$.

The steps will be shown in order to detail the algorithm for checking an initial condition. The algorithm applies when checking complex algebraic expressions. Abbreviated versions of the algorithm are used on simple expressions.

LHS $= y(0)$	Left side of the initial condition $y(0) = 0$.
$= (3e^x - 3)\|_{x=0}$	Notation $y(x_0)$ means substitute $x = x_0$ into the expression for y.
$= 3e^0 - 3$	Substitute $x = 0$ into the expression.
$= 0$	Because $e^0 = 1$.
= RHS	Initial condition verified.

River Crossing

A boat crosses a river at fixed speed with power applied perpendicular to the shoreline. Is it possible to estimate the boat's downstream location?

The answer is *yes*. The problem's variables are

x	Distance from shore,	w	Width of the river,
y	Distance downstream,	v_b	Boat velocity (dx/dt),
t	Time in hours,	v_r	River velocity (dy/dt).

The calculus chain rule $dy/dx = (dy/dt)/(dx/dt)$ is applied, using the symbols v_r and v_b instead of dy/dt and dx/dt, to give the *model equation*

$$\frac{dy}{dx} = \frac{v_r}{v_b}. \tag{3}$$

Stream Velocity. The downstream river velocity will be approximated by $v_r = kx(w-x)$, where $k > 0$ is a constant. This equation gives velocity $v_r = 0$ at the two shores $x = 0$ and $x = w$, while the **maximum stream velocity** at the center $x = w/2$ is (see page 79)

$$v_c = \frac{kw^2}{4}. \tag{4}$$

Special River-Crossing Model. The model equation (3) using $v_r = kx(w-x)$ and the constant k defined by (4) give the initial value problem

$$\frac{dy}{dx} = \frac{4v_c}{v_b w^2}x(w-x), \quad y(0) = 0. \tag{5}$$

The solution of (5) by the method of quadrature is

$$y = \frac{4v_c}{v_b w^2}\left(-\frac{1}{3}x^3 + \frac{1}{2}wx^2\right), \tag{6}$$

where w is the river's width, v_c is the river's midstream velocity and v_b is the boat's velocity. In particular, the boat's **downstream drift** on the opposite shore is $\frac{2}{3}w(v_c/v_b)$. See *Technical Details* page 79.

Example 2.3 (River Crossing)

A boat crosses a mile-wide river at 3 miles per hour with power applied perpendicular to the shoreline. The river's midstream velocity is 10 miles per hour. Find the transit time and the downstream drift to the opposite shore.

Solution: The answers, justified below, are 20 minutes and 20/9 miles.

Transit time. This is the time it takes to reach the opposite shore. The layman answer of 20 minutes is correct, because the boat goes 3 miles in one hour, hence 1 mile in 1/3 of an hour, perpendicular to the shoreline.

Downstream drift. This is the value $y(1)$, where y is the solution of equation (5), with $v_c = 10$, $v_b = 3$, $w = 1$, all distances in miles. The special model is

$$\frac{dy}{dx} = \frac{40}{3}x(1-x), \quad y(0) = 0.$$

The solution given by equation (6) is $y = \frac{40}{3}\left(-\frac{1}{3}x^3 + \frac{1}{2}x^2\right)$ and the downstream drift is then $y(1) = 20/9$ miles. This answer is 2/3 of the layman's answer of (1/3)(10) miles. The explanation is that the boat is pushed downstream at a variable rate from 0 to 10 miles per hour, depending on its position x.

Details and Proofs

Proof of Theorem 2.3:

Uniqueness. Let $y(x)$ be any solution of (1). It will be shown that $y(x)$ is given by the solution formula (2).

$$\begin{aligned} y(x) &= y(0) + \int_{x_0}^{x} y'(t)dt && \text{Fundamental theorem of calculus, page 1008.} \\ &= y_0 + \int_{x_0}^{x} F(t)dt && \text{Use (1). This verifies equation (2).} \end{aligned}$$

Answer Check. Let $y(x)$ be given by solution formula (2). It will be shown that $y(x)$ solves initial value problem (1).

$$\begin{aligned} y'(x) &= \left(y_0 + \int_{x_0}^{x} F(t)dt\right)' && \text{Compute the derivative from (2).} \\ &= F(x) && \text{Apply the fundamental theorem of calculus.} \end{aligned}$$

The initial condition is verified in a similar manner:

$$\begin{aligned} y(x_0) &= y_0 + \int_{x_0}^{x_0} F(t)dt && \text{Apply (2) with } x = x_0. \\ &= y_0 && \text{The integral is zero: } \int_a^a F(x)dx = 0. \end{aligned}$$

■

Technical Details for (4): The maximum of a continuously differentiable function $f(x)$ on $0 \le x \le w$ can be found by locating the critical points (i.e., where $f'(x) = 0$) and then testing also the endpoints $x = 0$ and $x = w$. The derivative $f'(x) = k(w - 2x)$ is zero at $x = w/2$. Then $f(w/2) = kw^2/4$. This value is the maximum of f, because $f = 0$ at the endpoints.

Technical Details for (6): Let $a = \dfrac{4v_c}{v_b w^2}$. Then

$$\begin{aligned} y &= y(0) + \int_0^x y'(t)dt && \text{Method of quadrature.} \\ &= 0 + a\int_0^x t(w - t)dt && \text{By (5), } y' = at(w - t). \\ &= a\left(-\tfrac{1}{3}x^3 + \tfrac{1}{2}wx^2\right). && \text{Integral table.} \end{aligned}$$

To compute the downstream drift, evaluate $y(w) = a\dfrac{w^3}{6}$ or $y(w) = \dfrac{2w}{3}\dfrac{v_c}{v_b}$.

Exercises 2.1

Quadrature

Find a candidate solution for each initial value problem and verify the solution. See Example 2.1 and Example 2.2, page 76.

1. $y' = 4e^{2x}$, $y(0) = 0$.

2. $y' = 2e^{4x}$, $y(0) = 0$.

3. $(1 + x)y' = x$, $y(0) = 0$.

4. $(1 - x)y' = x$, $y(0) = 0$.

5. $y' = \sin 2x$, $y(0) = 1$.

6. $y' = \cos 2x$, $y(0) = 1$.

7. $y' = xe^{x}$, $y(0) = 0$.

8. $y' = xe^{-x^2}$, $y(0) = 0$.

9. $y' = \tan x$, $y(0) = 0$.

10. $y' = 1 + \tan^2 x$, $y(0) = 0$.

11. $(1 + x^2)y' = 1$, $y(0) = 0$.

12. $(1+4x^2)y' = 1$, $y(0) = 0$.

13. $y' = \sin^3 x$, $y(0) = 0$.

14. $y' = \cos^3 x$, $y(0) = 0$.

15. $(1+x)y' = 1$, $y(0) = 0$.

16. $(2+x)y' = 2$, $y(0) = 0$.

17. $(2+x)(1+x)y' = 2$, $y(0) = 0$.

18. $(2+x)(3+x)y' = 3$, $y(0) = 0$.

19. $y' = \sin x \cos 2x$, $y(0) = 0$.

20. $y' = (1+\cos 2x)\sin 2x$, $y(0) = 0$.

River Crossing

A boat crosses a river of width w miles at v_b miles per hour with power applied perpendicular to the shoreline. The river's midstream velocity is v_c miles per hour. Find the transit time and the downstream drift to the opposite shore. See Example 2.3, page 78, and the details for (6).

21. $w = 1$, $v_b = 4$, $v_c = 12$

22. $w = 1$, $v_b = 5$, $v_c = 15$

23. $w = 1.2$, $v_b = 3$, $v_c = 13$

24. $w = 1.2$, $v_b = 5$, $v_c = 9$

25. $w = 1.5$, $v_b = 7$, $v_c = 16$

26. $w = 2$, $v_b = 7$, $v_c = 10$

27. $w = 1.6$, $v_b = 4.5$, $v_c = 14.7$

28. $w = 1.6$, $v_b = 5.5$, $v_c = 17$

Fundamental Theorem I

Verify the identity. Use the fundamental theorem of calculus part (b), page 75.

29. $\int_0^x (1+t)^3 dt = \frac{1}{4}\left((1+x)^4 - 1\right)$.

30. $\int_0^x (1+t)^4 dt = \frac{1}{5}\left((1+x)^5 - 1\right)$.

31. $\int_0^x te^{-t} dt = -xe^{-x} - e^{-x} + 1$.

32. $\int_0^x te^{t} dt = xe^{x} - e^{x} + 1$.

Fundamental Theorem II

Differentiate. Use the fundamental theorem of calculus part (b), page 75.

33. $\int_0^{2x} t^2 \tan(t^3) dt$.

34. $\int_0^{3x} t^3 \tan(t^2) dt$.

35. $\int_0^{\sin x} te^{t+t^2} dt$.

36. $\int_0^{\sin x} \ln(1+t^3) dt$.

Fundamental Theorem III

Integrate $\int_0^1 f(x)dx$. Use the fundamental theorem of calculus part (a), page 75. Check answers with computer or calculator assist. Some require a clever u-substitution or an integral table.

37. $f(x) = x(x-1)$

38. $f(x) = x^2(x+1)$

39. $f(x) = \cos(3\pi x/4)$

40. $f(x) = \sin(5\pi x/6)$

41. $f(x) = \dfrac{1}{1+x^2}$

42. $f(x) = \dfrac{2x}{1+x^4}$

43. $f(x) = x^2 e^{x^3}$

44. $f(x) = x(\sin(x^2) + e^{x^2})$

45. $f(x) = \dfrac{1}{\sqrt{-1+x^2}}$

46. $f(x) = \dfrac{1}{\sqrt{1-x^2}}$

47. $f(x) = \dfrac{1}{\sqrt{1+x^2}}$

48. $f(x) = \dfrac{1}{\sqrt{1+4x^2}}$

49. $f(x) = \dfrac{x}{\sqrt{1+x^2}}$

50. $f(x) = \dfrac{4x}{\sqrt{1-4x^2}}$

51. $f(x) = \dfrac{\cos x}{\sin x}$

52. $f(x) = \dfrac{\cos x}{\sin^3 x}$

53. $f(x) = \dfrac{e^x}{1+e^x}$

54. $f(x) = \dfrac{\ln|x|}{x}$

55. $f(x) = \sec^2 x$

56. $f(x) = \sec^2 x - \tan^2 x$

57. $f(x) = \csc^2 x$

58. $f(x) = \csc^2 x - \cot^2 x$

59. $f(x) = \csc x \cot x$

60. $f(x) = \sec x \tan x$

Integration by Parts

Integrate $\int_0^1 f(x)dx$ by parts, $\int udv = uv - \int vdu$. Check answers with computer or calculator assist.

61. $f(x) = xe^x$

62. $f(x) = xe^{-x}$

63. $f(x) = \ln|x|$

64. $f(x) = x\ln|x|$

65. $f(x) = x^2e^{2x}$

66. $f(x) = (1+2x)e^{2x}$

67. $f(x) = x\cosh x$

68. $f(x) = x\sinh x$

69. $f(x) = x\arctan(x)$

70. $f(x) = x\arcsin(x)$

Partial Fractions

Integrate f by partial fractions. Check answers with computer or calculator assist.

71. $f(x) = \dfrac{x+4}{x+5}$

72. $f(x) = \dfrac{x-2}{x-4}$

73. $f(x) = \dfrac{x^2+4}{(x+1)(x+2)}$

74. $f(x) = \dfrac{x(x-1)}{(x+1)(x+2)}$

75. $f(x) = \dfrac{x+4}{(x+1)(x+2)}$

76. $f(x) = \dfrac{x-1}{(x+1)(x+2))}$

77. $f(x) = \dfrac{x+4}{(x+1)(x+2)(x+5)}$

78. $f(x) = \dfrac{x(x-1)}{(x+1)(x+2)(x+3)}$

79. $f(x) = \dfrac{x+4}{(x+1)(x+2)(x-1)}$

80. $f(x) = \dfrac{x(x-1)}{(x+1)(x+2)(x-1)}$

Special Methods

Integrate f by using the suggested u-substitution or method. Check answers with computer or calculator assist.

81. $f(x) = \dfrac{x^2+2}{(x+1)^2}$, $u = x+1$.

82. $f(x) = \dfrac{x^2+2}{(x-1)^2}$, $u = x-1$.

83. $f(x) = \dfrac{2x}{(x^2+1)^3}$, $u = x^2+1$.

84. $f(x) = \dfrac{3x^2}{(x^3+1)^2}$, $u = x^3+1$.

85. $f(x) = \dfrac{x^3+1}{x^2+1}$, use long division.

86. $f(x) = \dfrac{x^4+2}{x^2+1}$, use long division.

2.2 Separable Equations

An equation $y' = f(x,y)$ is called **separable** provided algebraic operations, usually multiplication, division and factorization, allow it to be written in a **separable form** $y' = F(x)G(y)$ for some functions F and G. This class includes the *quadrature equations* $y' = F(x)$. Separable equations and associated solution methods were discovered by G. Leibniz in 1691 and formalized by J. Bernoulli in 1694.

Finding a Separable Form

Given differential equation $y' = f(x,y)$, invent values x_0, y_0 such that $f(x_0,y_0) \neq 0$. Define F, G by the formulas

$$(1) \qquad F(x) = \frac{f(x,y_0)}{f(x_0,y_0)}, \quad G(y) = f(x_0,y).$$

Because $f(x_0,y_0) \neq 0$, then (1) makes sense.

Theorem 2.4 (Separability Test)
Let F and G be defined by equation (1). Compute $F(x)G(y)$. Then

(a) $F(x)G(y) = f(x,y)$ implies $y' = f(x,y)$ is **separable**.

(b) $F(x)G(y) \neq f(x,y)$ implies $y' = f(x,y)$ is **not separable**.

Proof: Conclusion (b) follows from separability test I, *infra.* Conclusion (a) follows because two functions $F(x)$, $G(y)$ have been defined in equation (1) such that $f(x,y) = F(x)G(y)$ (definition of separable equation).

Invention and Application. Initially, let (x_0,y_0) be $(0,0)$ or $(1,1)$ or some suitable pair, for which $f(x_0,y_0) \neq 0$; then define F and G by (1). Multiply F and G to test the equation $FG = f$. The algebra will discover a factorization $f = F(x)G(y)$ without having to know algebraic tricks like factorizing multivariable equations. But if $FG \neq f$, then the algebra *proves* the equation is not separable.

Non-Separability Tests

Test I Equation $y' = f(x,y)$ is not separable if

$$(2) \qquad f(x,y_0)f(x_0,y) - f(x_0,y_0)f(x,y) \neq 0$$

for some pair of points (x_0,y_0), (x,y) in the domain of f.

Test II The equation $y' = f(x,y)$ is not separable if either $f_x(x,y)/f(x,y)$ is non-constant in y or $f_y(x,y)/f(x,y)$ is non-constant in x.

Illustration. Consider $y' = xy + y^2$. *Test I* implies it is not separable, because $f(x,1)f(0,y) - f(0,1)f(x,y) = (x+1)y^2 - (xy+y^2) = x(y^2 - y) \neq 0$. *Test II* implies it is not separable, because $f_x/f = 1/(x+y)$ is not constant as a function of y.

Test I details. Assume $f(x,y) = F(x)G(y)$, then equation (2) fails because each term on the left side of (2) evaluates to $F(x)G(y_0)F(x_0)G(y)$ for all choices of (x_0, y_0) and (x, y) (hence contradiction $0 \neq 0$).

Test II details. Assume $f(x,y) = F(x)G(y)$ and F, G are sufficiently differentiable. Then $f_x(x,y)/f(x,y) = F'(x)/F(x)$ is independent of y and the fraction $f_y(x,y)/f(x,y) = G'(y)/G(y)$ is independent of x.

Separated Form Test

A **separated equation** $y'/G(y) = F(x)$ is recognized by this test:

Left Side Test. The left side of the equation has factor y' and it is independent of symbol x.

Right Side Test. The right side of the equation is independent of symbols y and y'.

Variables-Separable Method

Determined by the method are the following kinds of solution formulas.

Equilibrium Solutions. They are the constant solutions $y = c$ of $y' = f(x,y)$. Find them by substituting $y = c$ in $y' = f(x,y)$, followed by solving for c, then report the list of answers $y = c$ so found.

Non-Equilibrium Solutions. For separable equation $y' = F(x)G(y)$, it is a solution y with $G(y) \neq 0$. It is found by dividing by $G(y)$ and applying the method of quadrature.

The term *equilibrium* is borrowed from kinematics. Alternative terms are **rest solution** and **stationary solution**; all mean $y' = 0$ in calculus terms.

Spurious Solutions. If $F(x)G(y) = 0$ is solved instead of $G(y) = 0$, then both x and y solutions might be found. The x-solutions are ignored: they are not equilibrium solutions. Only solutions of the form $y =$ constant are called equilibrium solutions.

It is important to *check the solution* to a separable equation, because certain steps used to arrive at the solution may not be reversible.

For most applications, the two kinds of solutions suffice to determine all possible solutions. In general, a separable equation may have non-unique solutions to some initial value problem. To prevent this from happening, it can be assumed that F, G and G' are continuous; see the Picard-Lindelöf theorem, page 68. If non-uniqueness does occur, then often the equilibrium and non-equilibrium solutions can be pieced together to represent all solutions.

Finding Equilibrium Solutions

The search for equilibria can be done without finding the separable form of $y' = f(x,y)$. It is enough to solve for y in the equation $f(x,y) = 0$, *subject to the condition that x is arbitrary.* An equilibrium solution y cannot depend upon x, because it is *constant.* If y turns out to depend on x, after solving $f(x,y) = 0$ for y, then this is sufficient evidence that $y' = f(x,y)$ is **not separable**. Some examples:

$y' = y\sin(x - y)$ — It is *not separable*. The solutions of $y\sin(x-y) = 0$ are $y = 0$ and $x - y = n\pi$ for any integer n. The solution $y = x - n\pi$ is non-constant, therefore the equation cannot be separable.

$y' = xy(1 - y^2)$ — It is *separable*. The equation $xy(1 - y^2) = 0$ has three equilibrium solutions $y = 0$, $y = 1$, $y = -1$. Equilibrium solutions must be constant solutions.

Algorithm. To find equilibrium solutions, formally substitute $y = c$ into the differential equation, then solve for c, and report all constant solutions $y = c$ so found. There can be zero solutions, or just one solution, or some finite number of solutions, or infinitely many solutions.

Shortcut. In a given problem, a formal substitution is not used, but instead y' is replaced by zero (the result when $y =$ constant). For $y' = f(x,y)$, the equation $f(x,y) = 0$ is to be solved for y. For example, $y' = (x+1)(y^2 - 4)$ becomes $0 = (x+1)(y^2-4)$, equivalent to $y^2 - 4 = 0$ or $y = 2$, $y = -2$. The spurious solution $x = -1$ is ignored, because we are looking for constant solutions of the form $y = c$, which in this example are $y = 2$ and $y = -2$.

The problem of finding all equilibrium solutions is known to be technically unsolvable, that is, there is no proven algorithm for finding all the solutions of $G(y) = 0$. However, there are some very good numerical methods that apply, including **Newton's method** and the **bisection method**. Modern computer algebra systems make it practical to find equilibrium solutions, both symbolic (like $y = \pi$) and numeric (like $y = 3.14159$), in a single effort.

Finding Non-Equilibrium Solutions

A given solution $y(x)$ satisfying $G(y(x)) \neq 0$ throughout its domain of definition is called a non-equilibrium solution. Then division by $G(y(x))$ is allowed in the differential equation $y'(x) = F(x)G(y(x))$. The *method of quadrature* applies to the separated equation $y'/G(y(x)) = F(x)$. Some details:

$\int_{x_0}^{x} \dfrac{y'(t)dt}{G(y(t))} = \int_{x_0}^{x} F(t)dt$ — Integrate both sides of the separated equation over $x_0 \le t \le x$.

$\int_{y_0}^{y(x)} \dfrac{du}{G(u)} = \int_{x_0}^{x} F(t)dt$ — Apply on the left the change of variables $u = y(t)$. Define $y_0 = y(x_0)$.

$y(x) = W^{-1}\left(\int_{x_0}^{x} F(t)dt\right)$	Define $W(y) = \int_{y_0}^{y} du/G(u)$. Take inverses to isolate $y(x)$.

The calculation produces a formula which is strictly speaking a *candidate solution* y. It does not prove that the formula works in the equation: *checking the solution* is required.

Theoretical Inversion

The function W^{-1} appearing in the last step above is generally not given by a formula. Therefore, W^{-1} rarely appears explicitly in applications or examples. It is the *method* that is memorized:

> Prepare a separable differential equation by transforming it to separated form. Then apply the method of quadrature.

The separated form $y' = F(x)G(y)$ is checked by the separated form test, page 83. For example, $y' = (1 + x^2)y^3$ has $F = 1 + x^2$ and $G = y^3$; quadrature is applied to the divided equation $y'/y^3 = 1 + x^2$.

The theoretical basis for using W^{-1} is a calculus theorem which says that *a strictly monotone continuous function has a continuous inverse.* The fundamental theorem of calculus implies that $W(y)$ is continuous with nonzero derivative $W'(y) = 1/G(y)$. Therefore, $W(y)$ is strictly monotone. The cited calculus theorem implies that $W(y)$ has a continuously differentiable inverse W^{-1}.

Explicit and Implicit Solutions

The variables-separable method gives equilibrium solutions which are already **explicit**, that is:

Definition 2.1 (Explicit Solution)
A solution of $y' = f(x, y)$ is called **explicit** provided it is given by an equation

$$y = \text{an expression independent of } y.$$

To elaborate, on the left side must appear exactly the symbol y followed by an equal sign. Symbols y and $=$ are followed by an expression which does not contain the symbol y. Examples of explicit equations are $y = 0$, $y = -1$, $y = x + 2\pi$, $y = \sin x + x^2 + 10$. The definition is strict, for example $y + 1 = 0$ is **not explicit** because it fails to have y isolated left. Yes, it can be converted into an explicit equation $y = -1$.

Definition 2.2 (Implicit Solution)
A solution of $y' = f(x, y)$ is called **implicit** provided it is not explicit.

Equations like $2y = x$ are not explicit (they are called *implicit*) because the coefficient of y on the left is not 1. Similarly, $y = x + y^2$ is not explicit because the right side contains symbol y. Equation $y = e^\pi$ is explicit because the right side fails to contain symbol y (symbol x may be absent). Applications can leave the non-equilibrium solutions in *implicit* form $\int_{y_0}^{y(x)} du/G(u) = \int_{x_0}^{x} F(t)dt$, with serious effort being expended to do the indicated integrations.

In special cases, it is possible to find an explicit solution from the implicit one by algebraic methods. The required algebraic methods might appear to be unmotivated tricks. Computer algebra systems can make this step look like science instead of art.

Examples

Example 2.4 (Non-separable Equation)
Explain why $yy' = x - y^2$ is not separable.

Solution: It is tempting to try manipulations like adding y^2 to both sides of the equation, in an attempt to obtain a separable form, but every such trick fails. The failure of such attempts is evidence that the equation is perhaps not separable. Failure of attempts does not *prove* non-separability.

Test I applies to verify that the equation is not separable. Let $f(x, y) = x/y - y$ and choose $x_0 = 0$, $y_0 = 1$. Then $f(x_0, y_0) \neq 0$. Compute as follows:

$$\begin{aligned}
\text{LHS} &= f(x, y_0)f(x_0, y) - f(x_0, y_0)f(x, y) && \text{Identity (2) left side.}\\
&= f(x, 1)f(0, y) - f(0, 1)f(x, y) && \text{Use } x_0 = 0,\ y_0 = 1.\\
&= (x - 1)(-y) - (-1)(x/y - y) && \text{Substitute } f(x, y) = x/y - y.\\
&= -xy + x/y && \text{Simplify.}
\end{aligned}$$

This expression fails to be zero for all (x, y) (e.g., $x = 1$, $y = 2$), therefore the equation is not separable, by *Test I*.

Test II also applies to verify the equation is not separable: $\dfrac{f_x}{f} = \dfrac{1/y}{f} = x - y^2$ is non-constant in x.

Example 2.5 (Separated Form Test Failure)
Given $yy' = 1 - y^2$, explain why the equivalent equation $yy' + y^2 = 1$, obtained by adding y^2 across the equation, fails the separated form test, page 83.

Solution: The *test* requires the left side of $yy' + y^2 = 1$ to contain the factor y'. It doesn't, so it fails the test. Yes, $yy' + y^2 = 1$ does pass the other checkpoints of the *test*: the left side is independent of x and the right side is independent of y and y'.

Example 2.6 (Separated Equation)
Find for $(x + 1)yy' = x - xy^2$ a separated equation using the *test*, page 83.

Solution: The equation usually reported is $\dfrac{yy'}{(1-y)(1+y)} = \dfrac{x}{x+1}$. It is found by factoring and division.

The given equation is factored into $(1+x)yy' = x(1-y)(1+y)$. To pass the *test*, the objective is to move all factors containing only y to the left and all factors containing only x to the right. This is technically accomplished using division by $(x+1)(1-y)(1+y)$.

To the result of the division is applied the *test* on page 83: the *left side* contains factor y' and otherwise involves the factor $y/(1-y^2)$, which depends only on y; the *right side* is $x/(x+1)$, which depends only on x. In short, the candidate separated equation passes the test.

There is another way to approach the problem, by writing the differential equation in standard form $y' = f(x,y)$ where $f(x,y) = x(1-y^2)/(1+x)$. Then $f(1,0) = 1/2 \neq 0$. Define $F(x) = f(x,0)/f(1,0)$, $G(y) = f(1,y)$. We verify $F(x)G(y) = f(x,y)$. A separated form is then $y'/G(y) = F(x)$ or $2y'/(1-y^2) = 2x/(1+x)$.

Example 2.7 (Equilibria)

Given $y' = x(1-y)(1+y)$, find all equilibria.

Solution: The constant solutions $y = -1$ and $y = 1$ are the equilibria, as will be justified.

Equilibria are found by substituting $y = c$ into the differential equation $y' = x(1-y)(1+y)$, which gives the equation

$$x(1-c)(1+c) = 0.$$

The formal college algebra solutions are $x = 0$, $c = -1$ and $c = 1$. However, we do not seek these college algebra answers! Equilibria are the solutions $y = c$ such that $x(1-c)(1+c) = 0$ *for all* x. The conditional *for all* x causes the algebra problem to reduce to just two equations: $0 = 0$ (from $x = 0$) and $(1-c)(1+c) = 0$ (from $x \neq 0$). We solve for $c = 1$ and $c = -1$, then report the two equilibrium solutions $y = 1$ and $y = -1$. Spurious algebraic solutions like $x = 0$ can appear, which must be removed from equilibrium solution reports.

Example 2.8 (Non-Equilibria)

Given $y' = x^2(1+y)$, $y(0) = y_0$, find all non-equilibrium solutions.

Solution: The unique solution is $y = (1+y_0)e^{x^3/3} - 1$. Details follow.

The separable form $y' = F(x)G(y)$ is realized for $F(x) = x^2$ and $G(y) = 1 + y$. Sought are solutions with $G(y) \neq 0$, or simply $1 + y \neq 0$.

$y' = x^2(1+y)$	Original equation.
$\dfrac{y'}{1+y} = x^2$	Divide by $1+y$. Separated form found.
$\int \dfrac{y'}{1+y} dx = \int x^2 dx$	Method of quadrature.
$\int \dfrac{du}{1+u} = \int x^2 dx$	Change variables $u = y(x)$ on the left.
$\ln\|1+y(x)\| = x^3/3 + c$	Evaluate integrals. Implicit solution found.

Applications might stop at this point and report the *implicit solution*. This illustration continues, to find the *explicit solution* $y = (1+y_0)e^{x^3/3} - 1$.

$\lvert 1+y(x)\rvert = e^{x^3/3+c}$	By definition, $\ln u = w$ means $u = e^w$.
$1+y(x) = ke^{x^3/3+c}$	Drop absolute value, $k = \pm 1$.
$y(x) = ke^{x^3/3+c} - 1$	Candidate solution. Constants unresolved.

The initial condition $y(0) = y_0$ is used to resolve the constants c and k. First, $\lvert 1+y_0\rvert = e^c$ from the first equation. Second, $1+y_0$ and $1+y(x)$ must have the same sign (they are never zero), so $k(1+y_0) > 0$. Hence, $1+y_0 = ke^c$, which implies the solution is $y = ke^ce^{x^3/3} - 1$ or $y = (1+y_0)e^{x^3/3} - 1$.

Example 2.9 (Equilibria)

Given $y' = x\sin(1-y)\cos(1+y)$, find all equilibrium solutions.

Solution: The infinite set of equilibria are justified below to be

$$y = 1 + n\pi, \quad y = -1 + (2n+1)\frac{\pi}{2}, \quad n = 0, \pm 1, \pm 2, \ldots$$

A separable form $y' = F(x)G(y)$ is verified by defining $F(x) = x$ and $G(y) = \sin(1-y)\cos(1+y)$. Equilibria $y = c$ are found by solving for c in the equation $G(c) = 0$, which is

$$\sin(1-c)\cos(1+c) = 0.$$

This equation is satisfied when the argument of the sine is an integer multiple of π or when the argument of the cosine is an odd integer multiple of $\pi/2$. The solutions are $c - 1 = 0, \pm\pi, \pm 2\pi, \ldots$ and $1 + c = \pm\pi/2, \pm 3\pi/2, \ldots$.

Multiple solutions and `maple`. Equations having multiple solutions may require **CAS** setup. Below, the first code fragment returns two solutions, $y = 1$ and $y = -1 + \pi/2$. The second returns all solutions.

```
# The default returns two solutions
G:=y->sin(1-y)*cos(1+y):
solve(G(y)=0,y);
# Special setup returns all solutions
_EnvAllSolutions := true:
G:=y->sin(1-y)*cos(1+y):
solve(G(y)=0,y);
```

Example 2.10 (Non-Equilibria)

Given $y' = x^2\sin(y)$, $y(0) = y_0$, justify that all non-equilibrium solutions are given by[2]

$$y = 2\,\mathsf{Arctan}\left(\tan(y_0/2)e^{x^3/3}\right) + 2n\pi.$$

Solution: A separable form $y' = F(x)G(y)$ is defined by $F(x) = x^2$ and $G(y) = \sin(y)$. A non-equilibrium solution will satisfy $G(y) \neq 0$, or simply $\sin(y) \neq 0$. Define n by $y_0/2 = \text{Arctan}(\tan(y_0/2)) + n\pi$, where $\lvert\text{Arctan}(u)\rvert < \pi/2$. Then

[2] While $\theta = \arctan u$ gives any angle, $\theta = \text{Arctan}(u)$ gives $\lvert\theta\rvert < \pi/2$.

$y' = x^2 \sin(y)$	The original equation.
$\csc(y)y' = x^2$	Separated form. Divided by $\sin(y) \neq 0$.
$\int \csc(y)y'dx = \int x^2 dx$	Quadrature using indefinite integrals.
$\int \csc(u)du = \int x^2 dx$	Change variables $u = y(x)$ on the left.
$\ln \lvert \csc y(x) - \cot y(x) \rvert = \frac{1}{3}x^3 + c$	Integral tables. Implicit solution found.

Trigonometric Identity. Integral tables make use of the identity $\tan(y/2) = \csc y - \cot y$, which is derived from the relations $2\theta = y$, $1 - \cos 2\theta = 2\sin^2\theta$, $\sin 2\theta = 2\sin\theta\cos\theta$. Tables offer an alternate answer for the last integral above, $\ln \lvert \tan(y/2) \rvert$.

The solution obtained at this stage is called an *implicit solution*, because y has not been isolated. It is possible to solve for y in terms of x, an *explicit solution*. The details:

$\lvert \csc y - \cot y \rvert = e^{x^3/3+c}$	By definition, $\ln u = w$ means $u = e^w$.
$\csc y - \cot y = ke^{x^3/3+c}$	Assign $k = \pm 1$ to drop absolute values.
$\dfrac{1-\cos y}{\sin y} = ke^{x^3/3+c}$	Then k has the same sign as $\sin(y)$, because $1 - \cos y \geq 0$.
$\tan(y/2) = ke^{x^3/3+c}$	Use $\tan(y/2) = \csc y - \cot y$.
$y = 2\text{Arctan}\left(ke^{x^3/3+c}\right) + 2n\pi$	Candidate solution, $n = 0, \pm 1, \pm 2, \ldots$

Resolving the Constants. Constants c and k are uniquely resolved for a given initial condition $y(0) = y_0$. Values $x = 0$ and $y = y_0$ determine constant c by the equation $\tan(y_0/2) = ke^c$ (two equations back). The condition $k\sin(y_0) > 0$ determines k, because $\sin y_0$ and $\sin y$ have identical signs. If n is defined by $y_0/2 = \text{Arctan}(\tan(y_0/2)) + n\pi$ and $K = ke^c = \tan(y_0/2)$, then the *explicit solution* is

$$y = 2\text{Arctan}\left(Ke^{x^3/3}\right) + 2n\pi, \quad K = \tan(y_0/2).$$

Trigonometric identities and `maple`. Using the identity $\csc y - \cot y = \tan(y/2)$, `maple` finds the same relation. Complications occur without it.

```
_EnvAllSolutions := true:
solve(csc(y)-cot(y)=k*exp(x^3/3+c),y);
solve(tan(y/2)=k*exp(x^3/3+c),y);
```

Example 2.11 (Independent of x)

Solve $y' = y(1 - \ln y)$, $y(0) = y_0$.

Solution: There is just one equilibrium solution $y = e \approx 2.718$. Not included is $y = 0$, because $y(1 - \ln y)$ is undefined for $y \leq 0$. Details appear below for the explicit solution (which includes $y = e$)

$$y = e^{1 - (1 - \ln y_0)e^{-x}}.$$

An equation $y' = f(x, y)$ independent of x has the form $y' = F(x)G(y)$ where $F(x) = 1$. Divide by $G(y)$ to obtain a separated form $y'/G(y) = 1$. In the present case, $G(y) = y(1 - \ln y)$ is defined for $y > 0$. To require $G(y) \neq 0$ means $y > 0$, $y \neq e$. Non-equilibrium solutions will satisfy $y > 0$ and $y \neq e$.

$$\frac{y'}{y(1-\ln y)} = 1$$

Separated form. Assume $y > 0$ and $y \neq e$.

$$\int \frac{y'}{y(1-\ln y)}dx = \int dx$$

Method of quadrature.

$$\int \frac{-du}{u} = \int dx$$

Substitute $u = 1 - \ln y$ on the left. Chain rule $(\ln y)' = y'/y$ applied; $du = -y'dx/y$.

$$-\ln|1-\ln y(x)| = x + c$$

Evaluate the integral using $u = 1 - \ln y$. Implicit solution found.

The remainder of the solution contains college algebra details, to find from the *implicit solution* all *explicit solutions* y.

$$|1-\ln y(x)| = e^{-x-c}$$

Use $\ln u = w$ equivalent to $u = e^w$.

$$1-\ln y(x) = ke^{-x-c}$$

Drop absolute value, $k = \pm 1$.

$$\ln y(x) = 1 - ke^{-x-c}$$

Solved for $\ln y$.

$$y(x) = e^{1 - ke^{-x-c}}$$

Candidate solution; c and k unresolved.

To resolve the constants, start with $y_0 > 0$ and $y_0 \neq e$. To determine k, use the requirement $G(y) \neq 0$ to deduce that $k(1-\ln y(x)) > 0$. At $x = 0$, it means $k|1-\ln y_0| = 1 - \ln y_0$. Then $k = 1$ for $0 < y_0 < e$ and $k = -1$ otherwise.

Let $y = y_0$, $x = 0$ to determine c through the equation $|1 - \ln y_0| = e^{-c}$. Combining with the value of k gives $1 - \ln y_0 = ke^{-c}$.

Assembling the answers for k and c produces the relations

$$y = e^{1 - ke^{-x-c}}$$

Candidate solution.

$$= e^{1 - ke^{-c}e^{-x}}$$

Exponential rule $e^{a+b} = e^a e^b$.

$$= e^{1 - (1-\ln y_0)e^{-x}}$$

Explicit solution. Used $ke^{-c} = 1 - \ln y_0$.

Even though the solution has been found through legal methods, it remains to *verify the solution*. See the exercises.

Exercises 2.2

Separated Form Test

Test the given equation by the separated form test on page 83.

Report whether or not the equation *passes* or *fails*, as written. In this test, algebraic operations on the equation are disallowed. See Examples 2.4 and 2.5, page 86.

1. $y' = 2$

2. $y' = x$

3. $y' + y = 2$

4. $y' + 2y = x$

5. $yy' = 2 - x$

6. $2yy' = x + x^2$

7. $yy' + \sin(y') = 2 - x$

8. $2yy' + \cos(y) = x$

9. $2yy' = y'\cos(y) + x$

10. $(2y + \tan(y))y' = x$

Separated Equation

Determine the separated form $y'/G(y) = F(x)$ for the given separable equation. See Example 2.6, page 86.

11. $(1+x)y' = 2+y$

12. $(1+y)y' = xy$

13. $y' = \dfrac{x+xy}{(x+1)^2-1}$

14. $y' = \sin(x)\dfrac{1+y}{(x+2)^2-4}$

15. $xy' = y\sin(y)\cos(x)$

16. $x^2y' = y\cos(y)\tan(x)$

17. $y^2(x-y)y' = \dfrac{x^2-y^2}{x+y}$

18. $xy^2(x+y)y' = \dfrac{y^2-x^2}{x-y}$

19. $xy^2y' = \dfrac{y-x}{x-y}$

20. $xy^2y' = \dfrac{x^2-xy}{x-y}$

Equilibrium solutions

Determine the equilibria for the given equation. See Examples 2.7 and 2.9.

21. $y' = xy(1+y)$

22. $xy' = y(1-y)$

23. $y' = \dfrac{1+y}{1-y}$

24. $xy' = \dfrac{y(1-y)}{1+y}$

25. $y' = (1+x)\tan(y)$

26. $y' = y(1+\ln y)$

27. $y' = xe^y(1+y)$

28. $xy' = e^y(1-y)$

29. $xy' = e^y(1-y^2)(1+y)^3$

30. $xy' = e^y(1-y^3)(1+y^3)$

Non-Equilibrium Solutions

Find the non-equilibrium solutions for the given separable equation. See Examples 2.8 and 2.10 for details.

31. $y' = (xy)^{1/3}$, $y(0) = y_0$.

32. $y' = (xy)^{1/5}$, $y(0) = y_0$.

33. $y' = 1+x-y-xy$, $y(0) = y_0$.

34. $y' = 1+x+2y+2xy$, $y(0) = y_0$.

35. $y' = \dfrac{(x+1)y^3}{x^2(y^3-y)}$, $y(1) = y_0 \neq 0$.

36. $y' = \dfrac{(x-1)y^2}{x^3(y^3+y)}$, $y(0) = y_0$.

37. $2yy' = x(1-y^2)$

38. $2yy' = x(1+y^2)$

39. $(1+x)y' = 1-y$

40. $(1-x)y' = 1+y$, $y(0) = y_0$.

41. $\tan(x)y' = y$, $y(\pi/2) = y_0$.

42. $\tan(x)y' = 1+y$, $y(\pi/2) = y_0$.

43. $\sqrt{x}y' = \cos^2(y)$, $y(1) = y_0$.

44. $\sqrt{1-x}y' = \sin^2(y)$, $y(0) = y_0$.

45. $\sqrt{x^2-16y}y' = x$, $y(5) = y_0$.

46. $\sqrt{x^2-1y}y' = x$, $y(2) = y_0$.

47. $y' = x^2(1+y^2)$, $y(0) = 1$.

48. $(1-x)y' = x(1+y^2)$, $y(0) = 1$.

Independent of x

Solve the given equation, finding all solutions. See Example 2.11.

49. $y' = \sin y$, $y(0) = y_0$.

50. $y' = \cos y$, $y(0) = y_0$.

51. $y' = y(1+\ln y)$, $y(0) = y_0$.

52. $y' = y(2+\ln y)$, $y(0) = y_0$.

53. $y' = y(y-1)(y-2)$, $y(0) = y_0$.

54. $y' = y(y-1)(y+1)$, $y(0) = y_0$.

55. $y' = y^2+2y+5$, $y(0) = y_0$.

56. $y' = y^2+2y+7$, $y(0) = y_0$.

Details in the Examples

Collected here are verifications for details in the examples.

57. **(Example 2.7)** The equation $x(1-y)(1+y) = 0$ was solved in the example, but $x = 0$ was ignored, and only $y = -1$ and $y = 1$ were reported. Why?

58. **(Example 2.8)** An absolute value equation $|u| = w$ was replaced by $u = kw$ where $k = \pm 1$. Justify the replacement using the *definition* $|u| = u$ for $u \geq 0$, $|u| = -u$ for $u < 0$.

59. **(Example 2.8)** Verify directly that $y = (1+y_0)e^{x^3/3} - 1$ solves the initial value problem $y' = x^2(1+y)$, $y(0) = y_0$.

60. **(Example 2.9)** The relation $y = 1 + n\pi$, $n = 0, \pm 1, \pm 2, \ldots$ describes the list $\ldots, 1-\pi, 1, 1+\pi, \ldots$. Write the list for the relation $y = -1 + (2n+1)\frac{\pi}{2}$.

61. **(Example 2.9)** Solve $\sin(u) = 0$ and $\cos(v) = 0$ for u and v. Supply graphs which show why there are infinity many solutions.

62. **(Example 2.10)** Explain why $y_0/2$ does not equal $\text{Arctan}(\tan(y_0/2))$. Give a calculator example.

63. **(Example 2.10)** Establish the identity $\tan(y/2) = \csc y - \cot y$.

64. **(Example 2.11)** Let $y_0 > 0$. Verify that $y = e^{1-(1-\ln y_0)e^{-x}}$ solves

$$y' = y(1 - \ln y), \quad y(0) = y_0.$$

2.3 Linear Equations

Definition 2.3 (Linear Differential Equation)
An equation $y' = f(x,y)$ is called **first-order linear** or a **linear equation** provided functions $p(x)$ and $r(x)$ can be defined to re-write the equation in the **standard form**

$$y' + p(x)y = r(x). \tag{1}$$

In most applications, p and r are assumed to be continuous. Function $p(x)$ is called the **coefficient of** y. Function $r(x)$ (r abbreviates *right side*) is called the **non-homogeneous term** or the **forcing term**. Engineering texts call $r(x)$ the **input** and the solution $y(x)$ the **output**.

In examples, a linear equation is identified by matching:

$$\frac{dy}{dx} + \left(\begin{array}{l} p(x)\text{, an expression} \\ \text{independent of } y \end{array} \right) y = \left(\begin{array}{l} r(x)\text{, another expression} \\ \text{independent of } y \end{array} \right)$$

.

Calculus Test:

> An equation $y' = f(x,y)$ with f continuously differentiable is **linear** provided $\dfrac{\partial f(x,y)}{\partial y}$ is independent of y.

If the test is passed, then standard linear form (1) is obtained by defining $r(x) = f(x,0)$ and $p(x) = -\partial f/\partial y(x,y)$. A brief calculation verifies this statement.

Key Examples

$L\dfrac{dI}{dt} + RI = E$	The LR-circuit equation. Symbols L, R and E are respectively inductance, resistance and electromotive force, while $I(t)$ = current in amperes and t = time. $\boxed{\mathbf{1}}$
$\dfrac{du}{dt} = -h(u - u_1)$	Newton's cooling equation. In the roast model, the oven temperature is u_1 and the meat thermometer reading is $u(t)$, with t = time. $\boxed{\mathbf{2}}$

Notes.
$\boxed{\mathbf{1}}$ Linear equation $y' + p(x)y = r(x)$ is realized with symbols y, x, p, r undergoing name changes. Define $x = t$, $y = I$, $p(x) = R/L$, $r(x) = E/L$.
$\boxed{\mathbf{2}}$ Linear equation $y' + p(x)y = r(x)$ is realized by re-defining symbols y, x, p, r. Start with the equation re-arranged algebraically to $\dfrac{du}{dt} + hu = hu_1$. Define $x = t$, $y = u$, $p(x) = h$, $r(x) = hu_1$.

Homogeneous Equation $y' + p(x)y = 0$

Homogeneous equations $y' + p(x)y = 0$ occur in applications devoid of external forces, like an LR-circuit with no battery in the circuit. Justified on page 101 is the fundamental result for such systems. See also the proof of Theorem 2.5 (a).

The general solution of $\dfrac{dy}{dx} + p(x)y = 0$ is the fraction

$$y(x) = \frac{\text{constant}}{\text{integrating factor}} = \frac{c}{W(x)}$$

where **integrating factor** $W(x)$ is defined by the equation

$$W(x) = e^{\int p(x)dx}.$$

An Illustration. The LR-circuit equation $\dfrac{dI}{dt} + 2I = 0$ is the model equation $y' + p(x)y = 0$ with $p(x) = 2$. Then $W(x) = e^{\int 2dx} = e^{2x}$, with integration constant set to zero. The general solution of $y' + 2y = 0$ is given by

$$y = \frac{c}{W(x)} = \frac{c}{e^{2x}} = ce^{-2x}.$$

The current is $I(t) = c\,e^{-2t}$, by the variable swap $x \to t$, $y \to I$.

Definition 2.4 (Integrating Factor)
An **integrating factor** $W(x)$ for equation $y' + p(x)y = r(x)$ is

$$W(x) = e^{\int p(x)dx}.$$

Lemma 2.1 (Integrating Factor Identity)
The integrating factor $W(x)$ satisfies the differential equation

$$W'(x) = p(x)W(x).$$

Lemma Details. Write $W = e^u$ where $u = \int p(x)dx$. By the fundamental theorem of calculus, $u' = p(x) =$ the integrand. Then the chain rule implies $W' = u'e^u = u'W = pW$.

A Shortcut. Factor $W(x)$ is generally expressed as a simplified expression, with integration constant set to zero and absolute value symbols removed. See the exercises for details about this simplification. For instance, integration in the special case $p(x) = 2$ formally gives $\int p(x)dx = \int 2dx = 2x + c_1$. Then the integrating factor becomes $W(x) = e^{\int 2dx} = e^{2x+c_1} = e^{2x}e^{c_1}$. Fraction $c/W(x)$ equals c_2/e^{2x}, where $c_2 = c/e^{c_1}$. The lesson is that we could have chosen $c_1 = 0$ to produce the same fraction. This is a shortcut, recognized as such, but it applies in examples to save effort.

Non-Homogeneous Equation $y' + p(x)y = r(x)$

Definition 2.5 (Homogeneous and Particular Solution)
Let $W(x)$ be an integrating factor constructed for $y'+p(x)y = r(x)$, that is, $W(x) = e^u$, where $u = \int p(x)dx$ is an antiderivative of $p(x)$.

Symbol y_h, called the **homogeneous solution**, is defined by the expression

$$y_h(x) = \frac{c}{W(x)}.$$

Symbol y_p, called a **particular solution**, is defined by the expression

$$y_p(x) = \frac{\int r(x)W(x)dx}{W(x)}$$

Theorem 2.5 (Homogeneous and Particular Solutions)
(a) Expression $y_h(x)$ is a solution of the homogeneous differential equation $y' + p(x)y = 0$.

(b) Expression $y_p(x)$ is a solution of the non-homogeneous differential equation $y' + p(x)y = r(x)$.

Proof:
(a) Define $y = c/W$. We prove $y' + py = 0$. Formula $y = c/W$ implies $(yW)' = (c)' = 0$. The product rule and the Lemma imply $(yW)' = y'W + yW' = y'W + y(pW) = (y' + py)W$. Then $(yW)' = 0$ implies $y' + py = 0$. The proof is complete.

(b) We prove $y' + py = r$ when y is replaced by the fraction y_p. Define $C(x) = \int r(x)W(x)dx$, so that $y = C(x)/W(x)$. The fundamental theorem of calculus implies $C'(x) = r(x)W(x)$. The product rule and the Lemma imply $C' = (yW)' = y'W + yW' = y'W + ypW = (y' + py)W$. Competition between the two equations for C' gives $rW = (y' + py)W)$. Cancel W to obtain $r = y' + py$. ■

Historical Note. The formula for $y_p(x)$ has the historical name **variation of constants** or **variation of parameters**. Both y_h and y_p have the same form C/W, with $C(x)$ constant for y_h and $C(x)$ equal to a function of x for y_p: **variation of constant** c in y_h produces the expression for y_p.

Experimental Viewpoint. The particular solution y_p depends on the forcing term $r(x)$, but the homogeneous solution y_h does not. Experimentalists view the computation of y_p as a *single experiment* in which the state y_p is determined by the forcing term $r(x)$ and zero initial data $y = 0$ at $x = x_0$. This particular experimental solution y_p^* is given by the definite integral formula

$$y_p^*(x) = \frac{1}{W(x)} \int_{x_0}^{x} r(x)W(x)dx. \tag{2}$$

Superposition. The sum of constant multiples of solutions to $y' + p(x)y = 0$ is again a solution. The next two theorems are **superposition** for $y'+p(x)y = r(x)$.

Theorem 2.6 (General Solution = Homogeneous + Particular)
Assume $p(x)$ and $r(x)$ are continuous on $a < x < b$ and $a < x_0 < b$. Let y be a solution of $y' + p(x)y = r(x)$ on $a < x < b$. Then y can be decomposed as $y = y_h + y_p$.

In short, a linear equation has the solution structure *homogeneous plus particular*.

The constant c in formula y_h and the integration constant in $\int W(x)rx)dx$ can always be selected to satisfy initial condition $y(x_0) = y_0$.

Theorem 2.7 (Difference of Solutions = Homogeneous Solution)
Assume $p(x)$ and $r(x)$ are continuous on $a < x < b$ and $a < x_0 < b$. Let y_1 and y_2 be two solutions of $y' + p(x)y = r(x)$ on $a < x < b$. Then $y = y_1 - y_2$ is a solution of the homogeneous differential equation

$$y' + p(x)y = 0.$$

In short, any two solutions of the non-homogeneous equation differ by some solution y_h of the homogeneous equation.

Integrating Factor Method

The technique called the **method of integrating factors** uses the replacement rule (justified on page 101)

$$(3) \qquad \text{Fraction } \frac{(YW)'}{W} \text{ replaces } Y' + p(x)Y, \text{ where } W = e^{\int p(x)dx}.$$

The fraction $(YW)'/W$ is called the **integrating factor fraction**.

The Integrating Factor Method

Standard Form Rewrite $y' = f(x,y)$ in the form $y' + p(x)y = r(x)$ where p, r are continuous. The method applies only in case this is possible.

Find W Find a simplified formula for $W = e^{\int p(x)dx}$. The antiderivative $\int p(x)dx$ can be chosen conveniently.

Prepare for Quadrature Obtain the new equation $\dfrac{(yW)'}{W} = r$ by replacing the left side of $y' + p(x)y = r(x)$ by equivalence (3).

Method of Quadrature Clear fractions to obtain $(yW)' = rW$. Apply the method of quadrature to get $yW = \int r(x)W(x)dx + C$. Divide by W to isolate the explicit solution $y(x)$.

In identity (3), functions p, Y and Y' are assumed continuous with p and Y *arbitrary* functions. Equation (3) is central to the method, because it collapses the two terms $y' + py$ into a single term $(Wy)'/W$; the method of quadrature applies to $(Wy)' = rW$. The literature calls the exponential factor W an **integrating factor** and equivalence (3) a **factorization** of $Y' + p(x)Y$.

Simplifying an Integrating Factor

Factor W is simplified by dropping constants of integration. To illustrate, if $p(x) = 1/x$, then $\int p(x)dx = \ln|x| + C$. The algebra rule $e^{A+B} = e^A e^B$ implies that $W = e^C e^{\ln|x|} = |x|e^C = (\pm e^C)x$, because $|x| = (\pm)x$. Let $c_1 = \pm e^C$. Then $W = c_1 W_1$ where $W_1 = x$. The fraction $(Wy)'/W$ reduces to $(W_1 y)'/W_1$, because c_1 cancels. In an application, we choose the simpler expression W_1. The illustration shows that exponentials in W can sometimes be eliminated.

Variation of Constants and $y' + p(x)y = r(x)$

Every solution of $y' + p(x)y = r(x)$ can be expressed as $y = y_h + y_p$, by choosing constants appropriately. The classical **variation of constants formula** puts initial condition zero on y_p and compresses all initial data into the constant c appearing in y_h. The general solution is given by

$$y(x) = \frac{y(x_0)}{W(x)} + \frac{\int_{x_0}^{x} r(x)W(x)dx}{W(x)}, \quad W(x) = e^{\int_{x_0}^{x} p(s)ds} \tag{4}$$

Classifying Linear and Non-Linear Equations

Definition 2.6 (Non-linear Differential Equation)
An equation $y' = f(x, y)$ that fails to be linear is called **non-linear**.

Algebraic Complexity. A linear equation $y' = f(x, y)$ may appear to be non-linear, e.g., $y' = (\sin^2(xy) + \cos^2(xy))y$ simplifies to $y' = y$.

Computer Algebra System. These systems classify an equation $y' = f(x, y)$ as linear provided the identity $f(x, y) = f(x, 0) + f_y(x, 0)y$ is valid. Equivalently, $f(x, y) = r(x) - p(x)y$, where $r(x) = f(x, 0)$ and $p(x) = -f_y(x, y)$.

Hand verification can use the same method. To illustrate, consider $y' = f(x, y)$ with $f(x, y) = (x - y)(x + y) + y(y - 2x)$. Compute $f(x, 0) = x^2$, $f_y(x, 0) = -2x$. Because f_y is independent of y, then $y' = f(x, y)$ is the linear equation $y' + p(x)y = r(x)$ with $p(x) = 2x$, $r(x) = x^2$.

Non-Linear Equation Tests. Elimination of an equation $y' = f(x, y)$ from the class of linear equations can be done from *necessary conditions.* The equality $f_y(x, y) = f_y(x, 0)$ implies two such conditions:

1. If $f_y(x, y)$ depends on y, then $y' = f(x, y)$ is not linear.
2. If $f_{yy}(x, y) \neq 0$, then $y' = f(x, y)$ is not linear.

For instance, either condition implies $y' = 1 + y^2$ is *not linear.*

Special Linear Equations

There are fast ways to solve certain linear differential equations that do not employ the linear integrating factor method.

Theorem 2.8 (Solving a Homogeneous Equation)
Assume $p(x)$ is continuous on $a < x < b$. Then the solution of the homogeneous differential equation $y' + p(x)y = 0$ is given by the formula

$$y(x) = \frac{\text{constant}}{\text{integrating factor}}. \tag{5}$$

Theorem 2.9 (Solving a Constant-Coefficient Equation)
Assume $p(x)$ and $r(x)$ are constants p, r with $p \neq 0$. Then the solution of the constant-coefficient differential equation $y' + py = r$ is given by the formula

$$\begin{aligned} y(x) &= \frac{\text{constant}}{\text{integrating factor}} + \text{equilibrium solution} \\ &= ce^{-px} + \frac{r}{p}. \end{aligned} \tag{6}$$

Proof: The homogeneous solution is a constant divided by the integrating factor, by Theorem 2.8. An equilibrium solution can be found by formally setting $y' = 0$, then solving for $y = r/p$. By superposition Theorem 2.6, the solution y must be the sum of these two solutions. The excluded case $p = 0$ results in a quadrature equation $y' = r$ which is routinely solved by the method of quadrature.

Examples

Example 2.12 (Shortcut: Homogeneous Equation)

Solve the homogeneous equation $2y' + x^2y = 0$.

Solution: By Theorem (2.8), the solution is a constant divided by the integrating factor. First, divide by 2 to get $y' + p(x)y = 0$ with $p(x) = \frac{1}{2}x^2$. Then $\int p(x)dx = x^3/6 + c$ implies $W = e^{x^3/6}$ is an integrating factor. The solution is $y = \dfrac{c}{e^{x^3/6}}$.

Example 2.13 (Shortcut: Constant-Coefficient Equation)

Solve the non-homogeneous constant-coefficient equation $2y' - 5y = -1$.

Solution: The method described here only works for first order constant coefficient differential equations. If $y' = f(x, y)$ is not linear or it fails to have constant coefficients, then the method fails.

The solution has two steps:

(1) Find the solution y_h of the homogeneous equation $2y' - 5y = 0$.
The answer is a constant divided by the integrating factor, which is $y = \dfrac{c}{e^{-5x/2}}$. First divide the equation by 2 to obtain the standard form $y' + (-5/2)y = 0$. Identify $p(x) = -5/2$, then $\int p(x)dx = -5x/2 + c$ and finally $W = e^{-5x/2}$ is the integrating factor. The answer is $y_h = c/W = ce^{5x/2}$.

(2) Find an equilibrium solution y_p for $2y' - 5y = -1$.
This answer is found by formally replacing y' by zero. Then $y_p = \frac{1}{5}$.

The answer is the sum of the answers from (1) and (2), by superposition, giving

$$y = y_h + y_p = ce^{5x/2} + \frac{1}{5}.$$

The method of this example is called the **superposition method shortcut**.

Example 2.14 (Integrating Factor Method)

Solve $2y' + 6y = e^{-x}$.

Solution: The solution is $y = \frac{1}{4}e^{-x} + ce^{-3x}$. An answer check appears in Example 2.16. The details:

$y' + 3y = 0.5e^{-x}$	Divide by 2 to get the standard form.
$W = e^{3x}$	Find the integrating factor $W = e^{\int 3dx}$.
$\dfrac{\left(e^{3x}y\right)'}{e^{3x}} = 0.5e^{-x}$	Replace the LHS of $y' + 3y = 0.5e^{-x}$ by the integrating factor quotient; see page 96.
$\left(e^{3x}y\right)' = 0.5e^{2x}$	Clear fractions. Prepared for quadrature
$e^{3x}y = 0.5\int e^{2x}dx$	Method of quadrature applied.
$y = 0.5\left(e^{2x}/2 + c_1\right)e^{-3x}$	Evaluate the integral. Divide by $W = e^{3x}$.
$= \frac{1}{4}e^{-x} + ce^{-3x}$	Final answer, $c = 0.5c_1$.

Example 2.15 (Superposition)

Find a particular solution of $y' + 2y = 3e^{x}$ with fewest terms.

Solution: The answer is $y = e^{x}$. The first step solves the equation using the integrating factor method, giving $y = e^{x} + ce^{-2x}$; details below. A particular solution with fewest terms, $y = e^{x}$, is found by setting $c = 0$.

Integrating factor method details:

$y' + 2y = 3e^{x}$	The standard form.
$W = e^{2x}$	Find the integrating factor $W = e^{\int 2dx}$.
$\dfrac{\left(e^{2x}y\right)'}{e^{2x}} = 3e^{x}$	Integrating factor identity applied to $y' + 2y = 3e^{x}$.
$e^{2x}y = 3\int e^{3x}dx$	Clear fractions and apply quadrature.
$y = \left(e^{3x} + c\right)e^{-2x}$	Evaluate the integral. Isolate y.

$= e^x + ce^{-2x}$	Solution found.

Remarks on Integral Formula (2). Computer algebra systems will compute the solution $y_p^* = e^x - e^{3x_0}e^{-2x}$ of equation (2). It has an extra term because of the condition $y = 0$ at $x = x_0$. The shortest particular solution e^x and the integral formula solution y_p^* differ by a homogeneous solution c_1e^{-2x}, where $c_1 = e^{3x_0}$. To shorten y_p^* to $y_p = e^x$ requires knowing the homogeneous solution, then apply superposition $y = y_p + y_h$ to extract a particular solution.

Example 2.16 (Answer Check)

Show the answer check details for $2y' + 6y = e^{-x}$ and candidate solution $y = \frac{1}{4}e^{-x} + ce^{-3x}$.

Solution: **Details**:

$\text{LHS} = 2y' + 6y$	Left side of the equation $2y'+6y = e^{-x}$.
$= 2(-\frac{1}{4}e^{-x} - 3ce^{-3x}) + 6(\frac{1}{4}e^{-x} + ce^{-3x})$	Substitute for y.
$= e^{-x} + 0$	Simplify terms.
$= \text{RHS}$	DE verified.

Example 2.17 (Finding y_h and y_p)

Find the homogeneous solution y_h and a particular solution y_p for the equation $2xy' + y = 4x^2$ on $x > 0$.

Solution: The solution by the integrating factor method is $y = 0.8x^2 + cx^{-1/2}$; details below. Then $y_h = cx^{-1/2}$ and $y_p = 0.8x^2$ give $y = y_h + y_p$.

The symbol y_p stands for *any* particular solution. It should be free of any arbitrary constants c.

Integral formula (2) gives a particular solution $y_p^* = 0.8x^2 - 0.8x_0^{5/2}x^{-1/2}$. It differs from the shortest particular solution $0.8x^2$ by a homogeneous solution $Kx^{-1/2}$.

Integrating factor method details:

$y' + 0.5y/x = 2x$	Standard form. Divided by $2x$.
$p(x) = 0.5/x$	Identify coefficient of y. Then $\int p(x)dx = 0.5\ln\lvert x\rvert + c$.
$W = e^{0.5\ln\lvert x\rvert + c}$	The integrating factor is $W = e^{\int p}$.
$W = e^{0.5\ln\lvert x\rvert}$	Choose integration constant zero.
$= \lvert x\rvert^{1/2}$	Used $\ln u^n = n\ln u$. Simplified W found.
$\dfrac{(x^{1/2}y)'}{x^{1/2}} = 2x$	Integrating factor identity applied on the left. Assumed $x > 0$.
$x^{1/2}y = 2\int x^{3/2}dx$	Clear fractions. Apply quadrature.
$y = \left(4x^{5/2}/5 + c\right)x^{-1/2}$	Evaluate the integral. Divide to isolate y.
$= \frac{4}{5}x^2 + cx^{-1/2}$	Solution found.

Example 2.18 (Classification)
Classify the equation $y' = x + \ln(xe^y)$ as linear or non-linear.

Solution: It's linear, with standard linear form $y' + (-1)y = x + \ln x$. To explain why, the term $\ln(xe^y)$ on the right expands into $\ln x + \ln e^y$, which in turn is $\ln x + y$, using logarithm rules. Because $e^y > 0$, then $\ln(xe^y)$ makes sense for only $x > 0$. Henceforth, assume $x > 0$.

Computer algebra test $f(x, y) = f(x, 0) + f_y(x, 0)y$. Expected is LHS$-$RHS $= 0$ after simplification. This example produced $\ln e^y - y$ instead of 0, evidence that limitations may exist.

```
assume(x>0):
f:=(x,y)->x+ln(x*exp(y)):
LHS:=f(x,y):
RHS:=f(x,0)+subs(y=0,diff(f(x,y),y))*y:
simplify(LHS-RHS);
```

If the test *passes*, then $y' = f(x, y)$ becomes $y' = f(x, 0) + f_y(x, 0)y$. This example gives $y' = x + \ln x + y$, which converts to the standard linear form $y' + (-1)y = x + \ln x$.

Details and Proofs

Justification of Homogeneous Solution $y = \dfrac{c}{W(x)}$:

Because $W = e^{\int p(x)dx}$, then $W' = p(x)W$ by the Fundamental Theorem of Calculus. Then $(e^u)' = u'e^u$ implies:

$$\frac{dy}{dx} + p(x)y = \frac{-cW'}{W^2} + \frac{cp(x)}{W} = \frac{-cp(x)W}{W^2} + \frac{cp(x)}{W} = 0$$

Justification of Factorization (3): It is assumed that $Y(x)$ is a given but otherwise arbitrary differentiable function. Equation (3) will be justified in its fraction-free form

$$(7) \qquad \left(Ye^{\mathbf{P}}\right)' = (Y' + pY)\,e^{\mathbf{P}}, \quad \mathbf{P}(x) = \int p(x)dx.$$

$$\begin{aligned}
\text{LHS} &= \left(Ye^{\mathbf{P}}\right)' && \text{The left side of equation (7).}\\
&= Y'e^{\mathbf{P}} + \left(e^{\mathbf{P}}\right)'Y && \text{Apply the product rule } (uv)' = u'v + uv'.\\
&= Y'e^{\mathbf{P}} + pe^{\mathbf{P}}Y && \text{Use the chain rule } (e^u)' = u'e^u \text{ and } \mathbf{P}' = p.\\
&= (Y' + pY)\,e^{\mathbf{P}} && \text{The common factor is } e^{\mathbf{P}}.\\
&= \text{RHS} && \text{The right hand side of equation (7).}
\end{aligned}$$

Justification of Formula (4):

Existence. Because the formula is $y = y_h + y_p$ for particular values of c and the constant of integration, then y is a solution by superposition Theorem (2.6) and existence Theorem (2.5).

Uniqueness. It remains to show that the solution given by (4) is the only solution. Start by assuming Y is another, subtract them to obtain $u = y - Y$. Then $u' + pu = 0$, $u(x_0) = 0$. To show $y \equiv Y$, it suffices to show $u \equiv 0$.

According to the integrating factor method, the equation $u' + pu = 0$ is equivalent to $(uW)' = 0$. Integrate $(uW)' = 0$ from x_0 to x, giving $u(x)W(x) = u(x_0)W(x_0)$. Since $u(x_0) = 0$ and $W(x) \neq 0$, it follows that $u(x) = 0$ for all x. ■

About Picard's Theorem. The Picard-Lindelöf theorem, page 68, implies existence-uniqueness, but only on a smaller interval, and furthermore it supplies no practical formula for the solution. Formula (4) is therefore an improvement over the results obtainable from the general theory.

Exercises 2.3

Integrating Factor Method

Apply the integrating factor method, page 96, to solve the given linear equation. See the examples starting on page 99 for details.

1. $y' + y = e^{-x}$

2. $y' + y = e^{-2x}$

3. $2y' + y = e^{-x}$

4. $2y' + y = e^{-2x}$

5. $2y' + y = 1$

6. $3y' + 2y = 2$

7. $2xy' + y = x$

8. $3xy' + y = 3x$

9. $y' + 2y = e^{2x}$

10. $2y' + y = 2e^{x/2}$

11. $y' + 2y = e^{-2x}$

12. $y' + 4y = e^{-4x}$

13. $2y' + y = e^{-x}$

14. $2y' + y = e^{-2x}$

15. $4y' + y = 1$

16. $4y' + 2y = 3$

17. $2xy' + y = 2x$

18. $3xy' + y = 4x$

19. $y' + 2y = e^{-x}$

20. $2y' + y = 2e^{-x}$

Superposition

Find a particular solution with fewest terms. See Example 2.15, page 99.

21. $3y' = x$

22. $3y' = 2x$

23. $y' + y = 1$

24. $y' + 2y = 2$

25. $2y' + y = 1$

26. $3y' + 2y = 1$

27. $y' - y = e^x$

28. $y' - y = xe^x$

29. $xy' + y = \sin x$ $(x > 0)$

30. $xy' + y = \cos x$ $(x > 0)$

31. $y' + y = x - x^2$

32. $y' + y = x + x^2$

General Solution

Find y_h and a particular solution y_p. Report the general solution $y = y_h + y_p$. See Example 2.17, page 100.

33. $y' + y = 1$

34. $xy' + y = 2$

35. $y' + y = x$

36. $xy' + y = 2x$

37. $y' - y = x + 1$

38. $xy' - y = 2x - 1$

39. $2xy' + y = 2x^2 \ (x > 0)$

40. $xy' + y = 2x^2 \ (x > 0)$

Classification

Classify as linear or non-linear. Use the test $f(x, y) = f(x, 0) + f_y(x, 0)y$ and a computer algebra system, when available, to check the answer. See Example 2.18, page 101.

41. $y' = 1 + 2y^2$

42. $y' = 1 + 2y^3$

43. $yy' = (1 + x)\ln e^y$

44. $yy' = (1 + x)\,(\ln e^y)^2$

45. $y' \sec^2 y = 1 + \tan^2 y$

46. $y' = \cos^2(xy) + \sin^2(xy)$

47. $y'(1 + y) = xy$

48. $y' = y(1 + y)$

49. $xy' = (x + 1)y - xe^{\ln y}$

50. $2xy' = (2x + 1)y - xye^{-\ln y}$

Shortcuts

Apply theorems for the homogeneous equation $y' + p(x)y = 0$ or for constant coefficient equations $y' + py = r$. Solutions should be done without paper or pencil, then write the answer and check it.

51. $y' - 5y = -1$

52. $3y' - 5y = -1$

53. $2y' + xy = 0$

54. $3y' - x^2y = 0$

55. $y' = 3x^4y$

56. $y' = (1 + x^2)y$

57. $\pi y' - \pi^2 y = -e^2$

58. $e^2y' + e^3y = \pi^2$

59. $xy' = (1 + x^2)y$

60. $e^xy' = (1 + e^{2x})y$

Proofs and Details

61. Prove directly without appeal to Theorem 2.6 that the difference of two solutions of $y' + p(x)y = r(x)$ is a solution of the homogeneous equation $y' + p(x)y = 0$.

62. Prove that y_p^* given by equation (2) and $y_p = W^{-1}\int r(x)W(x)dx$ given in the integrating factor method are related by $y_p = y_p^* + y_h$ for some solution y_h of the homogeneous equation.

63. The equation $y' = r$ with r constant can be solved by quadrature, without pencil and paper. Find y.

64. The equation $y' = r(x)$ with $r(x)$ continuous can be solved by quadrature. Find a formula for y.

2.4 Undetermined Coefficients

Studied here is the subject of undetermined coefficients for linear first order differential equations $y' + p(x)y = r(x)$. It finds a particular solution y_p *without* the integration steps present in variation of parameters (reviewed in an example and in exercises). The requirements and limitations:

> **1**. Coefficient $p(x)$ of $y' + p(x)y = r(x)$ is constant.
>
> **2**. The function $r(x)$ is a sum of constants times Euler solution atoms (defined below).

Definition 2.7 (Euler Solution Atom)
An **Euler base atom** is a term having one of the forms

$$1,\ e^{ax},\ \cos bx,\ \sin bx,\ e^{ax}\cos bx \quad \text{or} \quad e^{ax}\sin bx.$$

The symbols a and b are real constants, with $a \neq 0$ and $b > 0$.

An **Euler solution atom** equals x^n(Euler base atom). Symbol $n \geq 0$ is an integer.

Examples. The terms x^3, $x\cos 2x$, $\sin x$, e^{-x}, $x^6 e^{-\pi x}$ are Euler atoms. Conversely, if $r(x) = 4\sin x + 5xe^x$, then split the sum into terms and drop the coefficients 4 and 5 to identify Euler atoms $\sin x$ and xe^x; then $r(x)$ is a sum of constants times Euler solution atoms.

The Method

1. Repeatedly differentiate the Euler atoms in $r(x)$ until no new atoms appear. Multiply the distinct atoms so found by **undetermined coefficients** d_1, ..., d_k, then add to define a **trial solution** y.
2. **Correction rule**: if solution e^{-px} of $y' + py = 0$ appears in trial solution y, then replace in y matching Euler atoms e^{-px}, xe^{-px}, ... by xe^{-px}, x^2e^{-px}, ...(other Euler atoms in y are unchanged). The modified expression y is called the **corrected trial solution**.
3. Substitute y into the differential equation $y'+py = r(x)$. Match coefficients of Euler atoms left and right to write out linear algebraic equations for the undetermined coefficients d_1, ..., d_k.
4. Solve the equations. The trial solution y with evaluated coefficients d_1, ..., d_k becomes the particular solution y_p.

Undetermined Coefficients Illustrated

Solve

$$y' + 2y = xe^x + 2x + 1 + 3\sin x.$$

Solution:

Test Applicability. The right side $r(x) = xe^x + 2x + 1 + 3\sin x$ is a sum of terms constructed from the Euler atoms xe^x, x, 1, $\sin x$. The left side is $y' + p(x)y$ with $p(x) = 2$, a constant. Therefore, the method of undetermined coefficients applies to find y_p.

Trial Solution. The atoms of $r(x)$ are subjected to differentiation. The distinct Euler atoms so found are 1, x, e^x, xe^x, $\cos x$, $\sin x$ (split terms and drop coefficients to identify new atoms). Because the solution e^{-2x} of $y'+2y = 0$ does not appear in the list of atoms, then the correction rule does not apply. The corrected trial solution is the expression

$$y = d_1(1) + d_2(x) + d_3(e^x) + d_4(xe^x) + d_5(\cos x) + d_6(\sin x).$$

Equations for Undetermined Coefficients. To substitute the trial solution y into $y' + 2y$ requires a formula for y':

$$y' = d_2 + d_3e^x + d_4xe^x + d_4e^x - d_5 \sin x + d_6 \cos x.$$

Then

$$\begin{aligned} r(x) &= y' + 2y \\ &= d_2 + d_3e^x + d_4xe^x + d_4e^x - d_5 \sin x + d_6 \cos x \\ &\quad + 2d_1 + 2d_2x + 2d_3e^x + 2d_4xe^x + 2d_5 \cos x + 2d_6 \sin x \\ &= (d_2 + 2d_1)(1) + 2d_2(x) + (3d_3 + d_4)(e^x) + (3d_4)(xe^x) \\ &\quad + (2d_5 + d_6)(\cos x) + (2d_6 - d_5)(\sin x) \end{aligned}$$

Also, $r(x) \equiv 1 + 2x + xe^x + 3\sin x$. Coefficients of atoms on the left and right must match. For instance, constant term 1 in $r(x)$ matches the constant term in the expansion of $y' + 2y$, giving $1 = d_2 + 2d_1$. Writing out the matches, and swapping sides, gives the equations

$$\begin{array}{rrrrrrl} 2d_1 + & d_2 & & & & & = 1, \\ & 2d_2 & & & & & = 2, \\ & & 3d_3 + & d_4 & & & = 0, \\ & & & 3d_4 & & & = 1, \\ & & & & 2d_5 + & d_6 & = 0, \\ & & & & - \; d_5 + & 2d_6 & = 3. \end{array}$$

Solve. The first four equations can be solved by back-substitution to give $d_2 = 1$, $d_1 = 0$, $d_4 = 1/3$, $d_3 = -1/9$. The last two equations are solved by elimination or Cramer's rule (reviewed in Chapter 3) to give $d_6 = 6/5$, $d_5 = -3/5$.

Report y_p. The trial solution y with evaluated coefficients $d_1, \ldots, d_6$ becomes

$$y_p(x) = x - \frac{1}{9}e^x + \frac{1}{3}xe^x - \frac{3}{5}\cos x + \frac{6}{5}\sin x.$$

Remarks. The method of matching coefficients of atoms left and right is a subject of linear algebra, called *linear independence*. The method works because any finite list of atoms is known to be linearly independent. Further details for this technical topic appear in this text's linear algebra chapters.

A Correction Rule Illustration

Solve the equation

$$y' + 3y = 8e^x + 3x^2e^{-3x}$$

by the method of undetermined coefficients. Verify that the general solution $y = y_h + y_p$ is given by

$$y_h = ce^{-3x}, \quad y_p = 2e^x + x^3e^{-3x}.$$

Solution: The right side $r(x) = 8e^x + 3x^2e^{-3x}$ is constructed from atoms e^x, x^2e^{-3x}. Repeated differentiation of these atoms identifies the new list of atoms e^x, e^{-3x}, xe^{-3x}, x^2e^{-3x}. The correction rule applies because the solution e^{-3x} of $y' + 3y = 0$ appears in the list. The atoms of the form x^me^{-3x} are multiplied by x to give the new list of atoms e^x, xe^{-3x}, x^2e^{-3x}, x^3e^{-3x}. Readers should take note that atom e^x is unaffected by the correction rule modification. Then the corrected trial solution is

$$y = d_1e^x + d_2xe^{-3x} + d_3x^2e^{-3x} + d_4x^3e^{-3x}.$$

The trial solution expression y is substituted into $y' + 3y = 2e^x + x^2e^{-3x}$ to give the equation

$$4d_1e^x + d_2e^{-3x} + 2d_3xe^{-3x} + 3d_4x^2e^{-3x} = 8e^x + 3x^2e^{-3x}.$$

Coefficients of atoms on each side of the preceding equation are matched to give the equations

$$\begin{aligned} 4d_1 \quad\quad\quad\quad &= 8, \\ d_2 \quad\quad\quad &= 0, \\ 2d_3 \quad &= 0, \\ 3d_4 &= 3. \end{aligned}$$

Then $d_1 = 2$, $d_2 = d_3 = 0$, $d_4 = 1$ and the particular solution is reported to be $y_p = 2e^x + x^3e^{-3x}$.

Remarks on the Method of Undetermined Coefficients

A mystery for the novice is the construction of the trial solution. **Why should it work**? Explained here is the reason behind the method of repeated differentiation to find the Euler atoms in the trial solution.

The theory missing is that the general solution y of $y' + py = r(x)$ is a sum of constants times Euler atoms (under the cited **limitations**). We don't try to prove this result, but use it to motivate the method.

The theory reduces the question of finding a trial solution to finding a sum of constants times Euler atoms. The question is: *which atoms*?

Consider this example: $y' - 3y = e^{3x} + xe^x$. The answer for y is revealed by finding a sum of constants times atoms such that y' and $-3y$ add termwise to $e^{3x} + xe^x$. The requirement eliminates all atoms from consideration except those containing exponentials e^{3x} and e^x.

Initially, we have to consider infinitely many atoms e^{3x}, xe^{3x}, x^2e^{3x}, …and e^x, xe^x, x^2e^x, …. Such terms would also appear in y', but adding terms of this type to get $r(x) = e^{3x} + xe^x$ requires only the smaller list e^{3x}, xe^{3x}, e^x, xe^x. We have cut down the number of terms in y to four or less!

The algorithm presented here together with the correction rule strips down the number of terms to a minimum. Further details of the method appear in the chapter on scalar linear differential equations, page 430.

Examples

Example 2.19 (Variation of Parameters Method)

Solve the equation $2y' + 6y = 4xe^{-3x}$ by the method of variation of parameters, verifying $y = y_h + y_p$ is given by

$$y_h = ce^{-3x}, \quad y_p = x^2e^{-3x}.$$

Solution: Divide the equation by 2 to obtain the standard linear form

$$y' + 3y = 2xe^{-3x}.$$

Solution y_h. The homogeneous equation $y' + 3y = 0$ is solved by the shortcut formula $y_h = \frac{\text{constant}}{\text{integrating factor}}$ to give $y_h = ce^{-3x}$.

Solution y_p. Identify $p(x) = 3$, $r(x) = 2xe^{-3x}$ from the standard form. The mechanics: let $y' = f(x,y) \equiv 2xe^{-3x} - 3y$ and define $r(x) = f(x,0)$, $p(x) = -f_y(x,y) = 3$. The variation of parameters formula is applied as follows. First, compute the integrating factor $W(x) = e^{\int p(x)dx} = e^{3x}$. Then

$$\begin{aligned} y_p(x) &= (1/W(x)) \int r(x)W(x)dx \\ &= e^{-3x} \int 2xe^{-3x}e^{3x}dx \\ &= x^2e^{-3x}. \end{aligned}$$

It must be explained that all integration constants were set to zero, in order to obtain the shortest possible expression for y_p. Indeed, if $W = e^{3x+c_1}$ instead of e^{3x}, then the factors $1/W$ and W contribute constant factors $1/e^{c_1}$ and e^{c_1}, which multiply to one; the effect is to set $c_1 = 0$. On the other hand, an integration constant c_2 added to $\int r(x)W(x)dx$ adds the homogeneous solution c_2e^{-3x} to the expression for y_p. Because we seek the shortest expression which is a solution to the non-homogeneous differential equation, the constant c_2 is set to zero.

Example 2.20 (Undetermined Coefficient Method)

Solve the equation $2y' + 6y = 4xe^{-x} + 4xe^{-3x} + 5\sin x$ by the method of undetermined coefficients, verifying $y = y_h + y_p$ is given by

$$y_h = ce^{-3x}, \quad y_p = -\frac{1}{2}e^{-x} + xe^{-x} + x^2e^{-3x} - \frac{1}{4}\cos x + \frac{3}{4}\sin x.$$

Solution: The method applies, because the differential equation $2y' + 6y = 0$ has constant coefficients and the right side $r(x) = 4xe^{-x} + 4xe^{-3x} + 5\sin x$ is constructed from the list of atoms xe^{-x}, xe^{-3x}, $\sin x$.

List of Atoms. Differentiate the atoms in $r(x)$, namely xe^{-x}, xe^{-3x}, $\sin x$, to find the new list of atoms e^{-x}, xe^{-x}, e^{-3x}, xe^{-3x}, $\cos x$, $\sin x$. The solution e^{-3x} of $2y' + 6y = 0$ appears in the list: the correction rule applies. Then e^{-3x}, xe^{-3x} are replaced by xe^{-3x}, x^2e^{-3x} to give the corrected list of atoms e^{-x}, xe^{-x}, xe^{-3x}, x^2e^{-3x}, $\cos x$, $\sin x$. Please note that only two of the six atoms were corrected.

Trial solution. The corrected trial solution is

$$y = d_1e^{-x} + d_2xe^{-x} + d_3xe^{-3x} + d_4x^2e^{-3x} + d_5 \cos x + d_6 \sin x.$$

Substitute y into $2y' + 6y = r(x)$ to give

$$\begin{array}{rcl} r(x) & = & 2y' + 6y \\ & = & (4d_1 + 2d_2)e^{-x} + 4d_2xe^{-x} + 2d_3e^{-3x} + 4d_4xe^{-3x} \\ & & +(2d_6 + 6d_5) \cos x + (6d_6 - 2d_5) \sin x. \end{array}$$

Equations. Matching atoms on the left and right of $2y' + 6y = r(x)$, given $r(x) = 4xe^{-x} + 4xe^{-3x} + 5 \sin x$, justifies the following equations for the undetermined coefficients; the solution is $d_2 = 1$, $d_1 = -1/2$, $d_3 = 0$, $d_4 = 1$, $d_6 = 3/4$, $d_5 = -1/4$.

$$\begin{aligned} 4d_1 + 2d_2 \qquad\qquad\qquad\qquad &= 0, \\ 4d_2 \qquad\qquad\qquad\qquad &= 4, \\ 2d_3 \qquad\qquad\qquad &= 0, \\ 4d_4 \qquad\qquad &= 4, \\ 6d_5 + 2d_6 &= 0, \\ -\, 2d_5 + 6d_6 &= 5. \end{aligned}$$

Equations for variables d_5, d_6 were generated from trigonometric atoms. The 2×2 system has complex eigenvalues. The best method to find coefficients d_5, d_6 is not Gaussian elimination, but instead Cramer's Rule.

Report. The trial solution upon substitution of the values for the undetermined coefficients becomes

$$y_p = -\frac{1}{2}e^{-x} + xe^{-x} + x^2e^{-3x} - \frac{1}{4}\cos x + \frac{3}{4}\sin x.$$

Exercises 2.4

Variation of Parameters I

Report the shortest particular solution given by the formula

$$y_p(x) = \frac{\int rW}{W}, \quad W = e^{\int p(x)dx}$$

1. $y' = x + 1$

2. $y' = 2x - 1$

3. $y' + y = e^{-x}$

4. $y' + y = e^{-2x}$

5. $y' - 2y = 1$

6. $y' - y = 1$

7. $2y' + y = e^x$

8. $2y' + y = e^{-x}$

9. $xy' = x + 1$

10. $xy' = 1 - x^2$

Variation of Parameters II

Define $W(t) = e^{\int_{x_0}^t p(x)dx}$. Compute $y_p^*(x) = \frac{\int_{x_0}^x r(t)W(t)\,dt}{W(x)}$

11. $y' = x + 1$, $y(0) = 0$

12. $y' = 2x - 1$, $x_0 = 0$

13. $y' + y = e^{-x}$, $x_0 = 0$

14. $y' + y = e^{-2x}$, $x_0 = 0$

15. $y' - 2y = 1$, $x_0 = 0$

16. $y' - y = 1$, $x_0 = 0$

17. $2y' + y = e^{x}$, $x_0 = 0$

18. $2y' + y = e^{-x}$, $x_0 = 0$

19. $xy' = x + 2$, $x_0 = 1$

20. $xy' = 1 - x^2$, $x_0 = 1$

Euler Solution Atoms

Report the list L of distinct Euler solution atoms found in function $f(x)$. Then $f(x)$ is a sum of constants times the Euler atoms from L.

21. $x + e^x$

22. $1 + 2x + 5e^x$

23. $x(1 + x + 2e^x)$

24. $x^2(2 + x^2) + x^2e^{-x}$

25. $\sin x \cos x + e^x \sin 2x$

26. $\cos^2 x - \sin^2 x + x^2 e^x \cos 2x$

27. $(1 + 2x + 4x^5)e^x e^{-3x} e^{x/2}$

28. $(1 + 2x + 4x^5 + e^x \sin 2x)e^{-3x/4}e^{x/2}$

29. $\dfrac{x + e^x}{e^{-2x}} \sin 3x + e^{3x} \cos 3x$

30. $\dfrac{x + e^x \sin 2x + x^3}{e^{-2x}} \sin 5x$

Initial Trial Solution

Differentiate repeatedly $f(x)$ and report the list M of distinct Euler solution atoms which appear in f and all its derivatives. Then each of $f, f', \ldots$ is a sum of constants times Euler atoms in M.

31. $12 + 5x^2 + 6x^7$

32. $x^6/x^{-4} + 10x^4/x^{-6}$

33. $x^2 + e^x$

34. $x^3 + 5e^{2x}$

35. $(1 + x + x^3)e^x + \cos 2x$

36. $(x + e^x)\sin x + (x - e^{-x})\cos 2x$

37. $(x + e^x + \sin 3x + \cos 2x)e^{-2x}$

38. $(x^2e^{-x} + 4\cos 3x + 5\sin 2x)e^{-3x}$

39. $(1 + x^2)(\sin x \cos x - \sin 2x)e^{-x}$

40. $(8 - x^3)(\cos^2 x - \sin^2 x)e^{3x}$

Correction Rule

Given the homogeneous solution y_h and an initial trial solution y, determine the final trial solution according to the correction rule.

41. $y_h(x) = ce^{2x}$, $y = d_1 + d_2x + d_3e^{2x}$

42. $y_h(x) = ce^{2x}$, $y = d_1 + d_2e^{2x} + d_3xe^{2x}$

43. $y_h(x) = ce^{0x}$, $y = d_1 + d_2x + d_3x^2$

44. $y_h(x) = ce^{x}$, $y = d_1 + d_2x + d_3x^2$

45. $y_h(x) = ce^{x}$, $y = d_1 \cos x + d_2 \sin x + d_3e^x$

46. $y_h(x) = ce^{2x}$, $y = d_1e^{2x}\cos x + d_2e^{2x}\sin x$

47. $y_h(x) = ce^{2x}$, $y = d_1e^{2x} + d_2xe^{2x} + d_3x^2e^{2x}$

48. $y_h(x) = ce^{-2x}$, $y = d_1e^{-2x} + d_2xe^{-2x} + d_3e^{2x} + d_4xe^{2x}$

49. $y_h(x) = cx^2$, $y = d_1 + d_2x + d_3x^2$

50. $y_h(x) = cx^3$, $y = d_1 + d_2x + d_3x^2$

Trial Solution

Find the form of the **corrected** trial solution y but do not evaluate the undetermined coefficients.

51. $y' = x^3 + 5 + x^2e^x(3 + 2x + \sin 2x)$

52. $y' = x^2 + 5x + 2 + x^3e^x(2 + 3x + 5\cos 4x)$

53. $y' - y = x^3 + 2x + 5 + x^4e^x(2 + 4x + 7\cos 2x)$

54. $y' - y = x^4 + 5x + 2 + x^3e^x(2 + 3x + 5\cos 4x)$

55. $y' - 2y = x^3 + x^2 + x^3e^x(2e^x + 3x + 5\sin 4x)$

56. $y' - 2y = x^3e^{2x} + x^2e^x(3 + 4e^x + 2\cos 2x)$

57. $y' + y = x^2 + 5x + 2 + x^3e^{-x}(6x + 3\sin x + 2\cos x)$

58. $y' - 2y = x^5 + 5x^3 + 14 + x^3e^x(5 + 7xe^{-3x})$

59. $2y' + 4y = x^4 + 5x^5 + 2x^8 + x^3e^x(7 + 5xe^x + 5\sin 11x)$

60. $5y' + y = x^2 + 5x + 2e^{x/5} + x^3e^{x/5}(7 + 9x + 2\sin(9x/2))$

Undetermined Coefficients

Compute a particular solution y_p according to the method of undetermined coefficients. Expected details include:

(1) Initial trial solution
(2) Corrected trial solution
(3) Undetermined coefficient algebraic equations and solution
(4) Formula for y_p, coefficients evaluated

61. $y' + y = x + 1$

62. $y' + y = 2x - 1$

63. $y' - y = e^x + e^{-x}$

64. $y' - y = xe^x + e^{-x}$

65. $y' - 2y = 1 + x + e^{2x} + \sin x$

66. $y' - 2y = 1 + x + xe^{2x} + \cos x$

67. $y' + 2y = xe^{-2x} + x^3$

68. $y' + 2y = (2 + x)e^{-2x} + xe^x$

69. $y' = x^2 + 4 + xe^x(3 + \cos x)$

70. $y' = x^2 + 5 + xe^x(2 + \sin x)$

2.5 Linear Applications

This collection of applications for the linear equation $y' + p(x)y = r(x)$ includes mixing problems, especially brine tanks in single and multiple cascade, heating and cooling problems, radioactive isotope chains and elementary electric circuits.

The theory for brine cascades will be developed. Heating and cooling will be developed from Newton's cooling law. Radioactive decay theory appears on page 3. Electric LR or RC circuits appear on page 17.

Brine Mixing

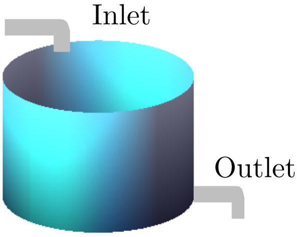


Figure 1. A Single Brine Tank.
The tank has one inlet and one outlet. The inlet supplies a brine mixture and the outlet drains the tank.

A given tank contains brine, which is a water and salt mixture. Input pipes supply other, possibly different brine mixtures at varying rates, while output pipes drain the tank. The problem is to determine the salt $x(t)$ in the tank at any time.

The basic chemical law to be applied is the **mixture law**

$$\frac{dx}{dt} = \textsf{input rate} - \textsf{output rate}.$$

The law is applied under a simplifying assumption: *the concentration of salt in the brine is uniform throughout the fluid.* Stirring is one way to meet this requirement. Because of the uniformity assumption, the amount $x(t)$ of salt in kilograms divided by the volume $V(t)$ of the tank in liters gives salt **concentration**[3] $x(t)/V(t)$ kilograms per liter.

One Input and One Output

Let the input be $a(t)$ liters per minute with concentration C_1 kilograms of salt per liter. Let the output empty $b(t)$ liters per minute. The tank is assumed to contain V_0 liters of brine at $t = 0$. The tank gains fluid at rate $a(t)$ and loses fluid at rate $b(t)$, therefore $V(t) = V_0 + \int_0^t [a(r) - b(r)]dr$ is the volume of brine in the tank at time t. The *mixture law* applies to obtain (derived on page 121) the model linear differential equation

$$\frac{dx}{dt} = a(t)\,C_1 - b(t)\,\frac{x(t)}{V(t)}. \tag{1}$$

[3] Concentration is defined as amount per unit volume: **concentration** $= \dfrac{\textbf{amount}}{\textbf{volume}}$.

This equation is solved by the *linear integrating factor method*, page 96.

Two-Tank Mixing

Two tanks A and B are assumed to contain A_0 and B_0 liters of brine at $t = 0$. Let the input for the first tank A be $a(t)$ liters per minute with concentration C_1 kilograms of salt per liter. Let tank A empty at $b(t)$ liters per minute into a second tank B, which itself empties at $c(t)$ liters per minute.

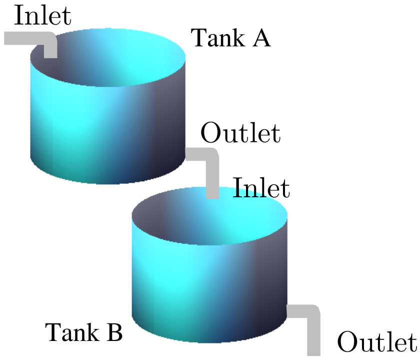


Figure 2. Two Brine Tanks.
Tank A has one inlet, which supplies a brine mixture. The outlet of Tank A cascades into Tank B. The outlet of Tank B drains the two-tank system.

Let $x(t)$ be the number of kilograms of salt in tank A at time t. Similarly, $y(t)$ is the amount of salt in tank B. The *objective* is to find differential equations for the unknowns $x(t)$, $y(t)$.

Fluid loses and gains in each tank give rise to the brine volume formulas $V_A(t) = A_0 + \int_0^t [a(r) - b(r)]dr$ and $V_B(t) = B_0 + \int_0^t [b(r) - c(r)]dr$, respectively, for tanks A and B, at time t.

The *mixture law* applies to obtain the model linear differential equations

$$\begin{aligned}\frac{dx}{dt} &= a(t)\,C_1 - b(t)\,\frac{x(t)}{V_A(t)},\\ \frac{dy}{dt} &= b(t)\,\frac{x(t)}{V_A(t)} - c(t)\,\frac{y(t)}{V_B(t)}.\end{aligned}$$

The first equation is solved for an explicit solution $x(t)$ by the linear integrating factor method. Substitute the expression for $x(t)$ into the second equation, then solve for $y(t)$ by the linear integrating factor method.

Residential Heating and Cooling

The internal temperature $u(t)$ in a residence fluctuates with the outdoor temperature, indoor heating and indoor cooling. Newton's law of cooling for linear convection can be written as

$$\frac{du}{dt} = k(a(t) - u(t)) + s(t) + f(t), \tag{2}$$

where the various symbols have the interpretation below.

k	The insulation constant (see **Remarks on Insulation Constants**, 119). Typically $1/2 \leq k < 1$, with $1 =$ no insulation, $0 =$ perfect insulation.
$a(t)$	The ambient outside temperature.
$s(t)$	Combined rate for all inside heat sources. Includes living beings, appliances and whatever uses energy.
$f(t)$	Inside heating or cooling rate.

Newton's cooling model applies to convection only, and not to heat transfer by radiation or conduction. A derivation of (2) appears on page 121. To solve equation (2), write it in standard linear form and use the integrating factor method on page 96.

No Sources

Assume the absence of heating inside the building, that is, $s(t) = f(t) = 0$. Let the outside temperature be constant: $a(t) = a_0$. Equation (2) simplifies to the Newton cooling equation on page 4:

$$\frac{du}{dt} + ku(t) = ka_0. \tag{3}$$

From Theorem 1.1, page 5, the solution is

$$u(t) = a_0 + (u(0) - a_0)e^{-kt}. \tag{4}$$

This formula represents *exponential decay* of the interior temperature from $u(0)$ to a_0.

Half-Time Insulation Constant

Suppose it's 50°F outside and 70°F initially inside when the electricity goes off. How long does it take to drop to 60°F inside? The answer is *about 1–3 hours*, depending on the insulation.

The importance of 60°F is that it is halfway between the inside and outside temperatures of 70°F and 50°F. The range 1–3 hours is found from (4) by solving $u(T) = 60$ for T, in the extreme cases of poor or perfect insulation.

The more general equation $u(T) = (a_0 + u(0))/2$ can be solved. The answer is $T = \ln(2)/k$, called the **half-time insulation constant** for the residence. It measures the insulation quality, larger T corresponding to better insulation. For most residences, the half-time insulation constant ranges from 1.4 ($k = 0.5$) to 14 ($k = 0.05$) hours.

Winter Heating

The introduction of a furnace and a thermostat set at temperature T_0 (typically, 68°F to 72°F) changes the source term $f(t)$ to the special form

$$f(t) = k_1(T_0 - u(t)),$$

according to Newton's law of cooling, where k_1 is a constant. The differential equation (2) becomes

$$\text{(5)} \qquad \frac{du}{dt} = k(a(t) - u(t)) + s(t) + k_1(T_0 - u(t)).$$

It is a first-order linear differential equation which can be solved by the integrating factor method.

Summer Air Conditioning

An air conditioner used with a thermostat leads to the same differential equation (5) and solution, because Newton's law of cooling applies to both heating and cooling.

Evaporative Cooling

In desert-mountain areas, where summer humidity is low, the **evaporative cooler** is a popular low-cost solution to cooling. The cooling effect is due to heat loss from the supply of outside air, caused by energy conversion during water evaporation. Cool air is pumped into the residence much like a furnace pumps warm air. An evaporative cooler may have no thermostat. The temperature $P(t)$ of the pumped air depends on the outside air temperature and humidity.

A Newton's cooling model for the inside temperature $u(t)$ requires a constant k_1 for the evaporative cooling term $f(t) = k_1(P(t) - u(t))$. If $s(t) = 0$ is assumed, then equation (2) becomes

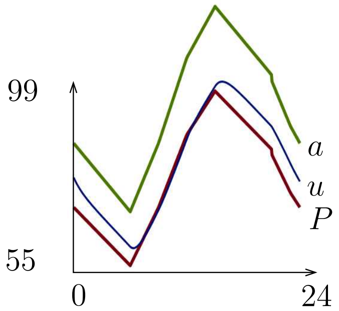
$$\text{(6)} \qquad \frac{du}{dt} = k(a(t) - u(t)) + k_1(P(t) - u(t)).$$

This is a first-order linear differential equation, solvable by the integrating factor method.

During hot summer days the relation $P(t) = 0.85a(t)$ could be valid, that is, the air pumped from the cooler vent is 85% of the ambient outside temperature $a(t)$. Extreme temperature variations can occur in the fall and spring. In July, the reverse is possible, e.g., $100 < a(t) < 115$. Assuming $P(t) = 0.85a(t)$, the solution of (6) is

$$u(t) = u(0)e^{-kt-k_1t} + (k + 0.85k_1)\int_0^t a(r)e^{(k+k_1)(r-t)}dr.$$

Figure 3 shows the solution for a 24-hour period, using a sample profile $a(t)$, $k = 1/4$, $k_1 = 2$ and $u(0) = 69$. The residence temperature $u(t)$ is expected to be approximately between $P(t)$ and $a(t)$.



$$a(t) = \begin{cases} 75 - 2t & 0 \le t \le 6 \\ 39 + 4t & 6 < t \le 9 \\ 30 + 5t & 9 < t \le 12 \\ 54 + 3t & 12 < t \le 15 \\ 129 - 2t & 15 < t \le 21 \\ 170 - 4t & 21 < t \le 23 \\ 147 - 3t & 23 < t \le 24 \end{cases}$$

Figure 3. A 24-hour plot of P, u and temperature profile $a(t)$.

Examples

Example 2.21 (Pollution)

When industrial pollution in Lake Erie ceased, the level was five times that of its inflow from Lake Huron. Assume Lake Erie has perfect mixing, constant volume V and equal inflow/outflow rates of $0.73V$ per year. Estimate the time required to reduce the pollution in half.

Solution: The answer is about 1.34 years. An overview of the solution will be given, followed by technical details.

Overview. The brine-mixing model applies to pollution problems, giving a differential equation model for the pollution concentration $x(t)$,

$$x'(t) = 0.73Vc - 0.73x(t), \quad x(0) = 5cV,$$

where c is the inflow pollution concentration. The model has solution

$$x(t) = x(0)\left(0.2 + 0.8e^{-0.73t}\right).$$

Solving for the time T at which $x(T) = \frac{1}{2}x(0)$ gives $T = \ln(8/3)/0.73 = 1.34$ years.

Model details. The rate of change of $x(t)$ equals the concentration rate in minus the concentration rate out. The in-rate equals c times the inflow rate, or $c(0.73V)$. The out-rate equals $x(t)$ times the outflow rate, or $\frac{0.73V}{V}x(t)$. This justifies the differential equation. The statement $x(0)$=“five times that of Lake Huron” means that $x(0)$ equals $5c$ times the volume of Lake Erie, or $5cV$.

Solution details. The differential equation can be re-written in equivalent form $x'(t) + 0.73x(t) = 0.73x(0)/5$. It has equilibrium solution $x_p = x(0)/5$. The homogeneous solution is $x_h = ke^{-0.73t}$, from the theory of growth-decay equations. Adding x_h and x_p gives the general solution x. To solve the initial value problem, substitute $t = 0$ and find $k = 4x(0)/5$. Substitute for k into $x = x(0)/5 + ke^{-0.73t}$ to obtain the reported solution.

Equation for T details. The equation $x(T) = \frac{1}{2}x(0)$ becomes $x(0)(0.2 + 0.8e^{-0.73T}) = x(0)/2$, which by algebra reduces to the exponential equation $e^{-0.73T} = 3/8$. Take logarithms to isolate $T = -\ln(3/8)/0.73 \approx 1.3436017$.

Example 2.22 (Brine Cascade)

Assume brine tanks A and B in Figure 4 have volumes 100 and 200 gallons, respectively. Let $A(t)$ and $B(t)$ denote the number of pounds of salt at time t, respectively, in tanks A and B. Pure water flows into tank A, brine flows out of tank A and into tank B, then brine flows out of tank B. All flows are at 4 gallons per minute. Given $A(0) = 40$ and $B(0) = 40$, find $A(t)$ and $B(t)$.

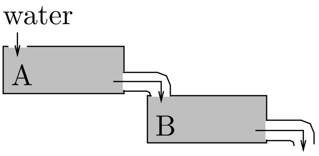


Figure 4. Cascade of two brine tanks.

Solution: The solutions for the brine cascade are (details below)

$$A(t) = 40e^{-t/25}, \quad B(t) = 120e^{-t/50} - 80e^{-t/25}.$$

Modeling. This is an instance of the two-tank mixing problem on page 112. The volumes in the tanks do not change and the input salt concentration is $C_1 = 0$. The equations are

$$\frac{dA}{dt} = -\frac{4A(t)}{100}, \quad \frac{dB}{dt} = \frac{4A(t)}{100} - \frac{4B(t)}{200}.$$

Solution $A(t)$ details.

$A' = -0.04A, \quad A(0) = 40$	Initial value problem to be solved.
$A = 40e^{-t/25}$	Solution found by the growth-decay model.

Solution $B(t)$ details.

$B' = 0.04A - 0.02B, \quad B(0) = 40$	Initial value problem to be solved.
$B' + 0.02B = 1.6e^{-t/25}$	Substitute for A. Get standard form.
$B' + 0.02B = 0, \quad B(0) = 40$	Homogeneous problem to be solved.
$B_h = 40e^{-t/50}$	Homogeneous solution. Growth-decay formula applied.
$B_p = e^{-t/50} \int_0^t 1.6e^{-r/25} e^{r/50} dr$	Variation of parameters solution.
$= 80e^{-t/50} - 80e^{-t/25}$	Evaluate integral.
$B = B_h + B_p$	Superposition.
$= 120e^{-t/50} - 80e^{-t/25}$	Final solution.

The solution can be checked in `maple` as follows.

```
de1:=diff(x(t),t)=-4*x(t)/100:
de2:=diff(y(t),t)=4*x(t)/100-4*y(t)/200:
ic:=x(0)=40,y(0)=40:
dsolve({de1,de2,ic},{x(t),y(t)});
```

Example 2.23 (Office Heating)

A worker shuts off the office heat and goes home at 5PM. It's 72°F inside and 60°F outside overnight. Estimate the office temperature at 8PM, 11PM and 6AM.

Solution:

The temperature estimates are 62.7-65.7°F, 60.6-62.7°F and 60.02-60.5°F. Details follow.

Model. The residential heating model applies, with no sources, to give $u(t) = a_0 + (u(0) - a_0)e^{-kt}$. Supplied are values $a_0 = 60$ and $u(0) = 72$. Unknown is constant k in the formula

$$u(t) = 60 + 12e^{-kt}.$$

Estimation of k. To make the estimate for k, assume the range $1/4 \le k \le 1/2$, which covers the possibilities of poor to excellent insulation.

Calculations. The estimates requested are for $t = 3$, $t = 6$ and $t = 13$. The formula $u(t) = 60 + 12e^{-kt}$ and the range $0.25 \le k \le 0.5$ gives the estimates

$$\begin{aligned} 62.68 \le 60 + 12e^{-3k} &\le 65.67, \\ 60.60 \le 60 + 12e^{-6k} &\le 62.68, \\ 60.02 \le 60 + 12e^{-13k} &\le 60.47. \end{aligned}$$

Example 2.24 (Spring Temperatures)

It's spring. The outside temperatures are between 45°F and 75°F and the residence has no heating or cooling. Find an approximation for the interior temperature fluctuation $u(t)$ using the estimate $a(t) = 60 - 15\cos(\pi(t-4)/12)$, $k = \ln(2)/2$ and $u(0) = 53$.

Solution: The approximation, justified below, is

$$u(t) \approx -8.5e^{-kt} + 60 + 1.5\cos\frac{\pi t}{12} - 12\sin\frac{\pi t}{12}.$$

Model. The residential model for no sources applies. Then

$$u'(t) = k(a(t) - u(t)).$$

Computation of $u(t)$. Let $\omega = \pi/12$ and $k = \ln(2)/2 \approx 0.35$ (poor insulation). The solution is

$$\begin{aligned} u &= u(0)e^{-kt} + \textstyle\int_0^t ka(r)e^{k(r-t)}dr && \text{Variation of parameters.} \\ &= 53e^{-kt} + \textstyle\int_0^t 15k(4 - \cos\omega(t-4))e^{k(r-t)}dr && \text{Insert } a(t) \text{ and } u(0). \\ &\approx -8.5e^{-kt} + 60 + 1.5\cos\omega t - 12\sin\omega t && \text{Used maple integration.} \end{aligned}$$

The `maple` code used for the integration appears below.

```
k:=ln(2)/2: u0:=53:
A:=r->k*(60-15*cos(Pi *(r-4)/12)):
U:=t->(u0+int(A(r)*exp(k*r),r=0..t))*exp(-k*t);
simplify(U(t));
```

Example 2.25 (Temperature Variation)

Justify that in the spring and fall, the interior of a residence might have temperature variation between 19% and 89% of the outside temperature variation.

Solution: The justification necessarily makes some assumptions, which are:

$a(t) = B - A\cos\omega(t-4)$	Assume $A > 0$, $B > 0$, $\omega = \pi/12$ and extreme temperatures at 4AM and 4PM.
$s(t) = 0$	No inside heat sources.
$f(t) = 0$	No furnace or air conditioner.
$0.05 \le k \le 0.5$	Vary from excellent ($k = 0.05$) to poor ($k = 0.5$) insulation.
$u(0) = B$	The average of the outside low and high.

Model. The residential model for no sources applies. Then

$$u'(t) = k(a(t) - u(t)).$$

Formula for u. Variation of parameters gives a compact formula:

$u = u(0)e^{-kt} + \int_0^t ka(r)e^{k(r-t)}dr$	See (4), page 97.
$= Be^{-kt} + \int_0^t k(B - A\cos\omega(t-4))e^{k(r-t)}dr$	Insert $a(t)$ and $u(0)$.
$= c_0Ae^{-kt} + B + c_1A\cos\omega t + c_2A\sin\omega t$	Evaluate. Values below.

The values of the constants in the calculation of u are

$$c_0 = 72k^2 - 6k\pi\sqrt{3}, \quad c_1 = \frac{6k\pi\sqrt{3} - 72k^2}{144k^2 + \pi^2}, \quad c_2 = \frac{-6k\pi - 72k^2\sqrt{3}}{144k^2 + \pi^2}.$$

The trigonometric formula $a\cos\theta + b\sin\theta = r\sin(\theta + \phi)$ where $r^2 = a^2 + b^2$ and $\tan\phi = a/b$ can be applied to the formula for u to rewrite it as

$$u = c_0Ae^{-kt} + B + A\sqrt{c_1^2 + c_2^2}\sin(\omega t + \phi).$$

The outside low and high are $B - A$ and $B + A$. The outside temperature variation is their difference $2A$. The exponential term contributes less than one degree after 12 hours. The inside low and high are therefore approximately $B - rA$ and $B + rA$ where $r = \sqrt{c_1^2 + c_2^2}$. The inside temperature variation is their difference $2rA$, which is r times the outside variation.

It remains to show that $0.19 \le r \le 0.89$. The equation for r has a simple representation:

$$r = \frac{12k}{\sqrt{144k^2 + \pi^2}}.$$

It has positive derivative dr/dk. Then extrema occur at the endpoints of the interval $0.05 \le k \le 0.5$, giving values $r = 0.19$ and $r = 0.89$, approximately. This justifies the estimates of 19% and 89%.

The `maple` code used for the integration appears below.

```
omega:=Pi/12:
F:=r->k*(B-A*cos(omega *(r-4))):
G:=t->(B+int(F(r)*exp(k*r),r=0..t))*exp(-k*t);
simplify(G(t));
```

Remarks on Insulation Constants. The insulation constant k in the Newton cooling model is usually between zero and one, with excellent insulation near zero and bad insulation near one. It is also called a **coupling constant**, because $k = 0$ means the temperature u is *decoupled* from the ambient temperature. The constant k depends in a complex way on geometry and insulation, therefore it is determined empirically, and not by a theoretical formula. Lab experiments with a thermocouple in an air-insulated vessel filled with about 300 ml of hot water (80 to 100 C) can determine insulation constants on the order of $k = 0.0003$ (units per second).

Printed on dual pane clear glass in the USA is a U-value of about 0.48. The U-value is equal to the reciprocal of the R-value (see below). You can think of it as the insulation constant k. The lower the U-value, the better the glass insulation quality.

For a solar water heater, $k = 0.00035$ is typical. This value is for an 80 gallon tank with R-15 insulation raised to 120 F during the day. Typically, the water temperature drops by only 3-4 F overnight.

The **thermal conductivity** symbol κ (Greek kappa) can be confused with the insulation constant symbol k, and it is a tragic error to substitute one for the other.

For USA R-values printed on insulation products, thermal conductivity is defined by the relation $U = \frac{1}{0.17611101838R} = \frac{\kappa}{L}$, where L is the material's thickness and U is the international U-factor in SI units. The U-factor value is the heat lost in Watts per square meter at a standard temperature difference of one degree Kelvin.

Example 2.26 (Radioactive Chain)

Let A, B and C be the amounts of three radioactive isotopes. Assume A decays into B at rate a, then B decays into C at rate b. Given $a \neq b$, $A(0) = A_0$ and $B(0) = 0$, find formulas for A and B.

Solution: The isotope amounts are (details below)

$$A(t) = A_0 e^{-at}, \quad B(t) = aA_0 \frac{e^{-at} - e^{-bt}}{b - a}.$$

Modeling. The reaction model will be shown to be

$$A' = -aA,\ A(0) = A_0, \quad B' = aA - bB,\ B(0) = 0.$$

The derivation uses the radioactive decay law on page 19. The model for A is simple decay $A' = -aA$. Isotope B is *created* from A at a rate equal to the disintegration rate of A, or aA. But B itself undergoes disintegration at rate bB. The rate of increase of B is not aA but the difference of aA and bB, which accounts for lost material. Therefore, $B' = aA - bB$.

Solution Details for A.

$A' = -aA, \quad A(0) = A_0$	Initial value problem to solve.
$A = A_0 e^{-at}$	Use the *growth-decay* formula on page 3.

Solution Details for B.

$B' = aA - bB, \quad B(0) = 0$	Initial value problem to solve.
$B' + bB = aA_0 e^{-at}, \quad B(0) = 0$	Insert $A = A_0 e^{-at}$. Standard form.
$B = e^{-bt} \int_0^t aA_0 e^{-ar} e^{br} dr$	Variation of parameters solution page 465, which already satisfies $B(0) = 0$.

$$= aA_0 \frac{e^{-at} - e^{-bt}}{b-a} \qquad \text{Evaluate the integral for } b \neq a.$$

Remark on radioactive chains. The sequence of radioactive decay processes creates at each stage a new element that may itself be radioactive. The chain ends when stable atoms are formed. For example, uranium-236 decays into thorium-232, which decays into radium-228, and so on, until stable lead-208 is created at the end of the chain. Analyzed here are 2 steps in such a chain.

Example 2.27 (Electric Circuits)

For the LR-circuit of Figure 5, show that $I_{\text{SS}} = E/R$ and $I_{\text{tr}} = I_0 e^{-Rt/L}$ are the steady-state and transient currents.

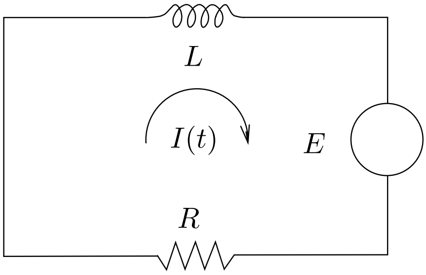


Figure 5. An LR-circuit with constant voltage E and zero initial current $I(0) = 0$.

Solution:

Model. The LR-circuit equation is derived from Kirchhoff's laws and the voltage drop formulas on page 17. The only new element is the added electromotive force term $E(t)$, which is set equal to the algebraic sum of the voltage drops, giving the model

$$LI'(t) + RI(t) = E(t), \quad I(0) = I_0.$$

General solution. The details:

$I' + (R/L)I = E/L$	Standard linear form.
$I_p = E/R$	Set I=constant, solve for a particular solution I_p.
$I' + (R/L)I = 0$	Homogeneous equation. Solve for $I = I_h$.
$I_h = I_0 e^{-Rt/L}$	Growth-decay formula, page 4.
$I = I_h + I_p$	Superposition.
$= I_0 e^{-Rt/L} + E/R$	General solution found.

Steady-state solution. The steady-state solution is found by striking out from the general solution all terms that approach zero at $t = \infty$. Remaining after strike-out is $I_{\text{SS}} = E/R$.

Transient solution. The term *transient* refers to the terms in the general solution which approaches zero at $t = \infty$. Therefore, $I_{\text{tr}} = I_0 e^{-Rt/L}$.

Example 2.28 (Time constant)

Show that the current $I(t)$ in the LR-circuit of Figure 5 is at least 95% of the steady-state current E/R after three time constants, i.e., after time $t = 3L/R$.

Solution: Physically, the **time constant** L/R for the circuit is found by an experiment in which the circuit is initialized to $I = 0$ at $t = 0$, then the current I is observed until it reaches 63% of its steady-state value.

Time to 95% of I_{SS}. The solution is $I(t) = E(1 - e^{-Rt/L})/R$. Solving the inequality $1 - e^{-Rt/L} \geq 0.95$ gives

$0.95 \leq 1 - e^{-Rt/L}$	Inequality to be solved for t.
$e^{-Rt/L} \leq 1/20$	Move terms across the inequality.
$\ln e^{-Rt/L} \leq \ln(1/20)$	Take the logarithm across the inequality.
$-Rt/L \leq \ln 1 - \ln 20$	Apply logarithm rules.
$t \geq L\ln(20)/R$	Isolate t on one side.

The value $\ln(20) = 2.9957323$ leads to the rule: *after three times the time constant has elapsed, the current has reached 95% of the steady-state current.*

Details and Proofs

Brine-Mixing One-tank Proof: Equation $x'(t) = C_1a(t) - b(t)x(t)/V(t)$, the brine-mixing equation, is justified for the one-tank model by applying the *mixture law* $dx/dt =$ input rate − output rate as follows.

$$\begin{aligned} \text{input rate} &= \left(a(t)\frac{\text{liters}}{\text{minute}}\right)\left(C_1\frac{\text{kilograms}}{\text{liter}}\right) \\ &= C_1a(t)\frac{\text{kilograms}}{\text{minute}}, \\ \text{output rate} &= \left(b(t)\frac{\text{liters}}{\text{minute}}\right)\left(\frac{x(t)}{V(t)}\frac{\text{kilograms}}{\text{liter}}\right) \\ &= \frac{b(t)x(t)}{V(t)}\frac{\text{kilograms}}{\text{minute}}. \end{aligned}$$

Residential Heating and Cooling Proof: Newton's law of cooling will be applied to justify the residential heating and cooling equation

$$\frac{du}{dt} = k(a(t) - u(t)) + s(t) + f(t).$$

Let $u(t)$ be the indoor temperature. The heat flux is due to three heat source rates:

$N(t) = k(a(t) - u(t))$	The Newton cooling rate.
$s(t)$	Combined rate for all inside heat sources.
$f(t)$	Inside heating or cooling rate.

The expected change in u is the sum of the rates N, s and f. In the limit, $u'(t)$ is on the left and the sum $N(t) + s(t) + f(t)$ is on the right. ■

Exercises 2.5

Concentration

A lab assistant collects a volume of brine, boils it until only salt crystals remain, then uses a scale to determine the crystal mass or weight.

Find the salt **concentration** of the brine in kilograms per liter.

1. One liter of brine, crystal mass 0.2275 kg

2. Two liters, crystal mass 0.32665 kg

3. Two liters, crystal mass 15.5 grams

4. Five pints, crystals weigh 1/4 lb

5. Eighty cups, crystals weigh 5 lb

6. Five gallons, crystals weigh 200 ounces

One-Tank Mixing

Assume one inlet and one outlet. Determine the amount $x(t)$ of salt in the tank at time t. Use the text notation for equation (1).

7. The inlet adds 10 liters per minute with concentration $C_1 = 0.023$ kilograms per liter. The tank contains 110 liters of distilled water. The outlet drains 10 liters per minute.

8. The inlet adds 12 liters per minute with concentration $C_1 = 0.0205$ kilograms per liter. The tank contains 200 liters of distilled water. The outlet drains 12 liters per minute.

9. The inlet adds 10 liters per minute with concentration $C_1 = 0.0375$ kilograms per liter. The tank contains 200 liters of brine in which 3 kilograms of salt is dissolved. The outlet drains 10 liters per minute.

10. The inlet adds 12 liters per minute with concentration $C_1 = 0.0375$ kilograms per liter. The tank contains 500 liters of brine in which 7 kilograms of salt is dissolved. The outlet drains 12 liters per minute.

11. The inlet adds 10 liters per minute with concentration $C_1 = 0.1075$ kilograms per liter. The tank contains 1000 liters of brine in which k kilograms of salt is dissolved. The outlet drains 10 liters per minute.

12. The inlet adds 14 liters per minute with concentration $C_1 = 0.1124$ kilograms per liter. The tank contains 2000 liters of brine in which k kilograms of salt is dissolved. The outlet drains 14 liters per minute.

13. The inlet adds 10 liters per minute with concentration $C_1 = 0.104$ kilograms per liter. The tank contains 100 liters of brine in which 0.25 kilograms of salt is dissolved. The outlet drains 11 liters per minute. Determine additionally the time when the tank is empty.

14. The inlet adds 16 liters per minute with concentration $C_1 = 0.01114$ kilograms per liter. The tank contains 1000 liters of brine in which 4 kilograms of salt is dissolved. The outlet drains 20 liters per minute. Determine additionally the time when the tank is empty.

15. The inlet adds 10 liters per minute with concentration $C_1 = 0.1$ kilograms per liter. The tank contains 500 liters of brine in which k kilograms of salt is dissolved. The outlet drains 12 liters per minute. Determine additionally the time when the tank is empty.

16. The inlet adds 11 liters per minute with concentration $C_1 = 0.0156$ kilograms per liter. The tank contains 700 liters of brine in which k kilograms of salt is dissolved. The outlet drains 12 liters per minute. Determine additionally the time when the tank is empty.

Two-Tank Mixing

Assume brine tanks A and B in Figure 4 have volumes 100 and 200 gallons, respectively. Let $x(t)$ and $y(t)$ denote the number of pounds of salt at time t, respectively, in tanks A and B. Distilled water flows into tank A, then brine flows out of tank A and into tank B, then out of tank B. All flows are at r gallons per minute. Given rate r and initial salt amounts $x(0)$ and $y(0)$, find $x(t)$ and $y(t)$.

17. $r = 4$, $x(0) = 40$, $y(0) = 20$.

18. $r = 3$, $x(0) = 10$, $y(0) = 15$.

19. $r = 5$, $x(0) = 20$, $y(0) = 40$.

20. $r = 5$, $x(0) = 40$, $y(0) = 30$.

21. $r = 8$, $x(0) = 10$, $y(0) = 12$.

22. $r = 8$, $x(0) = 30$, $y(0) = 12$.

23. $r = 9$, $x(0) = 16$, $y(0) = 14$.

24. $r = 9$, $x(0) = 22$, $y(0) = 10$.

25. $r = 7$, $x(0) = 6$, $y(0) = 5$.

26. $r = 7$, $x(0) = 13$, $y(0) = 26$

Residential Heating

Assume the Newton cooling model for heating and insulation values $1/4 \le k \le 1/2$. Follow Example 2.23, page 116.

27. The office heat goes off at 7PM. It's 74°F inside and 58°F outside overnight. Estimate the office temperature at 10PM, 1AM and 6AM.

28. The office heat goes off at 6:30PM. It's 73°F inside and 55°F outside overnight. Estimate the office temperature at 9PM, 3AM and 7AM.

29. The radiator goes off at 9PM. It's 74°F inside and 58°F outside overnight. Estimate the room temperature at 11PM, 2AM and 6AM.

30. The radiator goes off at 10PM. It's 72°F inside and 55°F outside overnight. Estimate the room temperature at 2AM, 5AM and 7AM.

31. The office heat goes on in the morning at 6:30AM. It's 57°F inside and 40° to 55°F outside until 11AM. Estimate the office temperature at 8AM, 9AM and 10AM. Assume the furnace provides a five degree temperature rise in 30 minutes with perfect insulation and the thermostat is set for 76°F.

32. The office heat goes on at 6AM. It's 55°F inside and 43° to 53°F outside until 10AM. Estimate the office temperature at 7AM, 8AM and 9AM. Assume the furnace provides a seven degree temperature rise in 45 minutes with perfect insulation and the thermostat is set for 78°F.

33. The hot water heating goes on at 6AM. It's 55°F inside and 50° to 60°F outside until 10AM. Estimate the room temperature at 7:30AM. Assume the radiator provides a four degree temperature rise in 45 minutes with perfect insulation and the thermostat is set for 74°F.

34. The hot water heating goes on at 5:30AM. It's 54°F inside and 48° to 58°F outside until 9AM. Estimate the room temperature at 7AM. Assume the radiator provides a five degree temperature rise in 45 minutes with perfect insulation and the thermostat is set for 74°F.

35. A portable heater goes on at 7AM. It's 45°F inside and 40° to 46°F outside until 11AM. Estimate the room temperature at 9AM. Assume the heater provides a two degree temperature rise in 30 minutes with perfect insulation and the thermostat is set for 90°F.

36. A portable heater goes on at 8AM. It's 40°F inside and 40° to 45°F outside until 11AM. Estimate the room temperature at 10AM. Assume the heater provides a two degree temperature rise in 20 minutes with perfect insulation and the thermostat is set for 90°F.

Evaporative Cooling

Define outside temperature (see Figure 3)

$$a(t) = \begin{cases} 75-2t & 0 \le t \le 6 \\ 39+4t & 6 < t \le 9 \\ 30+5t & 9 < t \le 12 \\ 54+3t & 12 < t \le 15. \\ 129-2t & 15 < t \le 21 \\ 170-4t & 21 < t \le 23 \\ 147-3t & 23 < t \le 24 \end{cases}$$

Given k, k_1, $P(t) = wa(t)$ and $u(0) = 69$, then plot $u(t)$, $P(t)$ and $a(t)$ on one graphic.

$$u(t) = u(0)e^{-kt-k_1t} + \\ (k + wk_1)\int_0^t a(r)e^{(k+k_1)(r-t)}dr.$$

37. $k = 1/4$, $k_1 = 2$, $w = 0.85$

38. $k = 1/4$, $k_1 = 1.8$, $w = 0.85$

39. $k = 3/8$, $k_1 = 2$, $w = 0.85$

40. $k = 3/8$, $k_1 = 2.4$, $w = 0.85$

41. $k = 1/4$, $k_1 = 3$, $w = 0.80$

42. $k = 1/4$, $k_1 = 4$, $w = 0.80$

43. $k = 1/2$, $k_1 = 4$, $w = 0.80$

44. $k = 1/2$, $k_1 = 5$, $w = 0.80$

45. $k = 3/8$, $k_1 = 3$, $w = 0.80$

46. $k = 3/8$, $k_1 = 4$, $w = 0.80$

Radioactive Chain

Let A, B and C be the amounts of three radioactive isotopes. Assume A decays into B at rate a, then B decays into C at rate b. Given a, b, $A(0) = A_0$ and $B(0) = B_0$, find formulas for A and B.

47. $a = 2$, $b = 3$, $A_0 = 100$, $B_0 = 10$

48. $a = 2$, $b = 3$, $A_0 = 100$, $B_0 = 100$

49. $a = 1$, $b = 4$, $A_0 = 100$, $B_0 = 200$

50. $a = 1$, $b = 4$, $A_0 = 300$, $B_0 = 100$

51. $a = 4$, $b = 3$, $A_0 = 100$, $B_0 = 100$

52. $a = 4$, $b = 3$, $A_0 = 100$, $B_0 = 200$

53. $a = 6$, $b = 1$, $A_0 = 600$, $B_0 = 100$

54. $a = 6$, $b = 1$, $A_0 = 500$, $B_0 = 400$

55. $a = 3$, $b = 1$, $A_0 = 100$, $B_0 = 200$

56. $a = 3$, $b = 1$, $A_0 = 400$, $B_0 = 700$

Electric Circuits

In the LR-circuit of Figure 5, assume $E(t) = A\cos wt$ and $I(0) = 0$. Solve for $I(t)$.

57. $A = 100$, $w = 2\pi$, $R = 1$, $L = 2$

58. $A = 100$, $w = 4\pi$, $R = 1$, $L = 2$

59. $A = 100$, $w = 2\pi$, $R = 10$, $L = 1$

60. $A = 100$, $w = 2\pi$, $R = 10$, $L = 2$

61. $A = 5$, $w = 10$, $R = 2$, $L = 3$

62. $A = 5$, $w = 4$, $R = 3$, $L = 2$

63. $A = 15$, $w = 2$, $R = 1$, $L = 4$

64. $A = 20$, $w = 2$, $R = 1$, $L = 3$

65. $A = 25$, $w = 100$, $R = 5$, $L = 15$

66. $A = 25$, $w = 50$, $R = 5$, $L = 5$

2.6 Kinetics

Studied are the following topics.

Newton's Laws	Free Fall with Constant Gravity
Linear Air Resistance	Nonlinear Air Resistance
Modeling	Parachutes
Lunar Lander	Escape Velocity

Newton's Laws

The ideal models of a particle or *point mass* constrained to move along the x-axis, or the motion of a projectile or satellite, have been studied from **Newton's second law**

$$F = ma. \tag{1}$$

In the *mks system* of units, F is the force in **Newtons**, m is the mass in kilograms and a is the acceleration in meters per second per second.

The closely-related **Newton universal gravitation law**

$$F = G\frac{m_1 m_2}{R^2} \tag{2}$$

is used in conjunction with (1) to determine the system's constant value g of gravitational acceleration. The masses m_1 and m_2 have centroids at a distance R. For the earth, $g = 9.8$ m/s^2 is commonly used; see Table 1.

Other commonly used unit systems are *cgs* and *fps*. Table 1 shows some useful equivalents.

Table 1. Units for *fps* and *mks* Systems

Unit name	*fps* unit	*mks* unit
Position	foot (ft)	meter (m)
Time	seconds (s)	seconds (s)
Velocity	feet/sec	meters/sec
Acceleration	feet/sec^2	meters/sec^2
Force	pound (lb)	Newton (N)
Mass	slug	kilogram (kg)
g	32.088 ft/s^2	9.7805 m/s^2

Other units in the various systems are in daily use. Table 2 shows some equivalents. An international synonym for **pound** is **libre**, with abbreviation **lb**. The origin of the word *pound* is migration of **libra pondo**, meaning *a pound in weight*. Dictionaries cite migrations *libra pondo* $\longrightarrow$ *pund* for German language, which is similar to English *pound*.

Table 2. Conversions for the *fps* and *mks* Systems

inch (in)	1/12 foot	2.54 centimeters
foot (ft)	12 inches	30.48 centimeters
centimeter (cm)	1/100 meter	0.39370079 inches
kilometer (km)	1000 meters	0.62137119 miles ($\approx 5/8$)
mile (mi)	5280 feet	1.609344 kilometers ($\approx 8/5$)
pound (lb)	$\approx$ 4.448 Newtons	
Newton (N)	$\approx$ 0.225 pounds	
kilogram (kg)	$\approx$ 0.06852 slugs	
slug	$\approx$ 14.59 kilograms	

Velocity and Acceleration

The position, velocity and acceleration of a particle moving along an axis are functions of time t. Notations vary; this text uses the following symbols, where primes denote t-differentiation.

$x = x(t)$	Particle **position** at time t.
$v = x'(t)$	Particle **velocity** at time t.
$a = x''(t)$	Particle **acceleration** at time t.
$x(0)$	**Initial position**.
$v(0)$	**Initial velocity**. Synonym $x'(0)$ is also used.

Free Fall with Constant Gravity

A body falling in a constant gravitational field might ideally move in a straight line, aligned with the gravitational vector. A typical case is the *lunar lander*, which falls freely toward the surface of the moon, its progress downward controlled by retrorockets. *Falling bodies*, e.g., an object launched up or down from a tall building, can be modeled similarly. For such ideal cases, in which air resistance and other external forces are ignored, the acceleration of the body is assumed to be a constant g and the differential equation model is

$$x''(t) = -g, \quad x(0) = x_0, \quad x'(0) = v_0. \tag{3}$$

The initial position x_0 and the initial velocity v_0 must be specified. The value of g in *mks* units is $g = 9.8$ m/s^2. The symbol x is the distance from the ground ($x = 0$); meters for *mks* units. The symbol t is the time in seconds. Falling body problems normally take $v_0 = 0$ and $x_0 > 0$, e.g., x_0 is the height of the building from which the body was dropped. Objects ejected downwards have $v_0 < 0$, which decreases the descent time. Objects thrown straight up satisfy $v_0 > 0$.

Equation (3) can be solved by the method of quadrature to give the explicit solution

$$x(t) = -\frac{g}{2}t^2 + x_0 + v_0 t. \tag{4}$$

See *Technical Details*, page 137, and the *method of quadrature*, page 74. Applications to free fall and the lunar lander appear in the examples, page 132.

Typical plots can be made by the following `maple` code.

```
X:=unapply(-9.8*t^2+100+(50)*t,t); #v(0)=50m/s,x(0)=100m
plot(X(t),t=0..7);
Y:=unapply(-9.8*t^2+100+(-5)*t,t); #v(0)=-5m/s,x(0)=100m
plot(Y(t),t=0..4);
```

Air Resistance Effects

The inclusion in a differential equation model of terms accounting for air resistance has historically two distinct models. The first is *linear resistance*, in which the force F due to air resistance is assumed to be proportional to the velocity v:

$$F \propto v. \tag{5}$$

It is known that linear resistance is appropriate only for slowly moving objects.[4] The second model is *nonlinear resistance*, modeled originally by Sir Isaac Newton himself as $F = kv^2$. The literature considers a generalized nonlinear resistance assumption

$$F \propto v|v|^p \tag{6}$$

where $0 < p \leq 1$ depends upon the *speed* of the object through the air; $p \approx 0$ is a low speed and $p \approx 1$ is a high speed. It will suffice for illustration purposes to treat just the two cases $F \propto v$ and $F \propto v|v|$.

Linear and Nonlinear Drag

For small spherical objects moving slowly through a viscous fluid, Sir George Gabriel Stokes derived an expression for the **linear drag force**:

$$-\frac{k}{m}v = \textbf{Stoke's drag force} = -6\,\pi\,\eta\,r\,v$$

The symbols: η = fluid viscosity and r = radius of the spherical object. References can use viscosity symbols ρ or μ instead of Stoke's symbol η.

Example: Falling raindrop
The radius is $r = 0.1$ to 0.3 mm and $\eta = 1.789x10^{-5}$ Kg/m/sec is the dynamic viscosity for 15 C air at sea level.

[4] More precisely, for Reynolds Number less than about 1000. The Reynolds Number is the ratio of inertial forces to viscous forces within a fluid.

Velocities $v = x'(t) >$ Mach 1 use **nonlinear drag force** $\pm kv^2$. Fluid theory gives $k = \frac{1}{2} C\,\eta\,A$ where $C =$ the **drag coefficient**, $\eta =$ **dynamic fluid viscosity**, $A =$ **frontal area** facing the fluid.

Example: 22 caliber high velocity long rifle bullet
Drag coefficient $C = 0.35$ to 0.4, air dynamic viscosity $\eta = 1.789x10^{-5}$ Kg/m/sec, frontal area $A = 0.25419304$ cm^2 . Nonlinear drag occurs for close targets. The bullet path below Mach 1 has a section of linear drag.

Linear Air Resistance

The model is determined by the sum of the forces due to air resistance and gravity, $F_{\text{air}} + F_{\text{gravity}}$, which by *Newton's second law* must equal $F = mx''(t)$, giving the differential equation

$$mx''(t) = -kx'(t) - mg. \tag{7}$$

In (7), the velocity is $v = x'(t)$ and k is a proportionality constant for the air resistance force $F \propto v$. The negative sign results from the assumed coordinates: x measures the distance from the ground ($x = 0$). We expect x to decrease, hence x' is negative. Equation (7) written in terms of the velocity $v = x'(t)$ becomes

$$v'(t) = -(k/m)v(t) - g. \tag{8}$$

This equation has a solution $v(t)$ which limits at $t = \infty$ to a **finite terminal velocity** $|v_\infty| = mg/k$; equation (9) below is justified in *Technical Details*, page 137. Physically, this limit is the **equilibrium solution** of (8), which is the observable steady state of the model. A quadrature applied to $x'(t) = v(t)$ using $v(t)$ in equation (9) solves (7). Then

$$\begin{aligned} v(t) &= -\frac{mg}{k} + \left(v(0) + \frac{mg}{k}\right) e^{-kt/m}, \\ x(t) &= x(0) - \frac{mg}{k}t + \frac{m}{k}\left(v(0) + \frac{mg}{k}\right)\left(1 - e^{-kt/m}\right). \end{aligned} \tag{9}$$

Nonlinear Air Resistance

The model applies primarily to rapidly moving objects. It is obtained by the same method as the linear model, replacing the linear resistance term $kx'(t)$ by the nonlinear term $kx'(t)|x'(t)|$. The resulting model is

$$mx''(t) = -kx'(t)|x'(t)| - mg. \tag{10}$$

Velocity substitution $v = x'(t)$ gives first order equation

$$v'(t) = -(k/m)v(t)|v(t)| - g. \tag{11}$$

The model applies in particular to parachute flight and to certain projectile problems, like an arrow or bullet fired straight up.

Upward Launch. Separable equation (11) in the case $v(0) > 0$ for a launch upward becomes $v'(t) = -(k/m)v^2(t) - g$. The solution for $v(0) > 0$ is given below in (12); see *Technical Details* page 137. The equation $x'(t) = v(t)$ can be solved by quadrature. Then for some constants c and d

$$\begin{aligned} v(t) &= \sqrt{\frac{mg}{k}}\,\tan\left(\sqrt{\frac{kg}{m}}(c-t)\right), \\ x(t) &= d + \frac{m}{k}\,\ln\left|\cos\left(\sqrt{\frac{kg}{m}}(c-t)\right)\right|. \end{aligned} \tag{12}$$

Downward Launch. The case $v(0) < 0$ for an object launched downward or dropped will use the equation $v'(t) = (k/m)v^2(t) - g$; see *Technical Details*, page 138. Then for some constants c and d

$$\begin{aligned} v(t) &= \sqrt{\frac{mg}{k}}\,\tanh\left(\sqrt{\frac{kg}{m}}(c-t)\right), \\ x(t) &= d - \frac{m}{k}\,\ln\left|\cosh\left(\sqrt{\frac{kg}{m}}(c-t)\right)\right|. \end{aligned} \tag{13}$$

The **hyperbolic functions** appearing in (13) are *defined by*

$\cosh u = \frac{1}{2}\left(e^u + e^{-u}\right)$	Hyperbolic cosine.
$\sinh u = \frac{1}{2}\left(e^u - e^{-u}\right)$	Hyperbolic sine.
$\tanh u = \dfrac{e^u - e^{-u}}{e^u + e^{-u}}$	Hyperbolic tangent. Identity $\tanh u = \sinh u / \cosh u$.

The model applies to parachute problems in particular. Equation (13) and the limit formula $\lim_{|x|\to\infty} \tanh x = 1$ imply a *terminal velocity*

$$|v_\infty| = \sqrt{\frac{mg}{k}}.$$

The value is exactly the square root of the linear model terminal velocity. The falling body model (3) without air resistance effects allows the velocity to increase to unrealistic speeds. For instance, the terminal velocity of a raindrop falling from 3000 meters is about $25 - 35$ km/h, whereas the no air resistance model predicts about 870 km/h.

Modeling Remarks

It can be argued from air resistance models that projectiles spend more time falling to the ground than they spend reaching maximum height[5]; see Example 2.32. Simplistic models ignoring air resistance tend to over-estimate the maximum height of the projectile and the flight time; see Example 2.31. Falling bodies are predicted by air resistance models to have a *terminal velocity*.

Significant effects are ignored by the models of this text. Real projectiles are affected by spin and a flight path that is not planar. The **corkscrew** path of a bullet can cause it to miss a target, while a planar model predicts it will hit the target. The spin of a projectile can drastically alter its flight path and flight characteristics, as is known by players of table tennis, squash, court tennis, archery enthusiasts and gun club members.

Gravitational effects assumed constant may in fact not be constant along the flight path. This can happen in the soft touchdown problem for a lunar lander which activates retrorockets high above the moon's surface.

External effects like wind or the gravitational forces of nearby celestial bodies, ignored in simplistic models, may indeed produce significant effects. On the freeway, is it possible to throw an ice cube out the window ahead of your vehicle? Is it feasible to use forces from the moon to **assist** in the launch of an orbital satellite?

Parachutes

In a typical parachute problem, the jumper travels in a parabolic arc to the ground, buffeted about by up and down drafts in the atmosphere, but always moving in the direction determined by the airplane's flight. In short, a parachutist does not *fall* to the ground. Their flight path more closely resembles the path of a projectile and it is generally not a planar path.

Important to skydivers is an absolute limit to their speed, called the **terminal velocity**. It depends upon a number of physical factors, the dominant factor being body shape affecting area variable A of the **drag force**. See page . A parachutist with excess loose clothing will dive more slowly than when equipped with a tight lycra jump suit. When the parachute opens, the flight characteristics are dominated by physical factors of the open parachute.

The constant $k/m > 0$ is called the **drag factor**, where m is the mass and $k > 0$, appears in the resistive force equation $F = kv|v|$. In order for the parachute model to give a terminal velocity of 15 miles per hour, the drag factor must be approximately $k/m = 3/2$. Without the parachute, the skydiver can reach speeds of over 45 miles per hour, which corresponds to a drag factor $k/m < 1/2$.

[5] Racquetball, badminton, Lacrosse, tennis, squash, pickleball and table tennis players know about this effect and they use it in their game tactics and timing.

Who falls the greatest distance after 30 seconds, a 250-pound or a 110-pound parachutist? The answer is not always the layman's answer, because the 110-pound parachutist has *less* air resistance due to less body surface area but also *less* mass, making it difficult to compare the two drag factors.

Lunar Lander

A lunar lander is falling toward the moon's surface, in the radial direction, at a speed of 1000 miles per hour. It is equipped with retrorockets to retard the fall. In free space outside the gravitational effects of the moon the retrorockets provide a retardation thrust of 9 miles per hour per second of activation, e.g., 11 seconds of retrorocket power will slow the lander down by about 100 miles per hour.

A **soft touchdown** is made when the lander contacts the moon's surface falling at a speed of zero miles per hour. This ideal situation can be achieved by turning on the retrorockets at the right moment.

The lander is greatly affected by the gravitational field of the moon. Ignoring this field gives a gross overestimate for the activation time, causing the lander to reverse its direction and never reach the surface. The layman answer of $1000/9 \approx 112$ seconds to touchdown from an altitude of about 16 miles is incorrect by about 10 miles, causing the lander to crash at substantial speed into the lunar surface.

Escape velocity

Is it possible to fire a projectile from the earth's surface and reach the moon? The science fiction author Jules Verne, in his 1865 novel *From the Earth to the Moon*, seems to believe it is possible. Modern calculations give the initial **escape velocity** v_0 as about $25,000$ miles per hour. There is no record of this actually being tested, so the number $25,000$ remains a theoretical estimate.

This is a different problem than powered rocket flight. All the power must be applied initially, and it is not allowed to apply power during flight to the moon. Imagine instead a deep hole, in which a rocket is launched, the power being turned off just as the rocket exits the hole. The rocket has to coast to the moon, using just the velocity gained during launch.

Newton's law of universal gravitation gives m_1m_2G/r^2 as the magnitude of the force of attraction between two point-masses m_1, m_2 separated by distance r. The equation $g = Gm_2/R^2$ gives the acceleration due to gravity at the surface of the planet. For the earth, $g = 9.8$ meters per second per second and $R = 6,370,000$ meters.

A spherical projectile of mass m_1 hurled straight up from the surface of a planet moves in the radial direction. Ignoring air resistance and external gravitational forces, Newton's law implies the distance $y(t)$ traveled by the projectile satisfies

$$(14) \qquad m_1 y''(t) = -\frac{m_1 m_2 G}{(y(t)+R)^2}, \quad y(0)=0, \quad y'(0)=v_0,$$

where R is the radius of the planet, m_2 is its mass and G is the experimentally measured universal gravitation constant. Using $gR^2 = Gm_2$ and canceling m_1 in (14) gives

$$(15) \qquad y''(t) = -\frac{gR^2}{(y(t)+R)^2}, \quad y(0)=0, \quad y'(0)=v_0.$$

The projectile **escapes** the planet if $y(t) \to \infty$ as $t \to \infty$. The **escape velocity problem** asks which minimal value of v_0 causes escape.

To solve the escape velocity problem, multiply equation (15) by $y'(t)$, then integrate over $[0,t]$ and use the initial conditions $y(0)=0$, $y'(0)=v_0$ to obtain

$$\frac{1}{2}\left((y'(t))^2 - (v_0)^2\right) = \frac{gR^2}{y(t)+R} - Rg.$$

The square term $(y'(t))^2$ being nonnegative gives the inequality

$$0 \le (v_0)^2 + \frac{2gR^2}{y(t)+R} - 2Rg.$$

If $y(t) \to \infty$, then $v_0^2 \ge 2Rg$, which gives the *escape velocity*

$$(16) \qquad v_0 = \sqrt{2gR}.$$

For the earth, $v_0 \approx 11,174$ meters per second, which is slightly more than $25,000$ miles per hour.

Examples

Example 2.29 (Free Fall)

A ball is thrown straight up from the roof of a 100-foot building and allowed to fall to the ground. Assume initial velocity $v_0 = 32$ miles per hour. Estimate the maximum height of the ball and its flight time to the ground.

Solution: The maximum height H and flight time T are given by

$$H = 134.41 \text{ ft}, \quad T = 4.36 \text{ sec}.$$

Details: In *fps* units, $v_0 = 32(5280)/(3600) = 46.93$ ft/sec. Using solution (4) gives for $x_0 = 100$ and $v_0 = 46.93$

$$x(t) = -16t^2 + 100 + 46.93t.$$

Then $x(t) = H = \max$ when $x'(t) = 0$, which happens at $t = 46.93/32$. Therefore, $H = x(46.93/32) = 134.41$. The flight time T is given by the equation $x(T) = 0$ (the ground is $x = 0$). Solving this quadratic equation for $T > 0$ gives $T = 4.36$ seconds.

Example 2.30 (Lunar Lander)

A lunar lander falls to the moon's surface at $v_0 = -960$ miles per hour. The retrorockets in free space provide a deceleration effect on the lander of $a = 18,000$ miles per hour per hour. Estimate the retrorocket activation height above the surface which will give the lander zero touch-down velocity.

Solution: Presented here are two models, one which assumes the moon's gravitational field is constant and another which assumes it is variable. The results obtained for the activation height are different: 93.3 miles for the constant field model and 80.1 miles for the variable field model. The flight times to touchdown are estimated to be 11.7 minutes and 10.4 minutes, respectively.

Calculations use *mks* units: $v_0 = -429.1584$ meters per second and $a = 2.2352$ meters per second per second.

Constant field model. Let's assume constant gravitational acceleration $\mathcal{G}$ due to the moon. Other gravitational effects are ignored.

The acceleration value $\mathcal{G}$ is found in *mks* units from the formula

$$\mathcal{G} = \frac{G\,m_1}{R^2}.$$

Symbols: $m_1 = 7.36 \times 10^{22}$ kilograms and $R = 1.74 \times 10^6$ meters (1740 kilometers, 1081 miles), which are the mass and radius of the moon. Newton's universal gravitation constant is $G \approx 6.6726 \times 10^{-11}$ $\mathrm{N(m/kg)^2}$. Then $\mathcal{G} = 1.622087990$.

The lander itself has mass m. Let $r(t)$ be the distance from the lander to the surface of the moon. The value $r(0)$ is the height above the moon when the retrorockets are activated for the soft landing at time t_0. Then force analysis and Newton's second law implies the differential equation model

$$mr''(t) = ma - m\mathcal{G}, \quad r(t_0) = 0, \quad r'(t_0) = 0, \quad r'(0) = v_0.$$

The objective is to find $r(0)$. Cancel m, then integrate twice to obtain the quadrature solution

$$\begin{aligned} r'(t) &= (a - \mathcal{G})t + v_0, \\ r(t) &= (a - \mathcal{G})t^2/2 + v_0 t + r(0). \end{aligned}$$

Then $r'(t_0) = 0$ and $r(t_0) = 0$ give the equations

$$(a - \mathcal{G})t + v_0 = 0, \quad r(0) = -v_0 t_0 - (a - \mathcal{G})t_0^2/2.$$

The symbols in *mks* units: $a = 2.2352$, $v_0 = -429.1584$, $\mathcal{G} = 1.622087990$. Solving simultaneously provides the numerical answers

$$t_0 = 11.66 \text{ minutes}, \quad r(0) = 150.16 \text{ kilometers} = 93.3 \text{ miles}.$$

The conversion uses 1 mile = 1.609344 kilometers.

Variable field model. The constant field model will be modified to obtain this model. All notation developed above applies. We will replace the constant acceleration $\mathcal{G}$ by the variable acceleration $G\,m_1/(R + r(t))^2$. Then the model is

$$mr''(t) = ma - \frac{G\,m_1\,m}{(R + r(t))^2}, \quad r(t_0) = 0, \quad r'(t_0) = 0, \quad r'(0) = v_0.$$

Multiply this equation by $r'(t)/m$ and integrate. Then

$$\frac{(r'(t))^2}{2} = ar(t) + \frac{G\,m_1}{R + r(t)} + c, \quad \text{where} \quad c \equiv -\frac{G\,m_1}{R}.$$

We will find $r(0)$, the height above the moon. The equation to solve for $r(0)$ is found by substitution of $t = 0$ into the previous equation:

$$\frac{(r'(0))^2}{2} = ar(0) + \frac{Gm_1}{R + r(0)} - \frac{Gm_1}{R}.$$

After substitution of known values, the quadratic equation for $x = r(0)$ is:

$$92088.46615 = 2.2352\,x + \frac{2822179.310}{1 + x/1740000} - 2822179.310$$

Solving for the positive root gives $r(0) \approx 127.23$ kilometers or 79.06 miles. The analysis does not give the flight time t_0 directly, but it is approximately 10.4 minutes: see the exercises.

Answer check. A similar analysis is done in Edwards and Penney [EP2] for the case $a = 4$ meters per second per second, $v_0 = -450$ meters per second, with result $r(0) \approx 41.87$ kilometers. In their example, the retrorocket thrust is nearly doubled, resulting in a lower activation height. Substitute $v_0 = -450$ and $a = 4$ in the variable field model to obtain agreement: $r(0) \approx 41.90$ kilometers. The constant field model gives $r(0) \approx 42.58$ kilometers and $t_0 \approx 3.15$ minutes.

Example 2.31 (Flight Time and Maximum Height)

Show that the maximum height and the ascent time of a projectile are over-estimated by a model that ignores air resistance.

Solution: Treated here is the case of a projectile launched straight up from the ground $x = 0$ with velocity $v_0 > 0$. The ascent time is denoted t_1 and the maximum height M is then $M = x(t_1)$.

No air resistance. Consider the velocity model $v' = -g$, $v(0) = v_0$. The solution is $v = -gt + v_0$, $x = -gt^2/2 + v_0 t$. Then maximum height M occurs at $v'(t_1) = 0$ which gives $t_1 = v_0/g$ and $M = x(t_1) = t_1(v_0 - gt_1/2) = gv_0^2/2$.

Linear air resistance. Consider the model $v' = -\rho v - g$, $v(0) = v_0$. This is a Newton cooling equation in disguise, with solution given by equation (9), where $\rho = k/m$. Then t_1 is a function of (ρ, v_0) satisfying $ge^{\rho t_1} = v_0\rho + g$, hence t_1 is given by the equation

$$t_1(\rho, v_0) = \frac{1}{\rho} \ln \left| \frac{v_0\rho + g}{g} \right|. \tag{17}$$

The limit of $t_1 = t_1(\rho, v_0)$ as $\rho \to 0$ is the ascent time v_0/g of the no air resistance model. Verified in the exercises are the following.

Lemma 2.2 (Linear Ascent Time) The ascent time t_1 for linear air resistance satisfies $t_1(\rho, v_0) < v_0/g$.

The lemma implies that the rise time for linear air resistance is less than the rise time for no air resistance.

The inequality $v' = -\rho v - g < -g$ holds for $v > 0$, therefore $v(t) < -gt + v_0$ and $x(t) < -gt^2/2 + v_0 t =$ height for the no air resistance model. Thus the maximum height $x(t_1)$ is less than the maximum height for the no air resistance model, by Lemma 2.2; see the exercises page 141.

Nonlinear air resistance. The example is technically done, because it has been shown that the answers for t_1 and M decrease when using the linear model. Similar results can be stated for the nonlinear model $v' = \rho v|v| - g$; see the exercises page 141.

Example 2.32 (Modeling)

Argue from nonlinear air resistance models that a projectile takes more time to fall to the ground than it takes to reach maximum height.

Solution: The model will be the nonlinear model of the text, which historically goes back to Isaac Newton. The linear air resistance model, appropriate for slowly moving projectiles, is not considered in this example.

Let t_1 and t_2 be the ascent and fall times, so that the total flight time from the ground to maximum height and then to the ground again is $t_1 + t_2$.

The times t_1, t_2 are functions of the initial velocity $v_0 > 0$. As v_0 limits to zero, both t_1 and t_2 limit to zero. Inequality $t_2 dt_2/dv_0 - t_1 dt_1/dv_0 > 0$ is derived in Lemma 2.7 below. Integrate the inequality on variable v_0, then $\frac{1}{2}(t_2^2 - t_1^2) > 0$, from which it follows that $t_2 > t_1$ for $v_0 > 0$. Meaning: the projectile takes more time to fall to the ground (t_2) than it takes to reach maximum height (t_1).

Define nonlinear functions

$$f_1(v) = -(k/m)v^2 - g, \quad f_2(v) = (k/m)v^2 - g$$

The **ascent** or **rise** is controlled with velocity $v_1 > 0$ satisfying $v_1' = f_1(v_1)$, $v_1(0) = v_0 > 0$, $v_1(t_1) = 0$. The maximum height reached is $y_0 = \int_0^{t_1} v_1(t)dt$. The **descent** of **fall** is controlled with velocity $v_2(t)$ satisfying $v_2' = f_2(v_2)$, $v_2(t_1) = 0$. The flight ends at time $T = t_1 + t_2$, determined by $0 = y_0 + \int_{t_1}^{T} v_2(t)dt$.

Details of proof involve a number of technical results, some of which depend upon the formulas $f_1(v) = -(k/m)v^2 - g$, $f_2(v) = (k/m)v^2 - g$.

Lemma 2.3 The solution v_2 satisfies $v_2(t) = w(t - t_1)$, where w is defined by $w' = f_2(w)$, $w(0) = 0$. The solution w does not involve variables v_0, t_1, t_2.

Lemma 2.4 Assume f is continuously differentiable. Let $v(t, v_0)$ be the solution of $v' = f(v)$, $v(0) = v_0$. Then

$$\frac{dv}{dv_0} = e^{\int_0^t f'(v(t,v_0))dt}.$$

The function $z = dv/dv_0$ solves the linear problem $z' = f'(v(t, v_0))z$, $z(0) = 1$.

Lemma 2.5

$$\frac{dt_1}{dv_0} = \frac{1}{g} e^{-2k \int_0^{t_1} v_1(t,v_0)dt/m}.$$

Lemma 2.6

$$\frac{dt_2}{dv_0} = \frac{-1}{v_2(t_1 + t_2)} \int_0^{t_1} e^{-2k \int_0^t v_1(r,v_0)dr/m} dt.$$

Lemma 2.7

$$t_2\frac{dt_2}{dv_0} - t_1\frac{dt_1}{dv_0} > 0.$$

Proof of Lemma 2.7. Lemmas 2.3 to 2.6 will be applied. Define $w(t)$ by Lemma 2.3. Because $w' = f_2(w) = (k/m)w^2 - g$, then $f_2(w) \geq -g$ which implies $w(t) \geq w(0) - gt$. Using $w(0) = 0$ implies $v_2(t_1 + t_2) = w(t_2) \geq -gt_2$ and finally, using $w(t) < 0$ for $0 < t \leq t_2$,

$$\frac{1}{gt_2} \leq \frac{-1}{v_2(t_1 + t_2)}.$$

Multiply this inequality by $e^{u(t)}$, $u(t) = -2k\int_0^t v_1(r, v_0)dr/m$. Integrate over $t = 0$ to $t = t_1$. Then Lemma 2.6 implies

$$\frac{1}{gt_2}\int_0^{t_1} e^{u(t)}dt \leq \frac{dt_2}{dv_0}.$$

Because $u(t) > u(t_1)$, then

$$\frac{1}{gt_2}\int_0^{t_1} e^{u(t_1)}dt < \frac{dt_2}{dv_0}.$$

This implies by Lemma 2.5 the inequality

$$\frac{t_1}{t_2}\frac{dt_1}{dv_0} = \frac{t_1}{gt_2}e^{u(t_1)} < \frac{dt_2}{dv_0},$$

or $t_2dt_2/dv_0 - t_1dt_1/dv_0 > 0$. ■

Proof of Lemma 2.3. The function $z(t) = v_2(t + t_1)$ satisfies $z' = f_2(z)$, $z(0) = 0$ (an answer check for the reader). Function $w(t)$ is defined to solve $w' = f_2(w)$, $w(0) = 0$. By uniqueness, $z(t) \equiv w(t)$, or equivalently, $w(t) = v_2(t + t_1)$. Replace t by $t - t_1$ to obtain $v_2(t) = w(t - t_1)$.

Proof of Lemma 2.4. The exponential formula for dv_2/dv_0 is the unique solution of the first order initial value problem. It remains to show that the initial value problem is satisfied. Instead of doing the answer check, we motivate how to find the initial value problem. First, differentiate across the equation $v_2' = f_2(v_2)$ with respect to variable v_0 to obtain $z' = f_2'(v_2)z$ where $z = dv_2/dv_0$. Secondly, differentiate the relation $v_2(0, v_0) = v_0$ on variable v_0 to obtain $z(0) = 1$. The details of the answer check focus on showing Newton quotients converge to the given answer.

Proof of Lemma 2.5. Start with the determining equation $v_1(t_1, v_0) = 0$. Differentiate using the chain rule on variable v_0 to obtain the relation

$$v_1'(t_1, v_0)\frac{dt_1}{dv_0} + \frac{dv_1}{dv_0}(t_1, v_0) = 0.$$

Because $f_1'(u) = -2ku/m$, then the preceding lemma implies that dv_1/dv_0 is the same exponential function as in this Lemma. Also, $v_1(t_1, v_0) = 0$ implies $v_1'(t_1, v_0) = f_1(0) = -g$. Substitution gives the formula for dt_1/dv_0.

Proof of Lemma 2.6. Start with $y_0 = \int_0^{t_1} v_1(t, v_0)dt$ and $y(t) = y_0 + \int_{t_1}^t v_2(t)dt$. Then $0 = y(t_2 + t_1)$ implies that

$$\begin{aligned} 0 &= y(t_1 + t_2) \\ &= \int_0^{t_1} v_1(t, v_0)dt + \int_0^{t_2} v_2(t + t_1)dt \\ &= \int_0^{t_1} v_1(t, v_0)dt + \int_0^{t_2} w(t)dt. \end{aligned}$$

Because $w(t)$ is independent of t_1, t_2, v_0 and $v_1(t_1, v_0) = 0$, then differentiation on v_0 across the preceding formula gives

$$\begin{aligned} 0 &= \tfrac{d}{dv_0}\int_0^{t_1} v_1(t,v_0)dt + w(t_2)\tfrac{dt_2}{dv_0} \\ &= v_1(t_1,v_0)\tfrac{dt_1}{dv_0} + \int_0^{t_1} \tfrac{dv_1}{dv_0}(t,v_0)dt + w(t_2)\tfrac{dt_2}{dv_0} \\ &= 0 + \int_0^{t_1} e^{u(t)}dt + w(t_2)\tfrac{dt_2}{dv_0} \end{aligned}$$

where $u(t) = -2k\int_0^t v_1(r,v_0)dr/m$. Use $w(t_2) = v_2(t_2+t_1)$ after division by $w(t_2)$ in the last display to obtain the formula.

Details and Proofs

Proof for Equation (4). The method of quadrature is applied as follows.

$x''(t) = -g$	The given differential equation.
$\int x''(t)dt = \int -gdt$	Quadrature step.
$x'(t) = -gt + c_1$	Fundamental theorem of calculus.
$\int x'(t)dt = \int(-gt+c_1)dt$	Quadrature step.
$x(t) = -g\frac{t^2}{2} + c_1t + c_2$	Fundamental theorem of calculus.

Using initial conditions $x(0) = x_0$ and $x'(0) = v_0$ it follows that $c_1 = v_0$ and $c_2 = x_0$. These steps verify the formula $x(t) = -gt^2/2 + x_0 + v_0t$.

Technical Details for Equation (9).

$v'(t) + (k/m)v(t) = -g$	Standard linear form.
$\frac{(Qv)'}{Q} = -g$	Integrating factor $Q = e^{kt/m}$.
$(Qv)' = -gQ$	Quadrature form.
$Qv = -mgQ/k + c$	Method of quadrature.
$v = -mg/k + c/Q$	Velocity equation.
$v = -\frac{mg}{k} + \left(v(0) + \frac{mg}{k}\right)e^{-kt/m}$	Evaluate c and use $Q = e^{kt/m}$.

The equation $x(t) = x(0) + \int_0^t v(r)dr$ gives the last relation in (9):

$$x(t) = x(0) - \frac{mg}{k}t + \frac{m}{k}\left(v(0) + \frac{mg}{k}\right)\left(1 - e^{-kt/m}\right).$$

Technical Details for Equation (12), $v(0) > 0$.

$v'(t) = -(k/m)v^2(t) - g$	The upward launch equation.
$u'(t) = \sqrt{\frac{kg}{m}}(1+u^2(t))$	Change of variables $u = \sqrt{\frac{k}{mg}}\,v$.
$\frac{u'(t)}{1+u^2(t)} = -\sqrt{\frac{kg}{m}}$	A separated form.
$\arctan(u(t)) = -\sqrt{\frac{kg}{m}}t + c_1$	Quadrature.
$u(t) = \tan\left(c_1 - \sqrt{\frac{kg}{m}}t\right)$	Take the tangent of both sides.

$$\begin{aligned}
v(t) &= \sqrt{\tfrac{mg}{k}}\,\tan\left(\sqrt{\tfrac{kg}{m}}(c-t)\right) && \text{Define } c_1 = \sqrt{\tfrac{kg}{m}}c. \\
x(t) &= \textstyle\int v(t)dt && \text{Quadrature method.} \\
&= d + \tfrac{m}{k}\ln\left|\cos\left(\sqrt{\tfrac{kg}{m}}(c-t)\right)\right| && \text{Integration constant } d.
\end{aligned}$$

Technical Details for Equation (13), $v(0) < 0$.

$$\begin{aligned}
v'(t) &= (k/m)v^2(t) - g && \text{Downward launch equation.} \\
u'(t) &= \sqrt{\tfrac{kg}{m}}\left(u^2(t) - 1\right) && \text{Change of variables } u = \sqrt{\tfrac{k}{mg}}\,v. \\
\tfrac{u'(t)}{u^2(t)-1} &= \sqrt{\tfrac{kg}{m}} && \text{A separated form.} \\
-\operatorname{arctanh}(u) &= 2t\sqrt{\tfrac{kg}{m}} + c_1 && \text{Quadrature method and tables.} \\
u &= \tanh\left(\sqrt{\tfrac{kg}{m}}(c-t)\right) && \text{Define } c \text{ by } \sqrt{\tfrac{kg}{m}}\,c = -c_1. \\
v(t) &= \sqrt{\tfrac{mg}{k}}\,\tanh\left(\sqrt{\tfrac{kg}{m}}(c-t)\right) && \text{Use } v = \sqrt{\tfrac{mg}{k}}\,u. \\
x(t) &= \textstyle\int v(t)dt && \text{Quadrature.} \\
&= d - \tfrac{m}{k}\ln\left|\cosh\left(\sqrt{\tfrac{kg}{m}}(c-t)\right)\right| && \text{Integration constant } d.
\end{aligned}$$

Exercises 2.6

Newton's Laws

Review of units and conversions.

1. An object weighs 100 pounds. Find its mass in slugs and kilograms.

2. An object has mass 50 kilograms. Find its mass in slugs and its weight in pounds.

3. Convert from **fps** to **mks** systems: position 1000, velocity 10, acceleration 2.

4. Derive $g = \dfrac{Gm}{R^2}$, where m is the mass of the earth and R is its radius.

Velocity and Acceleration

Find the velocity x' and acceleration x''.

5. $x(t) = 16t^2 + 100$

6. $x(t) = 16t^2 + 10t + 100$

7. $x(t) = t^3 + t + 1$

8. $x(t) = t(t-1)(t-2)$

Free Fall with Constant Gravity

Solve using the model $x''(t) = -g$, $x(0) = x_0$, $x'(0) = v_0$.

9. A brick falls from a tall building, straight down. Find the distance it fell and its speed at three seconds.

10. An iron ingot falls from a tall building, straight down. Find the distance it fell and its speed at four seconds.

11. A ball is thrown straight up from the ground with initial velocity 66 feet per second. Find its maximum height.

12. A ball is thrown straight up from the ground with initial velocity 88 feet per second. Find its maximum height.

13. An arrow is shot straight up from the ground with initial velocity 23 meters per second. Find the flight time back to the ground.

14. An arrow is shot straight up from the ground with initial velocity 44 meters per second. Find the flight time back to the ground.

15. A car travels 140 kilometers per hour. Brakes are applied, with deceleration 10 meters per second per second. Find the distance the car travels before stopping.

16. A car travels 120 kilometers per hour. Brakes are applied, with deceleration 40 feet per second per second. Find the distance the car travels before stopping.

17. An arrow is shot straight down from a height of 500 feet, with initial velocity 44 feet per second. Find the flight time to the ground and its impact speed.

18. An arrow is shot straight down from a height of 200 meters, with initial velocity 13 meters per second. Find the flight time to the ground and its impact speed.

Linear Air Resistance

Solve using the linear air resistance model $mx''(t) = -kx'(t) - mg$. An equivalent model is $x'' = -\rho x' - g$, where $\rho = k/m$ is the drag factor.

19. An arrow is shot straight up from the ground with initial velocity 23 meters per second. Find the flight time back to the ground. Assume $\rho = 0.035$.

20. An arrow is shot straight up from the ground with initial velocity 27 meters per second. Find the maximum height. Assume $\rho = 0.04$.

21. A parcel is dropped from an aircraft at $32,000$ feet. It has a parachute that opens automatically after 25 seconds. Assume drag factor $\rho = 0.16$ without the parachute and $\rho = 1.45$ with it. Find the descent time to the ground.

22. A first aid kit is dropped from a helicopter at $12,000$ feet. It has a parachute that opens automatically after 15 seconds. Assume drag factor $\rho = 0.12$ without the parachute and $\rho = 1.55$ with it. Find the impact speed with the ground.

23. A motorboat has velocity v satisfying $1100v'(t) = 6000 - 110v$, $v(0) = 0$. Find the maximum speed of the boat.

24. A motorboat has velocity v satisfying $1000v'(t) = 4000 - 90v$, $v(0) = 0$. Find the maximum speed of the boat.

25. A parachutist falls until his speed is 65 miles per hour. He opens the parachute. Assume parachute drag factor $\rho = 1.57$. About how many seconds must elapse before his speed is reduced to within 1% of terminal velocity?

26. A parachutist falls until his speed is 120 kilometers per hour. He opens the parachute. Assume drag factor $\rho = 1.51$. About how many seconds must elapse before his speed is reduced to within 2% of terminal velocity?

27. A ball is thrown straight up with initial velocity 35 miles per hour. Find the ascent time and the descent time. Assume drag factor 0.042

28. A ball is thrown straight up with initial velocity 60 kilometers per hour. Find the ascent time and the descent time. Assume drag factor 0.042

Linear Ascent and Descent Times

Find the ascent time t_1 and the descent time t_2 for the linear model $x'' = -\rho x' - g$, $x(0) = 0$, $x'(0) = v_0$ where $\rho = k/m$ is the drag factor. Unit system **fps**. Computer algebra system expected.

29. $\rho = 0.01$, $v_0 = 50$

30. $\rho = 0.015$, $v_0 = 30$

31. $\rho = 0.02$, $v_0 = 50$

32. $\rho = 0.018$, $v_0 = 30$

33. $\rho = 0.022$, $v_0 = 50$

34. $\rho = 0.025$, $v_0 = 30$

35. $\rho = 1.5$, $v_0 = 50$

36. $\rho = 1.55$, $v_0 = 30$

37. $\rho = 1.6$, $v_0 = 50$

38. $\rho = 1.65$, $v_0 = 30$

39. $\rho = 1.45$, $v_0 = 50$

40. $\rho = 1.48$, $v_0 = 30$

Nonlinear Air Resistance

Assume ascent velocity v_1 satisfies $v_1' = -\rho v_1^2 - g$. Assume descent velocity v_2 satisfies $v_2' = \rho v_2^2 - g$. Motion from the ground $x = 0$. Let t_1 and t_2 be the ascent and descent times, so that $t_1 + t_2$ is the flight time. Let $g = 9.8$, $v_1(0) = v_0$, $v_1(t_1) = v_2(t_1) = 0$, units *mks*. Define $M =$ maximum height and $v_f =$ impact velocity. Computer algebra system expected.

41. Let $\rho = 0.0012$, $v_0 = 50$. Find t_1, t_2.

42. Let $\rho = 0.0012$, $v_0 = 30$. Find t_1, t_2.

43. Let $\rho = 0.0015$, $v_0 = 50$. Find t_1, t_2.

44. Let $\rho = 0.0015$, $v_0 = 30$. Find t_1, t_2.

45. Let $\rho = 0.001$, $v_0 = 50$. Find M, v_f.

46. Let $\rho = 0.001$, $v_0 = 30$. Find M, v_f.

47. Let $\rho = 0.0014$, $v_0 = 50$. Find M, v_f.

48. Let $\rho = 0.0014$, $v_0 = 30$. Find M, v_f.

49. Find t_1, t_2, M and v_f for $\rho = 0.00152$, $v_0 = 60$.

50. Find t_1, t_2, M and v_f for $\rho = 0.00152$, $v_0 = 40$.

Terminal Velocity

Find the terminal velocity for (a) a linear air resistance $a(t) = \rho v(t)$ and (b) a nonlinear air resistance $a(t) = \rho v^2(t)$. Use the model equation $v' = a(t) - g$ and the given drag factor ρ, **mks** units.

51. $\rho = 0.15$

52. $\rho = 0.155$

53. $\rho = 0.015$

54. $\rho = 0.017$

55. $\rho = 1.5$

56. $\rho = 1.55$

57. $\rho = 2.0$

58. $\rho = 1.89$

59. $\rho = 0.001$

60. $\rho = 0.0015$

Parachutes

A skydiver has velocity v_0 and height $5,500$ feet when the parachute opens. Velocity $v(t)$ is given by (a) linear resistance model $v' = -\rho v - g$ or (b) nonlinear resistance downward model $v' = \rho v^2 - g$. Given the drag factor ρ and the parachute-open velocity v_0, compute the elapsed time until the parachutist slows to within 2% of terminal velocity. Then find the flight time from parachute open to the ground. Report two values for (a) and two values for (b).

61. $\rho = 1.446$, $v_0 = -116$ ft/sec.

62. $\rho = 1.446$, $v_0 = -84$ ft/sec.

63. $\rho = 1.2$, $v_0 = -116$ ft/sec.

64. $\rho = 1.2$, $v_0 = -84$ ft/sec.

65. $\rho = 1.01$, $v_0 = -120$ ft/sec.

66. $\rho = 1.01$, $v_0 = -60$ ft/sec.

67. $\rho = 0.95$, $v_0 = -10$ ft/sec.

68. $\rho = 0.95$, $v_0 = -5$ ft/sec.

69. $\rho = 0.8$, $v_0 = -66$ ft/sec.

70. $\rho = 0.8$, $v_0 = -33$ ft/sec.

Lunar Lander

A lunar lander falls to the moon's surface at v_0 miles per hour. The retrorockets in free space provide a deceleration effect on the lander of a miles per hour per hour. Estimate the retrorocket activation height above the surface which will give the lander zero touch-down velocity. Follow Example 2.30, page 133.

71. $v_0 = -1000$, $a = 18000$

72. $v_0 = -980$, $a = 18000$

73. $v_0 = -1000$, $a = 20000$

74. $v_0 = -1000$, $a = 19000$

75. $v_0 = -900$, $a = 18000$

76. $v_0 = -900$, $a = 20000$

77. $v_0 = -1100$, $a = 22000$

78. $v_0 = -1100$, $a = 21000$

79. $v_0 = -800$, $a = 18000$

80. $v_0 = -800$, $a = 21000$

Escape velocity

Find the escape velocity of the given planet, given the planet's mass m and radius R.

81. **(Planet A)** $m = 3.1 \times 10^{23}$ kilograms, $R = 2.4 \times 10^7$ meters.

82. **(Mercury)** $m = 3.18 \times 10^{23}$ kilograms, $R = 2.43 \times 10^6$ meters.

83. **(Venus)** $m = 4.88 \times 10^{24}$ kilograms, $R = 6.06 \times 10^6$ meters.

84. **(Mars)** $m = 6.42 \times 10^{23}$ kilograms, $R = 3.37 \times 10^6$ meters.

85. **(Neptune)** $m = 1.03 \times 10^{26}$ kilograms, $R = 2.21 \times 10^7$ meters.

86. **(Jupiter)** $m = 1.90 \times 10^{27}$ kilograms, $R = 6.99 \times 10^7$ meters.

87. **(Uranus)** $m = 8.68 \times 10^{25}$ kilograms, $R = 2.33 \times 10^7$ meters.

88. **(Saturn)** $m = 5.68 \times 10^{26}$ kilograms, $R = 5.85 \times 10^7$ meters.

Lunar Lander Experiments

89. **(Lunar Lander)** Verify that the variable field model for Example 2.30 gives a soft landing flight model in MKS units

$$\begin{aligned} u''(t) &= 2.2352 - \frac{c_1}{(c_2 + u(t))^2}, \\ u(0) &= 127254.1306, \\ u'(0) &= -429.1584, \end{aligned}$$

where $c_1 = 4911033599000$ and $c_2 = 1740000$.

90. **(Lunar Lander: Numerical Experiment)** Using a computer, solve the flight model of the previous exercise. Determine the flight time $t_0 \approx 625.6$ seconds by solving $u(t) = 0$ for t.

Details and Proofs

91. **(Linear Rise Time)** Using the inequality $e^u > 1 + u$ for $u > 0$, show that the ascent time t_1 in equation (17) satisfies

$$g(1 + \rho t_1) < g e^{\rho t_1} = v_0 \rho + g.$$

Conclude that $t_1 < v_0/g$, proving Lemma 2.2.

92. **(Linear Maximum)** Verify that Lemma 2.2 plus the inequality $x(t) < -gt^2/2 + v_0 t$ imply $x(t_1) < g v_0^2/2$. Conclude that the maximum for $\rho > 0$ is less than the maximum for $\rho = 0$.

93. **(Linear Rise Time)** Consider the ascent time $t_1(\rho, v_0)$ given by equation (17). Prove that

$$\frac{dt_1}{d\rho} = \frac{\ln \frac{g}{v0\rho + g}}{\rho^2} + \frac{v0}{\rho(v0\rho + g)}.$$

94. **(Linear Rise Time)** Consider $dt_1(\rho, v_0)/d\rho$ given in the previous exercise. Let $\rho = gx/v_0$. Show that $dt_1/d\rho < 0$ by considering properties of the function $-(x+1)\ln(x+1) + x$. Then prove Lemma 2.2.

95. **(Compare Rise Times)** The ascent time for nonlinear model $v' = -g - \rho v^2$ is less than the ascent time for linear model $u' = -g - \rho u$. Verify for $\rho = 1$, $g = 32$ ft/sec/sec and initial velocity 50 ft/sec.

96. **(Compare Fall Times)** The descent time for nonlinear model $v' = \rho v^2 - g$, $v(0) = 0$ is greater than the descent time for linear model $u' = -\rho u - g$, $u(0) = 0$. Verify for $\rho = 1$, $g = 32$ ft/sec/sec and maximum heights both 100 feet.

2.7 Logistic Equation

The 1845 work of Pierre Francois Verhulst (1804–1849), Belgian demographer and mathematician, modified the classical growth-decay equation $y' = ky$ by replacing k by $a - by$ to obtain the **logistic equation**

$$y' = (a - by)y. \tag{1}$$

The solution of the logistic equation (1) is (details on page 11)

$$y(t) = \frac{ay(0)}{by(0) + (a - by(0))e^{-at}}. \tag{2}$$

The logistic equation (1) applies not only to human populations but also to populations of fish, animals and plants, such as yeast, mushrooms or wildflowers. The y-dependent growth rate $k = a - by$ allows the model to have a finite *limiting population* a/b. The constant $M = a/b$ is called the **carrying capacity** by demographers. Verhulst introduced the terminology *logistic curves* for the solutions of (1).

To use the Verhulst model, a demographer must supply three population counts at three different times; these values determine the constants a, b and $y(0)$ in solution (2).

Logistic Models

Below are some variants of the basic logistic model known to researchers in medicine, biology and ecology.

Limited Environment. A container of $y(t)$ flies has a *carrying capacity* of N insects. A growth-decay model $y' = Ky$ with combined growth-death rate $K = k(N - y)$ gives the model $y' = k(N - y)y$.

Spread of a Disease. The initial size of the susceptible population is N. Then y and $N - y$ are the number of infectives and susceptibles. Chance encounters spread the incurable disease at a rate proportional to the infectives and the susceptibles. The model is $y' = ky(N - y)$. The spread of rumors has an identical model.

Explosion–Extinction. The number $y(t)$ of alligators in a swamp can satisfy $y' = Ky$ where the growth-decay symbol K is proportional to $y - N$ and N is a **threshold population**. The logistic model $y' = k(y - N)y$ gives **extinction** for initial populations smaller than M and a *doomsday* population **explosion** $y(t) \to \infty$ for initial populations greater than M. This model ignores harvesting.

Constant Harvesting. The number $y(t)$ of fish in a lake can satisfy a logistic model $y' = (a - by)y - h$, provided fish are **harvested** at a constant rate $h > 0$. This model can be written as $y' = k(M - y)(y - N)$ for small harvesting rates h, where M is the *carrying capacity* and N is the *threshold population.*

Variable Harvesting. The special logistic model $y' = (a-by)y-hy$ results by **harvesting** at a non-constant rate proportional to the present population y. The effect is to decrease the natural growth rate a by the constant amount h in the standard logistic model.

Restocking. The equation $y' = (a-by)y-h\sin(\omega t)$ models a logistic population that is periodically harvested and restocked with maximal rate $h > 0$. The period is $T = 2\pi/\omega$. The equation might model extinction for stocks less than some threshold population y_0, and otherwise a stable population that oscillates about an ideal carrying capacity a/b with period T.

Example 2.33 (Limited Environment)

Find the equilibrium solutions and the carrying capacity for the logistic equation $P' = 0.04(2 - 3P)P$. Then solve the equation.

Solution: The given differential equation can be written as the separable autonomous equation $P' = G(P)$ where $G(y) = 0.04(2-3P)P$. Equilibria are obtained as $P = 0$ and $P = 2/3$, by solving the equation $G(P) = 0.04(2 - 3P)P = 0$. The carrying capacity is the stable equilibrium $P = 2/3$; here we used the derivative $G'(P) = 0.04(2 - 6P)$ and evaluations $G'(0) > 0$, $G'(2/3) < 0$ to determine that $P = 2/3$ is a stable sink or funnel.

Example 2.34 (Spread of a Disease)

Find the number of infectives, the number of susceptibles and the rate of spread of the disease at $t = 4$ months for logistic model $y' = \frac{15}{1000}(10000 - y)y$, $y(0) = 200$.

Solution:

Answer: By month 4, about 8900 were infected, about 1100 were not infected and the disease was spreading at a rate of about 1450 per month.

Details: Write the differential equation in the form $y' = (a - by)y$ with $a = 15/10$, $b = \frac{15}{100000}$. Let $M = a/b = 10000$. The number of infectives after 4 months is $y(4)$ and the number of susceptibles is $M - y(4)$. The rate of spread of the disease is $y'(4)$.

Using formula (2) with $a = 15/10$, $b = \frac{15}{100000}$ and $y(0) = 200$ gives

$$y(t) = \frac{10000}{1 + 49e^{-3t/2}}.$$

Then the number of infectives at $t = 4$ is $y(4) = 8916.956640$. The number of susceptibles is $M - y(4) = 1083.043360$. The rate of spread of the disease is $y'(4) = 1448.617600$.

Example 2.35 (Explosion-Extinction)

Classify the model as **explosion** or **extinction**: $y' = 2(y - 100)y$, $y(0) = 200$.

Solution: Let $G(y) = 2(y - 100)y$, then $G(y) = 0$ exactly for equilibria $y = 100$ and $y = 0$, at which $G'(y) = 4y - 200$ satisfies $G'(200) > 0$, $G'(0) < 0$. The initial value $y(0) = 200$ is above the equilibrium $y = 100$. Because $y = 100$ is a source, then $y \to \infty$, which implies the model is **explosion**.

A second, direct analysis can be made from the differential equation $y' = 2(y - 100)y$: $y'(0) = 2(200 - 100)200 > 0$ means y increases from 200, causing $y \to \infty$ and explosion.

Example 2.36 (Constant Harvesting)

Find the carrying capacity M and the threshold population N for the harvesting equation $P' = (3 - 2P)P - 1$.

Solution: Carrying Capacity $M = 1$, Threshold Population $N = 1/2$.

Let $f(P) = -2(P - 1)(P - 1/2)$, which is the factored form of $(3 - 2P)P - 1$, the right side of $P' = (3-2P)P-1$. Solve equation $f(P) = 0$ for $P = 1$, $P = 1/2$, the equilibrium solutions.

Requirements for **carrying capacity** M and **threshold population** N:

1. M and N are equilibrium solutions
2. M is a stable sink, a funnel in the phase portrait
3. If $P(0) > N$, then $\lim_{t\to\infty} P(t) = M$.

Stability test 1.3 on page 55 applies: if $f(M) = 0$ and $f'(M) < 0$, then equilibrium $P = M$ is a stable sink (a funnel). Calculate $G'(P) = 3 - 4P$. Test $P = 1$ and $P = 1/2$: $P = 1$ is a stable sink. Define $M = 1$, $N = 1/2$. Requirements **1** and **2** hold. To verify limit requirement **3**, write $G(P) = -2(P-1)(P-1/2) = -2(P-M)(P-N)$ and make a phase line diagram. Then use the **Three Drawing Rules** page 52.

Example 2.37 (Variable Harvesting)

Re-model the variable harvesting equation $P' = (3 - 2P)P - P$ as $y' = (a - by)y$ and solve the equation by formula (2), page 142.

Solution: The equation is rewritten as $P' = 2P - 2P^2 = (2 - 2P)P$. This has the form of $y' = (a - by)y$ where $a = b = 2$. Then (2) implies

$$P(t) = \frac{2P_0}{2P_0 + (2 - 2P_0)e^{-2t}}$$

which simplifies to

$$P(t) = \frac{P_0}{P_0 + (1 - P_0)e^{-2t}}.$$

Example 2.38 (Restocking)

Make a direction field graphic by computer for the restocking equation $P' = (1 - P)P - 2\sin(2\pi t)$. Using the graphic, report (a) an estimate for the carrying capacity C and (b) approximations for the amplitude A and period T of a periodic solution which oscillates about $P = C$.

Solution: The computer algebra system `maple` is used with the code below to make Figure 6. An essential feature of the `maple` code is plotting of multiple solution curves. For instance, `[P(0)=1.3]` in the list `ics` of initial conditions causes the solution to the problem $P' = (1-P)P - 2\sin(2\pi t)$, $P(0) = 1.3$ to be added to the graphic.

The resulting graphic, which contains 13 solution curves, shows that all solution curves limit as $t \to \infty$ to what appears to be a unique periodic solution.

Using features of the `maple` interface, it is possible to determine by experiment estimates for the maxima $M = 1.26$ and minima $m = 0.64$ of the apparent periodic solution. Then (a) $C = (M+m)/2 = 0.95$, (b) $A = (M-m)/2 = 0.31$ and $T = 1$. The experimentally obtained period $T = 1$ matches the period of the term $-2\sin(2\pi t)$.

```
de:=diff(P(t),t)=(1-P(t))*P(t)-2*sin(2*Pi* t);
ics:=[[P(0)=1.4],[P(0)=1.3],[P(0)=1.2],[P(0)=1.1],[P(0)=0.1],
[P(0)=0.2],[P(0)=0.3],[P(0)=0.4],[P(0)=0.5],[P(0)=0.6],
[P(0)=0.7],[P(0)=0.8],[P(0)=0.9]];
opts:=stepsize=0.05,arrows=none:
DEtools[DEplot](de,P(t),t=-3..12,P=-0.1..1.5,ics,opts);
```

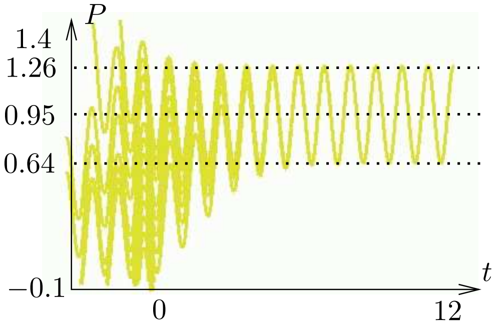


Figure 6. Solutions of $P' = (1-P)P - 2\sin(2\pi t)$.

The maximum is 1.26.
The minimum is 0.64.
Oscillation is about the line $P = 0.95$ with period 1.

Exercises 2.7

Limited Environment

Find the equilibrium solutions and the carrying capacity for each logistic equation.

1. $P' = 0.01(2-3P)P$

2. $P' = 0.2P - 3.5P^2$

3. $y' = 0.01(-3-2y)y$

4. $y' = -0.3y - 4y^2$

5. $u' = 30u + 4u^2$

6. $u' = 10u + 3u^2$

7. $w' = 2(2-5w)w$

8. $w' = -2(3-7w)w$

9. $Q' = Q^2 - 3(Q-2)Q$

10. $Q' = -Q^2 - 2(Q-3)Q$

Spread of a Disease

In each model, find the number of infectives and then the number of susceptibles at $t = 2$ months. Follow Example 2.34, page 143. A calculator is required for approximations.

11. $y' = (5/10 - 3y/100000)y$, $y(0) = 100$.

12. $y' = (13/10 - 3y/100000)y$, $y(0) = 200$.

13. $y' = (1/2 - 12y/100000)y$, $y(0) = 200$.

14. $y' = (15/10 - 4y/100000)y$, $y(0) = 100$.

15. $P' = (1/5 - 3P/100000)P$, $P(0) = 500$.

16. $P' = (5/10 - 3P/100000)P$, $P(0) = 600$.

17. $10P' = 2P - 5P^2/10000$, $P(0) = 500$.

18. $P' = 3P - 8P^2$, $P(0) = 10$.

Explosion–Extinction

Classify the model as **explosion** or **extinction**.

19. $y' = 2(y - 100)y$, $y(0) = 200$

20. $y' = 2(y - 200)y$, $y(0) = 300$

21. $y' = -100y + 250y^2$, $y(0) = 200$

22. $y' = -50y + 3y^2$, $y(0) = 25$

23. $y' = -60y + 70y^2$, $y(0) = 30$

24. $y' = -540y + 70y^2$, $y(0) = 30$

25. $y' = -16y + 12y^2$, $y(0) = 1$

26. $y' = -8y + 12y^2$, $y(0) = 1/2$

Constant Harvesting

Find the carrying capacity N and the threshold population M.

27. $P' = (3 - 2P)P - 1$

28. $P' = (4 - 3P)P - 1$

29. $P' = (5 - 4P)P - 1$

30. $P' = (6 - 5P)P - 1$

31. $P' = (6 - 3P)P - 1$

32. $P' = (6 - 4P)P - 1$

33. $P' = (8 - 5P)P - 2$

34. $P' = (8 - 3P)P - 2$

35. $P' = (9 - 4P)P - 2$

36. $P' = (10 - P)P - 2$

Variable Harvesting

Re-model the variable harvesting equation as $y' = (a - by)y$ and solve the equation by logistic solution (2) on page 142.

37. $P' = (3 - 2P)P - P$

38. $P' = (4 - 3P)P - P$

39. $P' = (5 - 4P)P - P$

40. $P' = (6 - 5P)P - P$

41. $P' = (6 - 3P)P - P$

42. $P' = (6 - 4P)P - P$

43. $P' = (8 - 5P)P - 2P$

44. $P' = (8 - 3P)P - 2P$

45. $P' = (9 - 4P)P - 2P$

46. $P' = (10 - P)P - 2P$

Restocking

Make a direction field graphic by computer following Example 2.38. Using the graphic, report (a) an estimate for the carrying capacity C and (b) approximations for the amplitude A and period T of a periodic solution which oscillates about $P = C$.

47. $P' = (2 - P)P - \sin(\pi t/3)$

48. $P' = (2 - P)P - \sin(\pi t/5)$

49. $P' = (2 - P)P - \sin(\pi t/7)$

50. $P' = (2 - P)P - \sin(\pi t/8)$

Richard Function

Ideas of L. von Bertalanffy (1934), A. Pütter (1920) and Verhulst were used by F. J. Richards (1957) to define a sigmoid function $Y(t)$ which generalizes the logistic function. It is suited for data-fitting models, for example forestry, tumor growth and stock-production problems. The Richard function is

$$Y(t) = A + \frac{K - A}{(1 + Qe^{-B(t-M)})^{1/\nu}},$$

where Y = weight, height, size, amount, etc., and t = time.

51. Differentiate for $\alpha > 0$, $\nu > 0$, the specialized Richard function

$$Y(t) = \frac{K}{(1 + Qe^{-\alpha\nu(t-t_0)})^{1/\nu}}$$

to obtain the sigmoid differential equation

$$Y'(t) = \alpha\left(1 - \left(\frac{Y}{K}\right)^{\nu}\right)Y.$$

The relation $Y(t_0) = \frac{K}{(1+Q)^{1/\nu}}$ implies $Q = -1 + \left(\frac{K}{Y(t_0)}\right)^{\nu}$.

52. Solve the differential equation $Y'(t) = \alpha\left(1 - \left(\frac{Y}{K}\right)^{\nu}\right)Y$ by means of the substitution $w = (Y/K)^{\nu}$, which gives a familiar logistic equation $w' = \alpha\nu(1 - w)w$.

2.8 Science and Engineering Applications

Assembled here are some classical applications of first order differential equations to problems of science and engineering.

Draining a Tank

Investigated here is a tank of water with orifice at the bottom emptying due to gravity; see Figure 7. The analysis applies to tanks of any geometrical shape.

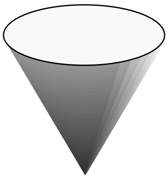


Figure 7. Draining a tank.
A tank empties from an orifice at the bottom. The fluid fills the tank to height y above the orifice, and it drains due to gravity.

Evangelista Torricelli (1608-1647), inventor of the **barometer**, investigated this physical problem using Newton's laws, obtaining the result in Lemma 2.8, proof on page 157.

Lemma 2.8 (Torricelli) A droplet falling freely from height h in a gravitational field with constant g arrives at the orifice with speed $\sqrt{2gh}$.

Tank Geometry. A simple but useful tank geometry can be constructed using *washers* of area $A(y)$, where y is the height above the orifice; see Figure 8.

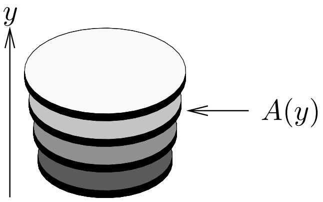


Figure 8. A tank constructed from washers.

Then the method of cross-sections in calculus implies that the *volume* $V(h)$ of the tank at height h is given by

$$(1) \qquad V(h) = \int_0^h A(y)dy, \quad \frac{dV}{dh} = A(h).$$

Torricelli's Equation. Torricelli's lemma applied to the tank fluid height $y(t)$ at time t implies, by matching drain rates at the orifice (see *Technical Details* page 156), that

$$(2) \qquad \frac{d}{dt}\left(V(y(t))\right) = -k\sqrt{y(t)}$$

for some proportionality constant $k > 0$. The *chain rule* gives the separable differential equation $V'(y(t))y'(t) = -k\sqrt{y(t)}$, or equivalently (see page 157), in terms of the **cross-sectional area** $A(y) = V'(y)$,

$$(3) \qquad y'(t) = -k\frac{\sqrt{y(t)}}{A(y(t))}.$$

Typical of the physical literature, the requirement $y(t) \geq 0$ is omitted in the model, but assumed implicitly. The model itself **exhibits non-uniqueness**: the tank can be drained hours ago or at instant $t = 0$ and result still in the solution $y(t) = 0$, interpreted as fluid height zero.

Stefan's Law

Heat energy can be transferred by **conduction**, **convection** or **radiation**. The following illustrations suffice to distinguish the three types of heat transfer.

Conduction. A soup spoon handle gains heat from the soup by exchange of kinetic energy at a molecular level.

Convection. A hot water radiator heats a room largely by *convection currents*, which move heated air upwards and denser cold air downwards to the radiator. In linear applications, **Newton's cooling law** applies.

Radiation. A car seat heated by the sun gets the heat energy from electromagnetic waves, which carry energy from the sun to the earth.

The rate at which an object emits or absorbs **radiant energy** is given by **Stefan's radiation law**

$$P = kT^4.$$

The symbol P is the power in watts (joules per second), k is a constant proportional to the surface area of the object and T is the temperature of the object in degrees Kelvin. Use $K = C + 273.15$ to convert Celsius to Kelvin.[6] The constant

[6] USA Fahrenheit F is Celsius $C = G + G/10 + G/100$, correct to 0.49 C for -40 to 120 F. Value $G = (F - 32)/2$ is a common **guess**. The idea uses $1/9 = 0.111\ldots$. **Example for** $F = 79$: Compute guess $G = (79 - 32)/2 = 23.5$. Then $C = 23.5 + 2.35 + 0.235 = 26.085$. The numbers added to G are decimal point shifts.

k in Stefan's law is decomposed as $k = \sigma A \mathcal{E}$. Here, $\sigma = 5.6696 \times 10^{-8} K^{-4}$ Watts per square meter (K=Kelvin), A is the surface area of the body in square meters and $\mathcal{E}$ is the **emissivity**, which is a constant between 0 and 1 depending on properties of the surface.

Constant room temperature. Suppose that a person with skin temperature T Kelvin sits unclothed in a room in which the thermometer reads T_0 Kelvin. The net heat flux P_{net} in joules per second (watts) is given by

$$P_{\text{net}} = k(T^4 - T_0^4). \tag{4}$$

If T and T_0 are constant, then $Q = kt(T^4 - T_0^4)$ can be used to estimate the total heat loss or gain in joules for a time period t. To illustrate, if the wall thermometer reads 20° Celsius, then $T_0 = 20 + 273.15$. Assume $A = 1.5$ square meters, $\mathcal{E} = 0.9$ and skin temperature 33° Celsius or $T = 33 + 273.15$. The total heat loss in 10 minutes is $Q = (10(60))(5.6696 \times 10^{-8})(1.5)(0.9)(305.15^4 - 293.15^4) = 64282$ joules. Over one hour, the total heat radiated is approximately $385,691$ joules, which is close to the total energy provided by a 6 ounce soft drink.[7]

Time-varying room temperature. Suppose that a person with skin temperature T degrees Kelvin sits unclothed in a room. Assume the thermometer initially reads 15° Celsius and then rises to 24° Celsius after t_1 seconds. The function $T_0(t)$ has values $T_0(0) = 15 + 273.15$ and $T_0(t_1) = 24 + 273.15$. In a possible physical setting, $T_0(t)$ reflects the reaction to the heating and cooling system, which is generally oscillatory about the thermostat setting. If the thermostat is off, then it is reasonable to assume a linear model $T_0(t) = at + b$, with $a = (T_0(t_1) - T_0(0))/t_1$, $b = T_0(0)$.

To compute the total heat radiated from the person's skin, we use the time-varying equation

$$\frac{dQ}{dt} = k(T^4 - T_0(t)^4). \tag{5}$$

The solution to (5) with $Q(0) = 0$ is formally given by the quadrature formula

$$Q(t) = k \int_0^t (T^4 - T_0(r)^4) dr. \tag{6}$$

For the case of a linear model $T_0(t) = at + b$, the total number of joules radiated from the person's skin is found by integrating (6), giving

$$Q(t_1) = kT^4 t_1 + k\frac{b^5 - (at_1 + b)^5}{5a}.$$

[7] American soft drinks are packaged in 12-ounce cans, twice the quantity cited. One calorie is *defined* to be 4.186 joules and one food **Calorie** is 1000 calories (a kilo-calorie) or 4186 joules. A boxed apple juice is about 6 ounces or 0.2 liters. Juice provides about 400 thousand joules in 0.2 liters. Product labels with 96 Calories mean 96 kilo-calories; it converts to $96(1000)(4.186) = 401,856$ joules.

Tsunami

A **seismic sea wave** due to an earthquake under the sea or some other natural event, called a **tsunami**, creates a wave on the surface of the ocean. The wave may have a height of less than 1 meter. These waves can have a very large wavelength, up to several hundred miles, depending upon the depth of the water where they were formed. The period is often more than one hour with wave velocity near 700 kilometers per hour. These waves contain a huge amount of energy. Their height increases as they crash upon the shore, sometimes to 30 meters high or more, depending upon water depth and underwater surface features. In the year 1737, a wave estimated to be 64 meters high hit Cape Lopatka, Kamchatka, in northeast Russia. The largest Tsunami ever recorded occurred in July of 1958 in Lituya Bay, Alaska, when a huge rock and ice fall caused water to surge up to 500 meters. For additional material on earthquakes, in particular the Sumatra and Chile earthquakes and resultant Tsunamis, see Chapter 11, **Systems of Differential Equations**.

Wave shape. A simplistic model for the shape $y(x)$ of a tsunami in the open sea is the differential equation [Zill-C, p. 81]

$$(y')^2 = 4y^2 - 2y^3. \tag{7}$$

This equation gives the *profile* $y(x)$ of one side of the 3D-wave, by cutting the 3D object with an xy-plane.

Equilibrium solutions. They are $y = 0$ and $y = 2$, corresponding to **no wave** and a **wall of water** 2 units above the ocean surface. There are *no solutions* for $y > 2$, because the two sides of (7) have in this case different signs.

Non-equilibrium solutions. They are given by

$$y(x) = 2 - 2\tanh^2(x + c). \tag{8}$$

The initial height of the wave is related to the parameter c by $y(0) = 2 - 2\tanh^2(c)$. Only initial heights $0 < y(0) < 2$ are physically significant. Due to the property $\lim_{u\to\infty} \tanh(u) = 1$ of the hyperbolic tangent, the wave height starts at $y(0)$ and quickly decreases to zero (sea level), as is evident from Figure 9.

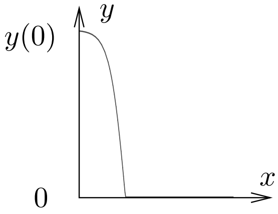


Figure 9. A tsunami profile.

Non-uniqueness. When $y(x_0) = 2$ for some $x = x_0$, then also $y'(x_0) = 0$, and this allows non-uniqueness of the solution y. An interesting solution different from equation (8) is the piecewise function

$$y(x) = \begin{cases} 2 - 2\tanh^2(x - x_0) & x > x_0, \\ 2 & x \le x_0. \end{cases} \tag{9}$$

This shape is an approximation to observed waves, in which the usual crest of the wave has been flattened. See Figure 12 on page 155.

Gompertz Tumor Equation

Researchers in tumor growth have shown that for some solid tumors the volume $V(t)$ of dividing cells at time t approaches a limiting volume M, even though the tumor volume may increase by 1000 fold. Gompertz is credited with an equation which fits the growth cycle of some solid tumors; the **Gompertzian relation** is

$$V(t) = V_0 e^{\frac{a}{b}\left(1-e^{-bt}\right)}. \tag{10}$$

The relation says that the doubling time for the total solid tumor volume *increases with time.* In contrast to a simple exponential model, which has a fixed doubling time and no volume limit, the limiting volume in the Gompertz model (10) is $M = V_0 e^{a/b}$.

Experts suggest to verify from Gompertz's relation (10) the formula

$$V' = ae^{-bt}V,$$

and then use this differential equation to argue why the tumor volume V approaches a limiting value M with a necrotic core; see *Technical Details for* (11), page 157.

A different approach is to make the substitution $y = V/V_0$ to obtain the differential equation

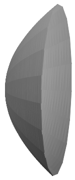
$$y' = (a - b\ln y)y, \tag{11}$$

which is almost a logistic equation, sometimes called the **Gompertz equation**. For details, see page 157. In analogy with logistic theory, low volume tumors should grow exponentially with rate a and then slow down like a population that is approaching the carrying capacity.

The exact mechanism for the slowing of tumor growth can be debated. One view is that the number of reproductive cells is related to available oxygen and nutrients present only near the surface of the tumor, hence this number decreases with time as the necrotic core grows in size.

Parabolic Mirror

Overhead projectors might use a high-intensity lamp located near a silvered reflector to provide a nearly parallel light source of high brightness. It is called a **parabolic mirror** because the surface of revolution is formed from a parabola, a fact which will be justified below.



The requirement is a shape $x = g(y)$ such that a light beam emanating from $C(0,0)$ reflects at point on the curve into a second beam parallel to the x-axis; see Figure 10. The **optical law of reflection** implies that the angle of incidence equals the angle of reflection, the straight reference line being the tangent to the curve $x = g(y)$.

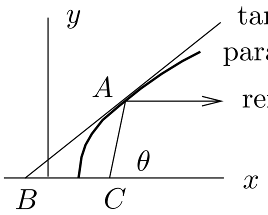


Figure 10. A parabolic mirror.

Symmetry suggests the restriction $y \ge 0$ will suffice to determine the shape. The assumption $y(0) = 1$ determines the y-axis scale.

The mirror shape $x = g(y)$ is shown in *Technical Details* page 157 to satisfy

$$\frac{dx}{dy} = \frac{x + \sqrt{x^2 + y^2}}{y}, \quad x(1) = 0. \tag{12}$$

This equation is equivalent for $y > 0$ to the separable equation $du/dy = \sqrt{u^2+1}$, $u(1) = 0$; see *Technical Details* page 157. Solving the separable equation (see page 157) gives the *parabola*

$$2x + 1 = y^2. \tag{13}$$

Logarithmic Spiral

The polar curve

$$r = r_0 e^{k\theta} \tag{14}$$

is called a **logarithmic spiral**. In equation (14), symbols r, θ are polar variables and r_0, k are constants. It will be shown that a logarithmic spiral has the following geometric characterization.

> A logarithmic spiral cuts each radial line from the origin at a constant angle.

The background required is the polar coordinate calculus formula

$$\tan(\alpha - \theta) = r\frac{d\theta}{dr} \tag{15}$$

where α is the angle between the x-axis and the tangent line at (r, θ); see Technical Details page 158. The angle α can also be defined from the calculus formula $\tan\alpha = dy/dx$.

The angle ϕ which a polar curve cuts a radial line is $\phi = \alpha - \theta$. By equation (15), the polar curve must satisfy the polar differential equation

$$r\frac{d\theta}{dr} = \frac{1}{k}$$

for constant $k = 1/\tan\phi$. This differential equation is separable with separated form

$$kd\theta = \frac{dr}{r}.$$

Solving gives $k\theta = \ln r + c$ or equivalently $r = r_0 e^{k\theta}$, for $c = -\ln r_0$. Hence equation (14) holds. All steps are reversible, therefore a logarithmic spiral is characterized by the geometrical description given above.

Examples

Example 2.39 (Conical Tank)

A conical tank with xy-projection given in Figure 11 is realized by rotation about the y-axis. An orifice at $x = y = 0$ is created at time $t = 0$. Find an approximation for the drain time and the time to empty the tank to half-volume, given 10% drains in 20 seconds.

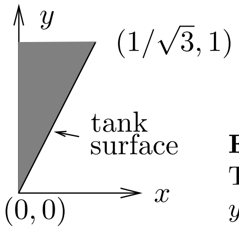


Figure 11. Conical tank xy-projection. The tank is obtained by rotation of the shaded triangle about the y-axis. The cone has height 1.

Solution: The answers are approximately 238 seconds and 104 seconds. The incorrect drain time estimate of ten times the given 20 seconds is wrong by 19 percent. Doubling the half-volume time to find the drain time is equally invalid (both 200 and 208 are incorrect).

Tank cross-section $A(y)$. From Figure 11, the line segment along the tank surface has equation $y = \sqrt{3}x$; the equation was found from the two points $(0, 0)$ and $(1/\sqrt{3}, 1)$ using the point-slope form of a line. A washer then has area $A(y) = \pi x^2$ or $A(y) = \pi y^2/3$.

Tank half-volume V_h. The half-volume is given by

$$\begin{aligned} V_h &= \frac{1}{2}V(1) && \text{Full volume is } V(1). \\ &= \frac{1}{2}\int_0^1 A(y)dy && \text{Apply } V(h) = \textstyle\int_0^h A(y)dy. \\ &= \frac{\pi}{18} && \text{Evaluate integral, } A(y) = \pi y^2/3. \end{aligned}$$

Torricelli's equation. The differential equation (3) becomes

$$y'(t) = -\frac{3k}{\pi\sqrt{y^3(t)}}, \quad y(0) = 1, \tag{16}$$

with k to be determined. The solution by separation of variables is

$$y(t) = \left(1 - \frac{15k}{2\pi}\,t\right)^{2/5}. \tag{17}$$

The details:

$$
\begin{aligned}
&y^{3/2}y' = -\frac{3k}{\pi} && \text{Separated form.}\\
&\frac{2}{5}y^{5/2} = -\frac{3kt}{\pi} + C && \text{Integrate both sides.}\\
&y^{5/2} = -\frac{15kt}{2\pi} + 1 && \text{Isolate } y\text{, then use } y(0) = 1.\\
&y = \left(1 - \frac{15kt}{2\pi}\right)^{2/5} && \text{Take roots.}
\end{aligned}
$$

Determination of k. Let $V_0 = V(1)/10$ be the volume drained after $t_0 = 20$ seconds. Then t_0, V_0 and k satisfy

$$
\begin{aligned}
V_0 &= V(1) - V(y(t_0)) && \text{Volume from height } y(t_0) \text{ to } y(0).\\
&= \frac{\pi}{9}\left(1 - y^3(t_0)\right) &&\\
&= \frac{\pi}{9}\left(1 - \left(1 - \frac{15k}{2\pi}t_0\right)^{6/5}\right) && \text{Substitute (17).}\\
k &= \frac{2\pi}{15t_0}\left(1 - \left(1 - \frac{9V_0}{\pi}\right)^{5/6}\right) && \text{Solve for } k.\\
&= \frac{2\pi}{15t_0}\left(1 - 0.9^{5/6}\right) &&
\end{aligned}
$$

Drain times. The volume is $V_h = \pi/18$ at time t_1 given by $\pi/18 = V(t_1)$ or in detail $\pi/18 = \pi y^3(t_1)/9$. This requirement simplifies to $y^3(t_1) = 1/2$. Then

$$
\begin{aligned}
&\left(1 - \frac{15kt_1}{2\pi}\right)^{6/5} = \frac{1}{2} && \text{Insert the formula for } y(t).\\
&1 - \frac{15kt_1}{2\pi} = \frac{1}{2^{5/6}} && \text{Take the } 5/6 \text{ power of both sides.}\\
&t_1 = \frac{2\pi}{15k}\left(1 - 2^{-5/6}\right) && \text{Solve for } t_1.\\
&\quad = t_0\frac{1 - 2^{-5/6}}{1 - 0.9^{5/6}} && \text{Insert the formula for } k.\\
&\quad \approx 104.4 && \text{Half-tank drain time in seconds.}
\end{aligned}
$$

The drain time t_2 for the full tank is not twice this answer but $t_2 \approx 2.28t_1$ or 237.9 seconds. The result is justified by solving for t_2 in the equation $y(t_2) = 0$, which gives $t_2 = \dfrac{2\pi}{15k} = \dfrac{t_1}{1 - 2^{-5/6}} = \dfrac{t_0}{1 - 0.9^{5/6}}$.

Example 2.40 (Stefan's Law)

An inmate sits unclothed in a room with skin temperature 33° Celsius. The Celsius room temperature is given by $C(r) = 14 + 11r/20$ for r in minutes. Assume in Stefan's law $k = \sigma A\mathcal{E} = 6.349952 \times 10^{-8}$. Find the number of joules lost through the skin in the first 20 minutes.

Solution: The theory implies that the answer is $Q(t_1)$ where $t_1 = (20)(60)$ is in seconds and $Q' = kT^4 - kT_0^4$. Equation $r = t/60$ converts seconds t to minutes r. Let $T = 33 + 273.15$ and $T_0(t) = C(t/60) + 273.15$. Then

$$Q(t_1) = k\int_0^{t_1} (T^4 - (T_0(t))^4)dt \approx 110,0095 \quad \text{joules.}$$

Example 2.41 (Tsunami)

Find a piecewise solution, which represents a Tsunami wave profile, similar to equation (9), on page 150. Graph the solution on $|x - x_0| \le 2$.

$$(y')^2 = 8y^2 - 4y^3, \quad x_0 = 1.$$

Solution: Equilibrium solutions $y = 0$ and $y = 2$ are found from the equation $8y^2 - 4y^3 = 0$, which has factored form $4y^2(2 - y) = 0$.

Non-equilibrium solutions with $y' \ge 0$ and $0 < y < 2$ satisfy the first order differential equation

$$y' = 2y\sqrt{2 - y}.$$

Consulting a computer algebra system gives the solution

$$y(x) = 2 - 2\tanh^2(\sqrt{2}(x - x_0)).$$

Treating $-y' = 2y\sqrt{2 - y}$ similarly results in exactly the same solution.

Hand solution. Start with the substitution $u = \sqrt{2 - y}$. Then $u^2 = 2 - y$ and $2uu' = -y' = -2yu = -2(2 - u^2)u$, giving the separable equation $u' = u^2 - 2$. Reformulate it as $u' = (u - a)(u + a)$ where $a = \sqrt{2}$. Normal partial fraction methods apply to find an implicit solution involving the inverse hyperbolic tangent. Some integral tables tabulate the integral involved, therefore partial fractions can be technically avoided. Solving for u in the implicit equation gives the hyperbolic tangent solution $u = \sqrt{2}\tanh(\sqrt{2}(x - x_0))$. Then $y = 2 - u^2$ produces the answer reported above. The piecewise solution, which represents an ocean Tsunami wave, is given by

$$y(x) = \begin{cases} 2 & x \le 1, & \textbf{back-wave} \\ 2 - 2\tanh^2(\sqrt{2}(x - 1)) & 1 < x < \infty. & \textbf{wave front} \end{cases}$$

The figure can be made by hand. A computer algebra graphic appears in Figure 12, with `maple` code as indicated.

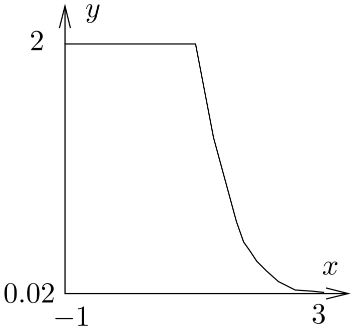


Figure 12. Tsunami wave profile.
The back-wave is at height 2. The front wave has height given by the hyperbolic tangent term, which approaches zero as $x \to \infty$. The `maple` code:

```
g:=x->2-2*tanh(sqrt(2)*(x-1))^2;
f:=x->piecewise(x<1,2,g(x));
plot(f,-1..3);
```

Example 2.42 (Gompertz Equation)

First, solve the Gompertz tumor equation, and then make (a) a phase line diagram and (b) a direction field.

$$y' = (8 - 2\ln y)y.$$

Solution: The only equilibrium solution computed from $G(y) \equiv (8 - 2\ln y)y = 0$ is $y = e^4 \approx 54.598$, because $y = 0$ is not in the domain of the right side of the differential equation.

Non-equilibrium solutions require integration of $1/G(y)$. Evaluation using a computer algebra system gives the implicit solution

$$-\frac{1}{2}\ln(8 - 2\ln(y)) = x + c.$$

Solving this equation for y in terms of x results in the explicit solution

$$y(x) = c_1 e^{-\frac{1}{2}e^{-2x}}, \quad c_1 = e^{4-\frac{1}{2}e^{-2c}}.$$

The `maple` code for these two independent tasks appears below.

```
p:=int(1/((8-2*ln(y))*y),y);
solve(p=x+c,y);
```

The phase line diagram in Figure 13 requires the equilibrium $y = e^4$ and formulas $G(y) = (8 - 2\ln y)y$, $G'(y) = 8 - 2\ln y - 2$. Then $G'(e^4) = -2$ implies G changes sign from positive to negative at $y = e^4$, making $y = e^4$ a stable sink or funnel.

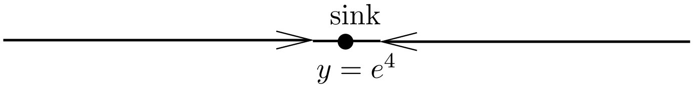


Figure 13. Gompertz phase line diagram.
The unique equilibrium at $y = e^4$ is a stable sink.

A computer-generated direction field appears in Figure 14, using the following `maple` code. Visible is the funnel structure at the equilibrium point.

```
de:=diff(y(x),x)=y(x)*(8-2*ln(y(x)));
with(DEtools):
DEplot(de,y(x),x=0..4,y=1..70);
```

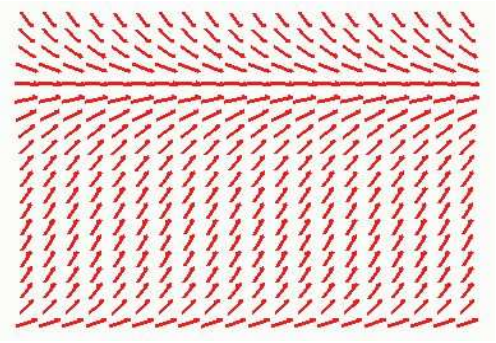


Figure 14. A Gompertz direction field.

Details and Proofs

Technical Details for (2): The derivation of $\frac{d}{dt}(V(y(t))) = -k\sqrt{y(t)}$ uses Torricelli's speed formula $|v| = \sqrt{2gy(t)}$. The volume change in the tank for an orifice of cross-sectional area a is $-av$. Therefore, $dV(y(t))/dt = -a\sqrt{2gy(t)}$. Succinctly, $dV(y(t))/dt = -k\sqrt{y(t)}$. This completes the verification.

Technical Details for (3): The equation $y'(t) = -k\dfrac{\sqrt{y(t)}}{A(y(t))}$ is equivalent to equation $A(y(t))\,y'(t) = -k\sqrt{y(t)}$. Equation $dV(y(t))/dt = V'(y(t))y'(t)$ obtained by the chain rule, definition $A(y) = V'(y)$, and equation (2) give result (3).

Technical Details for (2.8): To be verified is the Torricelli orifice equation $|v| = \sqrt{2gh}$ for the speed $|v|$ of a droplet falling from height h. Let's view the droplet as a point mass m located at the droplet's centroid. The distance $x(t)$ from the droplet to the orifice satisfies a falling body model $mx''(t) = -mg$. The model has solution $x(t) = -gt^2/2 + x(0)$, because $x'(0) = 0$. The droplet arrives at the orifice in time t given by $x(t) = 0$. Because $x(0) = h$, then $t = \sqrt{2h/g}$. The velocity v at this time is $v = x'(t) = -gt = -\sqrt{2gh}$. A technically precise derivation can be done using kinetic and potential energy relations; some researchers prefer energy method derivations for Torricelli's law. Formulas for the orifice speed depend upon the shape and size of the orifice. For common drilled holes, the speed is a constant multiple $c\sqrt{2gh}$, where $0 < c < 1$.

Technical Details for (11): Assume $V = V_0 e^{\mu(t)}$ and $\mu(t) = a(1 - e^{-bt})/b$. Then $\mu' = ae^{-bt}$ and

$$\begin{aligned} V' &= V_0\mu'(t)e^{\mu(t)} && \text{Calculus rule } (e^u)' = u'e^u. \\ &= \mu'(t)V && \text{Use } V = V_0e^{\mu(t)}. \\ &= ae^{-bt}V && \text{Use } \mu' = ae^{-bt}. \end{aligned}$$

The equation $V' = ae^{-bt}V$ is a growth equation $y' = ky$ where k decreases with time, causing the doubling time to increase. One biological explanation for the increase in the mean generation time of the tumor cells is aging of the reproducing cells, causing a slower dividing time. The correctness of this explanation is debatable.

Let $y = V/V_0$. Then

$$\begin{aligned} \frac{y'}{y} &= \frac{V'}{V} && \text{The factor } 1/V_0 \text{ cancels.} \\ &= ae^{-bt} && \text{Differential equation } V' = ae^{-bt}V \text{ applied.} \\ &= a - b\mu(t) && \text{Use } \mu(t) = a(1 - e^{-bt})/b. \\ &= a - b\ln(V/V_0) && \text{Take logs across } V/V_0 = e^{\mu(t)} \text{ to find } \mu(t). \\ &= a - b\ln y && \text{Use } y = V/V_0. \end{aligned}$$

Hence $y' = (a - b\ln y)y$. When $V \approx V_0$, then $y \approx 1$ and the growth rate $a - b\ln y$ is approximately a. Hence the model behaves like the exponential growth model $y' = ay$ when the tumor is small. The tumor grows subject to $a - b\ln y > 0$, which produces the volume restraint $\ln y = a/b$ or $V_{\max} = V_0e^{a/b}$.

Technical Details for (12): Polar coordinates r, θ will be used. The geometry in the parabolic mirror Figure 10 shows that triangle ABC is isosceles with angles α, α and

$\pi - 2\alpha$. Therefore, $\theta = 2\alpha$ is the angle made by segment CA with the x-axis (C is the origin $(0,0)$).

$$\begin{aligned}
y &= r\sin\theta && \text{Polar coordinates.}\\
&= 2r\sin\alpha\cos\alpha && \text{Use } \theta = 2\alpha \text{ and } \sin 2x = 2\sin x\cos x.\\
&= 2r\tan\alpha\cos^2\alpha && \text{Identity } \tan x = \sin x/\cos x \text{ applied.}\\
&= 2r\frac{dy}{dx}\cos^2\alpha && \text{Use calculus relation } \tan\alpha = dy/dx.\\
&= r\frac{dy}{dx}(1+\cos 2\alpha) && \text{Identity } 2\cos^2 x - 1 = \cos 2x \text{ applied.}\\
&= \frac{dy}{dx}(r+x) && \text{Use } x = r\cos\theta \text{ and } 2\alpha = \theta.
\end{aligned}$$

For $y > 0$, equation (12) can be solved as follows.

$$\begin{aligned}
&\frac{dx}{dy} = \frac{x}{y} + \sqrt{(x/y)^2+1} && \text{Divide by } y \text{ on the right side of (12).}\\
&y\frac{du}{dy} = \sqrt{u^2+1} && \text{Substitute } u = x/y \text{ (u cancels).}\\
&\int \frac{du}{\sqrt{u^2+1}} = \int\frac{dy}{y} && \text{Integrate the separated form.}\\
&\sinh^{-1} u = \ln y && \text{Integral tables. The integration constant is}\\
& && \text{zero because } u(1) = 0.\\
&\frac{x}{y} = \sinh(\ln y) && \text{Let } u = x/y \text{ and apply sinh to both sides.}\\
&\quad = \frac{1}{2}\left(e^{\ln y} - e^{-\ln y}\right) && \text{Definition } \sinh u = (e^u - e^{u})/2.\\
&\quad = \frac{1}{2}\left(y - 1/y\right) && \text{Identity } e^{\ln y} = y.
\end{aligned}$$

Clearing fractions in the last equality gives $2x + 1 = y^2$, a parabola of the form $X = Y^2$.

Technical Details for (15): Given polar coordinates r, θ and $\tan\alpha = dy/dx$, it will be shown that $r\,d\theta/dr = \tan(\alpha - \theta)$. Details require the formulas

$$\begin{aligned}
x &= r\cos\theta, & \frac{dx}{dr} &= \cos\theta - r\frac{d\theta}{dr}\sin\theta,\\
y &= r\sin\theta, & \frac{dy}{dr} &= \sin\theta + r\frac{d\theta}{dr}\cos\theta.
\end{aligned} \tag{18}$$

Then

$$\begin{aligned}
\tan\alpha &= \frac{dy}{dx} && \text{Definition of derivative.}\\
&= \frac{dy/dr}{dx/dr} && \text{Chain rule.}\\
&= \frac{\sin\theta + r\frac{d\theta}{dr}\cos\theta}{\cos\theta - r\frac{d\theta}{dr}\sin\theta} && \text{Apply equation (18).}\\
&= \frac{\tan\theta + r\frac{d\theta}{dr}}{1 - r\frac{d\theta}{dr}\tan\theta} && \text{Divide by } \cos\theta.
\end{aligned}$$

Let $X = rd\theta/dr$ and cross-multiply to eliminate fractions. Then the preceding relation implies $(1 - X\tan\theta)\tan\alpha = \tan\theta + X$ and finally

$$
\begin{aligned}
r\frac{d\theta}{dr} &= X && \text{Definition of } X. \\
&= \frac{\tan\alpha - \tan\theta}{1+\tan\alpha\tan\theta} && \text{Solve for } X \text{ in } (1 - X\tan\theta)\tan\alpha = \tan\theta + X. \\
&= \tan(\alpha-\theta) && \text{Apply identity } \tan(a-b) = \frac{\tan a - \tan b}{1+\tan a \tan b}.
\end{aligned}
$$

Physicists and engineers often justify formula (15) referring to Figure 15. Such diagrams are indeed the initial intuition required to *guess* formulas like (15).

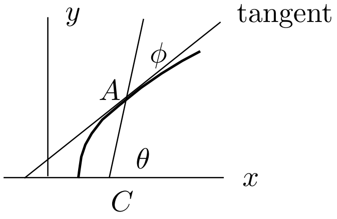


Figure 15. Polar differential triangle. Angle ϕ is the *signed angle* between the radial vector and the tangent line.

Exercises 2.8

Tank Draining

1. A cylindrical tank 6 feet high with 6-foot diameter is filled with gasoline. In 15 seconds, 5 gallons drain out. Find the drain times for the next 20 gallons and the half-volume.

2. A cylindrical tank 4 feet high with 5-foot diameter is filled with gasoline. The half-volume drain time is 11 minutes. Find the drain time for the full volume.

3. A conical tank is filled with water. The tank geometry is a solid of revolution formed from $y = 2x$, $0 \le x \le 5$. The units are in feet. Find the drain time for the tank, given the first 5 gallons drain out in 12 seconds.

4. A conical tank is filled with oil. The tank geometry is a solid of revolution formed from $y = 3x$, $0 \le x \le 5$. The units are in meters. Find the half-volume drain time for the tank, given the first 5 liters drain out in 10 seconds.

5. A spherical tank of diameter 12 feet is filled with water. Find the drain time for the tank, given the first 5 gallons drain out in 20 seconds.

6. A spherical tank of diameter 9 feet is filled with solvent. Find the half-volume drain time for the tank, given the first gallon drains out in 3 seconds.

7. A hemispherical tank of diameter 16 feet is filled with water. Find the drain time for the tank, given the first 5 gallons drain out in 25 seconds.

8. A hemispherical tank of diameter 10 feet is filled with solvent. Find the half-volume drain time for the tank, given the first gallon drains out in 4 seconds.

9. A parabolic tank is filled with water. The tank geometry is a solid of revolution formed from $y = 2x^2$, $0 \le x \le 2$. The units are in feet. Find the drain time for the tank, given the first 5 gallons drain out in 12 seconds.

10. A parabolic tank is filled with oil. The tank geometry is a solid of revolution formed from $y = 3x^2$, $0 \le x \le 2$. The units are in meters. Find the half-volume drain time for the tank, given the first 4 liters drain out in 16 seconds.

Torricelli's Law and Uniqueness

It it known that Torricelli's law gives a differential equation for which Picard's existence-uniqueness theorem is inapplicable for initial data $y(0) = 0$.

11. Explain why Torricelli's equation $y' = k\sqrt{y}$ plus initial condition $y(0) = 0$ fails to satisfy the hypotheses in Picard's theorem. Cite all failed hypotheses.

12. Consider a typical Torricelli's law equation $y' = k\sqrt{y}$ with initial condition $y(0) = 0$. Argue physically that the depth $y(t)$ of the tank for $t < 0$ can be zero for an arbitrary duration of time t near $t = 0$, even though $y(t)$ is not zero for all t.

13. Display infinitely many solutions $y(t)$ on $-5 \le t \le 5$ of Torricelli's equation $y' = k\sqrt{y}$ such that $y(t)$ is not identically zero but $y(t) = 0$ for $0 \le t \le 1$.

14. Does Torricelli's equation $y' = k\sqrt{y}$ plus initial condition $y(0) = 0$ have a solution $y(t)$ defined for $t \ge 0$? Is it unique? Apply Picard's theorem and Peano's theorem, if possible.

Clepsydra: Water Clock Design

A surface of revolution is used to make a container of height h feet for a water clock. An increasing curve $y = f(x)$ on $0 \le x \le 1$ is revolved around the y-axis to make the container shape, e.g., $y = x$ makes a conical tank. Water drains by gravity out of diameter d orifice at $(0, 0)$. The tank water level must fall at a constant rate of r inches per hour, important for marking a time scale on the tank. Find d and $f(x)$, given h and r.

15. $h = 5$ feet, $r = 4$ inches/hour. Answers: $f(x) = 5x^4$, $d = 0.05460241726 \approx 3/64$ inch.

16. $h = 4$, $r = 4$

17. $h = 3$, $r = 6$

18. $h = 4$, $r = 3$

19. $h = 3$, $r = 2$

20. $h = 4$, $r = 1$

Stefan's Law

An unclothed prison inmate is handcuffed to a chair. The inmate's skin temperature is 33° Celsius. Find the number of Joules of heat lost by the inmate's skin after t_0 minutes, given skin area A in square meters, Kelvin room temperature $T_0(r) = C(r/60) + 273.15$ and Celsius room temperature $C(t)$. Variables: t minutes, r seconds. Use equation $\frac{dQ}{dt} = k(T^4 - T_0(t)^4)$ page 149. Assume emissivity $\sigma = 5.6696 \times 10^{-8} K^{-4}$ Watts per square meter, K=degrees Kelvin.

21. $\mathcal{E} = 0.9$, $A = 1.5$, $t_0 = 10$, $C(t) = 24 + 7t/t_0$

22. $\mathcal{E} = 0.9$, $A = 1.7$, $t_0 = 12$, $C(t) = 21 + 10t/12$

23. $\mathcal{E} = 0.9$, $A = 1.4$, $t_0 = 10$, $C(t) = 15 + 15t/t_0$

24. $\mathcal{E} = 0.9$, $A = 1.5$, $t_0 = 12$, $C(t) = 15 + 14t/t_0$

On the next two exercises, use a computer algebra system (CAS). Same assumptions as Exercise 21.

25. $\mathcal{E} = 0.8$, $A = 1.4$, $t_0 = 15$, $C(t) = 15 + 15 \sin \pi(t - t_0)/12$

26. $\mathcal{E} = 0.8$, $A = 1.4$, $t_0 = 20$, $C(t) = 15 + 14 \sin \pi(t - t_0)/12$

Tsunami Wave Shape

Plot the piecewise solution

$$(19)\; y(x) = 2 - \begin{cases} 2\tanh^2(x - x_0) & x>x_0, \\ 0 & x\le x_0. \end{cases}$$

See Figure 12 page 155.

27. $x_0 = 2$, $|x - x_0| \le 2$

28. $x_0 = 3$, $|x - x_0| \le 4$.

Tsunami Wavefront

Find non-equilibrium solutions for the given differential equation.

29. $(y')^2 = 12y^2 - 10y^3$.

30. $(y')^2 = 13y^2 - 12y^3$.

31. $(y')^2 = 8y^2 - 2y^3$.

32. $(y')^2 = 7y^2 - 4y^3$.

Gompertz Tumor Equation

Solve the Gompertz tumor equation $y' = (a - b\ln y)y$.

33. $a = 1$, $b = 1$

34. $a = 1$, $b = 2$

35. $a = -1$, $b = 1$

36. $a = -1$, $b = 2$

37. $a = 4$, $b = 1$

38. $a = 5$, $b = 1$

2.9 Exact Equations and Level Curves

A **level curve** or a **conservation law** is an equation of the form

$$U(x, y) = c.$$

Hikers like to think of U as the *altitude* at position (x, y) on the map and $U(x, y) = c$ as the *curve* which represents the easiest walking path, that is, altitude does not change along that route. The altitude is **conserved** along the route, hence the terminology *conservation law.*

Other examples of level curves are *isobars* and *isotherms*. An **isobar** is a planar curve where the atmospheric pressure is constant. An **isotherm** is a planar curve along which the temperature is constant.

Definition 2.8 (Potential)
The function $U(x, y)$ in a conservation law is called a **potential**. The **dynamical equation** is the first order differential equation

$$(1) \qquad M dx + N dy = 0, \quad M = U_x(x, y), \quad N = U_y(x, y).$$

The *dynamics* or *changes* in x and y are described by (1). To **solve** $Mdx + Ndy = 0$ means this: find a conservation law $U(x, y) = c$ so that (1) holds. Formally, (1) is found by *implicit differentiation* of $U(x, y) = c$; see *Technical Details*, page 165.

The Potential Problem and Exactness

The **potential problem** assumes given a dynamical equation $Mdx + Ndy = 0$ and seeks to find a potential $U(x, y)$ from the set of equations

$$(2) \qquad \begin{aligned} U_x &= M(x, y), \\ U_y &= N(x, y). \end{aligned}$$

If some potential $U(x, y)$ satisfies equation (2), then $Mdx + Ndy = 0$ is said to be **exact**. It is a consequence of the mixed partial equality $U_{xy} = U_{yx}$ that the existence of a solution U implies $M_y = N_x$. Surprisingly, this condition is also sufficient.

Theorem 2.10 (Exactness)
Let $M(x, y)$, $N(x, y)$ and their first partials be continuous in a rectangle D. Assume $M_y(x, y) = N_x(x, y)$ in D and (x_0, y_0) is a point of D. Then the equation $Mdx + Ndy = 0$ is exact with potential U given by the formula

$$(3) \qquad U(x, y) = \int_{x_0}^{x} M(t, y)dt + \int_{y_0}^{y} N(x_0, s)ds.$$

The proof is on page 165.

The Method of Potentials

Formula (3) has technical problems because it requires two integrations. The integrands have a *parameter*: they are *parametric integrals*. Integration effort can be reduced by using the **method of potentials** for $Mdx + Ndy = 0$, which applies equation (3) with $x_0 = y_0 = 0$ in order to simplify integrations.

Test $M_y = N_x$	Compute the partials M_y and N_x, then test the equality $M_y = N_x$. Proceed if equality holds.
Trial Potential	Let $U = \int_0^x M(x, y)dx + \int_0^y N(0, y)dy$. Evaluate both integrals.
Test $U(x, y)$	Compute U_x and U_y, then test both $U_x = M$ and $U_y = N$. This step finds integration errors.

Examples

Example 2.43 (Exactness Test)

Test $Mdx + Ndy = 0$ for the existence of a potential U, given $M = 2xy + y^3 + y$ and $N = x^2 + 3xy^2 + x$,

Solution: Theorem 2.10 implies that $Mdx + Ndy = 0$ has a potential U exactly when $M_y = N_x$. It suffices to compute the partials and show they are equal.

$$\begin{aligned} M_y &= \partial_y(2xy + y^3 + y) \\ &= 2x + 3y^2 + 1, \end{aligned} \qquad \begin{aligned} N_x &= \partial_x(x^2 + 3xy^2 + x) \\ &= 2x + 3y^2 + 1. \end{aligned}$$

Example 2.44 (Conservation Law Test)

Test whether $U = x^2y + xy^3 + xy$ is a potential for $Mdx + Ndy = 0$, given $M = 2xy + y^3 + y$, $N = x^2 + 3xy^2 + x$.

Solution: By definition, it suffices to test the equalities $U_x = M$ and $U_y = N$.

$$\begin{aligned} U_x &= \partial_x(x^2y + xy^3 + xy) \\ &= 2xy + y^3 + y \\ &= M, \end{aligned} \qquad \begin{aligned} U_y &= \partial_y(x^2y + xy^3 + xy) \\ &= x^2 + 3xy^2 + x \\ &= N. \end{aligned}$$

Example 2.45 (Method of Potentials)

Solve $y' = -\dfrac{2xy + y^3 + y}{x^2 + 3xy^2 + x}$.

Solution: The implicit solution $x^2y + xy^3 + xy = c$ will be justified.

The equation has the form $Mdx+Ndy = 0$ where $M = 2xy+y^3+y$ and $N = x^2+3xy^2+x$. It is exact, by Theorem 2.10, because the partials $M_y = 2x+3y^2+1$ and $N_x = 2x+3y^2+1$ are equal.

The method of potentials applies to find the potential $U = x^2y + xy^3 + xy$ as follows.

$$\begin{aligned}
U &= \textstyle\int_0^x M(x,y)dx + \int_0^y N(0,y)dy && \text{Formula for } U\text{, Theorem 2.10.}\\
&= \textstyle\int_0^x \left(2xy + y^3 + y\right)dx + \int_0^y (0)dy && \text{Insert } M \text{ and } N.\\
&= x^2y + xy^3 + xy && \text{Evaluate integral.}
\end{aligned}$$

Observe that $N(x,y)$ simplifies to zero at $x=0$, which reduces the actual work in half. Any choice other than $x_0 = 0$ in Theorem 2.10 increases the labor.

To *test the solution*, compute the partials of U, then compare them to M and N; see Example 2.44.

Example 2.46 (Exact Equation)

Solve $\dfrac{x+y}{(1-x)^2}dx + \dfrac{x}{1-x}dy = 0.$

Solution: The implicit solution $\dfrac{xy+x}{1-x} + \ln|x-1| = c$ will be justified.

Assume given the exactness of the equation $Mdx+Ndy=0$, where $M = (x+y)/(1-x)^2$ and $N = x/(1-x)$. Apply Theorem 2.10:

$$\begin{aligned}
U &= \textstyle\int_0^x M(x,y)dx + \int_0^y N(0,y)dy && \text{Method of potentials.}\\
&= \textstyle\int_0^x \dfrac{x+y}{(1-x)^2}dx + \int_0^y (0)dy && \text{Substitute for } M,\ N.\\
&= \textstyle\int_0^x \left(\dfrac{y+1}{(x-1)^2} + \dfrac{1}{x-1}\right)dx && \text{Partial fractions.}\\
&= \dfrac{xy+x}{1-x} + \ln|x-1| && \text{Evaluate integral.}
\end{aligned}$$

Additional examples, including the context for the preceding example, appear in the next section.

Remarks on the Method of Potentials

Indefinite integrals $\int M(x,y)dx$ and $\int N(0,y)dy$ can be used, provided the two integration answers are zero at $x=0$ and $y=0$, respectively. Swapping the roles of x and y gives $U = \int_0^y N(x,y)dy + \int_0^x M(x,0)dx$, a form which may have easier integrations.

Can the test $M_y = N_x$ be skipped? True, it is enough to verify that the potential works (the last step). If the last step fails, then the first step must be done to resolve the error.

The equation $ydx + 2xdy = 0$ fails $M_y = N_x$ and the trial potential $U = xy$ fails $U_x = M$, $U_y = N$. In the equivalent form $x^{-1}dx + 2y^{-1}dy = 0$, the method of potentials does not apply directly, because $(0,0)$ is outside the domain of continuity. Nevertheless, the trial potential $U = \ln x + 2\ln y$ passes the test $U_x = M$, $U_y = N$. Such pleasant accidents account for the popularity of the method of potentials.

It is prudent in applications of Theorem 2.10 to test $x_0 = y_0 = 0$ in M and N, to detect a discontinuity. If detected, then another vertex x_0, y_0 of the unit square, e.g., $x_0 = y_0 = 1$, might suffice.

Details and Proofs

Justification of equation (1) uses the calculus *chain rule*

$$\frac{d}{dt}U(x(t),y(t)) = U_x(x(t),y(t))x'(t) + U_y(x(t),y(t))y'(t)$$

and differential notation $dx = x'(t)dt$, $dy = y'(t)dt$. To justify (1), let $(x(t),y(t))$ be some parameterization of the level curve, then differentiate on t across the equation $U(x(t),y(t)) = c$ and apply the chain rule.

Proof of Theorem 2.10

Background result. The proof assumes the following identity:

$$\frac{\partial}{\partial y}\int_{x_0}^{x} M(t,y)dt = \int_{x_0}^{x} M_y(t,y)dt.$$

The identity is obtained by forming the Newton quotient $(G(y+h)-G(y))/h$ for the derivative of $G(y) = \int_{x_0}^{x} M(t,y)dt$ and then taking the limit as h approaches zero. Technically, the limit must be taken inside an integral sign, which for success requires continuity of the partial M_y.

Details. It has to be shown that the implicit relation $U(x,y) = c$ with U defined by (3) is a solution of the exact equation $Mdx + Ndy = 0$, that is, the relations $U_x = M$, $U_y = N$ hold. The partials are calculated from the background result as follows.

$U_x = \partial_x \int_{x_0}^{x} M(t,y)dt$	Use (3), in which the second integral does not depend on x.
$= M(x,y)$,	Fundamental theorem of calculus.
$U_y = \partial_y \int_{x_0}^{x} M(t,y)dt$ $+ \partial_y \int_{y_0}^{y} N(x_0,s)ds$	Use (3).
$= \int_{x_0}^{x} M_y(t,y)dt + N(x_0,y)$	Apply the background result and the fundamental theorem.
$= \int_{x_0}^{x} N_x(t,y)dt + N(x_0,y)$	Substitute $M_y = N_x$.
$= N(x,y)$	Fundamental theorem of calculus.

■

Power Series Proof of Theorem 2.10 It will be assumed that M and N have power series expansions about $x = y = 0$. Let $U_1 = \int M(x,y)dx$ and $U_2 = \int N(x,y)dy$ with $U_1(0,y) = U_2(x,0) = 0$. The series forms of U_1 and U_2 will be

$$U_1 = \textstyle\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_{ij}x^i y^j + \sum_{i=1}^{\infty} a_i x^i,$$

$$U_2 = \textstyle\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} d_{ij}x^i y^j + \sum_{j=1}^{\infty} b_j y^j.$$

The identities $\partial_y\partial_x U_1 = M_y = N_x = \partial_x\partial_y U_2$ imply that $c_{ij} = d_{ij}$, using term-by-term differentiation. The trial potential is $U = U_1 + \sum_{j=1}^{\infty} b_j y^j$ or $U = U_2 + \sum_{i=1}^{\infty} a_i x^i$. From these relations it follows that $U_x = M$ and $U_y = N$. Therefore, $Mdx + Ndy = 0$ is exact with potential U.

Exercises 2.9

Exactness Test

Test the equality $M_y = N_x$ for the given equation, as written, and report *exact* when true. Do not try to solve the differential equation. See Example 2.43, page 163.

1. $(y - x)dx + (y + x)dy = 0$

2. $(y + x)dx + (x - y)dy = 0$

3. $(y + \sqrt{xy})dx + (-y)dy = 0$

4. $(y + \sqrt{xy})dx + xydy = 0$

5. $(x^2 + 3y^2)dx + 6xydy = 0$

6. $(y^2 + 3x^2)dx + 2xydy = 0$

7. $(y^3 + x^3)dx + 3xy^2dy = 0$

8. $(y^3 + x^3)dx + 2xy^2dy = 0$

9. $2xydx + (x^2 - y^2)dy = 0$

10. $2xydx + (x^2 + y^2)dy = 0$

Conservation Law Test

Test conservation law $U(x, y) = c$ for a solution to $Mdx + Ndy = 0$. See Example 2.44, page 163.

11. $2xydx + (x^2 + 3y^2)dy = 0$,
$x^2y + y^3 = c$

12. $2xydx + (x^2 - 3y^2)dy = 0$,
$x^2y - y^3 = c$

13. $(3x^2 + 3y^2)dx + 6xydy = 0$,
$x^3 + 3xy^2 = c$

14. $(x^2 + 3y^2)dx + 6xydy = 0$,
$x^3 + 3xy^2 = c$

15. $(y - 2x)dx + (2y + x)dy = 0$,
$xy - x^2 + y^2 = c$

16. $(y + 2x)dx + (-2y + x)dy = 0$,
$xy + x^2 - y^2 = c$

Exactness Theorem

Find an implicit solution $U(x, y) = c$. See Examples 2.45-2.46, page 163.

17. $(y - 4x)dx + (4y + x)dy = 0$

18. $(y + 4x)dx + (4y + x)dy = 0$

19. $(e^y + e^x)dx + (xe^y)dy = 0$

20. $(e^{2y} + e^x)dx + (2xe^{2y})dy = 0$

21. $(1 + ye^{xy})dx + (2y + xe^{xy})dy = 0$

22. $(1 + ye^{-xy})dx + (xe^{-xy} - 4y)dy = 0$

23. $(2x + \arctan y)dx + \dfrac{x}{1 + y^2}\,dy = 0$

24. $(2x + \arctan y)dx + \dfrac{x + 2y}{1 + y^2}\,dy = 0$

25. $\dfrac{2x^5 + 3y^3}{x^4y}dx - \dfrac{2y^3 + x^5}{x^3y^2}dy = 0$

26. $\dfrac{2x^4 + y^2}{x^3y}dx - \dfrac{2x^4 + y^2}{2x^2y^2}dy = 0$

27. $Mdx + Ndy = 0$, $M = e^x \sin y + \tan y$,
$N = e^x \cos y + x \sec^2 y$

28. $Mdx + Ndy = 0$, $M = e^x \cos y + \tan y$,
$N = -e^x \sin y + x \sec^2 y$

29. $\left(x^2 + \ln y\right) dx + \left(y^3 + x/y\right) dy = 0$

30. $\left(x^3 + \ln y\right) dx + \left(y^3 + x/y\right) dy = 0$

2.10 Special equations

Homogeneous-A Equation

A first order equation of the form $y' = F(y/x)$ is called a **homogeneous class A equation**. The substitution $u = y/x$ changes it into an equivalent first order separable equation $xu' + u = F(u)$. Solutions of $y' = F(y/x)$ and $xu' + u = F(u)$ are related by the equation $y = xu$.

Homogeneous-C Equation

Let $R(x, y)$ be a rational function constructed from *two affine functions*:

$$R(x,y) = \frac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}.$$

A first order equation of the form $y' = G(R(x, y))$ is called a **homogeneous class C equation** . If the system

$$a_1a + b_1b = c_1, \quad a_2a + b_2b = c_2$$

has a solution (a, b), then the change of variables $x = X - a$, $y = Y - b$ effectively eliminates the terms c_1 and c_2. Accordingly, the equation $y' = G(R(x, y))$ converts into a homogeneous class A equation

$$Y' = G\left(\frac{a_1 + b_1Y/X}{a_2 + b_2Y/X}\right).$$

This equation type was solved in the previous paragraph. Justification follows from $y' = Y'$ and $R(X - a, Y - b) = (a_1X + b_1Y)/(a_2X + b_2Y)$.

Bernoulli's Equation

The equation $y'+p(x)y = q(x)y^n$ is called the **Bernoulli differential equation**. If $n = 1$ or $n = 0$, then this is a linear equation. Otherwise, the substitution $u = y/y^n$ changes it into the linear first order equation $u'+(1-n)p(x)u = (1-n)q(x)$.

Integrating Factors and Exact Equations

An equation $\mathsf{M}dx + \mathsf{N}dy = 0$ is said to have an **integrating factor** $Q(x, y)$ if multiplication across the equation by Q produces an exact equation $Mdx+Ndy = 0$. The definition implies $M = Q\mathsf{M}$, $N = Q\mathsf{N}$ and $M_y = N_x$. The search for Q is only interesting when $\mathsf{M}_y \ne \mathsf{N}_x$.

A systematic approach to finding Q includes a list of **trial integrating factors**, which are known to work for special equations:

$Q = x^a y^b$	Require $xy\left(\mathbf{M}_y - \mathbf{N}_x\right) = ay\mathbf{N} - bx\mathbf{M}$. This integrating factor can introduce *extraneous solutions* $x = 0$ or $y = 0$.
$Q = e^{ax+by}$	Require $\mathbf{M}_y - \mathbf{N}_x = a\mathbf{N} - b\mathbf{M}$.
$Q = e^{\int \mu(x)dx}$	Require $\mu = \left(\mathbf{M}_y - \mathbf{N}_x\right)/N$ to be independent of y.
$Q = e^{\int \nu(y)dy}$	Require $\nu = \left(\mathbf{N}_x - \mathbf{M}_y\right)/M$ to be independent of x.

Examples

Example 2.47 (Homogeneous-A)

Solve $yy' = 2x + y^2/x$

Solution: The *implicit solution* will be shown to be

$$y^2 = cx^2 + 4x^2 \ln x.$$

The equation $yy' = 2x + y^2/x$ is not separable, linear nor exact. Division by y gives the homogeneous-A form $y' = 2/u + u$ where $u = y/x$. Then

$$\begin{aligned}
xu' + u &= 2/u + u && \text{Form } xu' + u = F(u).\\
xu' &= 2/u && \text{Separable form.}\\
u^2 &= c + 4\ln x && \text{Implicit solution } u.\\
y^2 &= x^2u^2 && \text{Change of variables } y = xu.\\
&= cx^2 + 4x^2 \ln x && \text{Substitute } u^2 = c + 4\ln x.
\end{aligned}$$

Check the implicit solution against $yy' = 2x + y^2/x$ as follows.

$$\begin{aligned}
\text{LHS} &= yy' && \text{Left side of } yy' = 2x + y^2/x.\\
&= \tfrac{1}{2}(y^2)' && \text{Calculus identity.}\\
&= \tfrac{1}{2}(cx^2 + 4x^2 \ln x)' && \text{Substitute.}\\
&= cx + 4x \ln x + 2x && \text{Differentiate.}\\
&= 2x + y^2/x && \text{Use } y^2 = cx^2 + 4x^2 \ln x.\\
&= \text{RHS}. && \text{Equality verified.}
\end{aligned}$$

Example 2.48 (Homogeneous-C)

Solve $y' = \dfrac{x + y + 3}{x - y + 5}$.

Solution: The *implicit solution* will be shown to be

$$2\ln(x+4) + \ln\left(\left(\frac{y-1}{x+4}\right)^2 + 1\right) - 2\arctan\left(\frac{y-1}{x+4}\right) = c.$$

The equation would be of type homogeneous-A, if not for the constants 3 and 5 in the fraction $(x+y+3)/(x-y+5)$. The method applies a translation of coordinates $x = X - a$, $y = Y - b$ as below.

$$\begin{aligned} x+y+3 &= X+Y, \\ x-y+5 &= X-Y \end{aligned}$$
Require the translation to remove the constant terms.

$$\begin{aligned} 3 &= a+b, \\ 5 &= a-b \end{aligned}$$
Substitute $X = x+a$, $Y = y+b$ and simplify.

$a = 4$, $b = -1$
Unique solution of the system.

$$\frac{dY}{dX} = \frac{X+Y}{X-Y}$$
Translated type homogeneous-A equation.

$$X\frac{du}{dX} + u = \frac{1+u}{1-u}$$
Use $u = Y/X$ to eliminate Y.

$$\frac{1-u}{1+u^2}\frac{du}{dX} = \frac{1}{X}$$
Separated form.

The separated form is integrated as $\int du/(1+u^2) - \int u du/(1+u^2) = \int dX/X$. Evaluation gives the implicit solution

$$\arctan(u) - \frac{1}{2}\ln\left(u^2+1\right) = C + \ln X.$$

Changing variables $x = X - 4$, $y = Y + 1$ and consolidating constants produces the announced solution.

To check the solution by `maple` assist, use the following code, which tests $U(x, y) = c$ against $y' = f(x, y)$. The test succeeds if `odetest` returns zero.

```
# Maple
U:=(x,y)->2*ln(x+4)+ln(((y-1)/(x+4))^2+1)-2*arctan((y-1)/(x+4));
f:=(x,y)->(x+y+3)/(x-y+5); DE:=diff(y(x),x)=f(x,y(x));
odetest(U(x,y(x))=c,DE);
```

Example 2.49 (Bernoulli Substitution)

Solve $y' + 2y = y^2$.

Solution: It will be shown that the solution is $y = \dfrac{1}{1+Ce^x}$.

The equation can be solved by other methods, notably separation of variables. Bernoulli's substitution $u = y/y^n$ will be applied to find the equivalent first order linear differential equation, as follows.

$u' = (y/y^2)'$ Bernoulli's substitution, $n = 2$.

$\quad = -y^{-2}y'$ Chain rule.

$\quad = -1 + y^{-1}$ Use $y' + 2y = y^2$.

$\quad = -1 + u$ Use $u = y/y^2$.

This linear equation $u' = -1 + u$ has equilibrium solution $u_p = 1$ and homogeneous solution $u_h = Ce^x$. Therefore, $u = u_h + u_p$ gives $y = u^{-1} = 1/(1 + Ce^x)$.

Example 2.50 (Integrating factor $Q = x^a y^b$)

Solve $(3y + 4xy^2)dx + (4x + 5x^2y)dy = 0$.

Solution: The implicit solution $x^3y^4 + x^4y^5 = c$ will be justified.

The equation is not exact as written. To explain why, let $\mathbf{M} = 3y + 4xy^2$ and $\mathbf{N} = 4x + 5x^2y$. Then $\mathbf{M}_y = 8xy + 3$, $\mathbf{N}_x = 10xy + 4$ which implies $\mathbf{M}_y \ne \mathbf{N}_x$ (not exact).

The factor $Q = x^a y^b$ will be an integrating factor for the equation provided a and b are chosen to satisfy $xy\,(\mathbf{M}_y - \mathbf{N}_x) = ay\mathbf{N} - bx\mathbf{M}$. This requirement becomes $xy\,(-2xy - 1) = ay(4x + 5x^2y) - bx(3y + 4xy^2)$. Comparing terms across the equation gives the 2×2 system of equations

$$\begin{array}{rcrcr} 4a & - & 3b & = & -1, \\ 5a & - & 4b & = & -2. \end{array}$$

The unique solution by Cramer's determinant rule is

$$a = \frac{\begin{vmatrix} -1 & -3 \\ -2 & -4 \end{vmatrix}}{\begin{vmatrix} 4 & -3 \\ 5 & -4 \end{vmatrix}} = 2, \quad b = \frac{\begin{vmatrix} 4 & -1 \\ 5 & -2 \end{vmatrix}}{\begin{vmatrix} 4 & -3 \\ 5 & -4 \end{vmatrix}} = 3.$$

Then $Q = x^2y^3$ is the required integrating factor. After multiplication by Q, the original equation becomes the exact equation

$$(3x^2y^4 + 4x^3y^5)dx + (4x^3y^3 + 5x^4y^4)dy = 0.$$

The method of potentials applied to $M = 3x^2y^4 + 4x^3y^5$ and $N = 4x^3y^3 + 5x^4y^4$ finds the potential U as follows.

$U = \int_0^x M(x,y)dx + \int_0^y N(0,y)dy$ — Method of potentials formula.

$\quad = \int_0^x (3x^2y^4 + 4x^3y^5)dx + \int_0^y (0)dy$ — Insert M and N.

$\quad = x^3y^4 + x^4y^5$ — Evaluate integral.

Example 2.51 (Integrating factor $Q = e^{ax+by}$)
Solve $(e^x + e^y)\,dx + (e^x + 2e^y)\,dy = 0$.

Solution: The implicit solution $2e^{3x+3y} + 3e^{2x+4y} = c$ will be justified. A constant 5/6 appears in the integrations below, mysteriously absent in the solution, because 5/6 has been absorbed into the constant c.

Let $\mathbf{M} = e^x + e^y$ and $\mathbf{N} = e^x + 2e^y$. Then $\mathbf{M}_y = e^y$ and $\mathbf{N}_x = e^x$ (not exact). The condition for $Q = e^{ax+by}$ to be an integrating factor is $\mathbf{M}_y - \mathbf{N}_x = a\mathbf{N} - b\mathbf{M}$, which becomes the requirement

$$e^y - e^x = a\,(e^x + 2e^y) - b\,(e^x + e^y)\,.$$

The equations are satisfied provided (a, b) is a solution of the 2×2 system of equations

$$\begin{array}{rcrcr} a & - & b & = & -1, \\ 2a & - & b & = & 1. \end{array}$$

The unique solution is $a = 2$, $b = 3$, by elimination. The original equation multiplied by the integrating factor $Q = e^{2x+3y}$ is the exact equation $Mdx + Ndy = 0$, where $M = e^{3x+3y} + e^{2x+4y}$ and $N = e^{3x+3y} + 2e^{2x+4y}$. The method of potentials applies to find the potential U, as follows.

$$\begin{aligned}
U &= \textstyle\int_0^x M(x,y)dx + \int_0^y N(0,y)dy && \text{Method of potentials.}\\
&= \textstyle\int_0^x \left(e^{3x+3y} + e^{2x+4y}\right)dx + \int_0^y \left(e^{3y} + 2e^{4y}\right)dy && \text{Insert } M \text{ and } N.\\
&= \tfrac{1}{3}e^{3x+3y} + \tfrac{1}{2}e^{2x+4y} - \tfrac{5}{6} && \text{Evaluate integral.}
\end{aligned}$$

Example 2.52 (Integrating factor $Q = Q(x)$)
Solve $(x + y)dx + (x - x^2)dy = 0$.

Solution: The implicit solution $\dfrac{xy + x}{1 - x} + \ln|x - 1| = c$ will be justified.

Let $\mathbf{M} = x + y$, $\mathbf{N} = x - x^2$. Then $\mathbf{M}_y = 1$ and $\mathbf{N}_x = 1 - 2x$ (not exact). Then

$$\begin{aligned}
\mu &= \frac{\mathbf{M}_y - \mathbf{N}_x}{\mathbf{N}} && \text{Hope } \mu \text{ depends on } x \text{ alone.}\\
&= 2/(1 - x) && \text{Substitute } \mathbf{M}, \mathbf{N}\text{; success.}\\
Q &= e^{\int \mu(x)dx} && \text{Integrating factor.}\\
&= e^{-2\ln|1-x|} && \text{Substitute for } \mu \text{ and integrate.}\\
&= (1 - x)^{-2} && \text{Simplified factor found.}
\end{aligned}$$

Multiplication of $\mathbf{M}dx + \mathbf{N}dy = 0$ by Q gives the corresponding exact equation

$$\frac{x + y}{(1 - x)^2}dx + \frac{x}{1 - x}dy = 0.$$

The method of potentials applied to $M = (x + y)/(1 - x)^2$, $N = x/(1 - x)$ finds the implicit solution as follows.

$$\begin{aligned}
U &= \textstyle\int_0^x M(x,y)dx + \int_0^y N(0,y)dy && \text{Method of potentials.}\\
&= \int_0^x \frac{x + y}{(1 - x)^2}dx + \textstyle\int_0^y (0)dy && \text{Substitute for } M, N.\\
&= \int_0^x \left(\frac{y + 1}{(x - 1)^2} + \frac{1}{x - 1}\right)dx && \text{Partial fractions.}\\
&= \frac{xy + x}{1 - x} + \ln|x - 1| && \text{Evaluate integral.}
\end{aligned}$$

Example 2.53 (Integrating factor $Q = Q(y)$)
Solve $(y - y^2)dx + (x + y)dy = 0$.

Solution: Interchange the roles of x and y, then apply the previous example, to obtain the implicit solution $\dfrac{xy + y}{1 - y} + \ln|y - 1| = c$.

This example happens to fit the case when the integrating factor is a function of y alone. The details parallel the previous example.

Details and Proofs

The exactness condition $M_y = N_x$ for $M = Q\mathbf{M}$ and $N = Q\mathbf{N}$ becomes in the case $Q = x^a y^b$ the relation

$$bx^a y^{b-1}\mathbf{M} + x^a y^b \mathbf{M}_y = ax^{a-1}y^b\mathbf{N} + x^a y^b \mathbf{N}_x$$

from which rearrangement gives $xy\left(\mathbf{M}_y - \mathbf{N}_x\right) = ay\mathbf{N} - bx\mathbf{M}$. The case $Q = e^{ax+by}$ is similar.

Consider $Q = e^{\int \mu(x)dx}$. Then $Q' = \mu Q$. The exactness condition $M_y = N_x$ for $M = Q\mathbf{M}$ and $N = Q\mathbf{N}$ becomes $Q\mathbf{M}_y = \mu Q\mathbf{N} + Q\mathbf{N}_x$ and finally

$$\mu = \frac{\mathbf{M}_y - \mathbf{N}_x}{\mathbf{N}}.$$

The similar case $Q = e^{\int \nu(y)dy}$ is obtained from the preceding case, by swapping the roles of x, y.

Exercises 2.10

Homogeneous-A Equations

Find f such that the equation can be written in the form $y' = f(y/x)$. Solve for y using a computer algebra system.

1. $xy' = y^2/x$

2. $x^2y' = x^2 + y^2$

3. $yy' = \dfrac{xy^2}{x^2+y^2}$

4. $yy' = \frac{2xy^2}{x^2+y^2}$

5. $y' = \dfrac{1}{x+y}$

6. $y' = y/x + x/y$

7. $y' = (1 + y/x)^2$

8. $y' = 2y/x + x/y$

9. $y' = 3y/x + x/y$

10. $y' = 4y/x + x/y$

Homogeneous-C Equations

Given $y' = f(x,y)$, decompose $f(x,y) = G(R(x,y))$ where $R(x,y) = \frac{a_1x+b_1y+c_1}{a_2x+b_2y+c_2}$, then convert to Homogeneous-A. Investigate solving $y' = f(x,y)$ by computer.

11. $y' = -\frac{(y+1)x}{y^2+2\,y+1+x^2}$

12. $y' = 2\,\dfrac{(1+y)\,x}{x^2+y^2+2\,y+1}$

13. $y' = \dfrac{(1+x)\,y}{x^2+4\,y^2+2\,x+1}$

14. $y' = \dfrac{1+x}{y+1+x}$

15. $y' = \dfrac{1+y}{x+y+1}$

16. $x(y+1)y' = x^2 + y^2 + 2y + 1$

17. $y' = \dfrac{x^2 - y^2 - 2\,y - 1}{(1+y)\,x}$

18. $y' = \dfrac{(y+2\,x)^2}{x^2}$

19. $y' = \dfrac{x^2 + xy + y^2 + 5\,x + 4\,y + 7}{(x+2)\,(3+y+x)}$

20. $y' = -\dfrac{x^2 - xy - y^2 + 5\,x - 5\,y + 5}{(3+x)\,(4+y+x)}$

Bernoulli's Equation

Identify the exponent n in Bernoulli's equation $y' + p(x)y = q(x)y^n$ and solve for $y(x)$.

21. $y^{-2}y' = 1 + x$

22. $yy' = 1 + x$

23. $y^{-2}y' + y^{-1} = 1 + x$

24. $yy' + y^2 = 1 + x$

25. $y' + y = y^{1/3}$

26. $y' + y = y^{1/5}$

27. $y' - y = y^{-1/2}$

28. $y' - y = y^{-1/3}$

29. $yy' + y^2 = e^x$

30. $y' + y = e^{2x}y^2$

Integrating Factor $x^a y^b$

Report an implicit solution for the given equation $M dx + N dy = 0$, using an integrating factor $Q = x^a y^b$. Follow Example 2.50, page 169. Computer assist expected.

31. $M = 3xy - 6y^2$, $N = 4x^2 - 15xy$

32. $M = 3xy - 10y^2$, $N = 4x^2 - 25xy$

33. $M = 2\,y - 12\,xy^2$, $N = 4\,x - 20\,x^2y$

34. $M = 2\,y - 21\,xy^2$, $N = 4\,x - 35\,x^2y$

35. $M = 3\,y - 32\,xy^2$, $N = 4\,x - 40\,x^2y$

36. $M = 3\,y - 20\,xy^2$, $N = 4\,x - 25\,x^2y$

37. $M = 12\,y - 30\,x^2y^2$, $N = 12\,x - 25\,x^3y$

38. $M = 12\,y + 90\,x^2y^2$, $N = 12\,x + 75\,x^3y$

39. $M = 15\,y + 90\,xy^2$, $N = 12\,x + 75\,x^2y$

40. $M = 35\,y + 30\,xy^2$, $N = 28\,x + 25\,x^2y$.

Integrating Factor e^{ax+by}

Report an implicit solution $U(x, y) = c$ for the given equation $M dx + N dy = 0$ using an integrating factor $Q = e^{ax+by}$. Follow Example 2.51, page 170.

41. $M = e^x + 2e^{2y}$, $N = e^x + 5e^{2y}$

42. $M = 3e^x + 2e^y$, $N = 4e^x + 5e^y$

43. $M = 12\,e^x + 2$, $N = 20\,e^x + 5$

44. $M = 12\,e^x + 2\,e^{-y}$, $N = 24\,e^x + 5\,e^{-y}$

45. $M = 12\,e^y + 2\,e^{-x}$, $N = 24\,e^y + 5\,e^{-x}$

46. $M = 12\,e^{-2\,y} + 2\,e^{-x}$, $N = 12\,e^{-2\,y} + 5\,e^{-x}$

47. $M = 16\,e^y + 2\,e^{-2\,x+3\,y}$, $N = 12\,e^y + 5\,e^{-2\,x+3\,y}$

48. $M = 16\,e^{-y} + 2\,e^{-2\,x-3\,y}$, $N = -12\,e^{-y} - 5\,e^{-2\,x-3\,y}$

49. $M = -16 - 2\,e^{2\,x+y}$, $N = 12 + 4\,e^{2\,x+y}$

50. $M = -16\,e^{-3\,y} - 2\,e^{2\,x}$, $N = 8\,e^{-3\,y} + 5\,e^{2\,x}$

Integrating Factor $Q(x)$

Report an implicit solution $U(x, y) = c$ for the given equation, using an integrating factor $Q = Q(x)$. Follow Example 2.52, page 171.

51. $(x + 2y)dx + (x - x^2)dy = 0$

52. $(x + 3y)dx + (x - x^2)dy = 0$

53. $(2x + y)dx + (x - x^2)dy = 0$

54. $(2x + y)dx + (x + x^2)dy = 0$

55. $(2x + y)dx + (-x - x^2)dy = 0$

56. $(x + y)dx + (-x - x^2)dy = 0$

57. $(x + y)dx + (-x - 2x^2)dy = 0$

58. $(x + y)dx + (x + 5x^2)dy = 0$

59. $(x + y)dx + (3x)dy = 0$

60. $(x + y)dx + (7x)dy = 0$

Integrating Factor $Q(y)$

61. $(y - y^2)dx + (x + y)dy = 0$

62. $(y - y^2)dx + (2x + y)dy = 0$

63. $(y - y^2)dx + (2x + 3y)dy = 0$

64. $(y + y^2)dx + (2x + 3y)dy = 0$

65. $(y + y^2)dx + (x + 3y)dy = 0$

66. $(y + 5y^2)dx + (x + 3y)dy = 0$

67. $(y + 3y^2)dx + (x + 3y)dy = 0$

68. $(2y + 5y^2)dx + (7x + 11y)dy = 0$

69. $(2y + 5y^2)dx + (x + 7y)dy = 0$

70. $(3y + 5y^3)dx + (7x + 9y)dy = 0$

Chapter 3

Linear Algebraic Equations No Matrices

Contents

This introduction to linear algebraic equations requires only a college algebra background. **Vector and matrix notation is not used**. The subject of linear algebra, using vectors, matrices and related tools, appears later in the text; see Chapter 5.

The topics studied are *linear equations*, *general solution*, *reduced echelon system*, *basis*, *nullity*, *rank* and *nullspace*. Introduced here are the **three possibilities**, a **toolkit sequence**, which uses the three rules **swap**, **combination** and **multiply**, and finally the method of **elimination**, in literature called **Gauss-Jordan elimination** or **Gaussian elimination**

3.1 Systems of Linear Equations

Background from college algebra includes systems of linear algebraic equations like

$$(1) \qquad \left\{\begin{array}{rcrcl} 3x & + & 2y & = & 1, \\ x & - & y & = & 2. \end{array}\right.$$

A **solution** (x, y) of **non-homogeneous system** (1) is a pair of values that simultaneously satisfy both equations. This example has unique solution $x = 1$, $y = -1$.

The **homogeneous system** corresponding to (1) is an auxiliary system invented by replacing the right sides of the equations by zero and symbols x, y by new symbols u, v:

$$\left\{\begin{array}{rcrcl} 3u & + & 2v & = & 0, \\ u & - & v & = & 0. \end{array}\right. \tag{2}$$

A short pause and computation verifies that system (2) has unique solution $u = 0$, $v = 0$.

It is unexpected, and also not true, that the original system (solution $x = 1, y = -1$) has any solutions in common with the invented homogeneous system (solution $u = 0, v = 0$). Theory provides **superposition** to relate the solutions of the two systems.

Unique solutions have emphasis in college algebra courses. In this chapter we study in depth the cases for **no solution** and **infinitely many solutions**. These two cases are illustrated by the examples

No Solution	**Infinitely Many Solutions**
$(3)\quad \left\{\begin{array}{rcrcl} x & - & y & = & 0, \\ & & 0 & = & 1. \end{array}\right.$	$(4)\quad \left\{\begin{array}{rcrcl} x & - & y & = & 0, \\ & & 0 & = & 0. \end{array}\right.$

Equations (3) cannot have a solution because of the **signal equation** $0 = 1$, a false equation. Equations (4) have one solution (x, y) for each point on the 45° line $x - y = 0$, therefore system (4) has infinitely many solutions.

The Three Possibilities

Solutions of general linear systems with m equations in n unknowns may be classified into exactly **three possibilities**:

1. No solution.
2. Infinitely many solutions.
3. A unique solution.

General Linear Systems

Given numbers $a_{11}, \ldots, a_{mn}, b_1, \ldots, b_m$, a **nonhomogeneous system** of m linear equations in n **unknowns** $x_1, x_2, \ldots, x_n$ is the system

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2, \\ & \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & b_m. \end{array} \tag{5}$$

Constants a_{11}, …, a_{mn} are called the **coefficients** of system (5). Constants b_1, …, b_m are collectively referenced as the **right hand side**, **right side** or **RHS**.

The associated **homogeneous system** corresponding to system (5) is **invented** by replacing the right side by zero:

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & 0, \\ & \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & 0. \end{array} \tag{6}$$

Convention dictates using the same variable list $x_1, \ldots, x_n$. This abuse of notation impacts casual readers: see example systems (1) and (2).

An assignment of possible values x_1, …, x_n which simultaneously satisfy all equations in (5) is called a **solution** of system (5). **Solving** system (5) refers to the process of finding all possible solutions of (5). The system (5) is called **consistent** if it has a solution and otherwise it is called **inconsistent**.

The Toolkit of Three Rules

Two systems (5) are said to be **equivalent** provided they have exactly the same solutions. For the purpose of solving systems, there is a toolkit of three reversible operations on equations which can be applied to obtain equivalent systems. These rules *neither create nor destroy solutions* of the original system:

Table 1. The Three Rules

Swap	Two equations can be interchanged without changing the solution set.
Multiply	An equation can be multiplied by $m \neq 0$ without changing the solution set.
Combination	A multiple of one equation can be added to a different equation without changing the solution set.

The last two rules replace an existing equation by a new one. A **swap** repeated reverses the swap operation. A **multiply** is reversed by multiplication by $1/m$, whereas the **combination** rule is reversed by subtracting the equation–multiple previously added. In short, the three operations are **reversible**.

Theorem 3.1 (Equivalent Systems)
A second system of linear equations, obtained from the first system of linear equations by a finite number of toolkit operations, has exactly the same solutions as the first system.

Exposition. Writing a set of equations and its equivalent system under toolkit rules demands that all equations be copied, not just the affected equation(s).

Generally, each displayed system changes just one equation, the single exception being a swap of two equations. Within an equation, variables appear left-to-right in variable list order. Equations that contain no variables, typically $0 = 0$, are displayed last.

Documenting the three rules. In blackboard and hand-written work, the acronyms **swap**, **mult** and **combo**, replace the longer terms *swap*, *multiply* and *combination.* They are placed next to the first changed equation. In cases where precision is required, additional information is supplied, namely the **source** and **target** equation numbers s, t and the multiplier $m \neq 0$ or c. Details:

Table 2. Documenting Toolkit Operations with `swap, mult, combo`.

`swap(s,t)`	Swap equations s and t.
`mult(t,m)`	Multiply target equation t by multiplier $m \neq 0$.
`combo(s,t,c)`	Multiply source equation s by multiplier c and add to target equation t.

The acronyms in Table 2 match usage in the computer algebra system `maple`, for package `linalg` and functions `swaprow`, `mulrow` and `addrow`.

Inverses of the Three Rules. Each toolkit operation `swap, mult, combo` has an inverse, which is documented in the following table. The facts can be used to back up several steps, unearthing a previous step to which a sequence of toolkit operations were performed.

Table 3. Inverses of Toolkit Operations `swap, mult, combo`.

Operation	Inverse
`swap(s,t)`	`swap(s,t)`
`mult(t,m)`	`mult(t,1/m)`
`combo(s,t,c)`	`combo(s,t,-c)`

To illustrate, suppose `swap(1,3)`, `combo(1,2,-3)` and `mult(2,4)` are used to obtain the current linear equations. Then the linear system three steps back can be obtained from the current system by applying the inverse steps in reverse order: `mult(2,1/4), combo(1,2,3), swap(1,3)`.

Solving Equations with Geometry

In the plane ($n = 2$) and in 3-space ($n = 3$), equations (5) have a geometric interpretation that can provide valuable intuition about possible solutions. College algebra courses might have omitted the case of *no solutions* or *infinitely many solutions*, discussing only the case of a single unique solution. In contrast, all cases are considered here.

Plane Geometry

A **straight line** may be represented as an equation $Ax + By = C$. Solving the system

$$\begin{array}{rcrcl} a_{11}x & + & a_{12}y & = & b_1 \\ a_{21}x & + & a_{22}y & = & b_2 \end{array} \tag{7}$$

is the geometrical equivalent of finding all possible (x, y)-intersections of the lines represented in system (7). The distinct geometrical possibilities appear in Figures 1, 2 and 3.

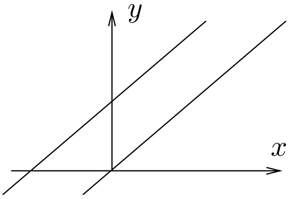


Figure 1. Parallel lines, no solution.

$$\begin{array}{rcl} -x + y & = & 1, \\ -x + y & = & 0. \end{array}$$

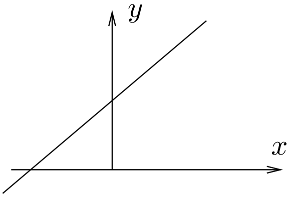


Figure 2. Identical lines, infinitely many solutions.

$$\begin{array}{rcl} -x + y & = & 1, \\ -2x + 2y & = & 2. \end{array}$$

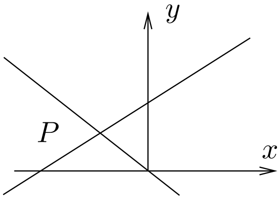


Figure 3. Non-parallel distinct lines, one solution at the unique intersection point P.

$$\begin{array}{rcl} -x + y & = & 2, \\ x + y & = & 0. \end{array}$$

Space Geometry

A **plane** in xyz-space is given by an equation $Ax + By + Cz = D$. The vector $A\vec{\imath} + B\vec{\jmath} + C\vec{k}$ is **normal** to the plane. An equivalent equation is $A(x - x_0) + B(y - y_0) + C(z - z_0) = 0$, where (x_0, y_0, z_0) is a given point in the plane. Solving system

$$\begin{array}{rcrcrcl} a_{11}x & + & a_{12}y & + & a_{13}z & = & b_1 \\ a_{21}x & + & a_{22}y & + & a_{23}z & = & b_2 \\ a_{31}x & + & a_{32}y & + & a_{33}z & = & b_3 \end{array} \tag{8}$$

is the geometric equivalent of finding all possible (x, y, z)-intersections of the planes represented by system (8). Illustrated in Figures 4–11 are some interesting geometrical possibilities.

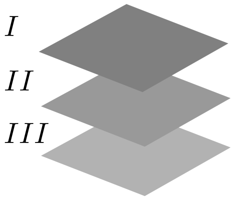


Figure 4. Three Parallel Shelves. Planes I, II, III are parallel. There is no intersection point.

$$I:\ z=2, \quad II:\ z=1, \quad III:\ z=0.$$

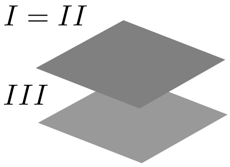


Figure 5. Two Parallel Shelves. Planes I, II are equal and parallel to plane III. There is no intersection point.

$$I:\ 2z=2, \quad II:\ z=1, \quad III:\ z=0.$$

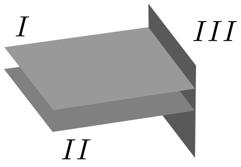


Figure 6. Book shelf. Two planes I, II are distinct and parallel. There is no intersection point.

$$I:\ z=2, \quad II:\ z=1, \quad III:\ y=0.$$

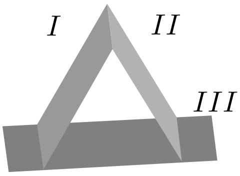


Figure 7. Pup tent. Two non-parallel planes I, II meet in a line which never meets plane III. There are no intersection points.

$$I:\ y+z=0, \quad II:\ y-z=0, \quad III:\ z=-1.$$

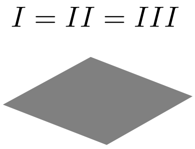


Figure 8. Three Identical Shelves. Planes I, II, III are equal. There are infinitely many intersection points.

$$I:\ z=1, \quad II:\ 2z=2, \quad III:\ 3z=3.$$

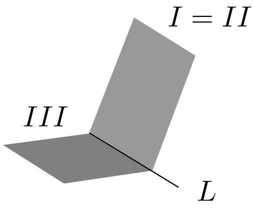


Figure 9. Open book. Equal planes I, II meet another plane III in a line L. There are infinitely many intersection points.

$$I: \; y + z = 0, \; II: \; 2y + 2z = 0, \; III: \; z = 0.$$

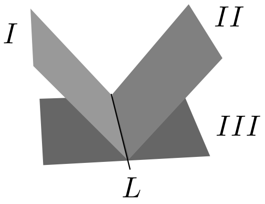


Figure 10. Saw Tooth. Two non-parallel planes I, II meet in a line L which lies in a third plane III. There are infinitely many intersection points.

$$I: \; -y + z = 0, \quad II: \; y + z = 0, \quad III: \; z = 0.$$

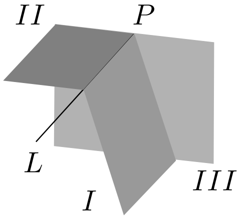


Figure 11. Knife Cuts an Open Book. Two non-parallel planes I, II meet in a line L not parallel to plane III. There is a *unique point* P of intersection of all three planes.

$$I: \; y + z = 0, \quad II: \; z = 0, \quad III: \; x = 0.$$

Examples and Methods

Example 3.1 (Toolkit)

Given system $\left|\begin{array}{ccccccc} x & & & + & 4z & = & 1 \\ x & + & y & + & 4z & = & 3 \\ & & & & z & = & 2 \end{array}\right|$, find the system that results from operation `swap(1,2)` followed by operation `combo(2,1,-1)`.

Solution: The steps are as follows, with the equivalent system equal to the last display.

$$\left|\begin{array}{ccccccc} x & & & + & 4z & = & 1 \\ x & + & y & + & 4z & = & 3 \\ & & & & z & = & 2 \end{array}\right| \quad \text{Original system.}$$

$$\left|\begin{array}{ccccccc} x & + & y & + & 4z & = & 3 \\ x & & & + & 4z & = & 1 \\ & & & & z & = & 2 \end{array}\right| \quad \texttt{swap(1,2)}$$

$$\left|\begin{array}{ccccccc} & & y & & & = & 2 \\ x & & & + & 4z & = & 1 \\ & & & & z & = & 2 \end{array}\right| \quad \texttt{combo(2,1,-1)}$$

Calculations for `combo(2,1,-1)` can be done on scratch paper. Experts do the arithmetic column-by-column, using no scratch paper. Here's the details for the scratch paper arithmetic:

$$\begin{array}{cccccccl} 1x & + & 0y & + & 4z & = & 1 & \text{Equation 2} \\ 1x & + & 1y & + & 4z & = & 3 & \text{Equation 1} \end{array}$$

$$\begin{array}{rcrcrcrl} -1x & + & 0y & - & 4z & = & -1 & \text{Equation 2 times -1} \\ 1x & + & 1y & + & 4z & = & 3 & \text{Equation 1} \end{array}$$

Add on the columns, replacing the second equation.

$$\begin{array}{rcrcrcrl} -1x & + & 0y & - & 4z & = & -1 & \text{Equation 2 times -1} \\ 0x & + & 1y & + & 0z & = & 2 & \text{Equation 1 + (-1)(Equation 2)} \end{array}$$

The last equation replaces equation 1 and the label `combo(2,1,-1)` is written next to the replacement. All of the scratch work is discarded.

Example 3.2 (Inverse Toolkit)

Let system $\left|\begin{array}{rcrcrcr} x & & & - & 3z & = & -1 \\ & & 2y & + & 6z & = & 4 \\ & & & & z & = & 3 \end{array}\right|$ be produced by toolkit operations, first `mult(2,2)` and then `combo(2,1,-1)`. Find the original system.

Solution: We begin by writing the given toolkit operation inverses, in reverse order, as `combo(2,1,1)` and `mult(2,1/2)`. The operations, in this order, are performed on the given system, to find the original system two steps back, in the last display.

$$\left|\begin{array}{rcrcrcr} x & & & - & 3z & = & -1 \\ & & 2y & + & 6z & = & 4 \\ & & & & z & = & 3 \end{array}\right|$$

Given system.

$$\left|\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 3 \\ & & 2y & + & 6z & = & 4 \\ & & & & z & = & 3 \end{array}\right|$$

`combo(2,1,1)`
One step back.

$$\left|\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 3 \\ & & y & + & 3z & = & 2 \\ & & & & z & = & 3 \end{array}\right|$$

`mult(2,1/2)`
Two steps back.

Example 3.3 (Planar System)

Classify the system geometrically as one of the three types displayed in Figures 1, 2, 3. Then solve for x and y.

$$\left|\begin{array}{rcrcr} x & + & 2y & = & 1, \\ 3x & + & 6y & = & 3. \end{array}\right| \tag{9}$$

Solution: The second equation, divided by 3, gives the first equation. In short, the two equations are proportional. The lines are geometrically **equal lines**, as in Figure 2. The two equations are equivalent to the system

$$\left|\begin{array}{rcrcr} x & + & 2y & = & 1, \\ & & 0 & = & 0. \end{array}\right|$$

To solve the system means to find all points (x, y) simultaneously common to both lines, which are all points (x, y) on $x + 2y = 1$.

A parametric representation of this line is possible, obtained by setting $y = t$ and then solving for $x = 1 - 2t$, $-\infty < t < \infty$. We report the solution as a parametric solution, but the first solution is also valid.

$$\begin{array}{l} x = 1 - 2t, \\ y = t. \end{array}$$

Example 3.4 (No Solution)

Classify the system geometrically as the type displayed in Figure 1. Explain why there is no solution.

$$\left|\begin{array}{rcrcl} x & + & 2y & = & 1, \\ 3x & + & 6y & = & 6. \end{array}\right| \tag{10}$$

Solution: The second equation, divided by 3, gives $x + 2y = 2$, a line parallel to the first line $x + 2y = 1$. The lines are geometrically **parallel lines**, as in Figure 1. The two equations are equivalent to the system

$$\left|\begin{array}{rcrcl} x & + & 2y & = & 1, \\ x & + & 2y & = & 2. \end{array}\right|$$

To solve the system means to find all points (x, y) simultaneously common to both lines, which are all points (x, y) on $x + 2y = 1$ and also on $x + 2y = 2$. If such a point (x, y) exists, then $1 = x + 2y = 2$ or $1 = 2$, a contradictory **signal equation**. Because $1 = 2$ is **false**, then no common point (x, y) exists and we report **no solution**.

Some readers will want to continue and write equations for x and y, a *solution* to the problem. We emphasize that this is not possible, because there is no solution at all.

The presence of a signal equation, which is a false equation used primarily to detect no solution, will appear always in the solution process for a system of equations that has no solution. Generally, this signal equation, if present, will be distilled to the single equation "$0 = 1$." For instance, $0 = 2$ can be distilled to $0 = 1$ by dividing equation $0 = 2$ by 2.

,

Exercises 3.1

Toolkit

Compute the equivalent system of equations. Definitions of `combo`, `swap` and `mult` on page 177.

1. Given $\left|\begin{array}{rcrcrcl} x & & & + & 2z & = & 1 \\ x & + & y & + & 2z & = & 4 \\ & & & & z & = & 0 \end{array}\right|$, find the system that results from `combo(2,1,-1)`.

2. Given $\left|\begin{array}{rcrcrcl} x & & & + & 2z & = & 1 \\ x & + & y & + & 2z & = & 4 \\ & & & & z & = & 0 \end{array}\right|$, find the system that results from `swap(1,2)` followed by `combo(2,1,-1)`.

3. Given $\left|\begin{array}{rcrcrcl} x & & & + & 3z & = & 1 \\ x & + & y & + & 3z & = & 4 \\ & & & & z & = & 1 \end{array}\right|$, find the system that results from `combo(1,2,-1)`.

4. Given $\left|\begin{array}{rcrcrcl} x & & & + & 3z & = & 1 \\ x & + & y & + & 3z & = & 4 \\ & & & & z & = & 1 \end{array}\right|$, find the system that results from `swap(1,2)` followed by `combo(1,2,-1)`.

5. Given $\left|\begin{array}{rcrcl} y & + & z & = & 2 \\ 3y & + & 3z & = & 6 \\ y & & & = & 0 \end{array}\right|$, find the system that results from `swap(2,3)`, `combo(2,1,-1)`.

6. Given $\left|\begin{array}{rcrcl} y & + & z & = & 2 \\ 3y & + & 3z & = & 6 \\ y & & & = & 0 \end{array}\right|$, find the system that results from `mult(2,1/3)`, `combo(1,2,-1)`, `swap(2,3)`, `swap(1,2)`.

Inverse Toolkit

Compute the equivalent system of equations.

7. If $\left|\begin{array}{rcr} -\ y &=& -3\\ x+y+2z &=& 4\\ z &=& 0\end{array}\right|$ resulted from `combo(2,1,-1)`, then find the original system.

8. If $\left|\begin{array}{rcl} y &=& 3\\ x\quad +2z &=& 1\\ z &=& 0\end{array}\right|$ resulted from `swap(1,2)` followed by `combo(2,1,-1)`, then find the original system.

9. If $\left|\begin{array}{rcl} x\quad +3z &=& 1\\ y-3z &=& 4\\ z &=& 1\end{array}\right|$ resulted from `combo(1,2,-1)`, then find the original system.

10. If $\left|\begin{array}{rcl} x\quad +3z &=& 1\\ x+y+3z &=& 4\\ z &=& 1\end{array}\right|$ resulted from `swap(1,2)` followed by `combo(2,1,2)`, then find the original system.

11. If $\left|\begin{array}{rcl} y+z &=& 2\\ 3y+3z &=& 6\\ y\quad &=& 0\end{array}\right|$ resulted from `mult(2,-1)`, `swap(2,3)`, `combo(2,1,-1)`, then find the original system.

12. If $\left|\begin{array}{rcl} 2y+z &=& 2\\ 3y+3z &=& 6\\ y\quad &=& 0\end{array}\right|$ resulted from `mult(2,1/3)`, `combo(1,2,-1)`, `swap(2,3)`, `swap(1,2)`, then find the original system.

Planar System

Solve the xy–system and interpret the solution geometrically as

(a) **parallel lines**

(b) **equal lines**

(c) **intersecting lines**.

13. $\left|\begin{array}{rcl} x+y &=& 1,\\ y &=& 1\end{array}\right|$

14. $\left|\begin{array}{rcr} x+y &=& -1\\ x\quad &=& 3\end{array}\right|$

15. $\left|\begin{array}{rcl} x+y &=& 1\\ x+2y &=& 2\end{array}\right|$

16. $\left|\begin{array}{rcl} x+y &=& 1\\ x+2y &=& 3\end{array}\right|$

17. $\left|\begin{array}{rcl} x+y &=& 1\\ 2x+2y &=& 2\end{array}\right|$

18. $\left|\begin{array}{rcl} 2x+y &=& 1\\ 6x+3y &=& 3\end{array}\right|$

19. $\left|\begin{array}{rcr} x-y &=& 1\\ -x-y &=& -1\end{array}\right|$

20. $\left|\begin{array}{rcl} 2x-y &=& 1\\ x-0.5y &=& 0.5\end{array}\right|$

21. $\left|\begin{array}{rcl} x+y &=& 1\\ x+y &=& 2\end{array}\right|$

22. $\left|\begin{array}{rcl} x-y &=& 1\\ x-y &=& 0\end{array}\right|$

System in Space

For each xyz–system:

(a) If no solution, then report **three identical shelves**, **pup tent**, **two parallel shelves** or **book shelf**.

(b) If infinitely many solutions, then report **one shelf**, **open book** or **saw tooth**.

(c) If a unique intersection point, then report the values of x, y and z.

23. $\left|\begin{array}{lcl} x-y+z &=& 2\\ x &=& 1\\ \quad\ y &=& 0\end{array}\right|$

24. $\left|\begin{array}{rcl} x+y-2z &=& 3\\ x\qquad &=& 2\\ z &=& 1\end{array}\right|$

25. $\left|\begin{array}{rcl} x-y &=& 2\\ x-y &=& 1\\ x-y &=& 0\end{array}\right|$

26. $\left|\begin{array}{rcl} x+y &=& 3\\ x+y &=& 2\\ x+y &=& 1\end{array}\right|$

27. $\left|\begin{array}{rcl} x+y+z &=& 3\\ x+y+z &=& 2\\ x+y+z &=& 1\end{array}\right|$

28. $$\left|\begin{array}{rcrcrcr} x & + & y & + & 2z & = & 2 \\ x & + & y & + & 2z & = & 1 \\ x & + & y & + & 2z & = & 0 \end{array}\right|$$

29. $$\left|\begin{array}{rcrcrcr} x & - & y & + & z & = & 2 \\ 2x & - & 2y & + & 2z & = & 4 \\ & & y & & & = & 0 \end{array}\right|$$

30. $$\left|\begin{array}{rcrcrcr} x & + & y & - & 2z & = & 3 \\ 3x & + & 3y & - & 6z & = & 6 \\ & & & & z & = & 1 \end{array}\right|$$

31. $$\left|\begin{array}{rcrcrcr} x & - & y & + & z & = & 2 \\ & & & & 0 & = & 0 \\ & & & & 0 & = & 0 \end{array}\right|$$

32. $$\left|\begin{array}{rcrcrcr} x & + & y & - & 2z & = & 3 \\ & & & & 0 & = & 0 \\ & & & & 1 & = & 1 \end{array}\right|$$

33. $$\left|\begin{array}{rcrcrcr} x & + & y & & & = & 2 \\ x & - & y & & & = & 2 \\ & & y & & & = & -1 \end{array}\right|$$

34. $$\left|\begin{array}{rcrcrcr} x & & & - & 2z & = & 4 \\ x & & & + & 2z & = & 0 \\ & & & & z & = & 2 \end{array}\right|$$

35. $$\left|\begin{array}{rcrcrcr} & & y & + & z & = & 2 \\ & & 3y & + & 3z & = & 6 \\ & & y & & & = & 0 \end{array}\right|$$

36. $$\left|\begin{array}{rcrcrcr} x & & & + & 2z & = & 1 \\ 4x & & & + & 8z & = & 4 \\ & & & & z & = & 0 \end{array}\right|$$

,

3.2 Filmstrips and Toolkit Sequences

Expert on Video. A linear algebra expert solves a system of equations with paper and pencil. A video records all the paper details, starting with the original system of equations and ending with the solution. Each application of one of the toolkit operations `swap`, `combo` or `mult` causes the system of equations to be re-written.

Filmstrip. The documentary video is edited into an ordered sequence of images, a **filmstrip** which eliminates all arithmetic details. The cropped images are the selected frames which record the result of each computation: only major toolkit steps appear (see Table 4).

Table 4. A Toolkit Sequence.
Each image is a cropped frame from a filmstrip, obtained by editing a video documentary of an expert solving the linear system.

Frame 1	**Frame 2**	**Frame 3**
Original System $\begin{cases} x - y = 2, \\ 3y = -3. \end{cases}$	Apply `mult(2,1/3)` $\begin{cases} x - y = 2, \\ y = -1. \end{cases}$	Apply `combo(2,1,1)` $\begin{cases} x = 1, \\ y = -1. \end{cases}$

Definition 3.1 (Toolkit Sequence)
Assume a video has been made of a person solving a linear system. A sequence of selected filmstrip images, presented in solution order, is called a **Toolkit Sequence**. The images are presumed cropped and devoid of arithmetic detail, but each toolkit step is documented.

> The cropped images of major toolkit steps make a filmstrip which represents the minimum set of solution steps to be written on paper.

Lead Variables

A variable chosen from the variable list x, y is called a **lead variable** provided it appears just once in the entire system of equations, and in addition, its appearance reading left-to-right is first, with coefficient one. The same definition applies to arbitrary variable lists x_1, x_2, $\ldots$, x_n.

Illustration. Symbol x is a lead variable in all three frames of the toolkit sequence in Table 4. But symbol y fails to be a lead variable in frames 1 and 2. In the final frame, both x and y are lead variables.

A **free variable** is a non-lead variable, detectable only from a frame in which every non-zero equation has a lead variable.

A consistent system in which every variable is a lead variable must have a unique solution. The system must look like the final frame of the sequence in Table 4. More precisely, the variables appear in variable list order to the left of the equal sign, each variable appearing just once, with numbers to the right of the equal sign.

Unique Solution

To solve a system with a unique solution, we apply the toolkit operations of swap, multiply and combination (acronyms `swap, mult, combo`), one operation per frame, until the last frame displays the unique solution.

Because all variables will be lead variables in the last frame, we seek to create a new lead variable in each frame. Sometimes, this is not possible, even if it is the general objective. Exceptions are swap and multiply operations, which are often used to prepare for creation of a lead variable. Listed in Table 5 are the rules and conventions that we use to create toolkit sequences.

Table 5. Conventions and Rules for Creating Toolkit Sequences.

Order of Variables. Variables in equations appear in variable list order to the left of the equal sign.

Order of Equations. Equations are listed in variable list order inherited from their lead variables. Equations without lead variables appear next. Equations without variables appear last. Multiple swap operations convert any system to this convention.

New Lead Variable. Select a new lead variable as the *first variable*, in variable list order, which appears among the equations without a lead variable.

An illustration:

$$\boxed{\begin{array}{rcrcrcr} & & y & + & 4z & = & 2, \\ x & + & y & & & = & 3, \\ x & + & 2y & + & 3z & = & 4. \end{array}}$$

Frame 1. Original system.

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ x & + & y & & & = & 3, \\ & & y & + & 4z & = & 2. \end{array}}$$

Frame 2.

`swap(1,3)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & y & + & 4z & = & 2. \end{array}}$$

Frame 3.
`combo(1,2,-1)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & & & z & = & 1. \end{array}}$$

Frame 4.

`combo(2,3,1)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & & y & + & 3z & = & 1, \\ & & & & z & = & 1. \end{array}}$$

Frame 5.
`mult(2,-1)`

$$\boxed{\begin{array}{rcrcrcr} x & & & - & 3z & = & 2, \\ & & y & + & 3z & = & 1, \\ & & & & z & = & 1. \end{array}}$$

Frame 6.
`combo(2,1,-2)`

$$\boxed{\begin{array}{rcrcrcr} x & & & - & 3z & = & 2, \\ & & y & & & = & -2, \\ & & & & z & = & 1. \end{array}}$$

Frame 7.
`combo(3,2,-3)`

$$\boxed{\begin{array}{rcrcrcr} x & & & & & = & 5, \\ & & y & & & = & -2, \\ & & & & z & = & 1. \end{array}}$$

Frame 8. `combo(3,1,3)`
Last Frame.
Unique solution.

No Solution

A special case occurs in a toolkit sequence, when a nonzero equation occurs having no variables. Called a **signal equation**, its occurrence signals **no solution**, because the equation is false. Normally, we halt the toolkit sequence at the point of first discovery, and then declare no solution. An illustration:

$$\boxed{\begin{array}{rcrcrcr} & & y & + & 3z & = & 2, \\ x & + & y & & & = & 3, \\ x & + & 2y & + & 3z & = & 4. \end{array}}$$

Frame 1. Original system.

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ x & + & y & & & = & 3, \\ & & y & + & 3z & = & 2. \end{array}}$$

Frame 2.

`swap(1,3)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & y & + & 3z & = & 2. \end{array}}$$

Frame 3.
`combo(1,2,-1)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & & & 0 & = & 1. \end{array}}$$

Frame 4.
Signal Equation $0 = 1$.
`combo(2,3,1)`

The signal equation $\boxed{0 = 1}$ is a false equation, therefore the last frame has no solution. Because the toolkit neither creates nor destroys solutions, then the original system in the first frame has **no solution**.

Readers who want to go on and write an answer for the system must be warned that **no such possibility exists**. Values cannot be assigned to any variables in the case of no solution. This can be perplexing, especially in a final frame like

$$\boxed{\begin{array}{rcr} x & = & 4, \\ z & = & -1, \\ 0 & = & 1. \end{array}}$$

While it is true that x and z were assigned values, the final signal equation $0 = 1$ is false, meaning any answer is impossible. There is no possibility to write equations for all variables. There is **no solution**. It is a **tragic error** to claim $x = 4$, $z = -1$ is a solution.

Infinitely Many Solutions

A system of equations having infinitely many solutions is solved from a toolkit sequence construction that parallels the unique solution case. The same quest for lead variables is made, hoping in the final frame to have just the variable list on the left and numbers on the right.

The stopping criterion which identifies the final frame, in either the case of a unique solution or infinitely many solutions, is exactly the same:

> **Last Frame Test**. A frame is the **last frame** when every nonzero equation has a lead variable. Remaining equations have the form $0 = 0$.

Any variables that are not lead variables, in the final frame, are called **free variables**, because their values are completely undetermined. Any **missing variable** must be a free variable.

$$\boxed{\begin{array}{rcrcrcr} & & y & + & 3z & = & 1, \\ x & + & y & & & = & 3, \\ x & + & 2y & + & 3z & = & 4. \end{array}}$$

Frame 1. Original system.

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ x & + & y & & & = & 3, \\ & & y & + & 3z & = & 1. \end{array}}$$

Frame 2.

`swap(1,3)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & y & + & 3z & = & 1. \end{array}}$$

Frame 3.
`combo(1,2,-1)`

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & - & y & - & 3z & = & -1, \\ & & & & 0 & = & 0. \end{array}}$$

Frame 4.

combo(2,3,1)

$$\boxed{\begin{array}{rcrcrcr} x & + & 2y & + & 3z & = & 4, \\ & & y & + & 3z & = & 1, \\ & & & & 0 & = & 0. \end{array}}$$

Frame 5.
mult(2,-1)

$$\boxed{\begin{array}{rcrcrcr} x & & & - & 3z & = & 2, \\ & & y & + & 3z & = & 1, \\ & & & & 0 & = & 0. \end{array}}$$

Frame 6. combo(2,1,-2)
Last Frame.
Lead=x, y, Free=z.

Last Frame to General Solution

Once the *last frame* of the toolkit sequence is obtained, then the general solution can be written by a fixed and easy-to-learn algorithm.

Last Frame Algorithm

This process applies only to the last frame in the case of infinitely many solutions.

(1) **Assign invented symbols** t_1, t_2, ... to the free variables.
(2) **Isolate** each lead variable.
(3) **Back-substitute** the free variable invented symbols.

To illustrate, assume the last frame of the toolkit sequence is

$$\boxed{\begin{array}{rcrcrcr} x & & & - & 3z & = & 2, \\ & & y & + & 3z & = & 1, \\ & & & & 0 & = & 0, \end{array}}$$

Last Frame.
Lead variables x, y.

then the general solution is written as follows.

$$\boxed{z = t_1}$$

The free variable z is assigned symbol t_1.

$$\boxed{\begin{array}{l} x = 2 + 3z, \\ y = 1 - 3z \end{array}}$$

The lead variables are x, y. Isolate them left.

$$\boxed{\begin{array}{l} x = 2 + 3t_1, \\ y = 1 - 3t_1, \\ z = t_1. \end{array}}$$

Back-substitute. Solution found.

In the **last frame**, variables appear left of the equal sign in variable list order. Only invented symbols[1] appear right of the equal sign. The expression is called a **standard general solution**. The meaning:

[1] Computer algebra system maple uses invented symbols t_1, t_2, t_3, ... and we follow the convention.

Nothing Skipped — Each solution of the system of equations can be obtained by specializing the invented symbols t_1, t_2, … to particular numbers.

It Works — The general solution expression satisfies the system of equations for all possible values of the symbols t_1, t_2, … .

General Solution and the Last Frame Algorithm

An additional illustration will be given for the last frame algorithm. Assume **variable list order** x, y, z, w, u, v for the **last frame**

$$\begin{array}{lrcl} (1) \quad & \boxed{x} + z + u + v & = & 1, \\ & \boxed{y} - u + v & = & 2, \\ & \boxed{w} + 2u - v & = & 0. \end{array}$$

Every nonzero equation above has a lead variable. The **lead variables** in (1) are the boxed symbols x, y, w. The **free variables** are z, u, v.

Assign invented symbols t_1, t_2, t_3 to the free variables and back-substitute in (1) to obtain a **standard general solution**

$$\left\{\begin{array}{lcl} x & = & 1 - t_1 - t_2 - t_3, \\ y & = & 2 + t_2 - t_3, \\ w & = & -2t_2 + t_3, \\ z & = & t_1, \\ u & = & t_2, \\ v & = & t_3. \end{array}\right. \quad \text{or} \quad \left\{\begin{array}{lcl} x & = & 1 - t_1 - t_2 - t_3, \\ y & = & 2 + t_2 - t_3, \\ z & = & t_1, \\ w & = & -2t_2 + t_3, \\ u & = & t_2, \\ v & = & t_3. \end{array}\right.$$

It is demanded by convention that general solutions be displayed in variable list order. This is why the above display bothers to re-write the equations in the new order on the right.

,

Exercises 3.2

Lead and free variables

For each system assume variable list x_1, …, x_5. List the lead and free variables.

1. $$\left|\begin{array}{rr} x_2 + 3x_3 & = 0 \\ x_4 & = 0 \\ 0 & = 0 \end{array}\right|$$

2. $$\left|\begin{array}{rr} x_2 \qquad\qquad & = 0 \\ x_3 \quad + 3x_5 & = 0 \\ x_4 + 2x_5 & = 0 \end{array}\right|$$

3. $$\left|\begin{array}{rr} x_1 + 3x_3 \qquad & = 0 \\ x_4 & = 0 \\ 0 & = 0 \end{array}\right|$$

4. $$\left|\begin{array}{rr} x_1 + 2x_2 + 3x_3 \qquad & = 0 \\ x_4 & = 0 \\ 0 & = 0 \end{array}\right|$$

5. $$\left|\begin{array}{rr} x_1 + 2x_2 + 3x_3 & = 0 \\ 0 & = 0 \\ 0 & = 0 \\ 0 & = 0 \end{array}\right|$$

6. $$\left|\begin{array}{rr} x_1 + x_2 \qquad & = 0 \\ x_3 & = 0 \\ 0 & = 0 \end{array}\right|$$

7. $\left|\begin{array}{r} x_1+x_2+3x_3+5x_4 \qquad =0\\ x_5=0\\ 0=0\end{array}\right|$

8. $\left|\begin{array}{r} x_1+2x_2 \qquad +3x_4+4x_5=0\\ x_3+\ \ x_4+\ \ x_5=0\\ 0=0\end{array}\right|$

9. $\left|\begin{array}{r} x_3+2x_4 \qquad =0\\ x_5=0\\ 0=0\\ 0=0\end{array}\right|$

10. $\left|\begin{array}{r} x_4+x_5=0\\ 0=0\\ 0=0\\ 0=0\end{array}\right|$

11. $\left|\begin{array}{r} x_2 \qquad +5x_4 \qquad =0\\ x_3+2x_4 \qquad =0\\ x_5=0\\ 0=0\end{array}\right|$

12. $\left|\begin{array}{r} x_1 \qquad +3x_3 \qquad\qquad =0\\ x_2 \qquad\qquad +x_4 \qquad =0\\ x_5=0\\ 0=0\end{array}\right|$

Elementary Operations

Consider the 3×3 system

$$\begin{array}{rcrcrcl} x & + & 2y & + & 3z & = & 2,\\ -2x & + & 3y & + & 4z & = & 0,\\ -3x & + & 5y & + & 7z & = & 3.\end{array}$$

Define symbols **combo**, **swap** and **mult** as in the textbook. Write the 3×3 system which results from each of the following operations.

13. `combo(1,3,-1)`

14. `combo(2,3,-5)`

15. `combo(3,2,4)`

16. `combo(2,1,4)`

17. `combo(1,2,-1)`

18. combo(1,2,-e^2)

19. `mult(1,5)`

20. `mult(1,-3)`

21. `mult(2,5)`

22. `mult(2,-2)`

23. `mult(3,4)`

24. `mult(3,5)`

25. mult(2,-π)

26. mult(2,π)

27. mult(1,e^2)

28. mult(1,-e^{-2})

29. `swap(1,3)`

30. `swap(1,2)`

31. `swap(2,3)`

32. `swap(2,1)`

33. `swap(3,2)`

34. `swap(3,1)`

Unique Solution

Create a toolkit sequence for each system, whose final frame displays the unique solution of the system of equations. Assume variable list order x_1, x_2, x_3, x_4, x_5 and the number of variables is the number of equations.

35. $\left|\begin{array}{r} x_1+3x_2=\ \ 0\\ x_2=-1\end{array}\right|$

36. $\left|\begin{array}{r} x_1+2x_2=\ \ 0\\ x_2=-2\end{array}\right|$

37. $\left|\begin{array}{r} x_1+3x_2=2\\ x_1-\ \ x_2=1\end{array}\right|$

38. $\left|\begin{array}{r} x_1+\ \ x_2=-1\\ x_1+2x_2=-2\end{array}\right|$

39. $\left|\begin{array}{r} x_1+3x_2+2x_3=1\\ x_2+4x_3=3\\ 4x_3=4\end{array}\right|$

40. $\left|\begin{array}{l} x_1 \qquad\qquad\qquad\quad =1\\ 3x_1+\ \ x_2 \qquad\quad\ =0\\ 2x_1+2x_2+3x_3=3\end{array}\right|$

41. $$\left|\begin{array}{rcl} x_1+x_2+3x_3 &=& 1 \\ x_2 &=& 2 \\ 3x_3 &=& 0 \end{array}\right|$$

42. $$\left|\begin{array}{rcl} x_1+3x_2+2x_3 &=& 1 \\ x_2 &=& 3 \\ 3x_3 &=& 0 \end{array}\right|$$

43. $$\left|\begin{array}{rcl} x_1 &=& 2 \\ x_1+2x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+2x_4 &=& 2 \end{array}\right|$$

44. $$\left|\begin{array}{rcl} x_1 &=& 3 \\ x_1-2x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+4x_4 &=& 2 \end{array}\right|$$

45. $$\left|\begin{array}{rcl} x_1+x_2 &=& 2 \\ x_1+2x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+2x_4 &=& 2 \end{array}\right|$$

46. $$\left|\begin{array}{rcl} x_1-2x_2 &=& 3 \\ x_1-x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+4x_4 &=& 1 \end{array}\right|$$

47. $$\left|\begin{array}{rcl} x_1 &=& 3 \\ x_1-x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+4x_4 &=& 1 \\ 3x_1+x_3+2x_5 &=& 1 \end{array}\right|$$

48. $$\left|\begin{array}{rcl} x_1 &=& 2 \\ x_1-x_2 &=& 0 \\ 2x_1+2x_2+x_3 &=& 1 \\ 3x_1+6x_2+x_3+3x_4 &=& 1 \\ 3x_1+x_3+3x_5 &=& 1 \end{array}\right|$$

49. $$\left|\begin{array}{rcl} x_1-x_2+x_3-x_4+x_5 &=& 0 \\ 2x_2-x_3+x_4-x_5 &=& 0 \\ 3x_3-x_4+x_5 &=& 0 \\ 4x_4-x_5 &=& 0 \\ 5x_5 &=& 20 \end{array}\right|$$

50. $$\left|\begin{array}{rcl} x_1-x_2 &=& 3 \\ x_1-2x_2 &=& 0 \\ 2x_1+2x_2+x_3 &=& 1 \\ 3x_1+6x_2+x_3+3x_4 &=& 1 \\ 3x_1+x_3+x_5 &=& 3 \end{array}\right|$$

No Solution

Develop a toolkit sequence for each system, whose final frame contains a signal equation (e.g., $0 = 1$), thereby showing that the system has no solution.

51. $$\left|\begin{array}{rcl} x_1+3x_2 &=& 0 \\ x_1+3x_2 &=& 1 \end{array}\right|$$

52. $$\left|\begin{array}{rcl} x_1+2x_2 &=& 1 \\ 2x_1+4x_2 &=& 2 \end{array}\right|$$

53. $$\left|\begin{array}{rcl} x_1+3x_2+2x_3 &=& 1 \\ x_2+4x_3 &=& 3 \\ x_2+4x_3 &=& 4 \end{array}\right|$$

54. $$\left|\begin{array}{rcl} x_1 &=& 0 \\ 3x_1+x_2+3x_3 &=& 1 \\ 2x_1+2x_2+6x_3 &=& 0 \end{array}\right|$$

55. $$\left|\begin{array}{rcl} x_1+x_2+3x_3 &=& 1 \\ x_2 &=& 2 \\ x_1+2x_2+3x_3 &=& 2 \end{array}\right|$$

56. $$\left|\begin{array}{rcl} x_1+3x_2+2x_3 &=& 1 \\ x_2+2x_3 &=& 3 \\ x_1+5x_3 &=& 5 \end{array}\right|$$

57. $$\left|\begin{array}{rcl} x_1 &=& 2 \\ x_1+2x_2 &=& 2 \\ x_1+2x_2+x_3+2x_4 &=& 0 \\ x_1+6x_2+x_3+2x_4 &=& 2 \end{array}\right|$$

58. $$\left|\begin{array}{rcl} x_1 &=& 3 \\ x_1-2x_2 &=& 1 \\ 2x_1+2x_2+x_3+4x_4 &=& 0 \\ 3x_1+6x_2+x_3+4x_4 &=& 2 \end{array}\right|$$

59. $$\left|\begin{array}{rcl} x_1 &=& 3 \\ x_1-x_2 &=& 1 \\ 2x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+4x_4-x_5 &=& 1 \\ -6x_2-x_3+4x_4+x_5 &=& 0 \end{array}\right|$$

60. $$\left|\begin{array}{rcl} x_1 &=& 3 \\ x_1-x_2 &=& 1 \\ 3x_1+2x_2+x_3 &=& 0 \\ 3x_1+6x_2+x_3+4x_4-x_5 &=& 1 \\ -6x_2-x_3-4x_4+x_5 &=& 2 \end{array}\right|$$

Infinitely Many Solutions

Display a toolkit sequence for each system, whose final frame has this property: *each* *nonzero equation has a lead variable.* Then apply the **last frame algorithm** to write out the standard general solution of the system. Assume in each system variable list x_1 to x_5.

61. $$\left|\begin{array}{rr} x_1+x_2+3x_3 & =0 \\ x_2 +x_4 & =0 \\ & 0=0 \end{array}\right|$$

62. $$\left|\begin{array}{rr} x_1 +x_3 & =0 \\ x_1+x_2+x_3 +3x_5 & =0 \\ x_4+2x_5 & =0 \end{array}\right|$$

63. $$\left|\begin{array}{rr} x_2+3x_3 & =0 \\ x_4 & =0 \\ & 0=0 \end{array}\right|$$

64. $$\left|\begin{array}{rr} x_1+2x_2+3x_3 & =0 \\ x_4 & =0 \\ & 0=0 \end{array}\right|$$

65. $$\left|\begin{array}{rr} x_1+2x_2+3x_3 & =0 \\ x_3+x_4 & 0=0 \end{array}\right|$$

66. $$\left|\begin{array}{rr} x_1+x_2 & =0 \\ x_2+x_3 & =0 \\ x_3 & 0=1 \end{array}\right|$$

67. $$\left|\begin{array}{rr} x_1+x_2+3x_3+5x_4+2x_5 & =0 \\ x_5 & =0 \end{array}\right|$$

68. $$\left|\begin{array}{rr} x_1+2x_2+x_3+3x_4+4x_5 & =0 \\ x_3+\ \ x_4+\ \ x_5 & =0 \end{array}\right|$$

69. $$\left|\begin{array}{rr} x_3+2x_4+\ \ x_5 & =0 \\ 2x_3+2x_4+2x_5 & =0 \\ x_5 & =0 \end{array}\right|$$

70. $$\left|\begin{array}{rr} x_4+x_5 & =0 \\ 0 & =0 \\ 0 & =0 \\ 0 & =0 \end{array}\right|$$

71. $$\left|\begin{array}{rr} x_2+x_3+5x_4 & =0 \\ x_3+2x_4 & =0 \\ x_5 & =0 \\ 0 & =0 \end{array}\right|$$

72. $$\left|\begin{array}{rr} x_1 +3x_3 & =0 \\ x_1+x_2 +x_4 & =0 \\ x_5 & =0 \\ 0 & =0 \end{array}\right|$$

Inverses of Elementary Operations

Given the final frame of a toolkit sequence is

$$\left|\begin{array}{rcrcrcr} 3x & + & 2y & + & 4z & = & 2 \\ x & + & 3y & + & 2z & = & -1 \\ 2x & + & y & + & 5z & = & 0 \end{array}\right|$$

and the given operations, find the original system in the first frame.

73. `combo(1,2,-1)`, `combo(2,3,-3)`, `mult(1,-2)`, `swap(2,3)`.

74. `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,2)`, `swap(3,2)`.

75. `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,4)`, `swap(1,3)`.

76. `combo(1,2,-1)`, `combo(2,3,4)`, `mult(1,3)`, `swap(3,2)`.

77. `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,4)`, `swap(1,3)`, `swap(2,3)`.

78. `swap(2,3)`, `combo(1,2,-1)`, `combo(2,3,4)`, `mult(1,3)`, `swap(3,2)`.

79. `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,4)`, `swap(1,3)`, `mult(2,3)`.

80. `combo(1,2,-1)`, `combo(2,3,4)`, `mult(1,3)`, `swap(3,2)`, `combo(2,3,-3)`.

,

3.3 General Solution Theory

Consider the nonhomogeneous system

$$(1)\qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

The **general solution** of system (1) is an expression which represents all possible solutions of the system.

The example above for infinitely many solutions contained an unmotivated algorithm which expressed the general solution in terms of invented symbols t_1, t_2, ..., which in mathematical literature are called **parameters**. We outline here some topics from calculus which form the assumed background for this subject.

Equations for Points, Lines and Planes

Background from analytic geometry appears in Table 6. In this table, t_1 and t_2 are **parameters**, which means they are allowed to take on any value between $-\infty$ and $+\infty$. The algebraic equations describing the geometric objects are called **parametric equations**.

Table 6. Parametric Equations with Geometrical Significance.

$x = d_1$, $y = d_2$, $z = d_3$.	**Point**. The equations have no parameters and describe a single point.
$x = d_1 + a_1t_1$, $y = d_2 + a_2t_1$, $z = d_3 + a_3t_1$.	**Line**. The equations with parameter t_1 describe a straight line through (d_1, d_2, d_3) with tangent vector $a_1\vec{\imath} + a_2\vec{\jmath} + a_3\vec{k}$.
$x = d_1 + a_1t_1 + b_1t_2$, $y = d_2 + a_2t_1 + b_2t_2$, $z = d_3 + a_3t_1 + b_3t_2$.	**Plane**. The equations with parameters t_1, t_2 describe a plane containing (d_1, d_2, d_3). The cross product $(a_1\vec{\imath} + a_2\vec{\jmath} + a_3\vec{k}) \times (b_1\vec{\imath} + b_2\vec{\jmath} + b_3\vec{k})$ is normal to the plane.

To illustrate, the parametric equations $x = 2 - 6t_1$, $y = -1 - t_1$, $z = 8t_1$ describe the unique line of intersection of the three planes (details in Example 3.5)

$$(2)\qquad \begin{array}{rcrcrcl} x &+& 2y &+& z &=& 0, \\ 2x &-& 4y &+& z &=& 8, \\ 3x &-& 2y &+& 2z &=& 8. \end{array}$$

General Solutions

Definition 3.2 (Parametric Equations)

Equations of the form

$$
\begin{aligned}
x_1 &= d_1 + c_{11}t_1 + \cdots + c_{1k}t_k, \\
x_2 &= d_2 + c_{21}t_1 + \cdots + c_{2k}t_k, \\
&\vdots \\
x_n &= d_n + c_{n1}t_1 + \cdots + c_{nk}t_k
\end{aligned}
\tag{3}
$$

are called **parametric equations** for the variables x_1, ..., x_n.

The numbers $d_1, \ldots, d_n, c_{11}, \ldots, c_{nk}$ are *known constants* and the symbols $t_1, \ldots, t_k$ are **parameters**, which are treated as variables that may be assigned any value from $-\infty$ to ∞.

Three cases appear often in examples and exercises, illustrated here for variables x_1, x_2, x_3:

No parameters	One parameter	Two parameters
$x_1 = d_1$	$x_1 = d_1 + a_1t_1$	$x_1 = d_1 + a_1t_1 + b_1t_2$
$x_2 = d_2$	$x_2 = d_2 + a_2t_1$	$x_2 = d_2 + a_2t_1 + b_2t_2$
$x_3 = d_3$	$x_3 = d_3 + a_3t_1$	$x_3 = d_3 + a_3t_1 + b_3t_2$

Definition 3.3 (General Solution)

A **general solution** of a linear algebraic system of equations (1) is a set of parametric equations (3) plus two additional requirements:

(4) Equations (3) satisfy (1) for all real values of t_1, ..., t_k.

(5) Any solution of (1) can be obtained from (3) by specializing values of the parameters t_1, t_2, ... t_k.

A general solution is sometimes called a **parametric solution**. Requirement (4) means that **the solution works**. Requirement (5) means that **no solution was skipped**.

Definition 3.4 (Standard General Solution)

Parametric equations (3) are called **standard** if they satisfy for distinct subscripts j_1, i_2, ..., j_k the equations

$$x_{j_1} = t_1, \quad x_{j_2} = t_2, \quad \ldots, \quad x_{j_k} = t_k. \tag{6}$$

The relations mean that the full set of parameter symbols t_1, t_2, ..., t_k were assigned to k distinct variable names (the **free variables**) selected from x_1, ..., x_n.

A **standard general solution** of system (1) is a special set of parametric equations (3) satisfying (4), (5) and additionally (6). Toolkit sequences always produce a standard general solution.

Theorem 3.2 (Standard General Solution)
A standard general solution has the fewest possible parameters and it represents each solution of the linear system by a unique set of parameter values.

The theorem supplies the theoretical basis for the method of toolkit sequences, which formally appears as an algorithm on page 197. The proof of Theorem 3.2 is delayed until page 220. It is unusual if this proof is a subject of a class lecture, due to its length; it is recommended reading for the mathematically inclined, after understanding the examples.

Reduced Echelon System

Consider a toolkit sequence. The last frame, from which we write the general solution, is called a reduced echelon system.

Definition 3.5 (Reduced Echelon System)
A linear system in which each nonzero equation has a **lead variable** is called a **reduced echelon system**. Implicitly assumed are the following definitions and rules.

- A **lead variable** is a variable which appears with coefficient one in the very first location, left to right, in *exactly one* equation.
- A variable not used as a lead variable is called a **free variable**. Variables that do not appear at all are free variables.
- The nonzero equations are listed in variable list order, inherited from their lead variables. Equations without variables are listed last.
- All variables in an equation are required to appear in variable list order. Therefore, within an equation, all free variables are to the right of the lead variable.

Detecting a Reduced Echelon System

A given system can be rapidly inspected to detect if it can be transformed into a reduced echelon system. We assume that within each equation, variables appear in variable list order.

> A nonhomogeneous linear system is recognized as a reduced echelon system when the first variable listed in each equation has coefficient one and that symbol appears nowhere else in the system of equations.[2]

Such a system can be re-written, by swapping equations and enforcing the rules above, so that the resulting system is a reduced echelon system.

[2] Children are better at such classifications than adults. A favorite puzzle among kids is a drawing which contains disguised figures, like a bird, a fire hydrant and Godzilla. Routinely, children find all the disguised figures.

Rank and Nullity

A reduced echelon system splits the variable names x_1, ..., x_n into the **lead variables** and the **free variables**. Because the entire variable list is exhausted by these two sets, then

$$\textbf{lead variables} + \textbf{free variables} = \textbf{total variables}.$$

Definition 3.6 (Rank and Nullity)
The **number of lead variables** in a reduced echelon system is called the **rank** of the system. The number of free variables in a reduced echelon system is called the **nullity** of the system.

Determining rank and nullity

First, display a toolkit sequence which starts with that system and ends in a reduced echelon system. Then the rank and nullity of the system are those determined by the final frame.

Theorem 3.3 (Rank and Nullity)
The following equation holds:

$$\textbf{rank} + \textbf{nullity} = \text{number of variables}.$$

Computers and Reduced Echelon Form

Computer algebra systems and computer numerical laboratories compute from a given linear system (5) a new equivalent system of identical size, which is called the **reduced row-echelon form**, abbreviated **rref**.

The computed **rref** will pass the *last frame test*, provided there is no signal equation, hence the **rref** is generally a reduced echelon system. This fact is the basis of answer checks with computer assist.

Computer assist requires **matrix input** of the data, a topic which is delayed until a later chapter. Popular commercial programs used to perform the computer assist are `maple`, `mathematica` and `matlab`.

Elimination

The elimination algorithm applies at each algebraic step one of the three toolkit rules defined in Table 1: **swap**, **multiply** and **combination**.

The objective of each algebraic step is to **increase the number of lead variables**. Equivalently, each algebraic step tries to **eliminate one repetition of**

a variable name, which justifies calling the process the **method of elimination**. The process of elimination stops when a signal equation (typically $0 = 1$) is found. Otherwise, elimination stops when no more lead variables can be found, and then the last system of equations is a **reduced echelon system**. A detailed explanation of the process has been given above in the discussion of toolkit sequences.

Reversibility of the algebraic steps means that no solutions are created nor destroyed during the algebra: the original system and all intermediate systems have *exactly the same solutions*.

The final reduced echelon system has either a unique solution or infinitely many solutions, in both cases we report the **general solution**. In the infinitely many solution case, the **last frame algorithm** on page 189 is used to write out a general solution.

Theorem 3.4 (Elimination)
Every linear system (5) has either no solution or else it has exactly the same solutions as an equivalent reduced echelon system, obtained by repeated use of toolkit rules **swap**, **multiply** and **combination**, page 176).

An Elimination Algorithm

An equation is said to be **processed** if it has a lead variable. Otherwise, the equation is said to be **unprocessed**.

The acronym `rref` abbreviates the phrase *reduced row echelon form*. This abbreviation appears in matrix literature, so we use it instead of creating an acronym for *reduced echelon form* (the word *row* is missing).

1. If an equation "$0 = 0$" appears, then move it to the end. If a signal equation "$0 = c$" appears ($c \neq 0$ required), then the system is inconsistent. In this case, the algorithm halts and we report **no solution**.
2. Identify the **first symbol** x_r, in variable list order $x_1, \ldots, x_n$, which appears in some unprocessed equation. Apply the **multiply** rule to insure x_r has leading coefficient one. Apply the **combination** rule to eliminate variable x_r from all other equations. Then x_r is a **lead variable**: the number of lead variables has been increased by one.
3. Apply the **swap** rule repeatedly to move this equation past all processed equations, but before the unprocessed equations. Mark the equation as **processed**, e.g., replace x_r by boxed symbol $\boxed{x_r}$.
4. Repeat steps 1–3, until all equations have been processed once. Then lead variables $x_{i_1}, \ldots, x_{i_m}$ have been defined and the last system is a reduced echelon system.

Uniqueness, Lead Variables and RREF

Elimination performed on a given system by two different persons will result in the same reduced echelon system. The answer is unique, because attention has been paid to the natural order $x_1, \ldots, x_n$ of the variable list. Uniqueness results from *critical step* **2**, also called the **rref step**:

> Always select a lead variable as the next possible variable name in the original list order $x_1, \ldots, x_n$, taken from all possible unprocessed equations.

This step insures that the final system is a **reduced echelon system**. Acronym **rref** abbreviates *reduced row echelon form*, where **row** refers to an encoding of one linear algebraic equation.

The wording **next possible** must be used, because once a variable name is used for a lead variable it may not be used again. The next variable following the last–used lead variable, from the list $x_1, \ldots, x_n$, might not appear in any unprocessed equation, in which case it is a **free variable**. The next variable name in the original list order is then tried as a lead variable.

Numerical Optimization

It is common for references to divide the effort for obtaining an **rref** into two stages, for which the second stage is **back-substitution**. This division of effort is motivated by numerical efficiency considerations, largely historical. The reader is advised to adopt the numerical point of view in hand calculations, as soon as possible. It changes the details of a toolkit sequence to the **rref**: most readers find the changes equally advantageous. The reason for the algorithm in the text is motivational: to become an expert, you have to first *know what you are trying to accomplish.* Exactly how to implement the toolkit to arrive at the **rref** will vary for each person. The recommendation can be phrased as follows:

> Don't bother to eliminate a lead variable from equations already assigned a lead variable. Go on to select the next lead variable and remove that variable from subsequent equations. Final elimination of lead variables from previous equations is saved for the end, then done in reverse variable list order (called **back-substitution**).

Avoiding Fractions

Integer arithmetic should be used, when possible, to speed up hand computation in elimination. To avoid fractions, the **rref** step **2** may be modified to read *with leading coefficient nonzero.* The final division to obtain leading coefficient one is then delayed until the last possible moment.

Examples and Methods

Example 3.5 (Line of Intersection)

Show that the parametric equations $x = 2 - 6t$, $y = -1 - t$, $z = 8t$ represent a line through $(2, -1, 0)$ with tangent $-6\vec\imath - \vec\jmath$ which is the line of intersection of the three planes

$$\begin{array}{rcrcrcl} x & + & 2y & + & z & = & 0, \\ 2x & - & 4y & + & z & = & 8, \\ 3x & - & 2y & + & 2z & = & 8. \end{array} \tag{7}$$

Solution: Using $t = 0$ in the parametric solution shows that $(2, -1, 0)$ is on the line. The tangent to the parametric curve is $x'(t)\vec\imath + y'(t)\vec\jmath + z'(t)\vec k$, which computes to $-6\vec\imath - \vec\jmath$. The details for showing the parametric solution satisfies the three equations simultaneously:

$$\begin{aligned}
\text{LHS} &= x + 2y + z && \text{First equation left side.} \\
&= (2 - 6t) + 2(-1 - t) + 8t && \text{Substitute parametric solution.} \\
&= 0 && \text{Matches the RHS in (7).} \\
\text{LHS} &= 2x - 4y + z && \text{Second equation left side.} \\
&= 2(2 - 6t) - 4(-1 - t) + 8t && \text{Substitute.} \\
&= 8 && \text{Matches (7).} \\
\text{LHS} &= 3x - 2y + 2z && \text{Third equation left side.} \\
&= 3(2 - 6t) - 2(-1 - t) + 16t && \text{Substitute.} \\
&= 8 && \text{Matches (7).}
\end{aligned}$$

Example 3.6 (Geometry of Solutions)

Solve the system and interpret the solution geometrically.

$$\begin{array}{rcrcl} x & + & 2z & = & 3, \\ y & + & z & = & 1. \end{array}$$

Solution: We begin by displaying the general solution, which is a **line**:

$$\begin{array}{rcll} x & = & 3 - 2t_1, & \\ y & = & 1 - t_1, & \\ z & = & t_1, & \qquad -\infty < t_1 < \infty. \end{array}$$

In standard xyz-coordinates, this line passes through $(3, 1, 0)$ with tangent direction $-2\vec\imath - \vec\jmath + \vec k$.

Details. To justify this solution, we observe that the first frame equals the last frame, which is a reduced echelon system in variable list order x, y, z. The standard general solution will be obtained from the last frame algorithm.

$$\boxed{\begin{array}{rcrcl} x & + & 2z & = & 3, \\ y & + & z & = & 1. \end{array}}$$

Frame 1 equals the last frame, a reduced echelon system The lead variables are x, y and the free variable is z.

$$\boxed{\begin{array}{rclcr} x & = & 3 & - & 2z, \\ y & = & 1 & - & z, \\ z & = & t_1. & & \end{array}}$$

Assign to z invented symbol t_1. Solve for lead variables x and y in terms of the free variable z.

$$\boxed{\begin{array}{rcrcr} x & = & 3 & - & 2t_1, \\ y & = & 1 & - & t_1, \\ z & = & t_1. & & \end{array}}$$

Back-substitute for free variable z. This is the standard general solution. It is geometrically a line, by Table 6.

Example 3.7 (Symbolic Answer Check)

Perform an answer check on

$$\begin{array}{rcrcr} x & + & 2z & = & 3, \\ y & + & z & = & 1, \end{array}$$

for the general solution

$$\begin{array}{rcl} x & = & 3 - 2t_1, \\ y & = & 1 - t_1, \\ z & = & t_1, \qquad -\infty < t_1 < \infty. \end{array}$$

Solution: The displayed answer can be checked manually by substituting the symbolic general solution into the equations $x + 2z = 3$, $y + z = 1$, as follows:

$$\begin{array}{rcl} x + 2z & = & (3 - 2t_1) + 2(t_1) \\ & = & 3, \\ y + z & = & (1 - t_1) + (t_1) \\ & = & 1. \end{array}$$

Therefore, the two equations are satisfied for all values of the symbol t_1.

Errors and Skipped Solutions. An algebraic error could lead to a claimed solution $x = 3$, $y = 1$, $z = 0$, which also passes the answer check. While it is *true* that $x = 3$, $y = 1$, $z = 0$ is a solution, it is not the general solution. Infinitely many solutions were skipped in the answer check.

General Solution and Free Variables. The number of lead variables is called the **rank**. The number of free variables is called the **nullity**. The basic relation is **rank + nullity = number of variables**. Computer algebra systems can compute the rank independently, as a double-check against hand computation. This check is useful for discovering skipped solution errors. The **rank** is unaffected by the ordering of variables.

Example 3.8 (Elimination)

Solve the system.

$$\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 1, \\ w & + & 3x & - & y & + & 2z & = & 0, \\ & & x & & & + & z & = & -1. \end{array}$$

Solution: The answer using the natural variable list order w, x, y, z is the standard general solution

$$\begin{array}{rcl} w & = & 3 + t_1 + t_2, \\ x & = & -1 - t_2, \\ y & = & t_1, \\ z & = & t_2, \qquad -\infty < t_1, t_2 < \infty. \end{array}$$

Details. Elimination will be applied to obtain a toolkit sequence whose last frame justifies the reported solution. The details amount to applying the three rules **swap**, **multiply** and **combination** for equivalent equations on page 176 to obtain a last frame which is a reduced echelon system. The standard general solution from the last frame algorithm matches the one reported above.

Let's mark processed equations with a box enclosing the lead variable (w is marked $\boxed{w}$).

$$\boxed{\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 1 \\ w & + & 3x & - & y & + & 2z & = & 0 \\ & & x & & & + & z & = & -1 \end{array}} \quad \boxed{\mathbf{1}}$$

$$\boxed{\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 1 \\ 0 & + & x & + & 0 & + & z & = & -1 \\ & & x & & & + & z & = & -1 \end{array}} \quad \boxed{\mathbf{2}}$$

$$\boxed{\begin{array}{rcrcrcrcr} \boxed{w} & + & 2x & - & y & + & z & = & 1 \\ & & x & & & + & z & = & -1 \\ & & & & & & 0 & = & 0 \end{array}} \quad \boxed{\mathbf{3}}$$

$$\boxed{\begin{array}{rcrcrcrcr} \boxed{w} & + & 0 & - & y & - & z & = & 3 \\ & & \boxed{x} & & & + & z & = & -1 \\ & & & & & & 0 & = & 0 \end{array}} \quad \boxed{\mathbf{4}}$$

$\boxed{\mathbf{1}}$ Original system. Identify the variable order as w, x, y, z.

$\boxed{\mathbf{2}}$ Choose w as a lead variable. Eliminate w from equation 2 by using `combo(1,2,-1)`.

$\boxed{\mathbf{3}}$ The w-equation is processed. Let x be the next lead variable. Eliminate x from equation 3 using `combo(2,3,-1)`.

$\boxed{\mathbf{4}}$ Eliminate x from equation 1 using `combo(2,1,-2)`. Mark the x-equation as processed. **Reduced echelon system** found.

The four frames make the **toolkit sequence** which takes the original system into a reduced echelon system. Basic exposition rules apply:

1. Variables in an equation appear in variable list order.
2. Equations inherit variable list order from the lead variables.

The last frame of the sequence, which must be a reduced echelon system, is used to write out the general solution, using the last frame algorithm.

$$\boxed{\begin{array}{rcrcrcr} \boxed{w} & = & 3 & + & y & + & z \\ \boxed{x} & = & -1 & - & z & & \\ y & = & t_1 & & & & \\ z & = & t_2 & & & & \end{array}}$$

Solve for the lead variables $\boxed{w}$, $\boxed{x}$. Assign invented symbols t_1, t_2 to the free variables y, z.

$$\boxed{\begin{array}{rcrcrcr} w & = & 3 & + & t_1 & + & t_2 \\ x & = & -1 & - & t_2 & & \\ y & = & t_1 & & & & \\ z & = & t_2 & & & & \end{array}}$$

Back-substitute free variables into the lead variable equations to get a standard general solution.

Answer check. The check will be performed according to the outline on page 218. The justification for this forward reference is to illustrate how to check answers without using the invented symbols t_1, t_2, ... in the details.

Step 1. The **nonhomogeneous trial solution** $w = 3$, $x = -1$, $y = z = 0$ is obtained by setting $t_1 = t_2 = 0$. It is required to satisfy the nonhomogeneous system

$$\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 1, \\ w & + & 3x & - & y & + & 2z & = & 0, \\ & & x & & & + & z & = & -1. \end{array}$$

Step 2. The partial derivatives ∂_{t_1}, ∂_{t_2} are applied to the parametric solution to obtain two homogeneous trial solutions $w = 1$, $x = 0$, $y = 1$, $z = 0$ and $w = 1$, $x = -1$, $y = 0$, $z = 1$, which are required to satisfy the homogeneous system

$$\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 0, \\ w & + & 3x & - & y & + & 2z & = & 0, \\ & & x & & & + & z & = & 0. \end{array}$$

Each trial solution from **Step 1** and **Step 2** is checked by direct substitution. The method uses **superposition** in order to eliminate the invented symbols from the answer check.

Example 3.9 (No solution)

Verify by applying elimination that the system has no solution.

$$\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 0, \\ w & + & 3x & - & y & + & 2z & = & 0, \\ & & x & & & + & z & = & 1. \end{array}$$

Solution: Elimination (page 198) will be applied, using the toolkit rules **swap**, **multiply** and **combination** (page 176).

$$\boxed{\begin{array}{rcrcrcrcr} w & + & 2x & - & y & + & z & = & 0 \\ w & + & 3x & - & y & + & 2z & = & 0 \\ & & x & & & + & z & = & 1 \end{array}} \qquad \boxed{\mathbf{1}}$$

$$\boxed{\begin{array}{rcrcrcrcr} \boxed{w} & + & 2x & - & y & + & z & = & 0 \\ 0 & + & x & + & 0 & + & z & = & 0 \\ & & x & & & + & z & = & 1 \end{array}} \qquad \boxed{\mathbf{2}}$$

$$\boxed{\begin{array}{rcrcrcrcr} \boxed{w} & + & 2x & - & y & + & z & = & 0 \\ & & x & & & + & z & = & 0 \\ & & & & & & 0 & = & 1 \end{array}} \qquad \boxed{\mathbf{3}}$$

$\boxed{\mathbf{1}}$ Original system. Select variable order w, x, y, z. Identify lead variable w.

$\boxed{\mathbf{2}}$ Eliminate w from other equations using `combo(1,2,-1)`. Mark the w-equation processed with $\boxed{w}$.

$\boxed{\mathbf{3}}$ Identify lead variable x. Then eliminate x from the third equation using operation `combo(2,3,-1)`. **Signal equation** found.

The appearance of the signal equation "$0 = 1$" means **no solution**. The logic: if the original system has a solution, then so does the present equivalent system, hence $0 = 1$, a contradiction. Elimination halts, because of the **inconsistent system** containing the false equation "$0 = 1$."

Example 3.10 (Reduced Echelon form)

Find an equivalent system in reduced echelon form.

$$\begin{array}{rcrcrcrcr} x_1 & + & 2x_2 & - & x_3 & + & x_4 & = & 1, \\ x_1 & + & 3x_2 & - & x_3 & + & 2x_4 & = & 0, \\ & & x_2 & & & + & x_4 & = & -1. \end{array}$$

Solution: The answer using the natural variable list order x_1, x_2, x_2, x_4 is the non-homogeneous system in **reduced echelon form** (briefly, **rref** form)

$$\begin{array}{rcrcrcrcr} x_1 & & & - & x_3 & - & x_4 & = & 3 \\ & & x_2 & & & + & x_4 & = & -1 \\ & & & & & & 0 & = & 0 \end{array}$$

The **lead variables** are x_1, x_2 and the **free variables** are x_3, x_4. The standard general solution of this system is

$$\begin{array}{rcll} x_1 & = & 3 + t_1 + t_2, & \\ x_2 & = & -1 - t_2, & \\ x_3 & = & t_1, & \\ x_4 & = & t_2, & \qquad -\infty < t_1, t_2 < \infty. \end{array}$$

The details are the same as Example 3.8, with $w = x_1$, $x = x_2$, $y = x_3$, $z = x_4$. The toolkit sequence has three frames and the last frame is used to display the general solution.

Answer check in `maple`. The output from the `maple` code below duplicates the reduced echelon system reported above and the general solution.

```
with(LinearAlgebra):
 eq1:=x[1]+2*x[2]-x[3]+x[4]=1:eq2:=x[1]+3*x[2]-x[3]+2*x[4]=0:
 eq3:=x[2]+x[4]=-1:eqs:=[eq1,eq2,eq3]:var:=[x[1],x[2],x[3],x[4]]:
 A:=GenerateMatrix(eqs,var,augmented);
 F:=ReducedRowEchelonForm(A);
 GenerateEquations(F,var);
 F,LinearSolve(F,free=t); # general solution answer check
 A,LinearSolve(A,free=t); # general solution answer check
```

,

Exercises 3.3

Classification

Classify the parametric equations as a point, line or plane, then compute as appropriate the tangent to the line or the normal to the plane.

1. $x = 0,\ y = 1,\ z = -2$

2. $x = 1,\ y = -1,\ z = 2$

3. $x = t_1,\ y = 1 + t_1,\ z = 0$

4. $x = 0,\ y = 0,\ z = 1 + t_1$

5. $x = 1 + t_1,\ y = 0,\ z = t_2$

6. $x = t_2 + t_1,\ y = t_2,\ z = t_1$

7. $x = 1,\ y = 1 + t_1,\ z = 1 + t_2$

8. $x = t_2 + t_1,\ y = t_1 - t_2,\ z = 0$

9. $x = t_2,\ y = 1 + t_1,\ z = t_1 + t_2$

10. $x = 3t_2 + t_1,\ y = t_1 - t_2,\ z = 2t_1$

Reduced Echelon System

Solve the xyz–system and interpret the solution geometrically.

11. $\left|\begin{array}{rcrcl} y &+& z &=& 1\\ x &+& 2z &=& 2 \end{array}\right|$

12. $\left|\begin{array}{rcrcl} x &+& z &=& 1\\ y &+& 2z &=& 4 \end{array}\right|$

13. $\left|\begin{array}{rcrcl} y &+& z &=& 1\\ x &+& 3z &=& 2 \end{array}\right|$

14. $\left|\begin{array}{rcrcl} x &+& z &=& 1\\ y &+& z &=& 5 \end{array}\right|$

15. $\left|\begin{array}{rcrcl} x &+& z &=& 1\\ 2x &+& 2z &=& 2 \end{array}\right|$

16. $\left|\begin{array}{rcrcl} x &+& y &=& 1\\ 3x &+& 3y &=& 3 \end{array}\right|$

17. $\left|\ x + y + z = 1.\ \right|$

18. $\left|\ x + 2y + 4z = 0.\ \right|$

19. $\left|\begin{array}{rcrcrcl} x &+& y && &=& 2\\ && && z &=& 1 \end{array}\right|$

20. $\left|\begin{array}{rcrcrcl} x && &+& 4z &=& 0\\ && y && &=& 1 \end{array}\right|$

Homogeneous System

Solve the xyz–system using elimination with variable list order x, y, z.

21. $\left|\begin{array}{rcrcrcl} && y &+& z &=& 0\\ 2x && &+& 2z &=& 0 \end{array}\right|$

22. $\left|\begin{array}{rcrcrcl} x && &+& z &=& 0\\ && 2y &+& 2z &=& 0 \end{array}\right|$

23. $\left|\begin{array}{rcrcl} x &+& z &=& 0\\ && 2z &=& 0 \end{array}\right|$

24. $\left|\begin{array}{rcrcl} y &+& z &=& 0\\ y &+& 3z &=& 0 \end{array}\right|$

25. $\left|\begin{array}{rcrcrcl} x &+& 2y &+& 3z &=& 0\\ && && 0 &=& 0 \end{array}\right|$

26. $\left|\begin{array}{rcrcl} x &+& 2y &=& 0\\ && 0 &=& 0 \end{array}\right|$

27. $\left|\begin{array}{rcrcrcl} && y &+& z &=& 0\\ 2x && &+& 2z &=& 0\\ x && &+& z &=& 0 \end{array}\right|$

28. $\left|\begin{array}{rcrcrcl} 2x &+& y &+& z &=& 0\\ x && &+& 2z &=& 0\\ x &+& y &-& z &=& 0 \end{array}\right|$

29. $\left|\begin{array}{rcrcrcl} x &+& y &+& z &=& 0\\ 2x && &+& 2z &=& 0\\ x && &+& z &=& 0 \end{array}\right|$

30. $\left|\begin{array}{rcrcrcl} x &+& y &+& z &=& 0\\ 2x && &+& 2z &=& 0\\ 3x &+& y &+& 3z &=& 0 \end{array}\right|$

Nonhomogeneous 3×3 System

Solve the xyz-system using elimination and variable list order x, y, z.

31. $\left|\begin{array}{rcl} y &=& 1\\ 2z &=& 2 \end{array}\right|$

32. $\left|\begin{array}{rcl} x &=& 1\\ 2z &=& 2 \end{array}\right|$

33. $\left|\begin{array}{rcrcrcr} & & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ x & & & + & z & = & 1 \end{array}\right|$

34. $\left|\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 1 \\ x & & & + & 2z & = & 2 \\ x & + & y & - & z & = & -1 \end{array}\right|$

35. $\left|\begin{array}{rcrcrcr} x & + & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ x & & & + & z & = & 1 \end{array}\right|$

36. $\left|\begin{array}{rcrcrcr} x & + & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ 3x & + & y & + & 3z & = & 3 \end{array}\right|$

37. $\left|\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 3 \\ 2x & & & + & 2z & = & 2 \\ 4x & + & y & + & 3z & = & 5 \end{array}\right|$

38. $\left|\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 2 \\ 6x & & y & + & 5z & = & 2 \\ 4x & + & y & + & 3z & = & 2 \end{array}\right|$

39. $\left|\begin{array}{rcrcrcr} 6x & + & 2y & + & 6z & = & 10 \\ 6x & & y & + & 6z & = & 11 \\ 4x & + & y & + & 4z & = & 7 \end{array}\right|$

40. $\left|\begin{array}{rcrcrcr} 6x & + & 2y & + & 4z & = & 6 \\ 6x & & y & + & 5z & = & 9 \\ 4x & + & y & + & 3z & = & 5 \end{array}\right|$

Nonhomogeneous 3×4 System

Solve the $yzuv$-system using elimination with variable list order y, z, u, v.

41. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 8v & = & 10 \\ & & 2z & - & u & + & v & = & 10 \\ 2y & & & - & u & + & 5v & = & 10 \end{array}\right|$

42. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 8v & = & 10 \\ & & 2z & - & 2u & + & 2v & = & 0 \\ y & + & 3z & + & 2u & + & 5v & = & 5 \end{array}\right|$

43. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 8v & = & 1 \\ & & 2z & - & 2u & + & 4v & = & 0 \\ y & + & 3z & + & 2u & + & 6v & = & 1 \end{array}\right|$

44. $\left|\begin{array}{rcrcrcrcr} y & + & 3z & + & 4u & + & 8v & = & 1 \\ & & 2z & - & 2u & + & 4v & = & 0 \\ y & + & 3z & + & 2u & + & 6v & = & 1 \end{array}\right|$

45. $\left|\begin{array}{rcrcrcrcr} y & + & 3z & + & 4u & + & 8v & = & 1 \\ & & 2z & - & 2u & + & 4v & = & 0 \\ y & + & 4z & + & 2u & + & 7v & = & 1 \end{array}\right|$

46. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 9v & = & 1 \\ & & 2z & - & 2u & + & 4v & = & 0 \\ y & + & 4z & + & 2u & + & 7v & = & 1 \end{array}\right|$

47. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 9v & = & 1 \\ & & 2z & - & 2u & + & 4v & = & 0 \\ y & + & 4z & + & 2u & + & 7v & = & 1 \end{array}\right|$

48. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 9v & = & 10 \\ & & 2z & - & 2u & + & 4v & = & 4 \\ y & + & 4z & + & 2u & + & 7v & = & 8 \end{array}\right|$

49. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 9v & = & 2 \\ & & 2z & - & 2u & + & 4v & = & 4 \\ y & + & 3z & + & 5u & + & 13v & = & 0 \end{array}\right|$

50. $\left|\begin{array}{rcrcrcrcr} y & + & z & + & 4u & + & 3v & = & 2 \\ & & 2z & - & 2u & + & 4v & = & 4 \\ y & + & 3z & + & 5u & + & 7v & = & 0 \end{array}\right|$

,

3.4 Basis, Dimension, Nullity and Rank

Studied here are the basic concepts of rank, nullity, basis and dimension for a system of linear algebraic equations.

Definition 3.7 (Rank and Nullity)
The **rank** of a system of linear algebraic equations is the number of lead variables appearing in its reduced echelon form. The **nullity** of a system of linear algebraic equations is the number of free variables.

> **rank** = number of lead variables
>
> **nullity** = number of free variables
>
> **rank + nullity = number of variables**

Definition 3.8 (Basis and Dimension)
Consider a homogeneous system of linear algebraic equations. A list of k solutions of the system is called a **basis** provided

1. The general solution of the system can be constructed from the list of k solutions.

2. The list size k cannot be decreased.

The **dimension** of the system of linear algebraic equations is the unique number k satisfying **1** and **2**. The dimension equals the minimum number of invented symbols used in any general solution, which also equals the nullity.

> A **basis** is an alternate representation of the general solution which has **no invented symbols**.

Basis Illustration

Consider the homogeneous system

$$\begin{aligned} x+2y+3z &= 0, \\ 0 &= 0, \\ 0 &= 0. \end{aligned}$$

It is a reduced echelon system with standard general solution

$$\begin{aligned} x &= -2t_1 - 3t_2, \\ y &= t_1, \\ z &= t_2. \end{aligned}$$

The formal partial derivatives ∂_{t_1}, ∂_{t_2} of the general solution are solutions of the homogeneous system, because they correspond exactly to setting $t_1 = 1$, $t_2 = 0$ and $t_1 = 0$, $t_2 = 1$, respectively:

$$\begin{array}{llll} x = -2, & y = 1, & z = 0, & (\text{partial on } t_1) \\ x = -3, & y = 0, & z = 1. & (\text{partial on } t_2) \end{array}$$

A **basis** for the homogeneous system is the list of two solutions displayed above. Calculus courses might write the two solutions as space vectors: $-2\vec{\imath} + \vec{\jmath}$ and $-3\vec{\imath} + \vec{k}$. See page 210 for more details.

A general solution of the homogeneous system can be re-constructed from this basis by multiplying the first solution by invented symbol t_1 and the second solution by invented symbol t_2, then add to obtain

$$\begin{aligned} x &= -2t_1 - 3t_2, \\ y &= t_1, \\ z &= t_2. \end{aligned}$$

This display is the original standard general solution, reconstructed from the list of solutions in the basis.

Non-uniqueness of a Basis

A given homogeneous linear system has a number of different standard general solutions, obtained, for example, by re-ordering the variable list. Therefore, a *basis is not unique*. Language like *the basis* is tragically incorrect.

To illustrate non-uniqueness, consider the homogeneous 3×3 system of equations

$$\begin{array}{rcl} x + y + z &=& 0, \\ 0 &=& 0, \\ 0 &=& 0. \end{array} \tag{1}$$

Equations (1) have two standard general solutions

$x = -t_1 - t_2$, $y = t_1$, $z = t_2$
and
$x = t_3$, $y = -t_3 - t_4$, $z = t_4$,

corresponding to two different orderings of the variable list x, y, z. Then **two different bases** for the system are given by the partial derivative relations

$$\partial_{t_1}, \partial_{t_2} : \quad \left\{ \begin{array}{llll} x = -1, & y = 1, & z = 0, & \text{Basis 1}, \\ x = -1, & y = 0, & z = 1, & \end{array} \right. \tag{2}$$

$$(3) \qquad \partial_{t_3}, \partial_{t_4} : \quad \begin{cases} x = 1, \quad y = -1, \quad z = 0, \quad \text{Basis 2}, \\ x = 0, \quad y = -1, \quad z = 1. \end{cases}$$

In general, there are *infinitely many bases* possible for a given linear homogeneous system.

Nullspace

Definition 3.9 (Nullspace)
Consider a system of linear homogeneous algebraic equations. The term **nullspace** refers to the set of all solutions to the system. The origin of the word **nullspace** is explained below.

Prefix null refers to the right side of the homogeneous system, which is zero, or *null*, for each equation. The main reason for introducing the term **nullspace** is to consider simultaneously *all possible* general solutions of the linear system, without regard to their representation in terms of invented symbols or the algorithm used to find the formulas.

Suffix space used in the term **nullspace** has meaning taken from the phrases **storage space** and **parking space** — it has no intended geometrical meaning whatsoever.

How to Find the Nullspace

A classical method for describing the nullspace is to form a toolkit sequence for the homogeneous system which ends with a reduced echelon system. The last frame algorithm applies to write the general solution in terms of invented symbols t_1, t_2, ... The meaning is that assignment of values to the symbols t_1, t_2, ... lists all possible solutions of the system. The general solution formula obtained by this method is one possible set of scalar equations that completely describes all solutions of the homogeneous equation, hence it describes completely the nullspace.

Basis for the Nullspace

A **basis** for the nullspace is found partial derivatives ∂_{t_1}, ∂_{t_2}, ... taken on the last frame algorithm general solution, giving k solutions. The general solution is reconstructed from these basis elements by multiplying them by the symbols t_1, t_2, ... and adding.

Common practise, an abuse of language, reports the answer for the problem **find the nullspace** as equations for variables $x_1, \ldots, x_n$ in terms of invented symbols. No such answer is a **set**: the equations are **not** the nullspace: they are an **algebraic representation** of the set of solutions to the homogeneous equation.

Geometers think of nullspace as an **object** like the plane or space.

Algebraists think of nullspace as a **set** consisting of value lists $x_1, \ldots, x_n$ that satisfy the homogeneous equation. There are no equal signs, no equations, no invented symbols. And no solutions are skipped!

Is there more than one answer for the *nullspace*? Technically **no**. By definition, the *nullspace* is a **set** of elements and it might be a geometric **object**.

An Illustration

Consider the system

$$\begin{array}{rcl} x+y+2z &=& 0, \\ 0 &=& 0, \\ 0 &=& 0. \end{array} \tag{4}$$

The nullspace is the set of all solutions of $x+y+2z=0$. Geometrically, it is the plane $x+y+2z=0$ through $x=y=z=0$ with normal vector $\vec{\imath}+\vec{\jmath}+2\vec{k}$. The nullspace has one possible algebraic representation given by the general solution formula

$$\begin{array}{rcl} x &=& -t_1-2t_2, \\ y &=& t_1, \\ z &=& t_2. \end{array}$$

There are infinitely many representations possible, e.g., replace t_1 by mt_1 where m is any nonzero integer.

The nullspace can be described geometrically as the plane generated by the basis

$$\begin{array}{c} x=-1, y=1, z=0, \\ x=-2, y=0, z=1. \end{array}$$

The basis elements are identified with points $(-1,1,0)$ and $(-2,0,1)$. Physics associates two free vectors with tail at $(0,0,0)$ and heads at $(-1,1,0)$ and $(-2,0,1)$, . Calculus courses represent the two basis elements as vectors $\vec{\mathbf{a}} = -\vec{\imath}+\vec{\jmath}$, $\vec{\mathbf{b}} = -2\vec{\imath}+\vec{k}$, which are two vectors in the plane $x+y+2z=0$. Their cross product $\vec{\mathbf{a}} \times \vec{\mathbf{b}}$ is normal to the plane, a multiple of normal vector $\vec{\imath}+\vec{\jmath}+2\vec{k}$ to the plane $x+y+2z=0$.

The Three Possibilities Revisited

We intend to justify the table below, which summarizes the three possibilities for a linear system, in terms of free variables, rank and nullity.

Table 7. Three Possibilities for an $m \times n$ Linear Algebraic System.

No solution	Signal equation	
∞-many solutions	One+ free variables	nullity ≥ 1 or rank $< n$
Unique solution	Zero free variables	nullity $= 0$ or rank $= n$

No Solution

There is no solution to a system of equations exactly when a signal equation $0 = 1$ occurs during the application of swap, multiply and combination rules. We report the system **inconsistent** and announce **no solution**.

Infinitely Many Solutions

The situation of infinitely many solutions occurs when there is no signal equation and **at least one free variable** to which an invented symbol, say t_1, is assigned. Since this symbol takes the values $-\infty < t_1 < \infty$, there are an infinity of solutions. The conditions **rank less than n** and **nullity positive** are the same.

Unique Solution

There is a unique solution to a consistent system of equations exactly when **zero free variables** are present. This is identical to requiring that the number n of variables equal the number of lead variables, or **rank = n**.

Existence of Infinitely Many Solutions

Homogeneous systems are always consistent[3], therefore if the number of variables exceeds the number of equations, then the equation **lead+free = variable count** implies there is always one free variable. This proves the following basic result of linear algebra.

Theorem 3.5 (Infinitely Many Solutions)
A system of $m \times n$ linear homogeneous equations (6) with fewer equations than unknowns ($m < n$) has at least one free variable, hence an infinite number of solutions. Therefore, such a system always has the **zero solution** and also a **nonzero solution**.

Non-homogeneous systems can be similarly analyzed by considering conditions under which there will be at least one free variable.

[3] All variables set to zero is always a solution of a homogeneous system.

Theorem 3.6 (Missing Variable and Infinitely Many Solutions)
A consistent system of $m \times n$ linear equations with one unknown missing has at least one free variable, hence an infinite number of solutions.

Theorem 3.7 (Rank, Nullity and Infinitely Many Solutions)
A consistent system of $m \times n$ linear equations with nonzero nullity or rank less than n has at least one free variable, hence an infinite number of solutions.

Examples and Methods

Example 3.11 (Rank and Nullity)
Determine using an abbreviated sequence of toolkit operations the rank and nullity of the homogeneous system

$$\begin{array}{rrrr} x_1 & + 4x_3 & + 8x_4 & = 0 \\ & - x_3 & + x_4 & = 0 \\ 2x_1 & - x_3 & + 5x_4 & = 0 \end{array}$$

Solution: The answer is three (3) lead variables and one (1) free variable, making rank=3 and nullity=1.

The missing variable x_2 implies that there is at least one free variable. The abbreviated steps are

$$\boxed{\begin{array}{rrrr} x_1 & + 4x_3 & + 8x_4 & = 0 \\ & - x_3 & + x_4 & = 0 \\ & - 9x_3 & - 11x_4 & = 0 \end{array}}$$

`combo(1,3,-2)`

$$\boxed{\begin{array}{rrrr} x_1 & + 4x_3 & + 8x_4 & = 0 \\ & - x_3 & + x_4 & = 0 \\ & & - 20x_4 & = 0 \end{array}}$$

`combo(2,3,-9)`

The triangular form implies that x_1, x_3, x_4 are lead variables and x_2 is a free variable.

Example 3.12 (Nullspace Basis or Kernel Basis)
Determine a nullspace basis by solving for the general solution of the homogeneous system

$$\begin{array}{rrrrr} x_1 & + x_2 & + 4x_3 & + 9x_4 & = 0 \\ & 2x_2 & - x_3 & + 4x_4 & = 0 \end{array}$$

Solution:

$$\boxed{\begin{array}{rrrrr} x_1 & + x_2 & + 4x_3 & + 9x_4 & = 0 \\ & 2x_2 & - x_3 & + 4x_4 & = 0 \end{array}}$$

Original system.

$$\boxed{\begin{array}{rrrrr} x_1 & + x_2 & + 4x_3 & + 9x_4 & = 0 \\ & x_2 & - \frac{1}{2}x_3 & + 2x_4 & = 0 \end{array}}$$

`mult(2,1/2)`

$$\boxed{\begin{array}{rcrcrl} x_1 & & + & \frac{9}{2}x_3 & + & 7x_4 = 0 \\ & x_2 & - & \frac{1}{2}x_3 & + & 2x_4 = 0 \end{array}}$$

`combo(2,1,-1)`

The lead variables are x_1, x_2 and the free variables are $x_3 = t_1, x_4 = t_2$ in terms of invented symbols t_1, t_2. Back-substitution implies the scalar general solution

$$\begin{array}{rcl} x_1 & = & -\frac{9}{2}t_1 - 7t_2, \\ x_2 & = & \frac{1}{2}t_1 - 2t_2, \\ x_3 & = & t_1, \\ x_4 & = & t_2. \end{array} \tag{5}$$

A suitable basis for the **nullspace**, also called the **kernel**, is found by substitution of $t_1 = 1, t_2 = 0$ and then $t_1 = 0, t_2 = 1$, to obtain the two vectors

Basis solution 1	Basis solution 2
$x_1 = -\frac{9}{2},$	$x_1 = -7,$
$x_2 = \frac{1}{2},$	$x_2 = -2,$
$x_3 = 1,$	$x_3 = 0,$
$x_4 = 0.$	$x_4 = 1.$

These two solutions are identical to the two solutions obtained by taking partial derivatives ∂_{t_1} and ∂_{t_2} on the scalar general solution displayed in equation (5).

Some references suggest to make the two basis answers fraction-free by choosing t_1, t_2 appropriately. In the present case, this amounts to multiplying the answers by 2. The result is a different basis.

Either answer is sufficient, because a basis is not unique: the only requirement is reconstruction of the general solution from the basis.

Example 3.13 (Three Possibilities with Symbol k)

Determine all values of the symbol k such that the system below has one of the **Three Possibilities** (1) *No solution*, (2) *Infinitely many solutions* or (3) *A unique solution*. Display all solutions found.

$$\begin{array}{rcrcl} x & + & ky & = & 2, \\ (2-k)x & + & y & = & 3. \end{array}$$

Solution: The Three Possibilities are detected by (1) A signal equation "0 = 1," (2) One or more free variables, (3) Zero free variables.

The solution of this problem involves construction of perhaps three toolkit sequences, the last frame of each resulting in one of the three possibilities (1), (2), (3).

$$\boxed{\begin{array}{rcrcr} x & + & ky & = & 2, \\ (2-k)x & + & y & = & 3. \end{array}}$$

Frame 1.

Original system.

$$\boxed{\begin{array}{rcrcr} x & + & ky & = & 2, \\ & & [1+k(k-2)]y & = & 2(k-2)+3. \end{array}}$$

Frame 2.

`combo(1,2,k-2)`

$$\boxed{\begin{array}{rcrcr} x & + & ky & = & 2, \\ & & (k-1)^2 y & = & 2k-1. \end{array}}$$

Frame 3.

Simplify.

The three expected toolkit sequences share these initial frames. At this point, we identify the values of k that split off into the three possibilities.

There will be a signal equation if the second equation of Frame 3 has no variables, but the resulting equation is not "$0 = 0$." This happens exactly for $k = 1$. The resulting signal equation is "$0 = 1$." We conclude that one of the three toolkit sequences terminates with the *no solution case*. This toolkit sequence corresponds to $k = 1$.

Otherwise, $k \neq 1$. For these values of k, there are zero free variables, which implies a unique solution. A by-product of the analysis is that the *infinitely many solutions* case never occurs!

The conclusion: The initially expected three toolkit sequences reduce to two toolkit sequences. One sequence gives no solution and the other sequence gives a unique solution.

The three answers:

(1) No solution occurs only for $k = 1$.

(2) Infinitely many solutions occurs for no value of k.

(3) A unique solution occurs for $k \neq 1$.

$$\begin{aligned} x &= 2 - \frac{k(2k-1)}{(k-1)^2}, \\ y &= \frac{(2k-1)}{(k-1)^2}. \end{aligned}$$

Example 3.14 (Symbols and the Three Possibilities)

Determine all values of the symbols a, b such that the system below has (1) No solution, (2) Infinitely many solutions or (3) A unique solution. Display all solutions found.

$$\begin{array}{rcrcrcr} x & + & ay & + & bz & = & 2, \\ & & y & + & z & = & 3, \\ & & by & + & z & = & 3b. \end{array}$$

Solution: The plan is to make three toolkit sequences, using swap, multiply and combination rules. Each sequence has last frame which is one of the three possibilities, the detection facilitated by (1) A signal equation "$0 = 1$," (2) At least one free variable, (3) Zero free variables. The initial three frames of each of the expected toolkit sequences is constructed as follows.

$$\boxed{\begin{array}{rcrcrcr} x & + & ay & + & bz & = & 2, \\ & & y & + & z & = & 3, \\ & & by & + & z & = & 3b. \end{array}}$$

Frame 1
Original system.

$$\boxed{\begin{array}{rcrcrcr} x & + & ay & + & bz & = & 2, \\ & & y & + & z & = & 3, \\ & & 0 & + & (1-b)z & = & 0. \end{array}}$$

Frame 2.

`combo(2,3,-b)`

$$\boxed{\begin{array}{ccccccc} x & + & 0 & + & (b-a)z & = & 2-3a, \\ & & y & + & z & = & 3, \\ & & 0 & + & (1-b)z & = & 0. \end{array}}$$

Frame 3. `combo(2,1,-a)`
Triangular form.
Lead variables determined.

The three toolkit sequences expected will share these initial frames. Frame 3 shows that there are either 2 lead variables or 3 lead variables, accordingly as the coefficient of z in the third equation is nonzero or zero. There will never be a signal equation. Consequently, the three expected toolkit sequences reduce to just two. We complete these two sequences to give the answer:

(1) There are no values of a, b that result in no solution.

(2) If $1-b=0$, then there are two lead variables and hence an infinite number of solutions, given by the general solution

$$\begin{cases} x = 2-3a-(b-a)t_1, \\ y = 3-t_1, \\ z = t_1. \end{cases}$$

(3) If $1-b \neq 0$, then there are three lead variables and there is a unique solution, given by

$$\begin{cases} x = 2-3a, \\ y = 3, \\ z = 0. \end{cases}$$

,

Exercises 3.4

Rank and Nullity

Compute an abbreviated sequence of `combo, swap, mult` steps which finds the value of the rank and nullity.

1. $\left|\begin{array}{ccccccc} x_1 & + & x_2 & + & 4x_3 & + & 8x_4 = 0 \\ & & 2x_2 & - & x_3 & + & x_4 = 0 \end{array}\right|$

2. $\left|\begin{array}{ccccc} x_1 & + & x_2 & + & 8x_4 = 0 \\ & & 2x_2 & + & x_4 = 0 \end{array}\right|$

3. $\left|\begin{array}{ccccccc} x_1 & + & 2x_2 & + & 4x_3 & + & 9x_4 = 0 \\ x_1 & + & 8x_2 & + & 2x_3 & + & 7x_4 = 0 \end{array}\right|$

4. $\left|\begin{array}{ccccccc} x_1 & + & x_2 & + & 4x_3 & + & 11x_4 = 0 \\ & & 2x_2 & - & 2x_3 & + & 4x_4 = 0 \end{array}\right|$

Nullspace

Solve using variable order y, z, u, v. Report the values of the **nullity** and **rank** in the equation **nullity+rank**=4.

5. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 8v = 0 \\ & & 2z & - & u & + & v = 0 \\ 2y & & & - & u & + & 5v = 0 \end{array}\right|$

6. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 8v = 0 \\ & & 2z & - & 2u & + & 2v = 0 \\ y & + & 3z & + & 2u & + & 5v = 0 \end{array}\right|$

7. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 8v = 0 \\ & & 2z & - & 2u & + & 4v = 0 \\ y & + & 3z & + & 2u & + & 6v = 0 \end{array}\right|$

8. $\left|\begin{array}{ccccccc} y & + & 3z & + & 4u & + & 8v = 0 \\ & & 2z & - & 2u & + & 4v = 0 \\ y & + & 3z & + & 2u & + & 6v = 0 \end{array}\right|$

9. $\left|\begin{array}{ccccccc} y & + & 3z & + & 4u & + & 8v = 0 \\ & & 2z & - & 2u & + & 4v = 0 \end{array}\right|$

10. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 9v = 0 \\ & & 2z & - & 2u & + & 4v = 0 \end{array}\right|$

11. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 9v = 0 \\ 3y & + & 4z & + & 2u & + & 5v = 0 \end{array}\right|$

12. $\left|\begin{array}{ccccccc} y & + & 2z & + & 4u & + & 9v = 0 \\ y & + & 8z & + & 2u & + & 7v = 0 \end{array}\right|$

13. $\left|\begin{array}{ccccccc} y & + & z & + & 4u & + & 11v = 0 \\ & & 2z & - & 2u & + & 4v = 0 \end{array}\right|$

14. $$\left|\begin{array}{rcrcrcl} y & + & z & + & 5u & + & 11v = 0 \\ & & 2z & - & 2u & + & 6v = 0 \end{array}\right|$$

Dimension of the nullspace

In the homogeneous systems, assume variable order x, y, z, u, v.

(a) Display an equivalent set of equations in reduced echelon form.

(b) Solve for the general solution and check the answer.

(c) Report the dimension of the nullspace.

15. $$\left|\begin{array}{rcrcrcrcl} x & + & y & + & z & + & 4u & + & 8v = 0 \\ -x & + & & & 2z & - & 2u & + & 2v = 0 \\ & & y & - & z & + & 6u & + & 6v = 0 \end{array}\right|$$

16. $$\left|\begin{array}{rcrcrcrcl} x & + & y & + & z & + & 4u & + & 8v = 0 \\ & & & - & 2z & - & u & + & v = 0 \\ & & 2y & & & - & u & + & 5v = 0 \end{array}\right|$$

17. $$\left|\begin{array}{rcrcrcrcl} & & y & + & z & + & 4u & + & 8v = 0 \\ x & & & + & 2z & - & 2u & + & 4v = 0 \\ 2x & + & y & + & 3z & + & 2u & + & 6v = 0 \end{array}\right|$$

18. $$\left|\begin{array}{rcrcrcrcl} x & + & y & + & 3z & + & 4u & + & 8v = 0 \\ 2x & & & + & 2z & - & 2u & + & 4v = 0 \\ x & - & y & + & 3z & + & 2u & + & 12v = 0 \end{array}\right|$$

19. $$\left|\begin{array}{rcrcrcl} y & + & 3z & + & 4u & + & 20v = 0 \\ & + & 2z & - & 2u & + & 10v = 0 \\ - y & + & 3z & + & 2u & + & 30v = 0 \end{array}\right|$$

20. $$\left|\begin{array}{rcrcl} y & + & 4u & + & 20v = 0 \\ & - & 2u & + & 10v = 0 \\ - y & + & 2u & + & 30v = 0 \end{array}\right|$$

21. $$\left|\begin{array}{rcrcrcrl} x & + & y & + & z & + & 4u & = 0 \\ & & & - & 2z & - & u & = 0 \\ & & 2y & & & - & u+ & = 0 \end{array}\right|$$

22. $$\left|\begin{array}{rcrcrcl} & + & z & + & 12u & + & 8v = 0 \\ x & + & 2z & - & 6u & + & 4v = 0 \\ 2x & + & 3z & + & 6u & + & 6v = 0 \end{array}\right|$$

23. $$\left|\begin{array}{rcrcrl} y & + & z & + & 4u & = 0 \\ & & 2z & - & 2u & = 0 \\ y & - & z & + & 6u & = 0 \end{array}\right|$$

24. $$\left|\begin{array}{rcrcl} x & + & z & + & 8v = 0 \\ & - & 2z & + & v = 0 \\ & & & & 5v = 0 \end{array}\right|$$

Three possibilities with symbols

Assume variables x, y, z. Determine the values of the constants (a, b, c, k, etc) such that the system has (1) *No solution*, (2) *A unique solution* or (3) *Infinitely many solutions.*

25. $$\left|\begin{array}{rcrcl} x & + & ky & = & 0 \\ x & + & 2ky & = & 0 \end{array}\right|$$

26. $$\left|\begin{array}{rcrcl} kx & + & ky & = & 0 \\ x & + & 2ky & = & 0 \end{array}\right|$$

27. $$\left|\begin{array}{rcrcl} ax & + & by & = & 0 \\ x & + & 2by & = & 0 \end{array}\right|$$

28. $$\left|\begin{array}{rcrcl} bx & + & ay & = & 0 \\ x & + & 2y & = & 0 \end{array}\right|$$

29. $$\left|\begin{array}{rcrcr} bx & + & ay & = & c \\ x & + & 2y & = & b-c \end{array}\right|$$

30. $$\left|\begin{array}{rcrcr} bx & + & ay & = & 2c \\ x & + & 2y & = & c+a \end{array}\right|$$

31. $$\left|\begin{array}{rcrcrcl} bx & + & ay & + & z & = & 0 \\ 2bx & + & ay & + & 2z & = & 0 \\ x & + & 2y & + & 2z & = & c \end{array}\right|$$

32. $$\left|\begin{array}{rcrcrcr} bx & + & ay & + & z & = & 0 \\ 3bx & + & 2ay & + & 2z & = & 2c, \\ x & + & 2y & + & 2z & = & c \end{array}\right|$$

33. $$\left|\begin{array}{rcrcrcl} 3x & + & ay & + & z & = & b \\ 2bx & + & ay & + & 2z & = & 0 \\ x & + & 2y & + & 2z & = & c \end{array}\right|$$

34. $$\left|\begin{array}{rcrcrcr} x & + & ay & + & z & = & 2b \\ 3bx & + & 2ay & + & 2z & = & 2c \\ x & + & 2y & + & 2z & = & c \end{array}\right|$$

Three Possibilities

Answer the following questions by using equivalents for the three possibilities in terms of lead and free variables, signal equations, rank and nullity.

35. Does there exist a homogeneous 3 × 2 system with a unique solution? Give an example or else prove that no such system exists.

36. Does there exist a homogeneous 2×3 system with a unique solution? Either give an example or else prove that no such system exists.

37. In a homogeneous 10×10 system, two equations are identical. Prove that the system has a nonzero solution.

38. In a homogeneous 5×5 system, each equation has a leading variable. Prove that the system has only the zero solution.

39. Suppose given two homogeneous systems A and B, with A having a unique solution and B having infinitely many solutions. Explain why B cannot be obtained from A by a sequence of swap, multiply and combination operations on the equations.

40. A 2×3 system cannot have a unique solution. Cite a theorem or explain why.

41. If a 3×3 homogeneous system contains no variables, then what is the general solution?

42. If a 3×3 non-homogeneous solution has a unique solution, then what is the nullity of the homogeneous system?

43. A 7×7 homogeneous system is missing two variables. What is the maximum rank of the system? Give examples for all possible ranks.

44. Suppose an $n \times n$ system of equations (homogeneous or non-homogeneous) has two solutions. Prove that it has infinitely many solutions.

45. What is the nullity and rank of an $n \times n$ system of homogeneous equations if the system has a unique solution?

46. What is the nullity and rank of an $n \times n$ system of non-homogeneous equations if the system has a unique solution?

47. Prove or else disprove by counter-example: A 4×3 nonhomogeneous system cannot have a unique solution.

48. Prove or disprove (by example): A 4×3 homogeneous system always has infinitely many solutions.

,

3.5 Answer Check, Proofs and Details

Answer Check Algorithm

A given general solution (3) can be tested for validity manually as in Example 3.6, page 200. It is possible to devise a **symbol-free answer check**. The technique checks a general solution (3) by testing constant trial solutions in systems (5) and (6).

Step 1. Set all invented symbols t_1, ..., t_k to zero in general solution (3) to obtain the nonhomogeneous trial solution $x_1 = d_1$, $x_2 = d_2$, ..., $x_n = d_n$. Test it by direct substitution into the nonhomogeneous system (5).

Step 2. Apply partial derivatives ∂_{t_1}, ∂_{t_2}, ..., ∂_{t_k} to the general solution (3), obtaining k homogeneous trial solutions. Verify that the trial solutions satisfy the homogeneous system (6), by direct substitution.

The trial solutions in **step 2** are obtained from the general solution (3) by setting one symbol equal to 1 and the others zero, followed by subtracting the nonhomogeneous trial solution of **step 1**. The partial derivative idea computes the same set of trial solutions, and it is easier to remember.

Theorem 3.8 (Answer Check)
The answer check algorithm described in steps 1–2 verifies a solution (3) for all values of the symbols. Please observe that this answer check cannot test for skipped solutions.

Proof of Theorem 3.8. To simplify notation and quickly communicate the ideas, a proof will be given for a 2×2 system. A proof for the $m \times n$ case can be constructed by the reader, using the same ideas. Consider the nonhomogeneous and homogeneous systems

$$(1) \qquad \begin{array}{rcl} ax_1 + by_1 &=& b_1, \\ cx_1 + dy_1 &=& b_2, \end{array}$$

$$(2) \qquad \begin{array}{rcl} ax_2 + by_2 &=& 0, \\ cx_2 + dy_2 &=& 0. \end{array}$$

Assume (x_1, y_1) is a solution of (1) and (x_2, y_2) is a solution of (2). Add corresponding equations in (1) and (2). Then collecting terms gives

$$(3) \qquad \begin{array}{rcl} a(x_1 + x_2) + b(y_1 + y_2) &=& b_1, \\ c(x_1 + x_2) + d(y_1 + y_2) &=& b_2. \end{array}$$

This proves that $(x_1 + x_2, y_1 + y_2)$ is a solution of the nonhomogeneous system. Similarly, a scalar multiple (kx_2, ky_2) of a solution (x_2, y_2) of system (2) is also a solution of (2) and the sum of two solutions of (2) is again a solution of (2).

Given each solution in **step 2** satisfies (2), then multiplying the first solution by t_1 and the second solution by t_2 and adding gives a solution (x_3, y_3) of (2). After adding (x_3, y_3) to the solution (x_1, y_1) of **step 1**, a solution of (1) is obtained, proving that the full parametric solution containing the symbols t_1, t_2 is a solution of (1). The proof for the 2×2 case is complete.

Failure of Answer Checks

An answer check only tests the given formulas against the equations. If too few parameters are present, then the answer check can be algebraically correct but the general solution check fails, because not all solutions can be obtained by specialization of the parameter values.

For example, $x = 1 - t_1$, $y = t_1$, $z = 0$ is a one-parameter solution for $x + y + z = 1$, as verified by an answer check. But the general solution $x = 1 - t_1 - t_2$, $y = t_1$, $z = t_2$ has two parameters t_1, t_2. Generally, an answer check decides if the formula supplied works in the equation. It does **not** decide if the given formula represents **all** solutions. This trouble, in which an error leads to a *smaller* value for the nullity of the system, is due largely to human error and not machine error.

Linear algebra workbenches have another kind of flaw: they may compute the **nullity** for a system incorrectly as an integer *larger* than the correct nullity. A parametric solution with nullity k might be obtained, checked to work in the original equations, then cross-checked by computing the nullity k independently. However, the computed nullity k could be greater than the actual nullity of the system. Here is a simple example, where ϵ is a *very* small positive number:

$$(4) \qquad \begin{array}{rcrcl} x & + & y & = & 0, \\ & & \epsilon y & = & \epsilon. \end{array}$$

On a limited precision machine, system (4) has internal machine representation[4]

$$(5) \qquad \begin{array}{rcrcl} x & + & y & = & 0, \\ & & 0 & = & 0. \end{array}$$

Representation (5) occurs because the coefficient ϵ is smaller than the smallest positive floating point number of the machine, hence it becomes zero during translation. System (4) has nullity zero and system (5) has nullity one. The parametric solution for system (5) is $x = -t_1$, $y = t_1$, with basis selected by setting $t_1 = 1$. The basis passes the answer check on system (4), because ϵ times 1 evaluates to ϵ. A second check for the nullity of system (5) gives 1, which supports the correctness of the parametric solution, but unfortunately there are not infinitely many solutions: for system (4) the correct answer is the unique solution $x = -1$, $y = 1$.

[4] For example, if the machine allows only 2-digit exponents (10^{99} is the maximum), then $\epsilon = 10^{-101}$ translates to zero.

Computer algebra systems (CAS) are supposed to avoid this kind of error, because they do not translate input into floating point representations. All input is supposed to remain in symbolic or in string form. In short, they don't change ϵ to zero. Because of this standard, CAS are safer systems in which to do linear algebra computations, albeit slower in execution.

The trouble reported here is not entirely one of input translation. An innocuous `combo(1,2,-1)` can cause an equation like $\epsilon y = \epsilon$ in the middle of a toolkit sequence. If floating point hardware is being used, and not symbolic computation, then the equation can translate to $0 = 0$, causing a false free variable appearance.

Minimal Parametric Solutions

Proof of Theorem 3.2: The proof of Theorem 3.2, page 196, will follow from the lemma and theorem below.

Lemma 3.1 (Unique Representation) If a set of parametric equations (3) satisfies (4), (5) and (6), then each solution of linear system (5) is given by (3) for exactly one set of parameter values.

Proof: Let a solution of system (5) be given by (3) for two sets of parameters $t_1, \ldots, t_k$ and $\bar{t}_1, \ldots, \bar{t}_k$. By (6), $t_j = x_{i_j} = \bar{t}_j$ for $1 \le j \le k$, therefore the parameter values are the same.

Definition 3.10 (Minimal Parametric Solution)
Given system (5) has a parametric solution $x_1, \ldots, x_n$ satisfying (3), (4), (5), then among all such parametric solutions there is one which uses the *fewest* possible parameters. A parametric solution with fewest parameters is called **minimal**. Parametric solutions with more parameters are called **redundant**.

To illustrate, the plane $x + y + z = 1$ has a minimal standard parametric solution $x = 1 - t_1 - t_2$, $y = t_1$, $z = t_2$. A redundant parametric solution of $x + y + z = 1$ is $x = 1 - t_1 - t_2 - 2t_3$, $y = t_1 + t_3$, $z = t_2 + t_3$, using three parameters t_1, t_2, t_3.

Theorem 3.9 (Minimal Parametric Solutions)
Let linear system (5) have a parametric solution satisfying (3), (4), (5). Then (3) has the fewest possible parameters if and only if each solution of linear system (5) is given by (3) for exactly one set of parameter values.

Proof: Suppose first that a general solution (3) is given with the least number k of parameters, but contrary to the theorem, there are two ways to represent some solution, with corresponding parameters $r_1, \ldots, r_k$ and also $s_1, \ldots, s_k$. Subtract the two sets of parametric equations, thus eliminating the symbols $x_1, \ldots, x_n$, to obtain:

$$\begin{aligned} c_{11}(r_1 - s_1) + \cdots + c_{1k}(r_k - s_k) &= 0, \\ &\vdots \\ c_{n1}(r_1 - s_1) + \cdots + c_{nk}(r_k - s_k) &= 0. \end{aligned}$$

Relabel the variables and constants so that $r_1 - s_1 \neq 0$, possible since the two sets of parameters are supposed to be different. Divide the preceding equations by $r_1 - s_1$ and solve for the constants $c_{11}, \ldots, c_{n1}$. This results in equations

$$\begin{array}{rcl} c_{11} & = & c_{12}w_2 + \cdots + c_{1k}w_k, \\ & \vdots & \\ c_{n1} & = & c_{n2}w_2 + \cdots + c_{nk}w_k, \end{array}$$

where $w_j = -\frac{r_j - s_j}{r_1 - s_1}$, $2 \leq j \leq k$. Insert these relations into (3), effectively eliminating the symbols $c_{11}, \ldots, c_{n1}$, to obtain

$$\begin{array}{rcl} x_1 & = & d_1 + c_{12}(t_2 + w_2t_1) + \cdots + c_{1k}(t_k + w_kt_1), \\ x_2 & = & d_2 + c_{22}(t_2 + w_2t_1) + \cdots + c_{2k}(t_k + w_kt_1), \\ & \vdots & \\ x_n & = & d_n + c_{n2}(t_2 + w_2t_1) + \cdots + c_{nk}(t_k + w_kt_1). \end{array}$$

Let $t_1 = 0$. The remaining parameters $t_2, \ldots, t_k$ are fewer parameters that describe all solutions of the system, a contradiction to the definition of k. This completes the proof of the first half of the theorem.

To prove the second half of the theorem, assume that a parametric solution (3) is given which represents all possible solutions of the system and in addition each solution is represented by exactly one set of parameter values. It will be established that the number k in (3) is the *least possible* parameter count.

Suppose not. Then there is a second parametric solution

$$\begin{array}{rcl} x_1 & = & e_1 + b_{11}v_1 + \cdots + b_{1\ell}v_\ell, \\ & \vdots & \\ x_n & = & e_n + b_{n1}v_1 + \cdots + b_{n\ell}v_\ell, \end{array} \tag{6}$$

where $\ell < k$ and $v_1, \ldots, v_\ell$ are the parameters. It is assumed that (6) represents all solutions of the linear system.

We shall prove that the solutions for zero parameters in (3) and (6) can be taken to be the same, that is, another parametric solution is given by

$$\begin{array}{rcl} x_1 & = & d_1 + b_{11}s_1 + \cdots + b_{1\ell}s_\ell, \\ & \vdots & \\ x_n & = & d_n + b_{n1}s_1 + \cdots + b_{n\ell}s_\ell. \end{array} \tag{7}$$

The idea of the proof is to substitute $x_1 = d_1, \ldots, x_n = d_n$ into (6) for parameters $r_1, \ldots, r_n$. Then solve for $e_1, \ldots, e_n$ and replace back into (6) to obtain

$$\begin{array}{rcl} x_1 & = & d_1 + b_{11}(v_1 - r_1) + \cdots + b_{1\ell}(v_\ell - r_\ell), \\ & \vdots & \\ x_n & = & d_n + b_{n1}(v_1 - r_1) + \cdots + b_{n\ell}(v_\ell - r_\ell). \end{array}$$

Replacing parameters $s_j = v_j - r_j$ gives (7).

From (3) it is known that $x_1 = d_1 + c_{11}, \ldots, x_n = d_n + c_{n1}$ is a solution. By (7), there are constants $r_1, \ldots, r_\ell$ such that (we cancel $d_1, \ldots, d_n$ from both sides)

$$\begin{array}{rcl} c_{11} & = & b_{11}r_1 + \cdots + b_{1\ell}r_\ell, \\ & \vdots & \\ c_{n1} & = & b_{n1}r_1 + \cdots + b_{n\ell}r_\ell. \end{array}$$

If r_1 through r_ℓ are all zero, then the solution just referenced equals $d_1, \ldots, d_n$, hence (3) has a solution that can be represented with parameters all zero or with $t_1 = 1$ and all other parameters zero, a contradiction. Therefore, some $r_i \neq 0$ and we can assume by renumbering that $r_1 \neq 0$. Return now to the last system of equations and divide by r_1 in order to solve for the constants $b_{11}, \ldots, b_{n1}$. Substitute the answers back into (7) in order to obtain parametric equations

$$\begin{aligned} x_1 &= d_1 + c_{11}w_1 + b_{12}w_2 + \cdots + b_{1\ell}w_\ell, \\ &\vdots \\ x_n &= d_n + c_{n1}w_1 + b_{n2}w_2 + \cdots + b_{n\ell}w_\ell, \end{aligned}$$

where $w_1 = s_1$, $w_j = s_j - r_j/r_1$. Given $s_1, \ldots, s_\ell$ are parameters, then so are $w_1, \ldots, w_\ell$.

This process can be repeated for the solution $x_1 = d_1 + c_{12}, \ldots, x_n = d_n + c_{n2}$. We assert that for some index j, $2 \le j \le \ell$, constants $b_{ij}, \ldots, b_{nj}$ in the previous display can be isolated, and the process of replacing symbols b by c continued. If not, then $w_2 = \cdots = w_\ell = 0$. Then solution $x_1, \ldots, x_n$ has two distinct representations in (3), first with $t_2 = 1$ and all other $t_j = 0$, then with $t_1 = w_1$ and all other $t_j = 0$. A contradiction results, which proves the assertion. After ℓ repetitions of this replacement process, we find a parametric solution

$$\begin{aligned} x_1 &= d_1 + c_{11}u_1 + c_{12}u_2 + \cdots + c_{1\ell}u_\ell, \\ &\vdots \\ x_n &= d_n + c_{n1}u_1 + c_{n2}u_2 + \cdots + c_{n\ell}u_\ell, \end{aligned}$$

in some set of parameters $u_1, \ldots, u_\ell$.

However, $\ell < k$, so at least the solution $x_1 = d_1 + c_{1k}, \ldots, x_n = d_n + c_{nk}$ remains unused by the process. Insert this solution into the previous display, valid for some parameters $u_1, \ldots, u_\ell$. The relation says that the solution $x_1 = d_1, \ldots, x_n = d_n$ in (3) has two distinct sets of parameters, namely $t_1 = u_1, \ldots, t_\ell = u_\ell$, $t_k = -1$, all others zero, and also all parameters zero, a contradiction. ■

,

Exercises 3.5

Parametric solutions

1. Is there a 2×3 homogeneous system with general solution having 2 parameters t_1, t_2?

2. Is there a 3×3 homogeneous system with general solution having 3 parameters t_1, t_2, t_3?

3. Give an example of a 4×3 homogeneous system with general solution having zero parameters, that is, $x = y = z = 0$ is the only solution.

4. Give an example of a 4×3 homogeneous system with general solution having exactly one parameter t_1.

5. Give an example of a 4×3 homogeneous system with general solution having exactly two parameters t_1, t_2.

6. Give an example of a 4×3 homogeneous system with general solution having exactly three parameters t_1, t_2, t_3.

7. Consider an $n \times n$ homogeneous system with parametric solution having parameters t_1 to t_k. What are the possible values of k?

8. Consider an $n \times m$ homogeneous system with parametric solution having parameters t_1 to t_k. What are the possible values of k?

Answer Checks

Assume variable list x, y, z and parameter t_1. (a) Display the answer check details. (b) Find the rank. (c) Report whether the given solution is a general solution.

9. $$\left|\begin{array}{ccccccc} & & y & & & = & 1 \\ & & & & 2z & = & 2 \end{array}\right|$$
$x = t_1, y = 1, z = 1.$

10. $$\left|\begin{array}{ccccccc} x & & & & & = & 1 \\ & & & & 2z & = & 2 \end{array}\right|$$
$x = 1, y = t_1, z = 1.$

11. $$\left|\begin{array}{ccccccc} & & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ x & & & + & z & = & 1 \end{array}\right|$$
$x = 0, y = 0, z = 1.$

12. $$\left|\begin{array}{ccccccr} 2x & + & y & + & z & = & 1 \\ x & & & + & 2z & = & 2 \\ x & + & y & - & z & = & -1 \end{array}\right|$$
$x = 2, y = -3, z = 0.$

13. $$\left|\begin{array}{ccccccc} x & + & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ x & & & + & z & = & 1 \end{array}\right|$$
$x = 1 - t_1, y = 0, z = t_1.$

14. $$\left|\begin{array}{ccccccc} x & + & y & + & z & = & 1 \\ 2x & & & + & 2z & = & 2 \\ 3x & + & y & + & 3z & = & 3 \end{array}\right|$$
$x = 1 - t_1, y = 0, z = t_1.$

Failure of Answer Checks

Find the unique solution for $\epsilon > 0$. Discuss how a machine might translate the system to obtain infinitely many solutions.

15. $x + \epsilon y = 1$, $x - \epsilon y = 1$

16. $x + y = 1$, $x + (1 + \epsilon)y = 1 + \epsilon$

17. $x + \epsilon y = 10\epsilon$, $x - \epsilon y = 10\epsilon$

18. $x + y = 1 + \epsilon$, $x + (1 + \epsilon)y = 1 + 11\epsilon$

Minimal Parametric Solutions

For each given system, determine if the expression is a minimal general solution.

19. $$\left|\begin{array}{ccccccccl} y & + & z & + & 4u & + & 8v & = 0 \\ & & 2z & - & u & + & v & = 0 \\ 2y & & & - & u & + & 5v & = 0 \end{array}\right|$$
$y = -3t_1, z = -t_1,$
$u = -t_1, v = t_1.$

20. $$\left|\begin{array}{ccccccccl} y & + & z & + & 4u & + & 8v & = 0 \\ & & 2z & - & 2u & + & 2v & = 0 \\ y & - & z & + & 6u & + & 6v & = 0 \end{array}\right|$$
$y = -5t_1 - 7t_2, z = t_1 - t_2,$
$u = t_1, v = t_2.$

21. $$\left|\begin{array}{ccccccccl} y & + & z & + & 4u & + & 8v & = 0 \\ & & 2z & - & 2u & + & 4v & = 0 \\ y & + & 3z & + & 2u & + & 6v & = 0 \end{array}\right|$$
$y = -5t_1 + 5t_2, z = t_1 - t_2,$
$u = t_1 - t_2, v = 0.$

22. $$\left|\begin{array}{ccccccccl} y & + & 3z & + & 4u & + & 8v & = 0 \\ & & 2z & - & 2u & + & 4v & = 0 \\ y & + & 3z & + & 2u & + & 12v & = 0 \end{array}\right|$$
$y = 5t_1 + 4t_2, z = -3t_1 - 6t_2,$
$u = -t_1 - 2t_2, v = t_1 + 2t_2.$

,

Chapter 4

Numerical Methods with Applications

Contents

4.1 Solving $y' = F(x)$ Numerically

Studied here is the creation of numerical tables and graphics for the solution of the initial value problem

(1) $$y' = F(x), \quad y(x_0) = y_0.$$

To illustrate, consider the initial value problem

$$y' = 3x^2 - 1, \quad y(0) = 2.$$

Quadrature gives the explicit **symbolic solution**

$$y(x) = x^3 - x + 2.$$

In Figure 1, evaluation of $y(x)$ from $x = 0$ to $x = 1$ in increments of 0.1 gives the xy-table, whose entries represent the **dots** for the **connect-the-dots** graphic.

x	y
0.0	2.000
0.1	1.901
0.2	1.808
0.3	1.727
0.4	1.664
0.5	1.625

x	y
0.6	1.616
0.7	1.643
0.8	1.712
0.9	1.829
1.0	2.000

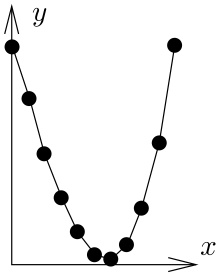


Figure 1. A table of xy-values for $y = x^3 - x + 2$.
The graphic represents table rows as *dots*, which are joined to make the *connect-the-dots* graphic.

The interesting case is when quadrature in (1) encounters an integral $\int_{x_0}^{x} F(t)dt$ that cannot be evaluated to provide an explicit symbolic equation for $y(x)$. Nevertheless, $y(x)$ can be computed numerically.

Applied here are numerical integration rules from calculus: *rectangular*, *trapezoidal* and *Simpson*; see page 232 for a review of the three rules. The ideas lead to the numerical methods of Euler, Heun and Runge-Kutta, which appear later in this chapter.

How to Make an xy-Table

Given $y' = F(x)$, $y(x_0) = y_0$, a table of xy-values is created as follows. The x-values are equally spaced a distance $h > 0$ apart. Each x, y pair in the table represents a *dot* in the *connect-the-dots* graphic of the explicit solution

$$y(x) = y_0 + \int_{x_0}^{x} F(t)dt.$$

First table entry. The *initial condition* $y(x_0) = y_0$ identifies two constants x_0, y_0 to be used for the first table pair X, Y. For example, $y(0) = 2$ identifies first table pair $X = 0$, $Y = 2$.

Second table entry. The second table pair X, Y is computed from the first table pair x_0, y_0 and a **recurrence**. The X-value is given by $X = x_0 + h$, while the Y-value is given by the numerical integration method being used, in accordance with Table 1. The table is justified on page 235. See Example 4.1 page 228 for a rectangular rule example.

Table 1. Three Numerical Integration Methods.

Rectangular Rule	$Y = y_0 + hF(x_0)$
Trapezoidal Rule	$Y = y_0 + \frac{h}{2}(F(x_0) + F(x_0 + h))$
Simpson's Rule	$Y = y_0 + \frac{h}{6}(F(x_0) + 4F(x_0 + h/2) + F(x_0 + h)))$

Third and higher table entries. They are computed by letting x_0, y_0 be the current table entry, then the next table entry X, Y is found exactly as outlined above for the second table entry.

It is expected, and normal, to compute the table entries using computer assist. In simple cases, a calculator will suffice. If F is complicated or Simpson's rule is used, then a computer algebra system or a numerical laboratory is recommended. See Example 4.2, page 229.

How to Make a Connect-The-Dots Graphic

To illustrate, consider the xy-pairs below, which are to represent the *dots* in the *connect-the-dots* graphic.

$$(0.0, 2.000), (0.1, 1.901), (0.2, 1.808), (0.3, 1.727), (0.4, 1.664),$$
$$(0.5, 1.625), (0.6, 1.616), (0.7, 1.643), (0.8, 1.712), (0.9, 1.829),$$
$$(1.0, 2.000).$$

Hand drawing. The method, unchanged from high school mathematics courses, is to plot the points as dots on an xy-coordinate system, then connect the dots with line segments. See Figure 2.

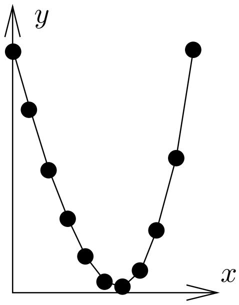


Figure 2. A Connect-the-Dots Graphic.
A computer-generated graphic simulating a hand-drawn graphic. The graphics engine draws straight lines between dots.

Computer Algebra System Graphic

Computer algebra system `maple`. It has a primitive syntax especially made for connect-the-dots graphics. Below, `Dots` is a list of xy-pairs.

```
Dots:=[0.0, 2.000], [0.1, 1.901], [0.2, 1.808],
      [0.3, 1.727], [0.4, 1.664], [0.5, 1.625],
      [0.6, 1.616], [0.7, 1.643], [0.8, 1.712],
      [0.9, 1.829], [1.0, 2.000]:
plot([Dots]);
```

The plotting of *points only* can be accomplished by adding options into the `plot` command: `type=point` and `symbol=circle` will suffice.

Computer algebra system xmaxima. The plot primitive can be invoked with x-array and y-array, or else pairs as above:

```
Dots:[[0.0, 2.000], [0.1, 1.901], [0.2, 1.808],
 [0.3, 1.727],[0.4, 1.664],[0.5, 1.625],
 [0.6, 1.616], [0.7, 1.643],[0.8, 1.712],
 [0.9, 1.829], [1.0,2.000]];
 plot2d([discrete,Dots]);
```

Numerical Laboratory Graphic

Computer programs `matlab`, `octave` and `scilab` provide primitive plotting facilities, as follows.

```
X=[0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1]
Y=[2.000, 1.901, 1.808, 1.727, 1.664, 1.625,
   1.616, 1.643, 1.712, 1.829, 2.000]
plot(X,Y)
```

Example 4.1 (Rectangular Rule)
Consider $y' = 3x^2 - 2x$, $y(0) = 0$. Apply the rectangular rule to make an xy-table for $y(x)$ from $x = 0$ to $x = 2$ in steps of $h = 0.2$. Graph the approximate solution and the exact solution $y(x) = x^3 - x^2$ for $0 \le x \le 2$.

Solution: The exact solution $y = x^3 - x^2$ is verified directly, by differentiation. It was obtained by quadrature applied to $y' = 3x^2 - 2x$, $y(0) = 0$.

The first table entry is 0, 0. It is decoded from $y(x_0) = y_0$ as entry x_0, y_0, applied to the present initial condition $y(0) = 0$. The first table row 0, 0 is used to obtain the second table row $X = 0.2$, $Y = 0$ as follows.

$x_0 = 0$, $y_0 = 0$	The current table entry, row 1.
$X = x_0 + h$	The next table entry, row 2.
$\quad = 0.2,$	Use $x_0 = 0$, $h = 0.2$.
$Y = y_0 + hF(x_0)$	Rectangular rule, $F(x) = 3x^2 - 2x$.
$\quad = 0 + 0.2(0).$	Use $y_0 = 0$, $h = 0.2$, $x_0 = 0$.

Row 3 starts with $x_0 = 0.2$, $y_0 = 0$ from row 2 to produce $X = 0.4$, $Y = 0 + 0.2F(0.2) = -0.056$. The remaining 8 rows of the table are completed by calculator, following the same pattern:

Table 2. Rectangular Rule Solution and Exact Values for $y' = 3x^2 - 2x$, $y(0) = 0$ on $0 \le x \le 2$, step size $h = 0.2$.

x	y-rect	y-exact	x	y-rect	y-exact
0.0	0.000	0.000	1.2	0.120	0.288
0.2	0.000	−0.032	1.4	0.504	0.784
0.4	−0.056	−0.096	1.6	1.120	1.536
0.6	−0.120	−0.144	1.8	2.016	2.592
0.8	−0.144	−0.128	2.0	3.240	4.000
1.0	−0.080	0.000			

The xy-values from the table are used to obtain the comparison plot in Figure 3.

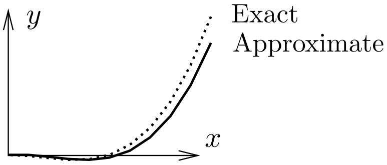


Figure 3. Comparison Plot. Rectangular rule numerical solution and the exact solution for $y = x^3 - x^2$ for $y' = 3x^2 - 2x$, $y(0) = 0$.

Example 4.2 (Trapezoidal Rule)

Consider $y' = \cos x + 2x$, $y(0) = 0$. Apply both the rectangular and trapezoidal rules to make an xy-table for $y(x)$ from $x = 0$ to $x = \pi$ in steps of $h = \pi/10$. Compare the two approximations in a graphic for $0 \le x \le \pi$.

Solution: The exact solution $y = \sin x + x^2$ is verified by differentiation. It will be seen that the trapezoidal solution is graphically nearly identical to the exact solution.

The table will have 11 rows. The three columns are x, y-rectangular and y-trapezoidal. The first table entry 0, 0, 0 is used to obtain the second table entry 0.1π, 0.31415927, 0.40516728 as follows.

Rectangular rule second entry.

$$\begin{aligned} Y &= y_0 + hF(x_0) && \text{Rectangular rule.} \\ &= 0 + h(\cos 0 + 2(0)) && \text{Use } F(x) = \cos x + 2x,\ x_0 = y_0 = 0. \\ &= 0.31415927. && \text{Use } h = 0.1\pi = 0.31415927. \end{aligned}$$

Trapezoidal rule second entry.

$$\begin{aligned} Y &= y_0 + 0.5h(F(x_0) + F(x_0 + h)) && \text{Trapezoidal rule.} \\ &= 0 + 0.05\pi(\cos 0 + \cos h + 2h) && \text{Use } x_0 = y_0 = 0,\ F(x) = \cos x + 2x. \\ &= 0.40516728. && \text{Use } h = 0.1\pi. \end{aligned}$$

The remaining 9 rows of the table are completed by calculator, following the pattern above for the second table entry. The result:

Table 3. Rectangular and Trapezoidal Solutions for $y' = \cos x + 2x$, $y(0) = 0$ on $0 \le x \le \pi$, step size $h = 0.1\pi$.

x	y-rect	y-trap	x	y-rect	y-trap
0.000000	0.000000	0.000000	1.884956	4.109723	4.496279
0.314159	0.314159	0.405167	2.199115	5.196995	5.638458
0.628319	0.810335	0.977727	2.513274	6.394081	6.899490
0.942478	1.459279	1.690617	2.827433	7.719058	8.300851
1.256637	2.236113	2.522358	3.141593	9.196803	9.869604
1.570796	3.122762	3.459163			

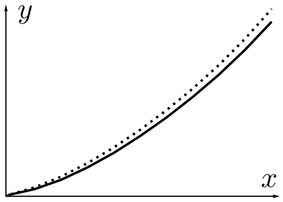


Figure 4. Comparison Plot. Rectangular (solid) and trapezoidal (dotted) numerical solutions for $y' = \cos x + 2x$, $y(0) = 0$ for $h = 0.1\pi$ on $0 \le x \le \pi$.

Computer algebra system. The `maple` implementation for Example 4.2 appears below. The code produces lists `Dots1` and `Dots2` which contain Rectangular (left panel) and Trapezoidal (right panel) approximations.

```
# Rectangular algorithm
# Group 1, initialize.
F:=x->evalf(cos(x) + 2*x):
x0:=0:y0:=0:h:=0.1*Pi:
Dots1:=[x0,y0]:

# Group 2, loop count = 10
for i from 1 to 10 do
Y:=y0+h*F(x0):
x0:=x0+h:y0:=evalf(Y):
Dots1:=Dots1,[x0,y0];
end do;

# Group 3, plot.
plot([Dots1]);
```

```
# Trapezoidal algorithm
# Group 1, initialize.
F:=x->evalf(cos(x) + 2*x):
x0:=0:y0:=0:h:=0.1*Pi:
Dots2:=[x0,y0]:

# Group 2, repeat 10 times
for i from 1 to 10 do
Y:=y0+h*(F(x0)+F(x0+h))/2:
x0:=x0+h:y0:=evalf(Y):
Dots2:=Dots2,[x0,y0];
end do;

# Group 3, plot.
plot([Dots2]);
```

Example 4.3 (Simpson's Rule)

Consider $y' = e^{-x^2}$, $y(0) = 0$. Apply both the rectangular and Simpson rules to make an xy-table for $y(x)$ from $x = 0$ to $x = 1$ in steps of $h = 0.1$. In the table, include values for the exact solution $y(x) = \frac{\sqrt{\pi}}{2}\,\mathrm{erf}(x)$. Compare the two approximations in a graphic for $0.8 \le x \le 1.0$.

Solution: The **error function** $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is a library function available in `maple`, `mathematica`, `matlab` and other computing platforms. It is known that the integral cannot be expressed in terms of elementary functions.

The xy-table. There will be 11 rows, for $x = 0$ to $x = 1$ in steps of $h = 0.1$. There are four columns: x, y-rectangular, y-Simpson, y-exact.

It will be shown how to obtain the first and second rows by calculator methods, for the two algorithms *rectangular* and *Simpson*.

Rectangular rule table row 1.
Initial condition $y(x_0) = y_0$ gives row 1 table pair x_0, y_0. For initial condition $y(0) = 0$, the pattern decodes into row 1 table pair $x_0 = 0$, $y_0 = 0$.

Rectangular rule table row 2. Label the second table pair (X, Y).

$$\begin{aligned}
X &= x_0 + h && \text{Equal divisions.}\\
Y &= y_0 + hF(x_0) && \text{Rectangular rule.}\\
&= 0 + h(e^0) && \text{Use } F(x) = e^{-x^2},\ x_0 = y_0 = 0.\\
&= 0.1. && \text{Use } h = 0.1 \text{ and } e^0 = 1.
\end{aligned}$$

Simpson rule table row 1.
Identical for all rules, therefore table row 1 is $x_0 = 0$, $y_0 = 0$, copied from the rectangular rule above.

Simpson rule row 2. Row 2 table pair is labeled (X, Y).

$$X = x_0 + h \qquad \text{Equal divisions.}$$

$Y = y_0 + \frac{h}{6}(F(x_0) + 4F(x_0 + h/2) + F(x_0 + h))$	Simpson rule.
$= 0 + \frac{0.1}{6}(e^0 + 4e^{.5} + e^{.1})$	Use $F(x) = e^{-x^2}$, $x_0 = y_0 = 0$, $h = 0.1$.
$= 0.09966770540.$	Calculator.

Exact solution table row 2.
The numerical work requires the tabulated function erf(x). The `maple` details:

```
x0:=0:y0:=0:h:=0.1:          Given.
c:=sqrt(Pi)/2                Conversion factor.
Exact:=x->y0+c*erf(x):       Exact solution y = y0 + ∫_0^x e^{-t^2} dt.
Y3:=Exact(x0+h);             Calculate exact answer.
# Y3 := .09966766428
```

Table 4. Rectangular and Simpson Rule.
Numerical solutions for $y' = e^{-x^2}$, $y(0) = 0$ on $0 \le x \le \pi$, step size $h = 0.1$.

x	y-rect	y-Simp	y-exact
0.0	0.00000000	0.00000000	0.00000000
0.1	0.10000000	0.09966771	0.09966766
0.2	0.19900498	0.19736511	0.19736503
0.3	0.29508393	0.29123799	0.29123788
0.4	0.38647705	0.37965297	0.37965284
0.5	0.47169142	0.46128114	0.46128101
0.6	0.54957150	0.53515366	0.53515353
0.7	0.61933914	0.60068579	0.60068567
0.8	0.68060178	0.65766996	0.65766986
0.9	0.73333102	0.70624159	0.70624152
1.0	0.77781682	0.74682418	0.74682413

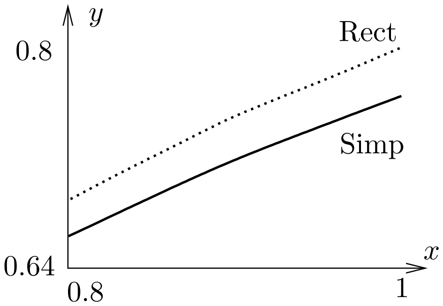


Figure 5. Comparison Plot.

Rectangular (dotted) and Simpson (solid) numerical solutions for $y' = e^{-x^2}$, $y(0) = 0$ for $h = 0.1$ on $0.8 \le x \le 1.0$.

Computer algebra system. The `maple` implementation for Example 4.3 appears below. The code produces two lists `Dots1` and `Dots2` which contain Rectangular (left panel) and Simpson (right panel) approximations.

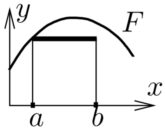
```
# Rectangular algorithm
# Group 1, initialize.
F:=x->evalf(exp(-x*x)):
x0:=0:y0:=0:h:=0.1:
Dots1:=[x0,y0]:

# Group 2, repeat 10 times
for i from 1 to 10 do
Y:=evalf(y0+h*F(x0)):
x0:=x0+h:y0:=Y:
Dots1:=Dots1,[x0,y0];
end do;


# Group 3, plot.
plot([Dots1]);
```

```
# Simpson algorithm
# Group 1, initialize.
F:=x->evalf(exp(-x*x)):
x0:=0:y0:=0:h:=0.1:
Dots2:=[x0,y0]:

# Group 2, loop count = 10
for i from 1 to 10 do
Y:=evalf(y0+h*(F(x0)+
   4*F(x0+h/2)+F(x0+h))/6):
x0:=x0+h:y0:=Y:
Dots2:=Dots2,[x0,y0];
end do;

# Group 3, plot.
plot([Dots2]);
```

Review of Numerical Integration

Reproduced here are calculus topics: the **rectangular rule**, the **trapezoidal rule** and **Simpson's rule**, which are tools for the numerical approximation of an integral $\int_a^b F(x)dx$. The approximations are valid for $b - a$ small. Larger intervals must be subdivided, then the rule applies to the small subdivisions.

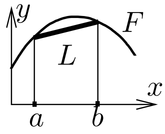
Rectangular Rule



The approximation uses Euler's idea of replacing the integrand by a constant. The value of the integral is approximately the area of a rectangle of width $b - a$ and height $F(a)$.

(2)
$$\int_a^b F(x)dx \approx (b-a)F(a).$$

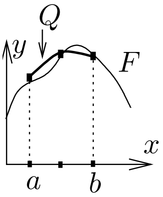
Trapezoidal Rule



The rule replaces the integrand $F(x)$ by a linear function $L(x)$ which connects the planar points $(a, F(a))$, $(b, F(b))$. The value of the integral is approximately the area under the curve L, which is the area of a trapezoid.

(3)
$$\int_a^b F(x)dx \approx \frac{b-a}{2}\left(F(a)+F(b)\right).$$

Simpson's Rule



The rule replaces the integrand $F(x)$ by a quadratic polynomial $Q(x)$ which connects the planar points $(a, F(a))$, $((a+b)/2, F((a+b)/2))$, $(b, F(b))$. The value of the integral is approximately the area under the quadratic curve Q.

$$\int_a^b F(x)dx \approx \frac{b-a}{6}\left(F(a) + 4F\left(\frac{a+b}{2}\right) + F(b)\right). \tag{4}$$

Simpson's Polynomial Rule

If $Q(x)$ is constant, or a linear, quadratic or cubic polynomial, then

$$\int_a^b Q(x)dx = \frac{b-a}{6}\left(Q(a) + 4Q\left(\frac{a+b}{2}\right) + Q(b)\right). \tag{5}$$

Integrals of linear, quadratic and cubic polynomials can be evaluated *exactly* using Simpson's polynomial rule (5). See Example 4.4, page 233 and the proof on page 234.

Remarks on Simpson's Rule

The right side of (4) is exactly the integral of $Q(x)$, which is evaluated by equation (5). The appearance of F instead of Q on the right in equation (4) is due to the relations $Q(a) = F(a)$, $Q((a+b)/2) = F((a+b)/2)$, $Q(b) = F(b)$, which arise from the requirement that Q connect three points along curve F.

The quadratic interpolation polynomial $Q(x)$ is determined uniquely from the three data points; see *Quadratic Interpolant*, page 234, for a formula for Q and a derivation. It is interesting that Simpson's rule depends only upon uniqueness and not upon an actual formula for Q!

Example 4.4 (Polynomial Quadrature)
Apply Simpson's polynomial rule (5) to verify $\int_1^2 (x^3 - 16x^2 + 4)dx = -355/12$.

Solution: The application proceeds as follows:

$$\begin{aligned}
I &= \textstyle\int_1^2 Q(x)dx && \text{Evaluate integral } I \text{ using } Q(x) = x^3 - 16x^2 + 4. \\
&= \frac{2-1}{6}\left(Q(1) + 4Q(3/2) + Q(2)\right) && \text{Apply Simpson's polynomial rule (5).} \\
&= \frac{1}{6}(-11 + 4(-229/8) - 52) && \text{Use } Q(x) = x^3 - 16x^2 + 4. \\
&= -\frac{355}{12}. && \text{Equality verified.}
\end{aligned}$$

Simpson's Polynomial Rule Proof. Let $Q(x)$ be a linear, quadratic or cubic polynomial. It will be verified that

$$\int_a^b Q(x)dx = \frac{b-a}{6}\left(Q(a) + 4Q\left(\frac{a+b}{2}\right) + Q(b)\right). \tag{6}$$

If the formula holds for polynomial Q and c is a constant, then the formula also holds for the polynomial cQ. Similarly, if the formula holds for polynomials Q_1 and Q_2, then it also holds for $Q_1 + Q_2$. Consequently, it suffices to show that the formula is true for the special polynomials 1, x, x^2 and x^3, because then it holds for all combinations $Q(x) = c_0 + c_1x + c_2x^2 + c_3x^3$.

Only the special case $Q(x) = x^3$ will be treated here. The other cases are left to the exercises. The details:

$\text{RHS} = \frac{b-a}{6}\left(Q(a) + 4Q\left(\frac{a+b}{2}\right) + Q(b)\right)$	Evaluate the right side of equation (6).
$= \frac{b-a}{6}\left(a^3 + \frac{1}{2}(a+b)^3 + b^3\right)$	Substitute $Q(x) = x^3$.
$= \frac{b-a}{6}\left(\frac{3}{2}\right)\left(a^3 + a^2b + ab^2 + b^3\right)$	Expand $(a+b)^3$. Simplify.
$= \frac{1}{4}\left(b^4 - a^4\right),$	Multiply and simplify.
$\text{LHS} = \int_a^b Q(x)dx$	Evaluate the left hand side (LHS) of equation (6).
$= \int_a^b x^3dx$	Substitute $Q(x) = x^3$.
$= \frac{1}{4}\left(b^4 - a^4\right)$	Evaluate.
$= \text{RHS}.$	Compare with the RHS.

■

Quadratic Interpolant Q

Given $a < b$ and the three data points (a, Y_0), $((a+b)/2, Y_1))$, $(b, Y_2))$, then there is a unique quadratic curve $Q(X)$ which connects the points, given by

$$\begin{aligned} Q(X) = Y_0 &+ (4Y_1 - Y_2 - 3Y_0)\frac{X-a}{b-a} \\ &+ (2Y_2 + 2Y_0 - 4Y_1)\frac{(X-a)^2}{(b-a)^2}. \end{aligned} \tag{7}$$

Proof: The term *quadratic* is meant loosely: it can be a constant or linear function as well.

Uniqueness of the interpolant Q is established by subtracting two candidates to obtain a polynomial P of degree at most two which vanishes at three distinct points. By Rolle's theorem, P' vanishes at two distinct points and hence P'' vanishes at one point. Writing

$P(X) = c_0 + c_1 X + c_2 X^2$ shows $c_2 = 0$ and then $c_1 = c_0 = 0$, or briefly, $P \equiv 0$. Hence the two candidates are identical.

It remains to verify the given formula (7). The details are presented as two lemmas.[1] The first lemma contains the essential ideas. The second simply translates the variables.

Lemma 4.1 Given y_1 and y_2, define $A = y_2 - y_1$, $B = 2y_1 - y_2$. Then the quadratic $y = x(Ax + B)$ fits the data items $(0, 0)$, $(1, y_1)$, $(2, 2y_2)$.

Lemma 4.2 Given Y_0, Y_1 and Y_2, define $y_1 = Y_1 - Y_0$, $y_2 = \frac{1}{2}(Y_2 - Y_0)$, $A = y_2 - y_1$, $B = 2y_1 - y_2$ and $x = 2(X - a)/(b - a)$. Then quadratic $Y(X) = Y_0 + x(Ax + B)$ fits the data items (a, Y_0), $((a + b)/2, Y_1)$, (b, Y_2).

To verify the first lemma, the formula $y = x(Ax + B)$ is tested to go through the given data points $(0, 0)$, $(1, y_1)$ and $(2, 2y_2)$. For example, the last pair is tested by the steps

$y(2) = 2(2A + B)$	Apply $y = x(Ax + B)$ with $x = 2$.
$= 4y_2 - 4y_1 + 4y_1 - 2y_2$	Use $A = y_2 - y_1$ and $B = 2y_1 - y_2$.
$= 2y_2$.	Therefore, the quadratic fits data item $(2, 2y_2)$.

The other two data items are tested similarly, details omitted here.

To verify the second lemma, observe that it is just a change of variables in the first lemma, $Y = Y_0 + y$. The data fit is checked as follows:

$Y(b) = Y_0 + y(2)$	Apply formulas $Y(X) = Y_0 + y(x)$, $y(x) = x(Ax+B)$ with $X = b$ and $x = 2$.
$= Y_0 + 2y_2$	Apply data fit $y(2) = 2y_2$.
$= Y_2$.	The quadratic fits the data item (b, Y_2).

The other two items are checked similarly, details omitted here. This completes the proof of the two lemmas. The formula for Q is obtained from the second lemma as $Q = Y_0 + Bx + Ax^2$ with substitutions for A, B and x performed to obtain the given equation for Q in terms of Y_0, Y_1, Y_2, a, b and X. ■

Justification of Table 1: The method of quadrature applied to $y' = F(x)$, $y(x_0) = y_0$ gives an explicit solution $y(x)$ involving the integral of F. Specialize this solution formula to $x = x_0 + h$ where $h > 0$. Then

$$y(x_0 + h) = y_0 + \int_{x_0}^{x_0+h} F(t)dt.$$

All three methods in Table 1 are derived by replacement of the integral above by the corresponding approximation taken from the rectangular, trapezoidal or Simpson method on page 232. For example, the trapezoidal method gives

$$\int_{x_0}^{x_0+h} F(t)dt \approx \frac{h}{2}\left(F(x_0) + F(x_0 + h)\right),$$

whereupon replacement into the formula for y gives the entry in Table 1 as

$$Y \approx y(x_0 + h) \approx y_0 + \frac{h}{2}\left(F(x_0) + F(x_0 + h)\right).$$

This completes the justification of Table 1.

[1] What's a lemma? It's a helper theorem, used to dissect long proofs into short pieces.

Exercises 4.1

Connect-the-Dots

Make a numerical table of 6 rows and a connect-the-dots graphic for exercises 1-10.

1. $y = 2x + 5$, $x = 0$ to $x = 1$

2. $y = 3x + 5$, $x = 0$ to $x = 2$

3. $y = 2x^2 + 5$, $x = 0$ to $x = 1$

4. $y = 3x^2 + 5$, $x = 0$ to $x = 2$

5. $y = \sin x$, $x = 0$ to $x = \pi/2$

6. $y = \sin 2x$, $x = 0$ to $x = \pi/4$

7. $y = x \ln|1 + x|$, $x = 0$ to $x = 2$

8. $y = x \ln|1 + 2x|$, $x = 0$ to $x = 1$

9. $y = xe^x$, $x = 0$ to $x = 1$

10. $y = x^2e^x$, $x = 0$ to $x = 1/2$

Rectangular Rule

Apply the rectangular rule to make an xy-table for $y(x)$ with 11 rows, $h = 0.1$. Graph the approximate solution and the exact solution. Follow example 4.1.

11. $y' = 2x$, $y(0) = 5$.

12. $y' = 3x^2$, $y(0) = 5$.

13. $y' = 3x^2 + 2x$, $y(0) = 4$.

14. $y' = 3x^2 + 4x^3$, $y(0) = 4$.

15. $y' = \sin x$, $y(0) = 1$.

16. $y' = 2 \sin 2x$, $y(0) = 1$.

17. $y' = \ln(1 + x)$, $y(0) = 1$. Exact $(1 + x) \ln|1 + x| + 1 - x$.

18. $y' = 2 \ln(1 + 2x)$, $y(0) = 1$. Exact $(1 + 2x) \ln|1 + 2x| + 1 - 2x$.

19. $y' = xe^x$, $y(0) = 1$. Exact $xe^x - e^x + 2$.

20. $y' = 2x^2e^{2x}$, $y(0) = 4$. Exact $2\,x^2e^x - 4\,xe^x + 4\,e^x$.

Trapezoidal Rule

Apply the trapezoidal rule to make an xy-table for $y(x)$ with 6 rows and step size $h = 0.2$. Graph the approximate solution and the exact solution. Follow example 4.2.

21. $y' = 2x$, $y(0) = 1$.

22. $y' = 3x^2$, $y(0) = 1$.

23. $y' = 3x^2 + 2x$, $y(0) = 2$.

24. $y' = 3x^2 + 4x^3$, $y(0) = 2$.

25. $y' = \sin x$, $y(0) = 4$.

26. $y' = 2 \sin 2x$, $y(0) = 4$.

27. $y' = \ln(1 + x)$, $y(0) = 1$. Exact $(1 + x) \ln|1 + x| + 1 - x$.

28. $y' = 2 \ln(1 + 2x)$, $y(0) = 1$. Exact $(1 + 2x) \ln|1 + 2x| + 1 - 2x$.

29. $y' = xe^x$, $y(0) = 1$. Exact $xe^x - e^x + 2$.

30. $y' = 2x^2e^{2x}$, $y(0) = 4$. Exact $2\,x^2e^x - 4\,xe^x + 4\,e^x$.

Simpson Rule

Apply Simpson's rule to make an xy-table for $y(x)$ with 6 rows and step size $h = 0.2$. Graph the approximate solution and the exact solution. Follow example 4.3.

31. $y' = 2x$, $y(0) = 2$.

32. $y' = 3x^2$, $y(0) = 2$.

33. $y' = 3x^2 + 2x$, $y(0) = 3$.

34. $y' = 3x^2 + 4x^3$, $y(0) = 3$.

35. $y' = \sin x$, $y(0) = 5$.

36. $y' = 2 \sin 2x$, $y(0) = 5$.

37. $y' = \ln(1 + x)$, $y(0) = 1$. Exact $(1 + x) \ln|1 + x| + 1 - x$.

38. $y' = 2 \ln(1 + 2x)$, $y(0) = 1$. Exact $(1 + 2x) \ln|1 + 2x| + 1 - 2x$.

39. $y' = xe^x$, $y(0) = 1$. Exact $xe^x - e^x + 2$.

40. $y' = 2x^2e^{2x}$, $y(0) = 4$. Exact $2\,x^2e^x - 4\,xe^x + 4\,e^x$.

Simpson's Rule

The following exercises use formulas and techniques found in the proof on page 234 and in Example 4.4, page 233.

41. Verify with Simpson's rule (5) for cubic polynomials the equality $\int_1^2 (x^3 + 16x^2 + 4)dx = 541/12$.

42. Verify with Simpson's rule (5) for cubic polynomials the equality $\int_1^2 (x^3 + x + 14)dx = 77/4$.

43. Let $f(x)$ satisfy $f(0) = 1$, $f(1/2) = 6/5$, $f(1) = 3/4$. Apply Simpson's rule with one division to verify that $\int_0^1 f(x)dx \approx 131/120$.

44. Let $f(x)$ satisfy $f(0) = -1$, $f(1/2) = 1$, $f(1) = 2$. Apply Simpson's rule with one division to verify that $\int_0^1 f(x)dx \approx 5/6$.

45. Verify Simpson's equality (5), assuming $Q(x) = 1$ and $Q(x) = x$.

46. Verify Simpson's equality (5), assuming $Q(x) = x^2$. Use college algebra identity $u^3 - v^3 = (u - v)(u^2 + uv + v^2)$.

Quadratic Interpolation

The following exercises use formulas and techniques from the proof on page 234.

47. Verify directly that the quadratic polynomial $y = x(7 - 4x)$ goes through the points $(0, 0)$, $(1, 3)$, $(2, -2)$.

48. Verify directly that the quadratic polynomial $y = x(8 - 5x)$ goes through the points $(0, 0)$, $(1, 3)$, $(2, -4)$.

49. Compute the quadratic interpolation polynomial $Q(x)$ which goes through the points $(0, 1)$, $(0.5, 1.2)$, $(1, 0.75)$.

50. Compute the quadratic interpolation polynomial $Q(x)$ which goes through the points $(0, -1)$, $(0.5, 1)$, $(1, 2)$.

51. Verify the remaining cases in Lemma 4.1, page 235.

52. Verify the remaining cases in Lemma 4.2, page 235.

4.2 Solving $y' = f(x, y)$ Numerically

The numerical solution of the initial value problem

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0 \tag{1}$$

is studied here by three basic methods. In each case, the current table entry x_0, y_0 plus step size h is used to find the next table entry X, Y. *Define* $X = x_0 + h$ and let Y be defined below, according to the algorithm selected (Euler, Heun, RK4)[2]. The *motivation* for the three methods appears on page 244.

Euler's Method

$$Y = y_0 + hf(x_0, y_0). \tag{2}$$

Heun's Method

$$\begin{aligned} y_1 &= y_0 + hf(x_0, y_0), \\ Y &= y_0 + \frac{h}{2}\left(f(x_0, y_0) + f(x_0 + h, y_1)\right). \end{aligned} \tag{3}$$

Runge-Kutta RK4 Method

$$\begin{aligned} k_1 &= hf(x_0, y_0), \\ k_2 &= hf(x_0 + h/2, y_0 + k_1/2), \\ k_3 &= hf(x_0 + h/2, y_0 + k_2/2), \\ k_4 &= hf(x_0 + h, y_0 + k_3), \\ Y &= y_0 + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}. \end{aligned} \tag{4}$$

The last quantity Y contains an average of six terms, where two appear in duplicate: $(k_1 + k_2 + k_2 + k_3 + k_3 + k_4)/6$. A similar average appears in Simpson's rule.

Relationship to Calculus Methods

If the differential equation (1) is specialized to the equation $y'(x) = F(x)$, $y(x_0) = y_0$, to agree with the previous section, then $f(x, y) = F(x)$ is independent of y and the three methods of Euler, Heun and RK4 reduce to the rectangular, trapezoidal and Simpson rules.

[2] Euler is pronounced *oiler*. Heun rhymes with *coin*. Runge rhymes with *run key*.

To justify the reduction in the case of Heun's method, start with the assumption $f(x,y) = F(x)$ and observe that by independence of y, variable y_1 is never used. Compute as follows:

$$\begin{aligned} Y &= y_0 + \tfrac{h}{2}\left(f(x_0,y_0) + f(x_0+h,y_1)\right) && \text{Apply equation (3).} \\ &= y_0 + \tfrac{h}{2}\left(F(x_0) + F(x_0+h)\right). && \text{Use } f(x,y) = F(x). \end{aligned}$$

The right side of the last equation is exactly the trapezoidal rule.

Examples and Methods

Example 4.5 (Euler's Method)
Solve $y' = -y + 1 - x$, $y(0) = 3$ by Euler's method for $x = 0$ to $x = 1$ in steps of $h = 0.1$. Produce a table of values which compares approximate and exact solutions. Graph both the exact solution $y = 2 - x + e^{-x}$ and the approximate solution.

Solution: **Exact solution**. The homogeneous solution is $y_h = ce^{-x}$. A particular solution $y_p = 2 - x$ is found by the method of undetermined coefficients or the linear integrating factor method. The general solution $y_h + y_p$ is then $y(x) = ce^{-x} + 2 - x$. Initial condition $y(0) = 3$ gives $c = 1$ and then $y = 2 - x + e^{-x}$.

Approximate Solution. The table of xy-values starts because of $y(0) = 3$ with the two values $X = 0$, $Y = 3$. Throughout, $f(x,y) = -y + 1 - x =$ RHS of the differential equation. The X-values will be $X = 0$ to $X = 1$ in increments of $h = 1/10$, making 11 rows total. The Y-values are computed from

$$\begin{aligned} Y &= y_0 + hf(x_0,y_0) && \text{Euler's method.} \\ &= y_0 + h(-y_0 + 1 - x_0) && \text{Use } f(x,y) = -y + 1 - x. \\ &= 0.9y_0 + 0.1(1 - x_0) && \text{Use } h = 0.1. \end{aligned}$$

The pair x_0, y_0 represents the two entries in the current row of the table. The next table pair X, Y is given by $X = x_0 + h$, $Y = 0.9y_0 + 0.1(1 - x_0)$. It is normal in a computation to do the *second pair* by hand, then use computing machinery to reproduce the hand result and finish the computation of the remaining table rows. Here's the second pair:

$$\begin{aligned} X &= x_0 + h && \text{Definition of } X\text{-values.} \\ &= 0.1, && \text{Substitute } x_0 = 0 \text{ and } h = 0.1. \\ Y &= 0.9y_0 + 0.1(1 - x_0), && \text{The simplified recurrence.} \\ &= 0.9(3) + 0.1(1 - 0) && \text{Substitute for row 1, } x_0 = 0,\ y_0 = 3. \\ &= 2.8. && \text{Second row found: } X = 0.1,\ Y = 2.8. \end{aligned}$$

By the same process, the third row is $X = 0.2$, $Y = 2.61$. This gives the xy-table below, in which the exact values from $y = 2 - x + e^{-x}$ are also tabulated.

Table 5. Euler's Method Applied with $h = 0.1$ on $0 \le x \le 1$ to the Problem $y' = -y + 1 - x$, $y(0) = 3$.

x	y-Euler	y-Exact
0.0	3.00000	3.0000000
0.1	2.80000	2.8048374
0.2	2.61000	2.6187308
0.3	2.42900	2.4408182
0.4	2.25610	2.2703200
0.5	2.09049	2.1065307
0.6	1.93144	1.9488116
0.7	1.77830	1.7965853
0.8	1.63047	1.6493290
0.9	1.48742	1.5065697
1.0	1.34868	1.3678794

See page 241 for `maple` code which automates Euler's method. The approximate solution graphed in Figure 6 is nearly identical to the exact solution $y = 2 - x + e^{-x}$. The `maple` plot code for Figure 6:

```
L:=[0.0,3.00000],[0.1,2.80000],[0.2,2.61000],[0.3,2.42900],
   [0.4,2.25610],[0.5,2.09049],[0.6,1.93144],[0.7,1.77830],
   [0.8,1.63047],[0.9,1.48742],[1.0,1.34868]:
plot({[L],2-x+exp(-x)},x=0..1);
```

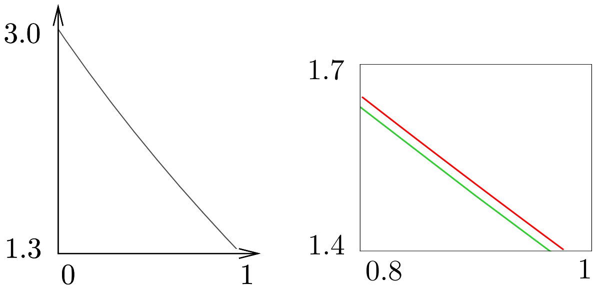


Figure 6. Numerical Solution of $y' = -y+1-x$, $y(0) = 3$
The Euler approximate solution on $[0, 1]$ is the black curve on the left. The exact solution $y = 2 - x + e^{-x}$ is the upper red curve on the right. The approximate solution is the lower green curve on the right.

Example 4.6 (Euler and Heun Methods)

Solve $y' = -y + 1 - x$, $y(0) = 3$ by both Euler's method and Heun's method for $x = 0$ to $x = 1$ in steps of $h = 0.1$. Produce a table of values which compares approximate and exact solutions.

Solution: **Table of xy-values**. The Euler method was applied in Example 4.5. Heun's method will be documented here. The first pair is 0, 3. The second pair X, Y will be computed by hand calculation below. Throughout, $f(x, y) = -y + 1 - x = \text{RHS}$ of the differential equation.

$X = x_0 + h$	Definition of X-values.
$\quad = 0.1,$	Substitute $x_0 = 0$ and $h = 0.1$.
$Y_1 = y_0 + hf(x_0, y_0)$	First Heun formula.
$\quad = y_0 + 0.1(-y_0 + 1 - x_0)$	Use $f(x, y) = -y + 1 - x$.
$\quad = 2.8,$	Row 1 gives x_0, y_0. Same as the Euler method value.
$Y = y_0 + h(f(x_0, y_0) + f(x_0 + h, Y_1))/2,$	Second Heun formula.
$\quad = 3 + 0.05(-3 + 1 - 0 - 2.8 + 1 - 0.1)$	Use $x_0 = 0$, $y_0 = 3$, $Y_1 = 2.8$.

$= 2.805.$

Therefore, the second row is $X = 0.1$, $Y = 2.805$. By the same process, the third row is $X = 0.2$, $Y = 2.619025$. This gives the xy-table below, in which the Euler approximate values and the exact values from $y = 2 - x + e^{-x}$ are also tabulated, taken from the preceding example.

Table 6. Euler and Heun methods Applied with $h = 0.1$ on $0 \le x \le 1$ to the Problem $y' = -y + 1 - x$, $y(0) = 3$.

x	y-Euler	y-Heun	y-Exact
0.0	3.00000	3.00000	3.0000000
0.1	2.80000	2.80500	2.8048374
0.2	2.61000	2.61903	2.6187308
0.3	2.42900	2.44122	2.4408182
0.4	2.25610	2.27080	2.2703200
0.5	2.09049	2.10708	2.1065307
0.6	1.93144	1.94940	1.9488116
0.7	1.77830	1.79721	1.7965853
0.8	1.63047	1.64998	1.6493290
0.9	1.48742	1.50723	1.5065697
1.0	1.34868	1.36854	1.3678794

Computer algebra system. The implementation for `maple` appears below. Part of the interface is execution of a group, which is used here to divide the algorithm into three distinct parts. The code produces a list `L` which contains Euler (left panel) or Heun (right panel) approximations.

```
# Euler algorithm
# Group 1, initialize.
f:=(x,y)->-y+1-x:
x0:=0:y0:=3:h:=.1:L:=[x0,y0]:
# Group 2, loop count = 10
for i from 1 to 10 do
Y:=y0+h*f(x0,y0):
x0:=x0+h:y0:=Y:L:=L,[x0,y0];
end do;
# Group 3, plot.
plot([L]);
```

```
# Heun algorithm
# Group 1, initialize.
f:=(x,y)->-y+1-x:
x0:=0:y0:=3:h:=.1:L:=[x0,y0]
# Group 2, loop count = 10
for i from 1 to 10 do
Y:=y0+h*f(x0,y0):
Y:=y0+h*(f(x0,y0)+f(x0+h,Y))/2:
x0:=x0+h:y0:=Y:L:=L,[x0,y0];
end do;
# Group 3, plot.
plot([L]);
```

Numerical laboratory. The implementation of the Heun method for `matlab`, `octave` and `scilab` will be described. The code is written into files `f.m` and `heun.m`, which must reside in a default directory. Then `[X,Y]=heun(0,3,1,10)` produces the xy-table. The graphic is made with `plot(X,Y)`.

File `f.m`:

```
function yp = f(x,y)
yp= -y+1-x;
```

File `heun.m`:

```
function [X,Y] = heun(x0,y0,x1,n)
h=(x1-x0)/n;X=x0;Y=y0;
for i=1:n;
y1= y0+h*f(x0,y0);
y0= y0+h*(f(x0,y0)+f(x0+h,y1))/2;
x0=x0+h;
X=[X;x0];Y=[Y;y0];
end
```

Example 4.7 (Euler, Heun and RK4 Methods)

Solve the initial value problem $y' = -y + 1 - x$, $y(0) = 3$ by Euler's method, Heun's method and the RK4 method for $x = 0$ to $x = 1$ in steps of $h = 0.1$. Produce a table of values which compares approximate and exact solutions.

Solution: **Table of xy-values**. The Euler and Heun methods were applied in Examples 4.5, 4.6. The Runge-Kutta method (RK4) will be illustrated here. The first pair is 0, 3. The second pair X, Y will be computed by hand calculator.

$X = x_0 + h$	Definition of X-values.
$= 0.1$,	Substitute $x_0 = 0$ and $h = 0.1$.
$k_1 = hf(x_0, y_0)$	First RK4 formula.
$= 0.1(-y_0 + 1 - x_0)$	Use $f(x, y) = -y + 1 - x$.
$= -0.2$,	Row 1 supplies $x_0 = 0$, $y_0 = 3$.
$k_2 = hf(x_0 + h/2, y_0 + k_1/2)$	Second RK4 formula.
$= 0.1f(0.05, 2.9)$	
$= -0.195$,	
$k_3 = hf(x_0 + h/2, y_0 + k_2/2)$	Third RK4 formula.
$= 0.1f(0.05, 2.9025)$	
$= -0.19525$,	
$k_4 = hf(x_0 + h, y_0 + k_3)$	Fourth RK4 formula.
$= 0.1f(0.1, 2.80475)$	
$= -0.190475$,	
$Y = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_2 + k_4)$,	Last RK4 formula.
$= 3 + \frac{1}{6}(-1.170975)$	Use $x_0 = 0$, $y_0 = 3$, $Y_1 = 2.8$.
$= 2.8048375$.	

Therefore, the second row is $X = 0.1$, $Y = 2.8048375$. Continuing, the third row is $X = 0.2$, $Y = 2.6187309$. The Euler and Heun steps were done in the previous example and recorded in Table 6. We have computed by hand calculator the first three rows of the computer-generated xy-table below, in which exact values $y = 2 - x + e^{-x}$ are also tabulated.

Table 7. Euler, Heun and RK4 methods Applied with $h = 0.1$ on $0 \le x \le 1$ to the Problem $y' = -y + 1 - x$, $y(0) = 3$.

x	y-Euler	y-Heun	y-RK4	y-Exact
0.0	3.00000	3.00000	3.0000000	3.0000000
0.1	2.80000	2.80500	2.8048375	2.8048374
0.2	2.61000	2.61903	2.6187309	2.6187308
0.3	2.42900	2.44122	2.4408184	2.4408182
0.4	2.25610	2.27080	2.2703203	2.2703200
0.5	2.09049	2.10708	2.1065309	2.1065307
0.6	1.93144	1.94940	1.9488119	1.9488116
0.7	1.77830	1.79721	1.7965856	1.7965853
0.8	1.63047	1.64998	1.6493293	1.6493290
0.9	1.48742	1.50723	1.5065700	1.5065697
1.0	1.34868	1.36854	1.3678798	1.3678794

Computer algebra system. The implementation of RK4 for `maple` appears below, as a modification of the code for Example 4.6.

```
# Group 2, loop count = 10
for i from 1 to 10 do
k1:=h*f(x0,y0):
k2:=h*f(x0+h/2,y0+k1/2):
k3:=h*f(x0+h/2,y0+k2/2):
k4:=h*f(x0+h,y0+k3):
Y:=y0+(k1+2*k2+2*k3+k4)/6:
x0:=x0+h:y0:=Y:L:=L,[x0,y0];
end do;
```

In the special case $f(x, y) = F(x)$ (independent of y), the computer code reduces to a poor implementation of Simpson's Rule for $\int_a^{a+h} F(x)dx$. The wasted effort is calculation of k_3, because k_2, k_3 are the same for $f(x, y) = F(x)$.

Numerical laboratory. The implementation of RK4 for `matlab`, `octave` and `scilab` appears below, to be added to the code for Example 4.6. The code is written into file `rk4.m`, which must reside in a default directory. The xy-table is produced by `[X,Y]=rk4(0,3,1,10)`.

```
function [X,Y] = rk4(x0,y0,x1,n)
h=(x1-x0)/n;X=x0;Y=y0;
for i=1:n;
 k1=h*f(x0,y0);
 k2=h*f(x0+h/2,y0+k1/2);
 k3=h*f(x0+h/2,y0+k2/2);
 k4=h*f(x0+h,y0+k3);
 y0=y0+(k1+2*k2+2*k3+k4)/6;
 x0=x0+h;
 X=[X;x0];Y=[Y;y0];
end
```

Motivation for the Three Methods

The entry point to the study is the equivalent integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt. \tag{5}$$

The ideas can be explained by replacement of the integral in (5) by the rectangular, trapezoidal or Simpson rule. Unknown values of y that appear are subsequently replaced by suitable approximations.

These approximations, originating with L. Euler, are known as **predictors** and **correctors**. They are defined as follows from the integral formula

$$y(b) = y(a) + \int_{a}^{b} f(x, y(x))dx, \tag{6}$$

by assuming the integrand is a constant C.

Predictor $Y = y(a) + (b-a)f(a, Y^*)$. Given an estimate or an exact value Y^* for $y(a)$, then variable Y predicts $y(b)$. The approximation assumes the integrand in (6) constantly $C = f(a, Y^*)$.

Corrector $Y = y(a) + (b-a)f(b, Y^{**})$. Given an estimate or an exact value Y^{**} for $y(b)$, then variable Y corrects $y(b)$. The approximation assumes the integrand in (6) constantly $C = f(b, Y^{**})$.

Euler's method. Replace in (5) $x = x_0 + h$ and apply the rectangular rule to the integral. The resulting approximation is known as **Euler's method**:

$$y(x_0 + h) \approx Y = y_0 + hf(x_0, y_0). \tag{7}$$

Heun's method. Replace in (5) $x = x_0 + h$ and apply the trapezoidal rule to the integral, to get

$$y(x_0 + h) \approx y_0 + \frac{h}{2}\left(f(x_0, y(x_0) + f(x_0 + h, y(x_0 + h))\right).$$

The troublesome expressions are $y(x_0)$ and $y(x_0 + h)$. The first is y_0. The second can be estimated by the **predictor** $y_0 + hf(x_0, y_0)$. The resulting approximation is known as **Heun's method** or the **Modified Euler method**:

$$\begin{aligned} &Y_1 = y_0 + hf(x_0, y_0), \\ &y(x_0 + h) \approx Y = y_0 + \frac{h}{2}\left(f(x_0, y_0) + f(x_0 + h, Y_1)\right). \end{aligned} \tag{8}$$

RK4 method. Replace in (5) $x = x_0 + h$ and apply Simpson's rule to the integral. This gives $y(x_0 + h) \approx y_0 + S$ where the Simpson estimate S is given by

$$S = \frac{h}{6}\left(f(x_0, y(x_0) + 4f(M, y(M)) + f(x_0 + h, y(x_0 + h))\right) \tag{9}$$

and $M = x_0 + h/2$ is the midpoint of $[x_0, x_0 + h]$. The troublesome expressions in S are $y(x_0)$, $y(M)$ and $y(x_0 + h)$. The work of Runge and Kutta shows that

- Expression $y(x_0)$ is replaced by y_0.
- Expression $y(M)$ can be replaced by either Y_1 or Y_2, where $Y_1 = y_0 + 0.5hf(x_0, y_0)$ is a **predictor** and $Y_2 = y_0 + 0.5hf(M, Y_1)$ is a **corrector**.
- Expression $y(x_0 + h)$ can be replaced by $Y_3 = y_0 + hf(M, Y_2)$. This replacement arises from the **predictor** $y(x_0 + h) \approx y(M) + 0.5hf(M, y(M))$ by using **corrector** $y(M) \approx y_0 + 0.5hf(M, y(M))$ and then replacing $y(M)$ by Y_2.

The formulas of Runge-Kutta result by using the above replacements for $y(x_0)$, $y(M)$ and $y(x_0 + h)$, with the caveat that $f(M, y(M))$ gets replaced by the **average** of $f(M, Y_1)$ and $f(M, Y_2)$. In detail,

$$\begin{aligned}
6S &= hf(x_0, y(x_0)) + 4hf(M, y(M)) + hf(x_0 + h, y(x_0 + h)) \\
&\approx hf(x_0, y_0) + 4h\frac{f(M, Y_1) + f(M, Y_2)}{2} + hf(x_0 + h, Y_3) \\
&= k_1 + 2k_2 + 2k_3 + k_4
\end{aligned}$$

where the RK4 quantities k_1, k_2, k_3, k_4 are defined by (4), page 238. The resulting approximation is known as the **RK4 method**. Justification uses multivariable Taylor remainder formulas. See Burden-Faires [BurFair] p 229 its references.

Exercises 4.2

Euler's Method

Apply Euler's method to make an xy-table for $y(x)$ with 11 rows and step size $h = 0.1$. Graph the approximate solution and the exact solution. Follow Example 4.5.

1. $y' = 2 + y$, $y(0) = 5$. Exact $y(x) = -2 + 7e^x$.

2. $y' = 3 + y$, $y(0) = 5$. Exact $y(x) = -3 + 8e^x$.

3. $y' = e^{-x} + y$, $y(0) = 4$. Exact $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$.

4. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact $y(x) = -e^{-2x} + 5e^x$.

5. $y' = y \sin x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos x}$.

6. $y' = 2y \sin 2x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos 2x}$.

7. $y' = y/(1 + x)$, $y(0) = 1$. Exact $y(x) = 1 + x$.

8. $y' = y(x)/(1 + 2x)$, $y(0) = 1$. Exact $y(x) = \sqrt{1 + 2x}$.

9. $y' = yxe^x$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = 1 + (x - 1)e^x$.

10. $y' = 2y(x^2 + x)e^{2x}$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = x^2e^{2x}$.

Heun's Method

Apply Heun's method to make an xy-table for $y(x)$ with 6 rows and step size $h = 0.2$. Graph the approximate solution and the exact solution. Follow Example 4.6.

11. $y' = 2 + y$, $y(0) = 5$. Exact $y(x) = -2 + 7e^x$.

12. $y' = 3 + y$, $y(0) = 5$. Exact $y(x) = -3 + 8e^x$.

13. $y' = e^{-x} + y$, $y(0) = 4$. Exact $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$.

14. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact $y(x) = -e^{-2x} + 5e^x$.

15. $y' = y\sin x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos x}$.

16. $y' = 2y\sin 2x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos 2x}$.

17. $y' = y/(1 + x)$, $y(0) = 1$. Exact $y(x) = 1 + x$.

18. $y' = y(x)/(1 + 2x)$, $y(0) = 1$. Exact $y(x) = \sqrt{1 + 2x}$.

19. $y' = yxe^x$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = 1 + (x - 1)e^x$.

20. $y' = 2y(x^2 + x)e^{2x}$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = x^2e^{2x}$.

RK4 Method

Apply the Runge-Kutta method (RK4) to make an xy-table for $y(x)$ with 6 rows and step size $h = 0.2$. Graph the approximate solution and the exact solution. Follow Example 4.7.

21. $y' = 2 + y$, $y(0) = 5$. Exact $y(x) = -2 + 7e^x$.

22. $y' = 3 + y$, $y(0) = 5$. Exact $y(x) = -3 + 8e^x$.

23. $y' = e^{-x} + y$, $y(0) = 4$. Exact $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$.

24. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact $y(x) = -e^{-2x} + 5e^x$.

25. $y' = y\sin x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos x}$.

26. $y' = 2y\sin 2x$, $y(0) = 1$. Exact $y(x) = e^{1-\cos 2x}$.

27. $y' = y/(1 + x)$, $y(0) = 1$. Exact $y(x) = 1 + x$.

28. $y' = y(x)/(1 + 2x)$, $y(0) = 1$. Exact $y(x) = \sqrt{1 + 2x}$.

29. $y' = yxe^x$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = 1 + (x - 1)e^x$.

30. $y' = 2y(x^2 + x)e^{2x}$, $y(0) = 1$. Exact $y(x) = e^{u(x)}$, $u(x) = x^2e^{2x}$.

Euler and RK4 Methods

Apply the Euler method and the Runge-Kutta method (RK4) to make a table with 6 rows and step size $h = 0.1$. The table columns are x, y_1, y_2, y where y_1 is the Euler approximation, y_2 is the RK4 approximation and y is the exact solution. Graph y_1, y_2, y.

31. $y' = \frac{1}{2}(y - 2)^2$, $y(0) = 3$. Exact $y(x) = \dfrac{2x - 6}{x - 2}$.

32. $y' = \frac{1}{2}(y - 3)^2$, $y(0) = 4$. Exact $y(x) = \dfrac{3x - 8}{x - 2}$.

33. $y' = x^3/y^2$, $y(2) = 3$. Exact $y(x) = \frac{1}{2}\sqrt[3]{6x^4 + 120}$.

34. $y' = x^5/y^2$, $y(2) = 3$. Exact $y(x) = \frac{1}{2}\sqrt[3]{4x^6 - 40}$.

35. $y' = 2x(1 + y^2)$, $y(0) = 1$. Exact $y(x) = \tan(x^2 + \pi/4)$.

36. $y' = 3y^{2/3}$, $y(0) = 1$. Exact $y(x) = (x + 1)^3$.

37. $y' = 1 + y^2$, $y(0) = 0$. Exact $y(x) = \tan x$.

38. $y' = 1 + y^2$, $y(0) = 1$. Exact $y(x) = \tan(x + \pi/4)$.

4.3 Error in Numerical Methods

Numerical Errors

Studied here are *cumulative error*, *local error*, *roundoff error* and *truncation error*. The Landau order notation is introduced.

Cumulative Error

This error measurement is commonly used in displays like Table 8, in which approximate and exact solution columns already appear. In such applications, the cumulative error is the difference of the approximate and exact columns. The **exact solution** refers to $y(x)$ defined by $y' = f(x,y)$, $y(x_0) = y_0$ ($x_0 = 0$, $y_0 = 3$ from line 1 of Table 8). The **approximate solution** refers to the y-values computed by the algorithm (column 2 in Table 8). A precise definition of the **cumulative error** E is given in terms of the exact solution $y(x)$: *given table entry* X, Y, *then* $E = |y(X) - Y|$.

Table 8. Cumulative Error.
A third column, cumulative error, is added to an existing xy-table of approximate and exact solutions. The cumulative error is computed by the formula $E = |y_2 - y_1|$, where y_1 is the approximation and y_2 is the exact value.

x	y-Approx	y-Exact	Error
0.0	3.00000	3.0000000	0.0000000
0.1	2.80000	2.8048374	0.0048374
0.2	2.61000	2.6187308	0.0087308
0.3	2.42900	2.4408182	0.0118182

Local Error

This error is made by one algorithm step in going from table entry x_1, y_1 to the next table entry x_2, y_2. It can be precisely defined in terms of the solution $u(x)$ to $u' = f(x,u)$, $u(x_1) = y_1$ by the formula

$$E_{\text{loc}} = |u(x_2) - y_2|.$$

Noteworthy is that $u(x) \neq y(x)$. To explain, the exact solution $y(x)$ solves $y' = f(x,y)$, $y(x_0) = y_0$ where x_0, y_0 is the *first table entry*, while $u(x)$ solves $u' = f(x,u)$ for a *different* set of initial conditions. In particular, an xy-table of approximate and exact solution values, like Table 8, does not contain enough information to determine the local error!

To illustrate the ideas, consider $y' = 2y$, $y(0) = 1$ with exact solution $y = e^{2x}$.

Using Euler's method with step size $h = 0.1$ gives the table

x	y-approx	y-exact
0	1	1
0.1	1.2	1.2214028
0.2	1.44	1.4918247

To find the local error for line 2 to line 3 requires solving $u' = 2u$, $u(0.1) = 1.2$, and then evaluating $E = |u(0.2) - 1.4918247|$. We find that $u(x) = 1.2e^{2(x-0.1)}$ and then $E = |1.2e^{0.2} - 1.4918247| = 0.026141390$.

Roundoff Error

Also called rounding error, the roundoff error is the difference between the calculated approximation of a number to finitely many digits and its exact value in terms of infinitely many digits. The technical error is made by computers due to the representation of floating point numbers, which limits the number of significant digits in any computation. *Integer arithmetic* will normally generate no errors, unless **integer overflow** occurs, i.e., $x + y$ or xy can result in an integer larger than the machine can represent. *Floating point arithmetic* usually generates errors because of results that must be rounded to give a machine representation. To illustrate, 8-digit precision requires $a = 1.00000005$ be represented as $\hat{a} = 1.0000001$ and $b = 1.00000004$ be represented as $\hat{b} = 1$. Then $2a + 2b = 4.00000018$, which rounds to 4.0000002, while $2\hat{a} + 2\hat{b} = 4.0000001$. The roundoff error in this example is 0.0000001.

For numerical methods, this translates into *fewer* roundoff errors for $h = 0.1$ than for $h = 0.001$, because the number of arithmetic operations increases 1000-fold for $h = 0.001$. The payoff in increased accuracy expected for a change in step size from $h = 0.1$ to $h = 0.001$ may be less than theoretically possible, because the roundoff errors accumulate to cancel the effects of decreased step size. Positive and negative roundoff errors tend to cancel, leading to situations where a thousand-fold step size change causes only a thirty-fold change in roundoff error.

Truncation Error

It is typical in numerical mathematics to use formulas like $\pi = 3.14159$ or $e = 2.718$. These formulas **truncate** the actual decimal expansion, causing an error. **Truncation** is the term used for reducing the number of digits to the right of the decimal point, by discarding all digits past a certain point, e.g., 0.123456789 truncated to 5 digits is 0.12345. Common truncation errors are caused by dropping higher order terms in a Taylor series, or by approximating a nonlinear term by its linearization. In general, a truncation error is made whenever a formula is replaced by an approximate formula, in which case the formula is wrong even if computed exactly.

Landau Symbol

German mathematician **Edmund Landau** introduced a convenient notation to represent truncation errors. If f and g are defined near $h = 0$, then $f = \mathbf{O}(g)$ means that $|f(h)| \le K|g(h)|$ as $h \to 0$, for some constant K. The **Landau notation** $f = \mathbf{O}(g)$ is vocalized as f *equals big owe of* g. The symbol $\mathbf{O}(h^n)$ therefore stands for terms of order h^n. Taylor series expansions can then be referenced succinctly, e.g., $\sin h = h + \mathbf{O}(h^3)$, $e^h = 1 + h + \mathbf{O}(h^2)$, and so on. Some simple rules for the Landau symbol:

$$\mathbf{O}(h^n) + \mathbf{O}(h^m) = \mathbf{O}(h^{\min(n,m)}), \quad \mathbf{O}(h^n)\mathbf{O}(h^m) = \mathbf{O}(h^{n+m}).$$

Finite Blowup of Solutions

The solution $y = (1 - x)^{-1}$ for $y' = y^2$, $y(0) = 1$ exists on $0 \le x < 1$, but it becomes infinite at $x = 1$. The finite value $x = 1$ causes blowup of the y-value. This event is called **finite blowup**. Attempts to solve $y' = y^2$, $y(0) = 1$ numerically will fail near $x = 1$, and these errors will propagate past $x = 1$, if the numerical problem is allowed to be solved over an interval larger than $0 \le x < 1$.

Unfortunately, finite blowup cannot be detected in advance from smoothness of $f(x, y)$ or the fact that the problem is *applied.* For example, logistic population models $y' = y(a - by)$ typically have solutions with finite blowup, because the solution y is a fraction which can have a zero denominator at some instant x . On the positive side, there are three common conditions which guarantee no finite blowup:

- A linear equation $y' + p(x)y = q(x)$ does not exhibit finite blowup on the domain of continuity of $p(x)$ and $q(x)$.
- An equation $y' = f(x, y)$ does not exhibit finite blowup if f is continuous and $\max|f_y(x, y)| < \infty$.
- An equation $y' = f(x, y)$ does not exhibit finite blowup if f is continuous and f satisfies a Lipschitz condition $|f(x, y_1) - f(x, y_2)| \le M|y_1 - y_2|$ for some constant $M > 0$ and all x, y_1, y_2.

Numerical Instability

The equation $y' = y + 1 - x$ has solution $y = x + ce^x$. Attempts to solve for $y(0) = 1$ will meet with failure, because errors will cause the numerical solution to lock onto some solution with $c \ne 0$ and small, which causes the numerical solution to grow like e^x. In this case, the instability was caused by the problem itself.

Numerical instability can result even though the solution is physically stable. An example is $y' = -50(y - \sin x) + \cos x$, $y(0) = 0$. The general solution is

$y = ce^{-50x} + \sin x$ and $y(0) = 0$ gives $c = 0$. The negative exponential term is *transient* and $\sin x$ is the unique periodic *steady-state* solution. The solution is insensitive to minor changes in the initial condition. For popular numerical methods, the value at $x = 1$ seems to depend greatly on the step size, as is shown by Table 9.

Table 9. Cumulative Error at $x = 1$
Euler, Heun and RK4 Methods applied to $y' = -50(y - \sin x) + \cos x$, $y(0) = 0$, for various step sizes.

	$h = 0.1$	$h = 0.05$	$h = 0.02$	$h = 0.01$
Euler	40701.23	0.183e7	0.00008	0.00004
Heun	0.328e12	0.430e14	0.005	0.00004
RK4	0.318e20	0.219e18	0.00004	0.000001

The sensitivity to step size is due to the *algorithm* and not to instability of the problem.

Stiff Problems

The differential equation $y' = -50(y - \sin x) + \cos x$, which has solution $y = ce^{-50x} + \sin x$, is called **stiff**, a technical term defined precisely in advanced numerical analysis references, e.g., Burden-Faires [BurFair] and Cheney-Kincaid [Cheney-K]. Characteristically, it means that the equation has a solution $y(x)$ containing a transient term $y_1(x)$ with derivative $y_1'(x)$ tending slowly to zero. For instance, if $y(x)$ has a term like $y_1(x) = ce^{-50x}$, then the derivative $y_1'(x)$ is approximately 50 times larger ($y_1'/y_1 \approx -50$). Applications with transient terms of Landau order e^{-at} are stiff when a is large. Stiff problems occupy an active branch of research in applied numerical analysis. Researchers call a problem **stiff** provided certain numerical methods for it are unstable (e.g., inaccurate) unless the step size is taken to be extremely small.

Cumulative Error Estimates

It is possible to give theoretical but not practical estimates for the cumulative error in the case of Euler's method, Heun's method and the RK4 method. Applied literature and computer documentation often contain references to these facts, typically in the following succinct form.

- Euler's method has order 1.
- Heun's method has order 2.
- The Runge-Kutta method (RK4) has order 4.

The *exact meaning* of these statements is given below in the theorems. The phrase **Order** n in this context refers to Edmund Landau's order notation $\mathbf{O}(h^n)$. In particular, *order* 2 means $\mathbf{O}(h^2)$.

In practical terms, the statements measure the quality and accuracy of the algorithms themselves, and hence establish an expectation of performance from each algorithm. They *do not mean* that step size $h = 0.001$ gives three digits of accuracy in the computed answer! The meaning is that repeated halving of the step size will result in three digits of accuracy, eventually. Most persons half the step size until the first three digits repeat, then they take this to be the optimal step size for three-digit accuracy. The theorems don't say that this practice is correct, only that for *some step size* it is correct.

Theorem 4.1 (Euler's Method Error)
Let the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ have a solution $y(x)$ in the region $x_0 \le x \le x_0 + H$, $|y - y_0| \le K$ and assume that f, f_x and f_y are continuous. Then the cumulative error $E(x_0 + nh)$ at step n, $nh \le H$, made by Euler's method using step size h satisfies $E(x_0 + nh) \le Ch$. The constant C depends only on x_0, y_0, H, K, f, f_x and f_y. See Cheney–Kinkaid [Cheney-K] and Burden–Faires [BurFair].

Theorem 4.2 (Heun Method Error)
Let the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ have a solution in the region $x_0 \le x \le x_0 + H$, $|y - y_0| \le K$. Assume f is continuous with continuous partials to order 3. Then the cumulative error $E(x_0 + nh)$ at step n, $nh \le H$, made by Heun's method using step size h, satisfies $E(x_0 + nh) \le Ch^2$. The constant C depends only on x_0, y_0, H, K, f and the partials of f to order 3.

Theorem 4.3 (RK4 Method Error)
Let the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ have a solution $y(x)$ in the region $x_0 \le x \le x_0 + H$, $|y - y_0| \le K$. Assume f is continuous with continuous partials to order 5. Then the cumulative error $E(x_0 + nh)$ at step n, $nh \le H$, made by the RK4 method using step size h, satisfies $E(x_0 + nh) \le Ch^4$. The constant C depends only on x_0, y_0, H, K, f, and the partials of f to order 5.

The last two results are implied by local truncation error estimates for Taylor's method of order n (section 5.3 in Burden-Faires [BurFair]).

Exercises 4.3

Cumulative Error

Make a table of 6 lines which has four columns x, y_1, y, E. Symbols y_1 and y are the approximate and exact solutions while $E = |y - y_1|$ is the cumulative error. Find y_1 using Euler's method in steps $h = 0.1$.

1. $y' = 2 + y$, $y(0) = 5$. Exact solution $y(x) = -2 + 7e^x$.

2. $y' = 3 + y$, $y(0) = 5$. Exact solution $y(x) = -3 + 8e^x$.

3. $y' = e^{-x} + y$, $y(0) = 4$. Exact solution $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$.

4. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact solution

$y(x) = -e^{-2x} + 5e^x$.

Local Error

Make a table of 4 lines which has four columns x, y_1, y, E. Symbols y_1 and y are the approximate and exact solutions while E is the local error. Find y_1 using Euler's method in steps $h = 0.1$. The general solution in each exercise is the solution for $y(0) = c$.

5. $y' = 2 + y$, $y(0) = 5$. General solution $y(x) = -2 + (2 + c)e^x$.

6. $y' = 3 + y$, $y(0) = 5$. General solution $y(x) = -3 + (3 + c)e^x$.

7. $y' = 2e^{-x} + y$, $y(0) = 4$. General solution $y(x) = -e^{-x} + (1 + c)e^x$.

8. $y' = 3e^{-2x} + y$, $y(0) = 4$. General solution $y(x) = -e^{-2x} + (1 + c)e^x$.

Roundoff Error

Compute the roundoff error for $y = 5a+4b$.

9. Assume 3-digit precision. Let $a = 0.0001$ and $b = 0.0003$.

10. Assume 3-digit precision. Let $a = 0.0002$ and $b = 0.0001$.

11. Assume 5-digit precision. Let $a = 0.000007$ and $b = 0.000003$.

12. Assume 5-digit precision. Let $a = 0.000005$ and $b = 0.000001$.

Truncation Error

Find the truncation error.

13. Truncate $x = 1.123456789$ to 3 digits right of the decimal point.

14. Truncate $x = 1.123456789$ to 4 digits right of the decimal point.

15. Truncate $x = 1.017171717$ to 7 digits right of the decimal point.

16. Truncate $x = 1.03939393939$ to 9 digits right of the decimal point.

Guessing the Step Size

Do a numerical experiment using the given method to estimate the number of steps needed to generate a numerical solution with 2-digit accuracy on $0 \le x \le 1$. The number reported, if increased, should not change the 2-digit accuracy.

17. $y' = 2 + y$, $y(0) = 5$. Exact solution $y(x) = -2 + 7e^x$. Euler's method.

18. $y' = 3 + y$, $y(0) = 5$. Exact solution $y(x) = -3 + 8e^x$. Euler's method

19. $y' = e^{-x} + y$, $y(0) = 4$. Exact solution $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$. Euler's method

20. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact solution $y(x) = -e^{-2x} + 5e^x$. Euler's method.

21. $y' = y/(1 + x)$, $y(0) = 1$. Exact solution $y(x) = 1 + x$. Euler's method.

22. $y' = y(x)/(1+2x)$, $y(0) = 1$. Exact solution $y(x) = \sqrt{1 + 2x}$. Euler's method.

23. $y' = 2 + y$, $y(0) = 5$. Exact solution $y(x) = -2 + 7e^x$. Heun's method.

24. $y' = 3 + y$, $y(0) = 5$. Exact solution $y(x) = -3 + 8e^x$. Heun's method

25. $y' = e^{-x} + y$, $y(0) = 4$. Exact solution $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$. Heun's method

26. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact solution $y(x) = -e^{-2x} + 5e^x$. Heun's method.

27. $y' = y/(1 + x)$, $y(0) = 1$. Exact solution $y(x) = 1 + x$. Heun's method.

28. $y' = y(x)/(1+2x)$, $y(0) = 1$. Exact solution $y(x) = \sqrt{1 + 2x}$. Heun's method.

29. $y' = 2 + y$, $y(0) = 5$. Exact solution $y(x) = -2 + 7e^x$. RK4 method.

30. $y' = 3 + y$, $y(0) = 5$. Exact solution $y(x) = -3 + 8e^x$. RK4 method

31. $y' = e^{-x} + y$, $y(0) = 4$. Exact solution $y(x) = -\frac{1}{2}e^{-x} + \frac{9}{2}e^x$. RK4 method

32. $y' = 3e^{-2x} + y$, $y(0) = 4$. Exact solution $y(x) = -e^{-2x} + 5e^{x}$. RK4 method.

33. $y' = y/(1 + x)$, $y(0) = 1$. Exact solution $y(x) = 1 + x$. RK4 method.

34. $y' = y(x)/(1 + 2x)$, $y(0) = 1$. Exact solution $y(x) = \sqrt{1 + 2x}$. RK4 method.

4.4 Computing π, $\ln 2$ and e

The approximations $\pi \approx 3.1415927$, $\ln 2 \approx 0.69314718$, $e \approx 2.7182818$ can be obtained by numerical methods applied to the following initial value problems:

$$(1) \qquad y' = \frac{4}{1+x^2}, \quad y(0) = 0, \quad \pi = y(1),$$

$$(2) \qquad y' = \frac{1}{1+x}, \quad y(0) = 0, \quad \ln 2 = y(1),$$

$$(3) \qquad y' = y, \quad y(0) = 1, \quad e = y(1).$$

Equations (1)–(3) *define* the constants π, $\ln 2$ and e through the corresponding initial value problems.

The third problem (3) requires a numerical method like RK4, while the other two can be solved using Simpson's quadrature rule. It is a fact that RK4 reduces to Simpson's rule for $y' = F(x)$, therefore, for simplicity, RK4 can be used for all three problems, ignoring speed issues. It will be seen that the choice of the DE-solver algorithm (e.g., RK4) affects computational accuracy.

Computing $\pi = \int_0^1 4(1+x^2)^{-1}dx$

The easiest method is Simpson's rule. It can be implemented in virtually every computing environment. The code below works in popular `matlab`-compatible numerical laboratories. It modifies easily to other computing platforms, such as `maple` and `mathematica`. To obtain the answer for $\pi = 3.1415926535897932385$ correct to 12 digits, execute the code on the right in Table 10, below the definition of f.

Table 10. Numerical Integration of $\int_0^1 4(1+x^2)^{-1}dx$.
Simpson's rule is applied, using `matlab`-compatible code. About 50 subdivisions are required.

```
function ans = simp(x0,x1,n,f)
h=(x1-x0)/n; ans=0;
for i=1:n;
ans1=f(x0)+4*f(x0+h/2)+f(x0+h);
ans=ans+(h/6)*ans1;
x0=x0+h;
end
```

```
function y = f(x)
y = 4/(1+x*x);





ans=simp(0,1,50,f)
```

It is convenient in some laboratories to display answers with `printf` or `fprintf`, in order to show 12 digits. For example, `scilab` prints 3.1415927 by default, but 3.141592653589800 using `printf`.

The results checked in `maple` give $\pi \approx 3.1415926535897932385$, accurate to 20 digits, regardless of the actual `maple` numerical integration algorithm chosen (three were possible). The checks are invoked by `evalf(X,20)` where `X` is replaced by `int(4/(1+x*x),x=0..1)`.

The results for an approximation to π using numerical solvers for differential equations varied considerably from one algorithm to another, although all were accurate to 5 rounded digits. A summary for `odepack` routines appears in Table 11, obtained from the `scilab` interface. A selection of routines supported by `maple` appear in Table 12. Default settings were used with no special attempt to increase accuracy.

The `Gear` routines refer to those in the 1971 textbook by C. F. Gear [Gear]. The Livermore stiff solver `lsode` can be found in Hindmarsh [ODEP]. The Runge-Kutta routine of order 7-8 called `dverk78` appears in the 1991 reference of Enright [Enright]. The multistep routines of Adams-Moulton and Adams-Bashforth are described in standard numerical analysis texts, such as Cheney–Kinkaid [Cheney-K]. Taylor series methods are described in the 1972 publication *Mathematical Software* [Rice1972]. The Fehlberg variant of RK4 is given in Forsythe, Malcolm and Moler [FMM].

Table 11. Differential Equation Numeric Solver Results
Package `odepack` applied to $y' = 4/(1 + x^2)$, $y(0) = 0$.

Exact value of π	3.14159265358979323850	20 digits
Runge-Kutta 4	3.1415926535910	10 digits
Adams-Moulton lsode	3.1415932355842	6 digits
Stiff Solver lsode	3.1415931587318	5 digits
Runge-Kutta-Fehlberg 45	3.1416249508084	4 digits

Table 12. Differential Equation Numeric Solver Results
Some `maple`-supported routines, applied to the problem $y' = 4/(1 + x^2)$, $y(0) = 0$.

Exact value of π	3.14159265358979323850	20 digits
Classical RK4	3.141592653589790	15 digits
Gear	3.141592653688446	11 digits
Dverk78	3.141592653607044	11 digits
Taylor Series	3.141592654	10 digits
Runge-Kutta-Fehlberg 45	3.141592674191119	8 digits
Multistep Gear	3.141591703761340	7 digits
Lsode stiff solver	3.141591733742521	6 digits

Computing $\ln 2 = \int_0^1 dx/(1 + x)$

Like the problem of computing π, the formula for $\ln 2$ arises from the method of quadrature applied to $y' = 1/(1+x)$, $y(0) = 0$. The solution is $y(x) = \int_0^x dt/(1+t)$. Application of Simpson's rule with 150 points gives $\ln 2 \approx 0.693147180563800$, which agrees with the exact value $\ln 2 = 0.69314718055994530942$ through 12 digits.

More robust numerical integration algorithms produce the exact answer for $\ln 2$, within the limitations of machine representation of numbers.

Differential equation methods, as in the case of computing π, have results accurate to at least 5 digits, as is shown in Tables 13 and 14. Lower order methods such as classical Euler will produce results accurate to three digits or less.

Table 13. Differential Equation Numeric Solver
Results for `odepack` routines, applied to the problem $y' = 1/(1+x)$, $y(0) = 0$.

Exact value of $\ln 2$	0.69314718055994530942	20 digits
Adams-Moulton lsode	0.69314720834637	7 digits
Stiff Solver lsode	0.69314702723982	6 digits
Runge-Kutta 4	0.69314718056011	11 digits
Runge-Kutta-Fehlberg 45	0.69314973055488	5 digits

Table 14. Differential Equation Numeric Solver
Results for `maple`-supported routines, applied to the problem $y' = 1/(1+x)$, $y(0) = 0$.

Exact value of $\ln 2$	0.69314718055994530942	20 digits
Classical Euler	0.6943987430550621	2 digits
Classical Heun	0.6931487430550620	5 digits
Classical RK4	0.6931471805611659	11 digits
Gear	0.6931471805646605	11 digits
Gear Poly-extr	0.6931471805664855	11 digits
Dverk78	0.6931471805696615	11 digits
Adams-Bashforth	0.6931471793736268	8 digits
Adams-Bashforth-Moulton	0.6931471806484283	10 digits
Taylor Series	0.6931471806	10 digits
Runge-Kutta-Fehlberg 45	0.6931481489496502	5 digits
Lsode stiff solver	0.6931470754312113	7 digits
Rosenbrock stiff solver	0.6931473787603164	6 digits

Computing e from $y' = y$, $y(0) = 1$

The initial attack on the problem uses classical RK4 with $f(x,y) = y$. After 300 steps, classical RK4 finds the correct answer for e to 12 digits: $e \approx 2.71828182846$. In Table 15, the details appear for how to accomplish the calculation using `matlab`-compatible code. Corresponding `maple` code appears in Table 16 and in Table 17. Additional code for `octave` and `scilab` appear in Tables 18 and 19.

Table 15. Numerical Solution of $y' = y$, $y(0) = 1$.
Classical RK4 with 300 subdivisions using matlab-compatible code.

```
function [x,y]=rk4(x0,y0,x1,n,f)
x=x0;y=y0;h=(x1-x0)/n;
for i=1:n;
 k1=h*f(x,y);
 k2=h*f(x+h/2,y+k1/2);
 k3=h*f(x+h/2,y+k2/2);
 k4=h*f(x+h,y+k3);
 y=y+(k1+2*k2+2*k3+k4)/6;
 x=x+h;
end
```

```
function yp = ff(x,y)
 yp= y;


[x,y]=rk4(0,1,1,300,ff)
```

Table 16. Numerical Solution of $y' = y$, $y(0) = 1$
using maple internal classical RK4 code.

```
de:=diff(y(x),x)=y(x):
ic:=y(0)=1:
Y:=dsolve({de,ic},y(x),
          type=numeric,method=classical[rk4]):
Y(1);
```

Table 17. Numerical Solution of $y' = y$, $y(0) = 1$
using classical RK4 with 300 subdivisions using maple-compatible code.

```
rk4 := proc(x0,y0,x1,n,f)
local x,y,k1,k2,k3,k4,h,i:
x=x0:  y=y0:  h=(x1-x0)/n:
for i from 1 to n do
 k1:=h*f(x,y):k2:=h*f(x+h/2,y+k1/2):
 k3:=h*f(x+h/2,y+k2/2):k4:=h*f(x+h,y+k3):
 y:=evalf(y+(k1+2*k2+2*k3+k4)/6,Digits+4):
 x:=x+h:
od:
RETURN(y):
end:

f:=(x,y)->y;
rk4(0,1,1,300,f);
```

A matlab *m*-file "rk4.m" is loaded into scilab-4.0 by getf("rk4.m"). Most scilab code is loaded by using default file extension .sci, e.g., rk4scilab.sci is a scilab file name. This code must obey scilab rules. An example appears below in Table 18.

Table 18. Numerical Solution of $y' = y$, $y(0) = 1$
using classical RK4 with 300 subdivisions with `scilab`-4.0 code.

```
function
[x,y]=rk4sci(x0,y0,x1,n,f)
x=x0,y=y0,h=(x1-x0)/n
 for i=1:n
 k1=h*f(x,y)
 k2=h*f(x+h/2,y+k1/2)
 k3=h*f(x+h/2,y+k2/2)
 k4=h*f(x+h,y+k3)
 y=y+(k1+2*k2+2*k3+k4)/6
 x=x+h
 end
endfunction
```

```
function yp = ff(x,y)
 yp= y
endfunction

[x,y]=rk4sci(0,1,1,300,ff)
```

The popularity of `octave` as a free alternative to `matlab` has kept it alive for a number of years. Writing code for `octave` is similar to `matlab` and `scilab`, however readers are advised to look at sample code supplied with `octave` before trying complicated projects. In Table 19 can be seen some essential agreements and differences between the languages. Versions of `scilab` after 4.0 have a `matlab` to `scilab` code translator.

Table 19. Numerical Solution of $y' = y$, $y(0) = 1$
using classical RK4 with 300 subdivisions with `octave`-2.1.

```
function
[x,y]=rk4oct(x0,y0,x1,n,f)
x=x0;y=y0;h=(x1-x0)/n;
 for i=1:n
 k1=h*feval(f,x,y);
 k2=h*feval(f,x+h/2,y+k1/2);
 k3=h*feval(f,x+h/2,y+k2/2);
 k4=h*feval(f,x+h,y+k3);
 y=y+(k1+2*k2+2*k3+k4)/6;
 x=x+h;
 endfor
endfunction
```

```
function yp = ff(x,y)
yp= y;
end

[x,y]=rk4oct(0,1,1,300,'ff')
```

Exercises 4.4

Computing π

Compute $\pi = y(1)$ from the initial value problem $y' = 4/(1+x^2)$, $y(0) = 0$, using the given method. Number 3.14159 with 3-digit precision is the rounded number 3.142.

1. Use the Rectangular integration rule. Determine the number of steps for 3-digit precision.

2. Use the Rectangular integration rule. Determine the number of steps for 4-digit precision.

3. Use the Trapezoidal integration rule. Determine the number of steps for 3-digit precision.

4. Use the Trapezoidal integration rule. Determine the number of steps for 4-

digit precision.

5. Use Simpson's rule. Determine the number of steps for 5-digit precision.

6. Use Simpson's rule. Determine the number of steps for 6-digit precision.

7. Use a computer algebra system library routine for RK4. Report the step size used and the number of steps for 5-digit precision.

8. Use a numerical workbench library routine for RK4. Report the step size used and the number of steps for 5-digit precision.

Computing $\ln(2)$

Compute $\ln(2) = y(1)$ from the initial value problem $y' = 1/(1+x)$, $y(0) = 0$, using the given method.

9. Use the Rectangular integration rule. Determine the number of steps for 3-digit precision.

10. Use the Rectangular integration rule. Determine the number of steps for 4-digit precision.

11. Use the Trapezoidal integration rule. Determine the number of steps for 5-digit precision.

12. Use the Trapezoidal integration rule. Determine the number of steps for 6-digit precision.

13. Use Simpson's rule. Determine the number of steps for 5-digit precision.

14. Use Simpson's rule. Determine the number of steps for 6-digit precision.

15. Use a computer algebra system library routine for RK4. Report the step size used and the number of steps for 5-digit precision.

16. Use a numerical workbench library routine for RK4. Report the step size used and the number of steps for 5-digit precision.

Computing e

Compute $e = y(1)$ from the initial value problem $y' = y$, $y(0) = 1$, using the given computer library routines. Report the approximate number of digits of precision attained with a computer algebra system or numerical workbench.

17. Improved Euler method, also known as Heun's method.

18. RK4 method.

19. RKF45 method.

20. Adams-Moulton method.

Stiff Differential Equation

The flame propagation equation $y' = y^2(1-y)$ is known to be **stiff** for small initial conditions $y(0) > 0$. Use classical rk4, then Runge-Kutta-Fehlberg rkf45 and finally a stiff solver to compute and plot the solution $y(t)$ in each case. Expect rk4 to fail, no matter the step size. Both rkf45 and a stiff solver will produce about the same plot, but at different speeds. Reference: `matlab` author Cleve Moler, blogs.mathworks.com 2014.

21. $y(0) = 0.01$

22. $y(0) = 0.005$

23. $y(0) = 0.001$

24. $y(0) = 0.0001$

4.5 Earth to the Moon

A projectile launched from the surface of the earth is attracted both by the earth and the moon. The altitude $r(t)$ of the projectile above the earth is known to satisfy the initial value problem (see *Technical Details* page 264)

$$(1)\qquad \begin{aligned} r''(t) &= -\frac{Gm_1}{(R_1+r(t))^2} + \frac{Gm_2}{(R_2-R_1-r(t))^2}, \\ r(0) &= 0, \quad r'(0) = v_0. \end{aligned}$$

The unknown initial velocity v_0 of the projectile is given in meters per second. The constants in (1) are defined as follows.

$G = 6.6726 \times 10^{-11}$ N-m^2/kg^2	Universal gravitation constant,
$m_1 = 5.975 \times 10^{24}$ kilograms	Mass of the earth,
$m_2 = 7.36 \times 10^{22}$ kilograms	Mass of the moon,
$R_1 = 6,378,000$ meters	Radius of the earth,
$R_2 = 384,400,000$ meters	Distance from the earth's center to the moon's center.

The Jules Verne Problem

In his 1865 novel *From the Earth to the Moon*, Jules Verne asked what initial velocity must be given to the projectile in order to reach the moon. The question in terms of equation (1) becomes:

> What minimal value of v_0 causes the projectile to have zero net acceleration at some point between the earth and the moon?

The projectile only has to travel a distance R equal to the surface-to-surface distance between the earth and the moon. The altitude $r(t)$ of the projectile must satisfy $0 \le r \le R$. Given v_0 for which the net acceleration is zero, $r''(t) = 0$ in (1), then the projectile has reached a critical altitude r^*, where gravitational effects of the moon take over and the projectile will fall to the surface of the moon.

Let $r''(t) = 0$ in (1) and substitute r^* for $r(t)$ in the resulting equation. Then[3]

$$(2)\qquad \begin{aligned} -\frac{Gm_1}{(R_1+r^*)^2} + \frac{Gm_2}{(R_2-R_1-r^*)^2} &= 0, \\ r^* = \frac{R_2}{1+\sqrt{m_2/m_1}} - R_1 &\approx 339,620,820 \text{ meters.} \end{aligned}$$

Using energy methods (see *Technical details*, page 264), it is possible to calculate exactly the *minimal* earth-to-moon velocity v_0^* required for the projectile to *just reach* critical altitude r^*:

$$(3)\qquad v_0^* \approx 11067.31016 \quad \text{meters per second.}$$

[3] Multiple values have been reported for the mass of the moon. Using $m_2 = 7.34767309 \times 10^{22}$ gives $r^* \approx 339,649,780$ meters.

A Numerical Experiment

The value $v_0^* \approx 11067.31016$ in (3) will be verified experimentally. As part of this experiment, the flight time is estimated.

Such a numerical experiment must adjust the initial velocity v_0 in initial value problem (1) so that $r(t)$ increases from 0 to R. Graphical analysis of a solution $r(t)$ for low velocities v_0 gives insight into the problem; see Figure 7.

The choice of numerical software solver makes for significant differences in this problem. Initial work used the Livermore Laboratory numerical stiff solver for ordinary differential equations (acronym `lsode`).

Computer algebra system `maple` has algorithms `lsode` or `rosenbrock`. The `dsolve` options are `method=lsode` or `stiff=true`. Other stiff solvers of equal quality can be used for nearly identical results. Experiments are necessary to determine the required accuracy.

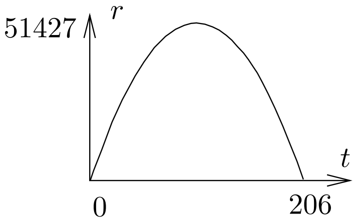


Figure 7. Jules Verne Problem.
The solution $r(t)$ of (1) for $v_0 = 1000$. The projectile rises to a maximum height of about 51,427 meters, then it falls back to earth. The trip time is 206 seconds.

The numerical experiment solves (1) using `rosenbrock`, then the solution is graphed to see if the projectile falls back to earth (as in Figure 7) or if it reaches an altitude near r^* and then falls to the moon. The first experiment might use velocity $v_0 = 1000$ and trip time $T = 210$ (see Figure 7). In this experiment the projectile falls back to earth. The projectile travels to the moon when the r-axis of the graphic has maximum greater than $r^* \approx 339,620,820$ meters. The logic is that $r(t) > r^*$ causes the gravitation effects of the moon to be strong enough to force the projectile to fall to the moon.

In Table 20, find the `maple` initialization code group 1. In Table 21, group 2 is executed a number of times, to refine estimates for the initial velocity v_0 and the trip time T. The graphics produced resemble Figure 7 or Figure 8. A successful trip to the moon is represented in Figure 8, which uses $v_0 = 11068$ meters per second and $T = 527000$ seconds.

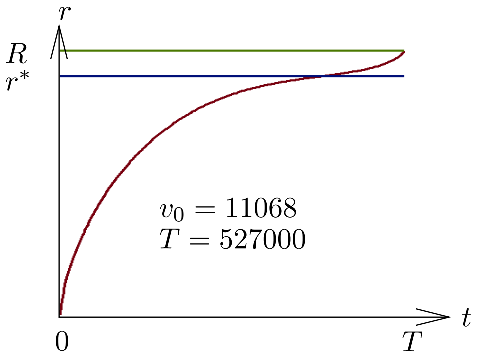


Figure 8. Experimental trip to the moon.
The initial velocity is $v_0 = 24,764$ miles per hour and the trip time is 147 hours.

Table 20. Initialization Code in `maple` for the Trip to the Moon Numerical Experiment.
Group 1 defines seven constants G, m_1, m_2, R_1, R_2, R_3, R and computes values $r^* \approx 339,620,820$ and $v_0^* \approx 11067.31016$.

```
# Group 1:  Constants, rstar and v0star
G:=6.6726e-11:  m1:=5.975e24:  m2:=7.36e22:
R1:=6.378e6:  R2:=3.844e8:  R3:=1.74e6:
R:=R2-R1-R3:
ans:=[solve(-G*m1/(r+R1)^2 + G*m2/(R2-R1-r)^2=0,r)]:
rstar:=ans[1];
FF:=r->G*m1/(R1+r)+G*m2/(R2-R1-r):
v0star:=sqrt(2*(FF(0)-FF(rstar)));# v0star=11067.31016
```

Two utility functions are used: `report(), makePlot()`.

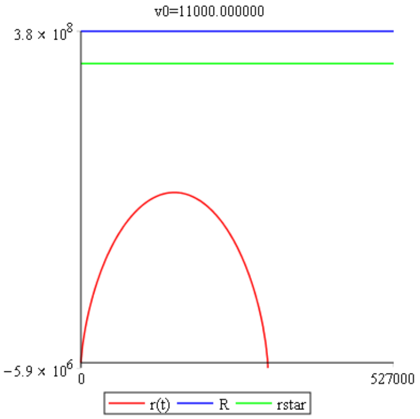
```
# Trip to the Moon Numerical Experiment
report:=proc() local s,hit;global R,rstar,v0,T;
 printf("v0=%a, T=%.2f\n",v0,T);
 printf("Moon at distance R=%.2f (blue)\n",R);
 printf("Acceleration=0 at r=rstar (green)\n");
end proc:

makePlot:=proc() local opt;global T,Y,R,rstar,v0;
 opt:=legend=["r(t)","R","rstar"],color=[red,blue,green],
 title=sprintf("v0=%f",v0);
 plot([Y(t),R,rstar],t=0..T,opt);
end proc:
```

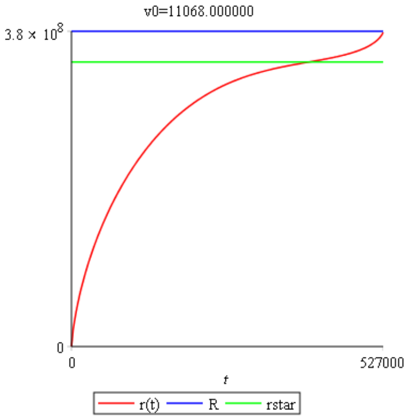
Table 21. Iteration Code in `maple` for the Trip to the Moon Numerical Experiment.
Group 2 plots two graphics for given v_0 and T. A successful trip to the moon uses velocity $v_0 > v_0^* \approx 11067.31016$. Curve $Y(t)$ should cross $r^* \approx 339,620,820$ and $Y(T) \geq R$ must hold.

```
# Group 2:  Iteration code
v0:=11068; # v0>v0star.  Projectile falls to the moon.
T:=527000:  # Guess the trip time T
de:=diff(r(t),t,t)=-G*m1/(r(t)+R1)^2+G*m2/(R2-R1-r(t))^2:
ic:=r(0)=0,D(r)(0)=v0:
NS:=numeric,stiff=true,output=listprocedure:
p:=dsolve([de,ic],r(t),NS); Y:=eval(r(t),p):
makePlot();report();
# Plot done.  Change v0, T and re-execute group 2.
```

Two typical experiments appear below for $v_0 = 11000$ (falls to earth) and $v_0 = 11068$ (falls to the moon). They have a `report` like this:

```
v0=11068, T=527000.00
Moon at distance R=376282000.00 (blue)
Acceleration=0 at r=rstar (green)
r(401326.1134)=rstar=339620820.00
```





Exact trip time. The time T for a trip with velocity $v_0 = 11068$ can be computed if an approximate value for the trip time is known.

```
# Group 2 extra code for trip time T
v0:=11068;
fsolve(Y(t)=R,t=526000); # T = 5.274409891*10^5
```

Details for (1) and (3)

Technical details for (1): To derive (1), it suffices to write down a competition between the Newton's second law force relation $mr''(t)$ and the sum of two forces due to gravitational attraction for the earth and the moon. Here, m stands for the mass of the projectile.

Gravitational force for the earth. This force, by Newton's universal gravitation law, has magnitude

$$F_1 = \frac{Gm_1 m}{\mathcal{R}_3^2}$$

where m_1 is the mass of the earth, G is the universal gravitation constant and $\mathcal{R}_3$ is the distance from the projectile to the center of the earth: $\mathcal{R}_3 = R_1 + r(t)$.

Gravitational force for the moon. Similarly, this force has magnitude

$$F_2 = \frac{Gm_2 m}{\mathcal{R}_4^2}$$

where m_2 is the mass of the moon and $\mathcal{R}_4$ is the distance from the projectile to the center of the moon: $\mathcal{R}_4 = R_2 - R_1 - r(t)$.

Competition between forces. The force equation is

$$mr''(t) = -F_1 + F_2$$

due to the directions of the force vectors. Simplifying the relations and cancelling m gives equation (1). ■

Technical details for (3): To justify the value for v_0, multiply equation (1) by r' and integrate the new equation from $t = 0$ to $t = t_0$ to get

$$\text{(4)} \qquad \begin{aligned} &\frac{1}{2}(r'(t_0))^2 = F(r(t_0)) - F(0) + \frac{1}{2}v_0^2, \quad \text{where} \\ &F(r) = \frac{Gm_1}{R_1 + r} + \frac{Gm_2}{R_2 - R_1 - r}. \end{aligned}$$

The expression $F(r)$ is minimized when $F'(r) = 0$ or else at $r = 0$ or $r = R$. The right side of (1) is $F'(r)$, hence $F(r)$ has unique critical point $r = r^*$. Compute $F(0) = 62522859.35$, $F(r^*) = 1280168.523$ and $F(R) = 3864318.458$. Then the minimum of $F(r)$ is at $r = r^*$ and $F(r^*) \le F(r(t_0))$.

The left side of the first equality in (4) is nonnegative, therefore also the right side is nonnegative, giving $\frac{1}{2}v_0^2 \ge F(0) - F(r(t_0))$. If the projectile ever reaches altitude r^*, then $r(t_0) = r^*$ is allowed and $v_0 \ge \sqrt{2F(0) - 2F(r^*)} \approx 11067.31016$. Restated, $v_0 < 11067.31016$ implies the projectile *never reaches altitude* r^*, hence it falls back to earth. On the other hand, if $v_0 > 11067.31016$, then by (4) and $F(r^*) \le F(r)$ it follows that $r'(t) > 0$ and therefore the projectile cannot return to earth. That is, $r(t) = 0$ for some $t > 0$ can't happen.

In summary, the least launch velocity v_0^* which allows $r(t) = r^*$ for some $t > 0$ is given by the formulas

$$v_0^* = \sqrt{2F(0) - 2F(r^*)}, \quad F(r) = \frac{Gm_1}{R_1 + r} + \frac{Gm_2}{R_2 - R_1 - r}.$$

This completes the proof of equation (3). ■

Exercises 4.5

Critical Altitude r^*

The symbol r^* is the altitude $r(t)$ at which gravitational effects of the moon take over, causing the projectile to fall to the moon.

1. Justify from the differential equation that $r''(t) = 0$ at $r^* = r(t)$ implies the first relation in (2):

$$\frac{Gm_2}{(R_2 - R_1 - r^*)^2} - \frac{Gm_1}{(R_1 + r^*)^2} = 0.$$

2. Solve symbolically the relation of the previous exercise for r^*, to obtain the second equation of (2):

$$r^* = \frac{R_2}{1 + \sqrt{m_2/m_1}} - R_1.$$

3. Use the previous exercise and values for the constants R_1, R_2, m_1, m_2 to obtain the approximation

$$r^* = 339,649,780 \text{ meters}.$$

4. Determine the effect on r^* for a one percent error in measurement m_2. Replace m_2 by $0.99m_2$ and $1.01m_2$ in the formula for r^* and report the two estimated critical altitudes.

Escape Velocity v_0^*

The symbol v_0^* is the velocity $r'(0)$ such that $\lim_{t\to\infty} r(t) = \infty$, but smaller launch velocities will cause the projectile to fall back to the earth. Throughout, define

$$F(r) = \frac{Gm_1}{R_1 + r} + \frac{Gm_2}{R_2 - R_1 - r}.$$

5. Let $v_0 = r'(0)$, $r^* = r(t_0)$. Derive the formula

$$\frac{1}{2}(r'(t_0))^2 = F(r^*) - F(0) + \frac{1}{2}v_0^2$$

which appears in the proof details.

6. Verify using the previous exercise that $r'(t_0) = 0$ implies

$$v_0^* = \sqrt{2(F(0) - F(r^*))}.$$

7. Verify by hand calculation that $v_0^* \approx 11067.31016$ meters per second.

8. Argue by mathematical proof that $F(r)$ is not minimized at the endpoints of the interval $0 \le r \le R$.

Numerical Experiments

Assume values given in the text for physical constants. Perform the given experiment with numerical software on initial value problem (1), page 260. The cases when $v_0 > v_0^*$ escape the earth, while the others fall back to earth.

9. RKF45 solver, $v_0 = 11068$, $T = 515000$. Plot the solution on $0 \le t \le T$.

10. Stiff solver, $v_0 = 11068$, $T = 515000$. Plot the solution on $0 \le t \le T$.

11. RKF45 solver, $v_0 = 11067.2$, $T = 800000$. Plot the solution on $0 \le t \le T$.

12. Stiff solver, $v_0 = 11067.2$, $T = 800000$. Plot the solution on $0 \le t \le T$.

13. RKF45 solver, $v_0 = 11067$, $T = 1000000$. Plot the solution on $0 \le t \le T$.

14. Stiff solver, $v_0 = 11067$, $T = 1000000$. Plot the solution on $0 \le t \le T$.

15. RKF45 solver, $v_0 = 11066$, $T = 800000$. Plot the solution on $0 \le t \le T$.

16. Stiff solver, $v_0 = 11066$, $T = 800000$. Plot the solution on $0 \le t \le T$.

17. RKF45 solver, $v_0 = 11065$. Find a suitable value T which shows that the projectile falls back to earth, then plot the solution on $0 \le t \le T$.

18. Stiff solver, $v_0 = 11065$. Find a suitable value T which shows that the projectile falls back to earth, then plot the solution on $0 \le t \le T$.

19. RKF45 solver, $v_0 = 11070$. Find a suitable value T which shows that the projectile falls to the moon, then plot the solution on $0 \le t \le T$.

20. Stiff solver, $v_0 = 11070$. Find a suitable value T which shows that the projectile falls to the moon, then plot the solution on $0 \le t \le T$.

4.6 Skydiving

A skydiver of 160 pounds jumps from a hovercraft at $15,000$ feet. The fall is mostly vertical from zero initial velocity, but there are significant effects from air resistance until the parachute opens at $5,000$ feet. The resistance effects are determined by the skydiver's clothing and body shape.

Velocity Model

Assume the skydiver's air resistance is modeled in terms of velocity v by a force equation

$$F(v) = av + bv^2 + cv^3.$$

The constants a, b, c are given by the formulas

$$a = 0.009, \quad b = 0.0008, \quad c = 0.0001.$$

In particular, the force $F(v)$ is positive for v positive. According to Newton's second law, the velocity $v(t)$ of the skydiver satisfies $mv'(t) = mg - F(v)$. We assume $mg = 160$ pounds and $g \approx 32$ feet per second per second. The **Velocity model** is

$$v'(t) = 32 - \frac{32}{160}\left(0.009v(t) + 0.0008v^2(t) + 0.0001v^3(t)\right), \quad v(0) = 0.$$

Distance Model

The distance $x(t)$ traveled by the skydiver, measured from the hovercraft, is given by the **Distance model**

$$x'(t) = v(t), \quad x(0) = 0.$$

The velocity is expected to be positive throughout the flight. Because the parachute opens at 5000 feet, at which time the velocity model must be replaced the open parachute model (not discussed here), the distance $x(t)$ increases with time from 0 feet to its limiting value of 10000 feet. Values of $x(t)$ from 10000 to 15000 feet make sense only for the open parachute model.

Terminal Velocity

The **terminal velocity** is an equilibrium solution $v(t) = v_\infty$ of the velocity model, therefore constant v_∞ satisfies

$$32 - \frac{32}{160}\left(0.009v_\infty + 0.0008v_\infty^2 + 0.0001v_\infty^3\right) = 0.$$

A numerical solver is applied to find the value $v_\infty = 114.1$ feet per second, which is about 77.8 miles per hour. For the solver, we define $f(v) = 32 - F(v)$ and solve $f(v) = 0$ for v. Some maple details:

```
f:=v->32 - (32/160)*(0.009*v+0.0008*v^2+0.0001*v^3);
fsolve(f(v)=0,v);             # 114.1032777 ft/sec
60*60*fsolve(f(v)=0,v)/5280;  # 77.79768934 mi/hr
```

A Numerical Experiment

The Runge-Kutta method will be applied to produce a table which contains the elapsed time t, the skydiver velocity $v(t)$ and the distance traveled $x(t)$, up until the distance reaches nearly 10000 feet, whereupon the parachute opens.

The objective here is to illustrate practical methods of table production in a computer algebra system or numerical laboratory. It is efficient in these computational systems to phrase the problem as a system of two differential equations with two initial conditions.

System Conversion. The velocity substitution $v(t) = x'(t)$ used in the velocity model gives us two differential equations in the unknowns $x(t)$, $v(t)$:

$$x'(t) = v(t),\ v'(t) = g - \frac{1}{m}F(v(t)).$$

Define $f(v) = g - (1/m)F(v)$. The path we follow is to execute the maple code below, which produces the table that follows using the default Runge-Kutta-Fehlberg algorithm.

```
eq:=32 - (32/160)*(0.009*v+0.0008*v^2+0.0001*v^3:
f:=unapply(eq,v);
de1:=diff(x(t),t)=v(t); de2:=diff(v(t),t)=f(v(t));
ic:=x(0)=0,v(0)=0;opts:=numeric,output=listprocedure:
p:=dsolve({de1,de2,ic},[x(t),v(t)],opts);
X:=eval(x(t),p); V:=eval(v(t),p);
fmt:="%10.2f  %10.2f  %10.2f\n";
seq(printf(fmt,5*t,X(5*t),V(5*t)),t=0..18);
```

t	$x(t)$	$v(t)$	t	$x(t)$	$v(t)$
5.00	331.26	106.84	50.00	5456.76	114.10
10.00	892.79	113.97	55.00	6027.28	114.10
15.00	1463.15	114.10	60.00	6597.80	114.10
20.00	2033.67	114.10	65.00	7168.31	114.10
25.00	2604.18	114.10	70.00	7738.83	114.10
30.00	3174.70	114.10	75.00	8309.35	114.10
35.00	3745.21	114.10	80.00	8879.86	114.10
40.00	4315.73	114.10	85.00	9450.38	114.10
45.00	4886.25	114.10	90.00	10020.90	114.10

The table says that the flight time to parachute open at 10,000 feet is about 90 seconds and the terminal velocity 114.10 feet/sec is reached in about 15 seconds.

More accurate values for the flight time 89.82 to 10,000 feet and time 14.47 to terminal velocity can be determined as follows.

```
fsolve(X(t)=10000,t,80..95);
fsolve(V(t)=114.10,t,2..20);
```

Alternate Method. Another way produce the table is to solve the velocity model numerically, then determine $x(t) = \int_0^t v(r)dr$ by numerical integration. Due to accuracy considerations, a variant of Simpson's rule is used, called the **Newton-cotes rule**. The `maple` implementation of this idea follows.

The first method of conversion into two differential equations is preferred, even though the alternate method reproduces the table using only the textbook material presented in this chapter.

```
f:=unapply(32-(32/160)*(0.009*v+0.0008*v^2+0.0001*v^3),v);
de:=diff(v(t),t)=f(v(t)); ic:=v(0)=0;
q:=dsolve({de,ic},v(t),numeric,output=listprocedure);
V:=eval(v(t),q);
X:=u->evalf(Int(V,0..u,continuous,_NCrule));
fmt:="%10.2f  %10.2f  %10.2f\n";
seq(printf(fmt,5*t,X(5*t),V(5*t)),t=0..18);
```

Ejected Baggage

Much of what has been done here applies as well to an ejected parcel, instead of a skydiver. What changes is the force equation $F(v)$, which depends upon the parcel exterior and shape. The distance model remains the same, but the restraint $0 \le x \le 10000$ no longer applies, since no parachute opens. We expect the parcel to reach terminal velocity in 5 to 10 seconds and hit the ground at that speed.

Variable Mass

The mass of a skydiver can be time-varying. For instance, the skydiver lets water leak from a reservoir. This kind of problem assumes mass $m(t)$, position $x(t)$ and velocity $v(t)$ for the diver. Then Newton's second law gives a position-velocity model

$$\begin{aligned} &x'(t) = v(t), \\ &(m(t)v(t))' = G(t, x(t), v(t)). \end{aligned}$$

The problem is similar to rocket propulsion, in which expended fuel decreases the in-flight mass of the rocket. Simplifying assumptions make it possible to present formulas for $m(t)$ and $G(t, x, v)$, which can be used by the differential equation solver.

Exercises 4.6

Terminal Velocity

Assume force $F(v) = av + bv^2 + cv^3$ and $g = 32$, $m = 160/g$. Using computer assist, find the terminal velocity v_∞ from the velocity model $v' = g - \frac{1}{m}F(v)$, $v(0) = 0$.

1. $a = 0$, $b = 0$ and $c = 0.0002$.

2. $a = 0$, $b = 0$ and $c = 0.00015$.

3. $a = 0$, $b = 0.0007$ and $c = 0.00009$.

4. $a = 0$, $b = 0.0007$ and $c = 0.000095$.

5. $a = 0.009$, $b = 0.0008$ and $c = 0.00015$.

6. $a = 0.009$, $b = 0.00075$ and $c = 0.00015$.

7. $a = 0.009$, $b = 0.0007$ and $c = 0.00009$.

8. $a = 0.009$, $b = 0.00077$ and $c = 0.00009$.

9. $a = 0.009$, $b = 0.0007$ and $c = 0$.

10. $a = 0.009$, $b = 0.00077$ and $c = 0$.

Numerical Experiment

Let $F(v) = av + bv^2 + cv^3$ and $g = 32$. Consider the skydiver problem $mv'(t) = mg - F(v)$ and constants m, a, b, c supplied below. Using computer assist, apply a numerical method to produce a table for the elapsed time t, the velocity $v(t)$ and the distance $x(t)$. The table must end at $x(t) \approx 10000$ feet, which determines the flight time.

11. $m = 160/g$, $a = 0$, $b = 0$ and $c = 0.0002$.

12. $m = 160/g$, $a = 0$, $b = 0$ and $c = 0.00015$.

13. $m = 130/g$, $a = 0$, $b = 0.0007$ and $c = 0.00009$.

14. $m = 130/g$, $a = 0$, $b = 0.0007$ and $c = 0.000095$.

15. $m = 180/g$, $a = 0.009$, $b = 0.0008$ and $c = 0.00015$.

16. $m = 180/g$, $a = 0.009$, $b = 0.00075$ and $c = 0.00015$.

17. $m = 170/g$, $a = 0.009$, $b = 0.0007$ and $c = 0.00009$.

18. $m = 170/g$, $a = 0.009$, $b = 0.00077$ and $c = 0.00009$.

19. $m = 200/g$, $a = 0.009$, $b = 0.0007$ and $c = 0$.

20. $m = 200/g$, $a = 0.009$, $b = 0.00077$ and $c = 0$.

Flight Time

Let $F(v) = av + bv^2 + cv^3$ and $g = 32$. Consider the skydiver problem $mv'(t) = mg - F(v)$ with constants m, a, b, c supplied below. Using computer assist, apply a numerical method to find accurate values for the flight time to 10,000 feet and the time required to reach terminal velocity.

21. $mg = 160$, $a = 0.0095$, $b = 0.0007$ and $c = 0.000092$.

22. $mg = 160$, $a = 0.0097$, $b = 0.00075$ and $c = 0.000095$.

23. $mg = 240$, $a = 0.0092$, $b = 0.0007$ and $c = 0$.

24. $mg = 240$, $a = 0.0095$, $b = 0.00075$ and $c = 0$.

Ejected Baggage

Baggage of 45 pounds is dropped from a hovercraft at $15,000$ feet. Assume air resistance force $F(v) = av + bv^2 + cv^3$, $g = 32$

and $mg = 45$. Using computer assist, find accurate values for the flight time to the ground and the terminal velocity. Estimate the time required to reach 99.95% of terminal velocity.

25. $a = 0.0095$, $b = 0.0007$, $c = 0.00009$

26. $a = 0.0097$, $b = 0.00075$, $c = 0.00009$

27. $a = 0.0099$, $b = 0.0007$, $c = 0.00009$

28. $a = 0.0099$, $b = 0.00075$, $c = 0.00009$

4.7 Lunar Lander

A lunar lander goes through free fall to the surface of the moon, its descent controlled by retrorockets that provide a constant deceleration to counter the effect of the moon's gravitational field.

The retrorocket control is supposed to produce a **soft touchdown**, which means that the velocity $v(t)$ of the lander is zero when the lander touches the moon's surface. To be determined:

H = height above the moon's surface for retrorocket activation,

T = flight time from retrorocket activation to soft touchdown.

Investigated here are two models for the lunar lander problem. In both cases, it is assumed that the lander has mass m and falls in the direction of the moon's gravity vector. The initial speed of the lander is assumed to be v_0. The retrorockets supply a constant thrust deceleration g_1. Either the fps or mks unit system will be used. Expended fuel ejected from the lander during thrust will be ignored, keeping the lander mass constantly m.

The distance $x(t)$ traveled by the lander t time units after retrorocket activation is given by

$$x(t) = \int_0^t v(r)dr, \quad 0 \le t \le T.$$

Therefore, H and T are related by the formulas

$$v(T) = 0, \quad x(T) = H.$$

Constant Gravitational Field

Let g_0 denote the constant acceleration due to the moon's gravitational field. Assume given initial velocity v_0 and the retrorocket thrust deceleration g_1. Define $A = g_1 - g_0$, the effective thrust. Set the origin of coordinates at the center of mass of the lunar lander. Let vector $\vec{\imath}$ have tail at the origin and direction towards the center of the moon. The force on the lander is $mv'(t)\vec{\imath}$ by Newton's second law. The forces $mg_0\vec{\imath}$ and $-mg_1\vec{\imath}$ add to $-mA\vec{\imath}$. Force competition $mv'(t)\vec{\imath} = -mA\vec{\imath}$ gives the velocity model

$$mv'(t) = -mA, \quad v(0) = v_0.$$

This quadrature-type equation is solved routinely to give

$$v(t) = -At + v_0, \quad x(t) = -A\frac{t^2}{2} + v_0 t.$$

The equation $v(T) = 0$ gives $T = v_0/A$ and $H = x(T) = v_0^2/(2A)$.

Numerical illustration. Let $v_0 = 1200$ miles per hour and $A = 30000$ miles per hour per hour. We compute values $T = 1/25$ hours $= 2.4$ minutes and $H = x(T) = 24$ miles. A `maple` answer check appears below.

```
v0:=1200; A:=30000;
X:=t->-A*t^2/2+v0*t;
T:=(v0/A): (T*60.0).'min',X(T).'miles'; # 2.4 min,24 miles
A1:=A*2.54*12*5280/100/3600/3600; # mks units 3.725333334
v1:=v0*12*2.54*5280/100/3600;     # mks units 536.448
evalf(convert(X(T),units,miles,meters)); # 38624.256
```

The constant field model predicts that the retrorockets should be turned on 24 miles above the moon's surface with soft landing descent time of 2.4 minutes. It turns out that a different model predicts that 24 miles is too high, but only by a small amount. We investigate now this alternative model, based upon replacing the constant gravitational field by a variable field.

Variable Gravitational Field

The system of units will be the *mks* system. Assume the lunar lander is located at position P above the moon's surface. Define symbols:

$m =$ mass of the lander in kilograms,

$M = 7.35 \times 10^{22}$ kilograms is the mass of the moon,

$R = 1.74 \times 10^{6}$ meters is the mean radius of the moon,

$G = 6.6726 \times 10^{-11}$ is the universal gravitation constant, in *mks* units,

$H =$ height in meters of position P above the moon's surface,

$v_0 =$ lander velocity at P in meters per second,

$g_0 = GM/R^2 =$ constant acceleration due to the moon's gravity in meters per second per second,

$g_1 =$ constant retrorocket thrust deceleration in meters per second per second,

$A = g_1 - g_0 =$ effective retrorocket thrust deceleration in meters per second per second, constant field model,

$t =$ time in seconds,

$x(t) =$ distance in meters from the lander to position P,

$v(t) = x'(t) =$ velocity of the lander in meters per second.

The project is to find the height H above the moon's surface and the descent time T for a soft landing, using fixed retrorockets at time $t = 0$.

The origin of coordinates will be P and $\vec{\imath}$ is directed from the lander to the moon. Then $x(t)\vec{\imath}$ is the lander position at time t. The initial conditions are $x(0) = 0$, $v(0) = v_0$. Let $g_0(t)$ denote the variable acceleration of the lander due to the moon's gravitational field. Newton's universal gravitation law applied to point masses representing the lander and the moon gives the expression

$$\text{Force} = mg_0(t)\vec{\imath} = \frac{GmM}{(R + H - x(t))^2}\vec{\imath}.$$

The force on the lander is $mx''(t)\vec{\imath}$ by Newton's second law. The force is also $mg_0(t)\vec{\imath} - mg_1\vec{\imath}$. Force competition gives the second order distance model

$$mx''(t) = -mg_1 + \frac{mMG}{(R + H - x(t))^2}, \quad x(0) = 0, \quad x'(0) = v_0.$$

The technique from the Jules Verne problem applies: multiply the differential equation by $x'(t)$ and integrate from $t = 0$ to the soft landing time $t = T$. The result:

$$\left.\frac{(x'(t))^2}{2}\right|_{t=0}^{t=T} = -g_1(x(T) - x(0)) + \left.\frac{GM}{R + H - x(t)}\right|_{t=0}^{t=T}.$$

Using the relations $x(0) = 0$, $x'(0) = v_0$, $x'(T) = 0$ and $x(T) = H$ gives a simplified implicit equation for H:

$$-\frac{v_0^2}{2} = -g_1 H + \frac{GM}{R} - \frac{GM}{R + H}.$$

Numerical illustration. Use $v_0 = 536.448$, $g_1 = 5.3452174$ to mimic the constant field example of initial velocity 1200 miles per hour and effective retrorocket thrust 30000 miles per hour per hour. A soft landing is possible from height $H = 23.7775$ miles with a descent time of $T = 2.385$ minutes. These results compare well with the constant field model, which had results of $H = 24$ miles and $T = 2.4$ minutes. Some `maple` details follow.

```
M:=7.35* 10^(22);R:=1.74* 10^6;G:=6.6726* 10^(-11);
v0_CFM:=1200: A_CFM:=30000: # Constant field model values
cf:=1*5280*12*2.54/100/3600: # miles/hour to meters/second
v0:=v0_CFM*cf; g0:=G*M/R^2: g1:=A_CFM*cf/3600+g0;
eq:= -(v0^2/2) + g1*H + G*M/(R+H) - G*M/R=0:
HH:=[solve(eq,H)][1];  # HH := 38266 meters
de:=diff(x(t),t,t) = -g1 + M*G/(R+HH-x(t))^2;
ic:= x(0)=0, D(x)(0)=v0;
p:=dsolve({de,ic},x(t),numeric):
X:=t->evalf(rhs(p(t)[2])):
V:=t-> evalf(rhs(p(t)[3])):
```

```
plot(V,0..300);# V=0 at approx 140 sec
TT1:=fsolve('V(t)'=0,t=140): TT:=TT1/60:
TT1.'seconds', TT.'minutes';
X(TT1).'meters', ((X(TT1)*100/2.54)/12/5280).'miles';
# 2.385 min, 23.78 miles
```

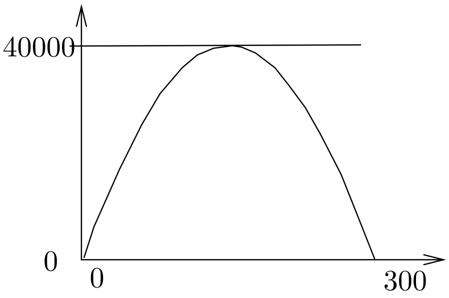


Figure 9. A `maple` plot used to determine the descent time $T = 2.385$ minutes.

Modeling

The field of the earth has been ignored in both models, which is largely justified because the universal gravitation law term for the lander and the earth is essentially zero for lander locations near the moon.

The field for the lander and the moon is not constant, and therefore it can be argued that conditions exist when assuming it is constant will produce invalid and obviously incorrect results.

Are there cases when the answers for the two models differ greatly? Yes, but the height H of retrorocket activation has to be large. This question is re-visited in the exercises.

Control problems. The descent problem for a lunar lander is a control problem in which the **controller** is the retrorocket plus the duration of time in which it is active. All we have done here is to decide that the descent should be controlled by retrorockets well in advance of 24 miles above the moon's surface. The methods used here can be applied to gain insight into the **bang-bang control problem** of turning on the retrorockets for n intervals of time of durations $\Delta t_1, \ldots, \Delta t_n$ to make an *almost* soft landing.

Primitive numerical methods. The predictions made here using the computer algebra system `maple` can be replaced by primitive RK4 methods and graphing. No practising scientist or engineer would do *only* that, however, because they want to be confident of the calculations and the results. The best idea is to use a **black box** of numerical and graphical methods which have little chance of failure, e.g., a computer algebra system or a numerical laboratory.

Exercises 4.7

Lunar Lander Constant Field

Find the retrorocket activation time T and the activation height $x(T)$. Assume the constant gravitational field model. Units are miles/hour and miles/hour per hour.

1. $v_0 = 1210$, $A = 30020$.

2. $v_0 = 1200$, $A = 30100$.

3. $v_0 = 1300$, $A = 32000$.

4. $v_0 = 1350$, $A = 32000$.

5. $v_0 = 1500$, $A = 45000$.

6. $v_0 = 1550$, $A = 45000$.

7. $v_0 = 1600$, $A = 53000$.

8. $v_0 = 1650$, $A = 53000$.

9. $v_0 = 1400$, $A = 40000$.

10. $v_0 = 1450$, $A = 40000$.

Lunar Lander Variable Field

Find the retrorocket activation time T and the activation height $x(T)$. Assume the variable gravitational field model and *mks* units.

11. $v_0 = 540.92$, $g_1 = 5.277$.

12. $v_0 = 536.45$, $g_1 = 5.288$.

13. $v_0 = 581.15$, $g_1 = 5.517$.

14. $v_0 = 603.504$, $g_1 = 5.5115$.

15. $v_0 = 625.86$, $g_1 = 5.59$.

16. $v_0 = 603.504$, $g_1 = 5.59$.

17. $v_0 = 581.15$, $g_1 = 5.59$.

18. $v_0 = 670.56$, $g_1 = 6.59$.

19. $v_0 = 670.56$, $g_1 = 6.83$.

20. $v_0 = 715.26$, $g_1 = 7.83$.

Distinguishing Models

The constant field model (**1**) page 272 and the variable field model (**2**) page 273 are verified here to be distinct, by example. Find the retrorocket activation times T_1, T_2 and the activation heights $x_1(T_1)$, $x_2(T_2)$ for the two models (**1**), (**2**). Relations $A = g_1 - g_0$ and $g_0 = GM/R^2$ apply to compute g_1 for the variable field model.

21. $v_0 = 1200$ mph, $A = 10000$ mph/h. Answer: 72, 66.91 miles.

22. $v_0 = 1200$ mph, $A = 12000$ mph/h. Answer: 60, 56.9 miles.

23. $v_0 = 1300$ mph, $A = 10000$ mph/h. Answer: 84.5, 74.23 miles.

24. $v_0 = 1300$ mph, $A = 12000$ mph/h. Answer: 76.82, 71.55 miles.

4.8 Comets

Planet Mercury

Its elliptical orbit has major semi-axis $a = 0.3871$ AU (astronomical units) and eccentricity $e = 0.2056$. The ellipse can be described by the equations

$$\begin{aligned} x(t) &= a\cos(E(t)), \\ y(t) &= a\sqrt{1-e^2}\,\sin(E(t)), \end{aligned}$$

where t is the mean anomaly ($0 \le t \le 2\pi$) and $E(t)$ is the eccentric anomaly determined from Kepler's equation $E = t + e\sin(E)$.

The path of mercury is an ellipse, yes. Like the earth, the path is essentially circular, due to eccentricity near zero.

Halley's Comet

The Kepler theory for mercury applies to Halley's comet, which has a highly elliptical orbit of eccentricity $e = 0.967$. The major semi-axis is $a = 17.8$ astronomical units (AU), the minor semi-axis is $b = a\sqrt{1-e^2} = 4.535019431$ AU, with period about 76 earth-years.

Our project is to determine $E(t)$ numerically for Halley's comet and plot an animation of the elliptical path of the comet.

History

Kepler's laws of planetary motion were published in 1609 and 1618. The laws are named after Johannes Kepler (1571-1630), a German mathematician and astronomer, who formulated the laws after years of calculation based upon excellent observational data of the Danish astronomer Tycho Brahe (1546-1601). The three laws:

I. The orbit of each planet is an ellipse with the sun at one focus.

II. The line joining the sun to a planet sweeps out equal areas in equal time.

III. The square of the planet's period of revolution is proportional to the cube of the major semi-axis of its elliptical orbit.

These laws apply not only to planets, but to satellites and comets. A proof of Kepler's first two laws, assuming Newton's laws and a vector analysis background, can be found in this text, page 546, *infra.*

The elliptical orbit can be written as

$$\begin{array}{rcl} x(M) & = & a\cos(E(M)), \\ y(M) & = & b\sin(E(M)), \end{array}$$

where a and b are the semi-axis lengths of the ellipse. Astronomers call function E the planet's **eccentric anomaly** and M the planet's **mean anomaly**.

The minor semi-axis of the ellipse is given by

$$b = a\sqrt{1-e^2},$$

where e is the **eccentricity** of the elliptical orbit. The mean anomaly satisfies $M = 2\pi t/T$, where t=time and T is the period of the planet.

It is known that the first two laws of Kepler imply **Kepler's equation**

$$E = M + e\sin(E).$$

Kepler's Initial Value Problem

The equation $E = M + e\sin E$, called Kepler's equation, is the unique implicit solution of the separable differential equation

$$(1) \qquad \left\{ \begin{array}{rcl} \dfrac{dE}{dM} & = & \dfrac{1}{1-e\cos(E)}, \\ E(0) & = & 0. \end{array} \right.$$

The initial value problem (1) *defines* the eccentric anomaly $E(M)$. We are able to compute values of E by suitable first order numerical methods, especially RK4.

It is routine to compute dE/dM by implicit differentiation of Kepler's equation. The idea works on many implicit equations: find an initial value problem replacement by implicit differentiation.

Eccentric Anomaly and Elliptical Orbit

The solution for comet Halley uses `maple` in a direct manner, basing the solution on Kepler's equation. Details:

```
# Kepler's equation E = M + e sin(E)
 e:=0.967:EE := unapply(RootOf(_Z-M-e*sin(_Z)),M);
 Ex:=cos(EE(M)):Ey:=sqrt(1-e^2)*sin(EE(M)):
 plot(EE(M),M=0..2*Pi);
 plot([Ex,Ey,M=0..2*Pi]);
```

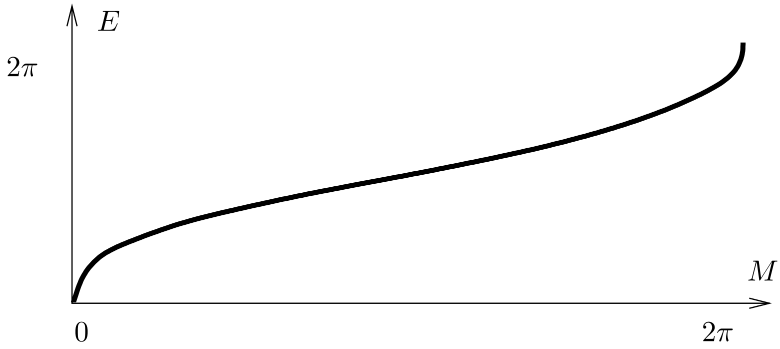


Figure 10. Eccentric anomaly plot for Halley's comet.

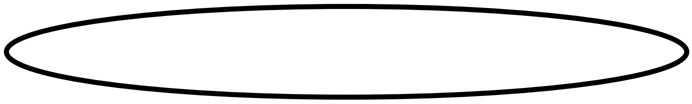


Figure 11. Elliptic trace plot of Halley's comet.

Comet Halley's Positions each Year

The elliptic trace plot can be modified to display a circle for each comet position from year 0 to year 75; see Figure 12. Implemented here is an approach to evaluation of the eccentric anomaly $E(M)$ by numerical differential equation methods. This method is orders of magnitude faster than the `RootOf` method of the previous illustration.

The lack of circles near the focus on the right is explained by the increased speed of the comet near the sun, which is at this focus.

```
# Comet positions each year
 e:=0.967:de:=diff(y(x),x)=1/(1-e*cos(y(x))); ic:=y(0)=0;
 p:=dsolve({de,ic},numeric,output=listprocedure);
 EE := eval(y(x),p):
 Ex:=unapply(cos(EE(M)),M):
 Ey:=unapply(sqrt(1-e^2)*sin(EE(M)),M):
 snapshots:=seq([Ex(2*n*Pi/56),Ey(2*n*Pi/56)],n=0..56):
 opts:=scaling=constrained,axes=boxed,style=point,
       symbolsize=20,symbol=circle,thickness=3:
 plot([snapshots],opts);
```

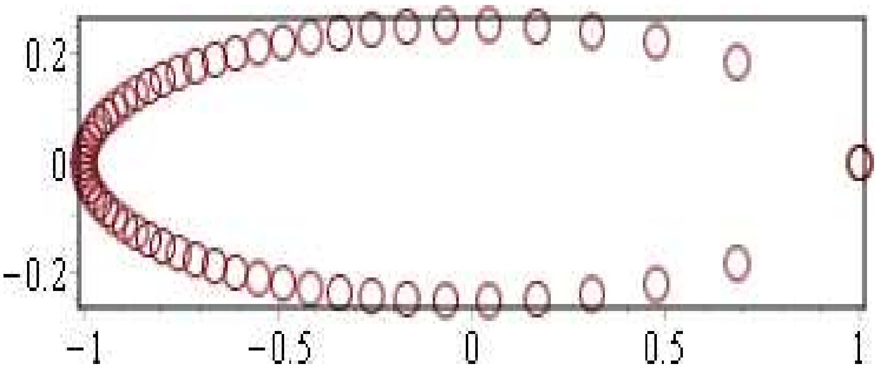


Figure 12. Halley's comet positions each earth-year. On the axes, one unit equals 17.8 **AU.**

Halley's Comet Animation

Computer algebra system `maple` will be used to produce a simple animation of Halley's comet as it traverses its 76-year orbit around the sun. The plan is to solve Kepler's initial value problem in order to find the value of the eccentric anomaly $E(M)$, then divide the orbit into 76 frames and display each in succession to obtain the animation. Defining E by Kepler's equation $E = M + e \sin E$ is too slow for most computer equipment, therefore differential equations are used.

While each comet position in Figure 13 represents an equal block of time, about one earth-year, the amount of path traveled varies. This is because the speed along the path is not constant, the comet traveling fastest near the sun. The most detail is shown for an animation at 2 frames per second. The orbit graph uses one unit equal to about 17.8 astronomical units, to simplify the display.

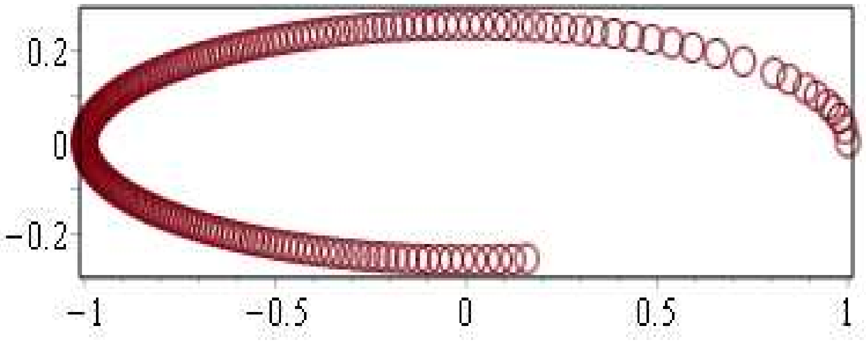


Figure 13. A simple Halley's comet animation.

```
# Simple Halley's comet animation
e:=0.967:de:=diff(y(x),x)=1/(1-e*cos(y(x))); ic:=y(0)=0;
p:=dsolve({de,ic},numeric,output=listprocedure);
EE := eval(y(x),p):
xt:=cos(EE(M)):yt:=sqrt(1-e^2)*sin(EE(M)):
```

```
opts:=view=[-1..1,-0.28..0.28],frames=56,axes=none,
      scaling=constrained,axes=boxed,style=point,
      symbolsize=20,symbol=circle,thickness=3:
plots[animatecurve]([xt,yt,M=0..2*Pi],opts);
```

Animation Video

A video of the comet moving along the ellipse will be produced. The comet position for $t = 2.4516$ earth-years ($M \approx 2\pi t/76$) is shown in Figure 14. During the animation, the comet travels at varying speeds along the ellipse.

```
# Video animation of Halley's comet
 e:=0.967:de:=diff(y(x),x)=1/(1-e*cos(y(x))); ic:=y(0)=0;
 p:=dsolve({de,ic},numeric,output=listprocedure);
 EE := eval(y(x),p):
 comet:=unapply([cos(EE(M)),sqrt(1-e^2)*sin(EE(M))],M):
 options1:=view=[-1..1,-0.28..0.28]:
 options2:=scaling=constrained,axes=none,thickness=3:
 options3:=style=point,symbolsize=40,symbol=solidcircle:
 opts1:=options1,options2,color=blue:
 opts:=options1,options2,options3:
 COMET:=[[comet(2*Pi*t/(76))],opts]:
 ellipse:=plot([cos(x),sqrt(1-e^2)*sin(x),x=0..2*Pi],opts1):
 with(plots):
 F:=animate( plot,COMET,t=0..4,frames=32,background=ellipse):
 G:=animate( plot,COMET,t=5..75,frames=71,background=ellipse):
 H:=animate( plot,COMET,t=75..76,frames=16,background=ellipse):
 display([F,G,H],insequence=true);
```

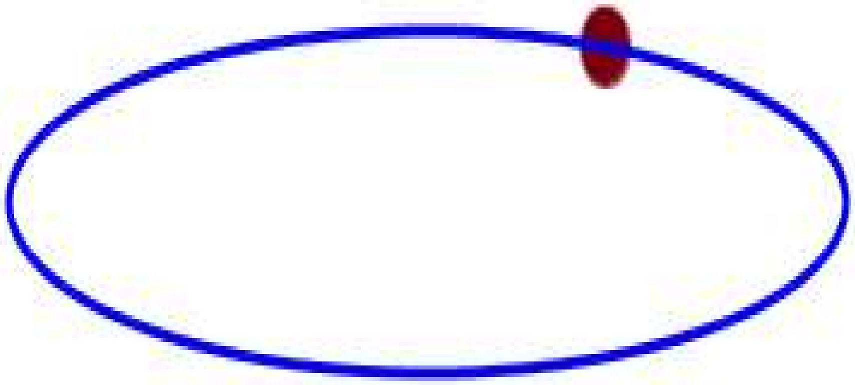


Figure 14. Halley's comet animation video. The frame shown is for $t = 3.0968$ earth-years, mean anomaly $M = 2.4516$ ($M \approx 2\pi t/76$).

Exercises 4.8

Eccentric Anomaly for the Planets

Make a plot of the eccentric anomaly $E(M)$ on $0 \leq M \leq 2\pi$.

1. Mercury, $e = 0.2056$

2. Venus, $e = 0.0068$

3. Earth, $e = 0.0167$

4. Mars, $e = 0.0934$

5. Jupiter, $e = 0.0483$

6. Saturn, $e = 0.0560$

7. Uranus, $e = 0.0461$

8. Neptune, $e = 0.0097$

Elliptic Path of the Planets

Make a plot of the elliptic path of each planet, using constrained scaling with the given major semi-axis A (in astronomical units AU). The equations:

$$\begin{aligned} x(M) &= A\cos(E(M)), \\ y(M) &= A\sqrt{1-e^2}\sin(E(M)) \end{aligned}$$

9. Mercury, $e = 0.2056$, $A = 0.39$

10. Venus, $e = 0.0068$, $A = 0.72$

11. Earth, $e = 0.0167$, $A = 1$

12. Mars, $e = 0.0934$, $A = 1.52$

13. Jupiter, $e = 0.0483$, $A = 5.20$

14. Saturn, $e = 0.0560$, $A = 9.54$

15. Uranus, $e = 0.0461$, $A = 19.18$

16. Neptune $e = 0.0097$, $A = 30.06$

Planet Positions

Make a plot with at least 8 planet positions displayed. Use constrained scaling with major semi-axis A in the plot. Display the given major semi-axis A and period T in the legend.

17. Mercury, $e = 0.2056$, $A = 0.39$ AU, $T = 0.24$ earth-years

18. Venus, $e = 0.0068$, $A = 0.72$ AU, $T = 0.62$ earth-years

19. Earth, $e = 0.0167$, $A = 1$ AU, $T = 1$ earth-years

20. Mars, $e = 0.0934$, $A = 1.52$ AU, $T = 1.88$ earth-years

21. Jupiter, $e = 0.0483$, $A = 5.20$ AU, $T = 11.86$ earth-years

22. Saturn, $e = 0.0560$, $A = 9.54$ AU, $T = 29.46$ earth-years

23. Uranus, $e = 0.0461$, $A = 19.18$ AU, $T = 84.01$ earth-years

24. Neptune $e = 0.0097$, $A = 30.06$ AU, $T = 164.8$ earth-years

Comet Positions

Make a plot with at least 8 comet positions displayed. Use constrained scaling with major-semiaxis 1 in the plot. Display the given eccentricity e and period T in the legend.

25. Churyumov-Gerasimenko orbits the sun every 6.57 earth-years. Discovered in 1969. Eccentricity $e = 0.632$.

26. Comet Wirtanen was the original target of the Rosetta space mission. This comet was discovered in 1948. The comet orbits the sun once every 5.46 earth-years. Eccentricity $e = 0.652$.

27. Comet Wild 2 was discovered in 1978. The comet orbits the sun once every 6.39 earth-years. Eccentricity $e = 0.540$.

28. Comet Biela was discovered in 1772. It orbits the sun every 6.62 earth-years. Eccentricity $e = 0.756$.

29. Comet Encke was discovered in 1786. It orbits the sun each 3.31 earth-years. Eccentricity $e = 0.846$.

30. Comet Giacobini-Zinner, discovered in 1900, orbits the sun each 6.59 earth-years. Eccentricity $e = 0.708$.

31. Comet Schwassmann-Wachmann, discovered in 1930, orbits the sun every 5.36 earth-years. Eccentricity $e = 0.694$.

32. Comet Swift-Tuttle was discovered in 1862. It orbits the sun each 120 earth-years. Eccentricity $e = 0.960$.

Comet Animations

Make an animation plot of comet positions. Use constrained scaling with major-semiaxis 1 in the plot. Display the given period T and eccentricity e in the legend.

33. Comet Churyumov-Gerasimenko
$T = 6.57$, $e = 0.632$.

34. Comet Wirtanen
$T = 5.46$, $e = 0.652$.

35. Comet Wild 2
$T = 6.39$, $e = 0.540$.

36. Comet Biela
$T = 6.62$, $e = 0.756$.

37. Comet Encke
$T = 3.31$, $e = 0.846$.

38. Comet Giacobini-Zinner
$T = 6.59$, $e = 0.708$.

39. Comet Schwassmann-Wachmann
$T = 5.36$, $e = 0.694$.

40. Comet Swift-Tuttle
$T = 120$, $e = 0.960$.

4.9 Fish Farming

Discussed are logistic models for population dynamics in fish farms. The models are suitable for Pangasius and Tilapia populations. The focus will be on species *tilapia.*

Pangasius. In America, both USA-produced and imported fresh-water catfish can be sold with the labels **Swai**, **Basa** or the subgenus label **Pangasius**, which is the predominant generic label in Europe, with more than 20 varieties. Basa and Swai are different catfish, with different texture and flavor. USA production of farmed catfish increased after 2002, when Vietnam Basa imports were stopped by labeling laws and tariffs. USA channel catfish (four barbels) are harvested after 18 months, at 10 pounds weight. Pangasius varieties are harvested after 4–6 months, at about 2 pounds or less, to produce fillets of 3–12 ounces.

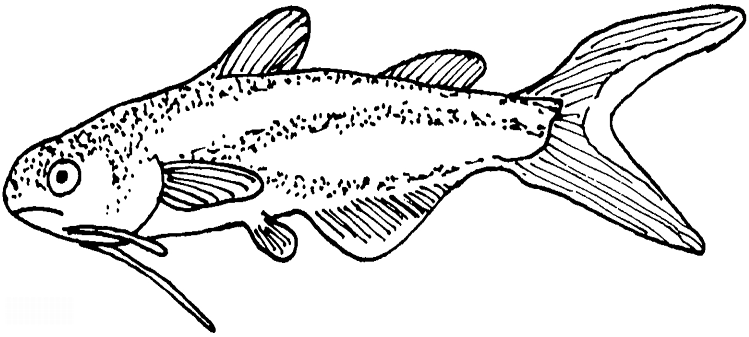


Figure 15. Pangasius, a fresh water catfish with two barbels.

Tilapia. This fresh-water fish originated in Africa 2500 years ago. The popular varieties sold in the USA are marketed under the label **Tilapia** (both dark and light flesh). They are produced in the USA at fish farms in Arizona, California and Florida. Imported Tilapia at 600-900 grams market weight (30% fillets) make up the bulk of USA-consumed Tilapia.

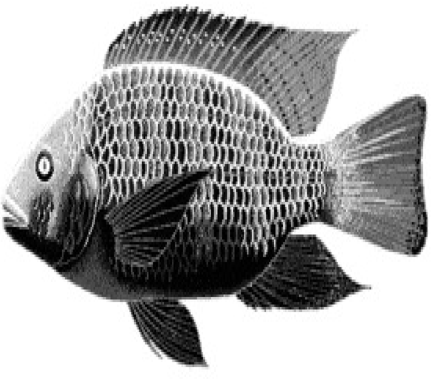


Figure 16. Tilapia.
A fresh water fish from the river Nile. Tilapia are farmed around the world in temperate climates.

Population Dynamics of Fisheries

Fisheries can be wild or farmed. One example is a fish hatchery using concrete tanks. Tilapia freshwater farms can use earthen ponds, canvas tanks, concrete tanks, river cages, pens and old mining quarries.

Tilapia Farming

Detailed life history data for Tilapia is as follows:

- Age at sexual maturity: 5–6 months
- Size at sexual maturity: 28–350 grams
- Stocking ratio for spawning: 7–10 broods/year using 2–5 females per male
- Spawning success: 20–30% spawns per week
- Eggs per female fish: 1–4 eggs per gram of fish
- Survival of egg to fry: 70–90% (fry less than 5 grams)
- Survival of fry to fingerling: 60–90% (fingerling 5–30 grams)
- Survival of fingerling to market: 70–98% (market is 30 to 680 grams)

Tilapia fry might be produced from an initial stock of 1000 female ND-2 and 250 male ND-1. Hatched ND-21 fry will be all male, which have higher market weight. Egg production per female averages from 300 to 500 fry per month, with about 10% lost before reaching 5 gram weight. The marketed Tilapia are about 900 grams in Central America plants (Belize, El Salvador). In Arizona, California and Florida plants, Tilapia market weights vary from 600 to 800 grams, or 1.5–1.75 pounds.

In commercial secondary tanks, fingerlings grow in water temperatures 76–84 degrees Fahrenheit with a death rate of about 0.05%. One fingerling grows to market size on less than 3 pounds of food.

Logistic Harvesting on a Time Interval

The Logistic equation for a **constant harvesting** rate $h \geq 0$ is

$$\frac{dx}{dt} = kx(t)(M - x(t)) - h.$$

The Logistic equation for a non-constant harvesting rate $h(t) \geq 0$ is

$$\frac{dx}{dt} = kx(t)(M - x(t)) - h(t).$$

A simplified situation is constant harvesting $h(t) = c > 0$ on a given time interval $a \le t \le b$, but zero otherwise.

In a more sophisticated setting, $h(t)$ is a positive constant c_i on given time interval $a_i \le t \le b_i$, $i = 1, \ldots, n$, but zero otherwise. Harvesting can also depend on the population size, which replaces $h(t)$ by $h(t)x(t)$ in the differential equation. Modelling need not be for an individual tank or pond, but the aggregate of all tanks, ponds and cages of an enterprise, viewed from the prospect of so many fish grown to market weight.

Logistic Periodic Harvesting

The periodic harvest Logistic equation is

$$\frac{dx}{dt} = kx(t)(M - x(t)) - h(t)$$

where $h(t) \ge 0$ is the rate of harvest, usually a positive constant c_i on a given time interval $a_i \le t \le b_i$, $i = 1, \ldots, n$, but zero otherwise. The equation $h(t+T) = h(t)$ might hold for some value of T, in which case $h(t)$ is a classical periodic function.

Tank harvests can be periodic, in order to reduce the density of fish per volume of water, or to remove fingerlings. Harvested fish can be assumed to be live, and sent either to slaughter or else to another tank, to grow bigger. This model fits Tilapia fry production in ponds, for which it is typical that ND-2 females produce more and more eggs as they mature (then $c_1 < c_2 < c_3 < \cdots$). The time intervals for Tilapia are about a month apart.

Malaysian Tilapia Example

Described here is the 2012 work of M. F. Laham, et al, [Laham], in which a logistic model is used to study harvesting strategies for tilapia fish farming. This work is elementary, in the sense that it treats an ideal example, with no intentional application to management of a Tilapia farm. It illustrates general expectations for fish production, based on gross estimates of a pond scenario.

The data was obtained from the Department of Fisheries of Malaysia and from the Malaysian fish owner of selected ponds situated at Gombak, Selangor. The fisheries department claims (2008) that a fish pond can sustain 5 tilapia fish for every square meter of surface area.[4] The selected pond has an area of 15.61 Hectors, which is equivalent to 156100 square meters, 38 acres or 25000 square feet. The pond carrying capacity is $M = 780500$ fish. According to a Malaysian study in 1999 (see [Laham]), Tilapia mature in 6 months and at least 80 percent will survive to maturity.

[4] Normal stocking is 1.6 fish per square meter, from which reproduction allows fish population growth to carrying capacity (a theoretical number).

The Logistic Growth Model, in the absence of harvesting, can be written in the form

$$\frac{dx}{dt} = rx(t)(1 - x(t)/M), \quad r = 0.8, \quad M = 780500.$$

In terms of the alternate model $P' = kP(M - P)$, the constant k equals $rM = 624400$. The 2012 work [Laham] focuses on harvesting strategies, considering the constant harvesting model

$$(1) \qquad \frac{dx}{dt} = rx(t)(1 - x(t)/M) - H_0$$

and the periodic harvesting model

$$(2) \qquad \frac{dy}{dt} = ry(t)(1 - y(t)/M) - H(t), \quad H(t) = \begin{cases} H_0 & 0 \le t \le 6, \\ 0 & 6 < t \le 12. \end{cases}$$

The constant $H_0 = 156100$ is explained below.The discontinuous harvesting function $H(t)$ is extended to be 12-month periodic: $H(t + 12) = H(t)$.

Constant Harvesting. The parameters in the model are $r = 0.8$, an estimate of the fraction of fish that will survive to market age, and the pond carrying capacity $M = 780500$. The periodic harvesting value $H_0 = 156100$ arises from the constant harvesting model, by maximizing population size at the equilibrium point for the constant harvesting model. Briefly, the value H_0 is found by requiring $\frac{dx}{dt} = 0$ in the constant harvesting model, replacing $x(t)$ by constant P. This implies

$$(3) \qquad rP\left(1 - \frac{P}{M}\right) - H_0 = 0.$$

The mysterious value H_0 is the one that makes the discriminant zero in the quadratic formula for P. Then $H_0 = \frac{rM}{4} = 156100$ and $P = 389482$. This **bifurcation point** separates the global behavior of the constant harvesting model as in Table 22. We use the notation P_1, P_2 for the two real equilibrium roots of the quadratic equation (3), assuming $H_0 < 156100$ and $P_1 < P_2$.

Table 22. Constant Harvesting Model

Harvest Constant	Initial Population	Behavior
$H_0 = 156100$	$x(0) \ge 389482$	$x(t) \to 389482$,
$H_0 = 156100$	$x(0) < 389482$	$x(t) \to 0$, extinction,
$H_0 > 156100$	any $x(0)$	$x(t) \to 0$, extinction,
$H_0 < 156100$	$x(0) < P_1$	$x(t) \to 0$, extinction,
$H_0 < 156100$	$P_1 < x(0) < P_2$	$x(t) \to P_2$, sustainable,
$H_0 < 156100$	$x(0) \ge P_2$	$x(t) \to P_2$, sustainable.

Periodic Harvesting. The model is an initial value problem (2) with initial population $y(0)$ equal to the number of Tilapia present, where $t = 0$ is an artificial time representing the current time after some months of growth. The plan is to harvest H_0 fish in the first 6 months.

Direct inspection of the two models shows that $x(t) = y(t)$ for the first six months, regardless of the choice of H_0. Because the constant harvesting model shows that harvesting rates larger than 156100 lead to extinction, then it is clear that the harvesting rate can be $H_0 = 156100$.

The harvesting constant H_0 can be larger than 156100, because the population of fish is allowed to recover for six months after the harvest. Assuming $H_0 > 156100$, then the solution $y(t)$ decreases for 6 months to value $y(6)$, which if positive, allows recovery of the population in the following 6 non-harvest months. There is a catch: the population could fail to grow to harvest size in the following 6 months, causing a reduced production in subsequent years.

To understand the problem more clearly, we present an example where $H_0 > 156100$ and the harvest is sustainable for 3 years, then another example where $H_0 > 156100$ and the harvest fails in the second year.

Example 4.8 (Sustainable Harvest $H_0 > 156100$)
Choose $H_0 = 190000$ and $y(0) = 390250 = M/2$. Computer assist gives 6-month population size decreasing to $y(6) = 16028.6$. Then for $6 < t < 12$ the population increases to $y(12) = 560497.2$, enough for a second harvest. The population continues to rise and fall, $y(18) = 320546.6$, $y(24) = 771390.7$, $y(30) = 391554.0$, $y(36) = 774167.6$, a sustainable harvest for the first three years.

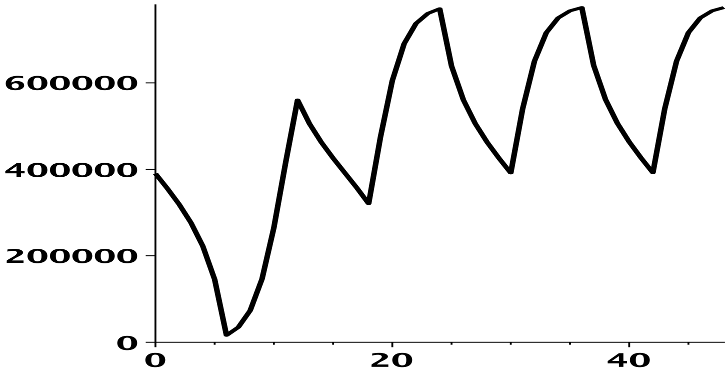


Figure 17. Sustainable harvest for 3 years, $H_0 = 190000$, $y(0) = M/2$. Abcissa t in months. Ordinate $y(t)$ is population size.

Example 4.9 (Unsustainable Harvest $H_0 > 156100$)
Choose $H_0 = 190500$ and $y(0) = 390250 = M/2$. Computer assist gives 6-month population size decreasing to $y(6) = 5263.1$. Then for $6 < t < 12$ the population increases to $y(12) = 352814$, enough for a second harvest. At $t = 16.95$ the population $y(t)$ decreases to zero (extinction), meaning the harvest fails in the second year.

The same example with $y(0) = (M/2)(1.02) = 398055$ (2 percent larger) happens to be sustainable for three years. Sustainable harvest is sensitive to both harvesting constant and initial population.

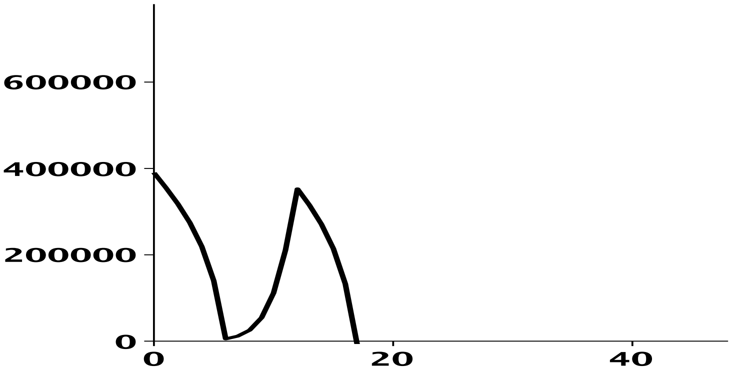


Figure 18. Unsustainable harvest, failure in year two.
$H_0 = 190500$, $y(0) = M/2$. Abcissa t in months. Ordinate $y(t)$ is population size.

Logistic Systems

The Lotka-Volterra equations, also known as the predator-prey equations, are a pair of first order nonlinear differential equations frequently used to describe the dynamics of biological systems in which two species interact, one a predator and one its prey (e.g., foxes and rabbits). They evolve in time according to the pair of equations:

$$\begin{aligned} \frac{dx}{dt} &= x(\alpha - \beta y), \\ \frac{dy}{dt} &= -y(\gamma - \delta x) \end{aligned}$$

where:
x is the number of prey,
y is the number of some predator,
t is time,
$\frac{dy}{dt}$ and $\frac{dx}{dt}$ are population growth rates,
Parameter α is a growth rate for the prey while parameter γ is a death rate for the predator.
Parameters β and δ describe species interaction, with $-\beta xy$ decreasing prey population and δxy increasing predator population.

A. J. Lotka (1910, 1920) used the predator-prey model to study autocatalytic chemical reactions and organic systems such as plants and grazing animals. In 1926, V. Volterra made a statistical analysis of fish catches in the Adriatic Sea, publishing at age 22 the same equations, an independent effort.

Walleye on Lake Erie

The one-dimensional theory of the logistic equation can be applied to fish populations in which there is a predator fish and a prey fish. This problem was studied by A. L. Jensen in 1988. Using the Canadian model of P. A. Larkin 1966, Jensen invented a mathematical model for walleye populations in the western basin of Lake Erie. The examples for **Prey** are Rainbow Smelt (*Osmerus mordax*) in Lake Superior and **Yellow Perch** (*Perca flavescens*) from Minnesota lakes. The **predator** is Walleye (*Sander vitreus*).

Figure 19. Yellow Perch.
The prey, from Shagawa Lake in Northeast Minnesota.

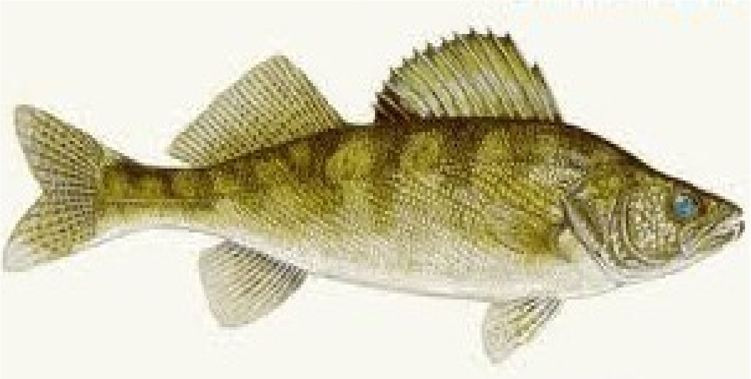


Figure 20. Walleye.
The predator, also called Yellow Pike, or Pickerel.

The basis for the simulation model is the Lotka-Volterra predator-prey model. The following assumptions were made.

- A decrease in abundance results in an increase in food concentration.
- An increase in food concentration results in an increase in growth and size.
- An increase in growth and size results in a decrease in mortality because mortality is a function of size.

The relation between prey abundance N_1 and predator abundance N_2 is given by the equations

$$\begin{aligned}\frac{dN_1}{dt} &= r_l N_1(1 - N_1/K_1) - b_1 N_1 N_2,\\ \frac{dN2}{dt} &= r_2 N_2(1 - N_2/K_2) - b_2 N_1 N_2.\end{aligned}$$

If $b_1 = b_2 = 0$, then there is no interaction of predator and prey, and the two populations N_1, N_2 grow and decay independently of one another. The carrying capacities are K_1, K_2, respectively, because each population N satisfies a logistic equation

$$\frac{dN}{dt} = rN(1 - N/K).$$

The literature below has further details. Solution methods for systems like (20) are largely numeric. Qualitative methods involving equilibrium points and phase diagrams have an important role in the analysis.

Jensen, A. L.: *Simulation of the potential for life history components to regulate Walleye population size*, Ecological Modelling 45(1), pp 27-41, 1989.

Larkin, P.A., 1966: *Exploitation in a type of predator-prey relationship. J. Fish. Res. Board Can.*, 23, pp 349-356, 1966.

Maple Code for Figures 17 and 18

The following sample `maple` code plots the solution on $0 < t < 24$ months with data $H_0 = 190000$, $P_0 = 390250$.

```
de:=diff(P(t),t)=r*(1-P(t)/M)*P(t)-H(t);
r:=0.8:M:=780500:H0:=190000:P0:=M/2:
H:=t->H0*piecewise(t<6,1,t<12,0,t<18,1,0);
DEtools[DEplot](de,P(t),t=0..24,P=0..M,[[P(0)=P0]]);
```

Exercises 4.9

Constant Logistic Harvesting

The model

$$x'(t) = kx(t)(M - x(t)) - h$$

can be converted to the logistic model

$$y'(t) = (a - by(t))y(t)$$

by a change of variables. Find the change of variables $y = x + c$ for the following pairs of equations.

1. $x' = -3x^2 + 8x - 5$,
$y' = (2 - 3y)y$

2. $x' = -2x^2 + 11x - 14$,
$y' = (3 - 2y)y$

3. $x' = -5x^2 - 19x - 18$,
$y' = (1 - 5y)y$

4. $x' = -x^2 + 3x + 4$,
$y' = (5 - y)y$

Periodic Logistic Harvesting

The periodic harvesting model

$$x'(t) = 0.8x(t)\left(1 - \frac{x(t)}{780500}\right) - H(t)$$

is considered with H defined by

$$H(t) = \begin{cases} 0 & 0 < t < 5, \\ H_0 & 5 < t < 6, \\ 0 & 6 < t < 17, \\ H_0 & 17 < t < 18, \\ 0 & 18 < t < 24. \end{cases}$$

This project makes as computer graph of the solution on $0 < t < 24$ for various values of H_0 and $x(0)$. See Figures 17 and 18 and the corresponding examples.

5. $H_0 = 156100$, $P(0) = 300000$

6. $H_0 = 156100$, $P(0) = 800000$

7. $H_0 = 800100$, $P(0) = 90000$

8. $H_0 = 800100$, $P(0) = 100000$

von Bertalanffy Equation

Karl Ludwig von Bertalanffy (1901-1972) derived in 1938 the equation

$$\frac{dL}{dt} = r_B(L_\infty - L(t))$$

from simple physiological arguments. It is a widely used growth curve, especially important in fisheries studies. The symbols:

t	time,
$L(t)$	length,
r_B	growth rate,
L_∞	expected length for zero growth.

9. Solve $\frac{dL}{dt} = 2(10-L)$, $L(0) = 0$. The answer is the length in inches of a fish over time, with final adult size 10 inches.

10. Solve von Bertalanffy's equation to obtain the algebraic model

$$L(t) = L_\infty\left(1 - e^{-r_B(t-t_0)}\right).$$

11. Assume von Bertalanffy's model. Suppose field data $L(0) = 0$, $L(1) = 5$, $L(2) = 7$. Display details using Exercise 10 to arrive for $t_0 = 0$ at values $L_\infty = 25/3$ and $r_B = \ln(5/2)$.

12. Assume von Bertalanffy's model with field data $L(0) = 0$, $L(1) = 10$, $L(2) = 13$. Find the expected length L_∞ of the fish.

Chapter 5

Linear Algebra

Contents

Linear algebra topics specific to **linear algebraic equations** were presented earlier in this text as an extension of college algebra topics, without the aid of vector-matrix notation.

The project before us introduces **specialized vector-matrix notation** in order to extend methods for solving linear algebraic equations. Enrichment includes a full study of rank, nullity, basis and independence from a vector-matrix viewpoint.

Engineering science views linear algebra as an essential language interface between an application and a computer algebra system or a computer numerical laboratory. Without the language interface provided by vectors and matrices, computer assist would be impossibly tedious.

Linear algebra with computer assist is advantageous in the study of **mechanical systems** and **electrical networks**, in which the notation and methods of linear algebra play an important and essential role.

5.1 Vectors and Matrices

The advent of computer algebra systems and computer numerical laboratories has precipitated a common need among engineers and scientists to learn the language of vectors and matrices, which is used heavily for theoretical analysis and computation in applications.

Fixed Vector Model

A **fixed vector** $\vec{X}$ is a one-dimensional array called a **column vector** or a **row vector**, denoted correspondingly by

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{or} \quad \vec{X} = \begin{pmatrix} x_1, x_2, \ldots, x_n \end{pmatrix}. \tag{1}$$

The **entries** or **components** x_1, …, x_n are numbers and n is correspondingly called the **column dimension** or the **row dimension** of the vector in (1). The set of all n-vectors (1) is denoted $\mathcal{R}^n$.

Practical matters. A fixed vector is a **package** of application data items. The term **vector** means **data item package** and the collection of all data item packages is the **data set**. Data items are usually numbers. A fixed vector imparts an implicit ordering to the package. To illustrate, a fixed vector might have $n = 6$ components x, y, z, p_x, p_y, p_z, where the first three are space position and the last three are momenta, with respective associated units meters and kilogram-meters per second.

Vector addition and **vector scalar multiplication** are defined by componentwise operations:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad k \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} kx_1 \\ kx_2 \\ \vdots \\ kx_n \end{pmatrix}.$$

The Mailbox Analogy

Fixed vectors can be visualized as in Table 1. Fixed vector entries x_1, …, x_n are numbers written individually onto papers $1, 2, \ldots, n$ deposited into mailboxes with names 1, 2, …, n.

Table 1. The Mailbox Analogy. Box i has contents x_i.

x_1	mailbox 1
x_2	mailbox 2
$\vdots$	$\vdots$
x_n	mailbox n

Free Vector Model

In the model, **rigid motions** from geometry are applied to directed line segments. A line segment $\overline{PQ}$ is represented as an **arrow** with head at Q and tail at P. Two such arrows are considered **equivalent** if they can be **rigidly translated** to the same arrow whose tail is at the origin. The arrows are called **free vectors**. They are denoted by the symbol $\overrightarrow{PQ}$, or sometimes $\vec{A} = \overrightarrow{PQ}$, which assigns label $\vec{A}$ to the arrow with tail at P and head at Q.

The parallelogram rule defines **free vector addition**, as in Figure 1. To define **free vector scalar multiplication** $k\vec{A}$, we change the location of the head of vector $\vec{A}$; see Figure 2. If $0 < k < 1$, then the head shrinks to a location along the segment between the head and tail. If $k > 1$, then the head moves in the direction of the arrowhead. If $k < 0$, then the head is reflected along the line and then moved.

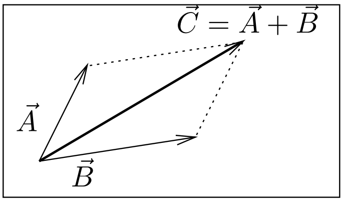


Figure 1. Free vector addition. The diagonal of the parallelogram formed by free vectors $\vec{A}$, $\vec{B}$ is the sum vector $\vec{C} = \vec{A} + \vec{B}$.

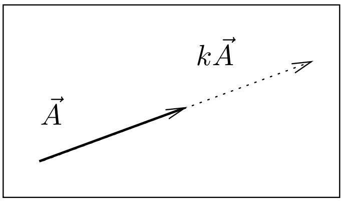


Figure 2. Free vector scalar multiplication. To form $k\vec{A}$, the head of free vector $\vec{A}$ is moved to a new location along the line formed by the head and tail.

Physics Vector Model

This model is also called the **$\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ vector model** and the **orthogonal triad model**. The model arises from the free vector model by inventing symbols $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ for a mutually orthogonal triad of free vectors. Usually, these three vectors represent free vectors of unit length along the coordinate axes, although use in the literature is not restricted to this specialized setting; see Figure 3.

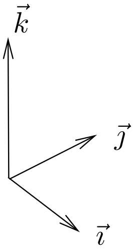


Figure 3. Fundamental triad. The free vectors $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ are 90° apart and of unit length.

The advantage of the model is that any free vector can be represented as $a\vec{\imath} + b\vec{\jmath} + c\vec{k}$ for some constants a, b, c, which gives an immediate connection to the free vector with head at (a, b, c) and tail at $(0, 0, 0)$, as well as to the fixed vector whose components are a, b, c.

Vector addition and scalar multiplication are defined **componentwise**: if $\vec{A} = a_1\vec{\imath} + a_2\vec{\jmath} + a_3\vec{k}$, $\vec{B} = b_1\vec{\imath} + b_2\vec{\jmath} + b_3\vec{k}$ and c is a constant, then

$$\begin{aligned}\vec{A} + \vec{B} &= (a_1 + b_1)\vec{\imath} + (a_2 + b_2)\vec{\jmath} + (a_3 + b_3)\vec{k},\\ c\vec{A} &= (ca_1)\vec{\imath} + (ca_2)\vec{\jmath} + (ca_3)\vec{k}.\end{aligned}$$

Formally, computations involving the **physics model** amount to fixed vector computations and the so-called *equalities* between free vectors and fixed vectors: $\vec{\imath} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\vec{\jmath} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\vec{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Gibbs Vector Model

The model assigns physical properties to vectors, thus avoiding the pitfalls of free vectors and fixed vectors. Gibbs defines a vector as a **linear motion** that takes a point A into a point B. Visualize this idea as a workman who carries material from A to B: the material is loaded at A, transported along a straight line to B, and then deposited at B. Arrow diagrams arise from this idea by representing a motion from A to B as an arrow with tail at A and head at B.

Vector addition is defined as composition of motions: material is loaded at A and transported to B, then loaded at B and transported to C. Gibbs' idea in the plane is the parallelogram law; see Figure 4.

Vector scalar multiplication is defined so that 1 times a motion is itself, 0 times a motion is no motion and -1 times a motion loads at B and transports to A (the reverse motion). If $k > 0$, then k times a motion from A to B causes the load to be deposited at C instead of B, where k is the ratio of the lengths of segments $\overline{AC}$ and $\overline{AB}$. If $k < 0$, then the definition is applied to the reverse motion from B to A using instead of k the constant $|k|$. Briefly, the load to be deposited along the direction to B is dropped earlier if $0 < |k| < 1$ and later if $|k| > 1$.

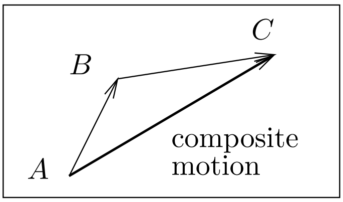


Figure 4. Planar composition of motions. The motion A to C is the composition of two motions or the *sum* of vectors AB and BC.

Comparison of Vector Models

In free vector diagrams it is possible to use free, physics and Gibbs vector models almost interchangeably. In the Gibbs model, the negative of a vector and the zero vector are natural objects, whereas in the other models they can be problematic. To understand the theoretical difficulties, try to answer these questions:

1. What is the zero vector?
2. What is the meaning of the negative of a vector?

Some working rules which connect the free, physics and Gibbs models to the fixed model are the following.

Conversion A fixed vector $\vec{X}$ with components a, b, c is realized as a free vector by drawing an arrow from $(0, 0, 0)$ to (a, b, c).

Addition To add two free vectors, $\vec{Z} = \vec{X} + \vec{Y}$, place the tail of $\vec{Y}$ at the head of $\vec{X}$, then draw vector $\vec{Z}$ to form a triangle, from the tail of $\vec{X}$ to the head of $\vec{Y}$.

Subtraction To subtract two free vectors, $\vec{Z} = \vec{Y} - \vec{X}$, place the tails of $\vec{X}$ and $\vec{Y}$ together, then draw $\vec{Z}$ between the heads of $\vec{X}$ and $\vec{Y}$, with the heads of $\vec{Z}$ and $\vec{Y}$ together.

Head Minus Tail A free vector $\vec{X}$ converts to a fixed vector whose components are the componentwise differences between the point at the head and the point at the tail. This statement is called the **head minus tail rule**.

Vector Spaces and the Toolkit

Consider any vector model: fixed, free, physics or Gibbs. Let V denote the **data set** of one of these models. The data set consists of packages of data items, called **vectors**.[1] Assume a particular dimension, n for fixed, 2 or 3 for the others. Let k, k_1, k_2 be constants. Let $\vec{X}$, $\vec{Y}$, $\vec{Z}$ represent three vectors in V. The following **toolkit** of eight (8) vector properties can be verified from the definitions.

[1] If you think vectors are arrows, then re-tool your thoughts. Think of vectors as **data item packages**. A technical word, **vector** can also mean a graph, a matrix for a digital photo, a sequence, a signal, an impulse, or a differential equation solution .

Closure	The operations $\vec{X} + \vec{Y}$ and $k\vec{X}$ are defined and result in a new data item package [a vector] which is also in V.	
Addition	$\vec{X} + \vec{Y} = \vec{Y} + \vec{X}$	commutative
	$\vec{X} + (\vec{Y} + \vec{Z}) = (\vec{Y} + \vec{X}) + \vec{Z}$	associative
	Vector $\vec{0}$ is defined and $\vec{0} + \vec{X} = \vec{X}$	zero
	Vector $-\vec{X}$ is defined and $\vec{X} + (-\vec{X}) = \vec{0}$	negative
Scalar multiply	$k(\vec{X} + \vec{Y}) = k\vec{X} + k\vec{Y}$	distributive I
	$(k_1 + k_2)\vec{X} = k_1\vec{X} + k_2\vec{X}$	distributive II
	$k_1(k_2\vec{X}) = (k_1 k_2)\vec{X}$	distributive III
	$1\vec{X} = \vec{X}$	identity

Definition 5.1 (Vector Space)
A data set V equipped with $\boxed{+}$ and $\boxed{\cdot}$ operations satisfying the closure law and the eight toolkit properties is called an **abstract vector space**.

What's a *space*? There is no intended geometrical implication in this term. The usage of **space** originates from phrases like **parking space** and **storage space**. An abstract vector space is a data set for an application, organized as packages of data items, together with $\boxed{+}$ and $\boxed{\cdot}$ operations, which satisfy the eight toolkit manipulation rules. The packaging of individual data items is structured, or organized, by some scheme, which amounts to a *storage space*, hence the term *space.*

What does *abstract* mean? The technical details of the packaging and the organization of the data set are invisible to the toolkit rules. The toolkit acts on the formal packages, which are called **vectors**. Briefly, the toolkit is used **abstractly**, devoid of any details of the storage scheme. **Bursting** data packages into data items is generally counterproductive for algebraic manipulations. Resist the temptation to burst vectors.

A variety of data sets. The following key examples are a basis for initial intuition about vector spaces.

> **Coordinate space** $\mathcal{R}^n$ is the set of all fixed n-vectors. Sets $\mathcal{R}^n$ are structured packaging systems which organize data sets from calculations, geometrical problems and physical vector diagrams.
>
> **Function spaces** are structured packages of graphs, such as solutions to differential equations.
>
> **Infinite sequence spaces** are suited to organize the coefficients of numerical approximation sequences. Additional applications are coefficients of Fourier series and Taylor series.
>
> A **Matrix space** is a structured system which can organize two-dimensional data sets. Examples are the array of pixels for a digital photograph and robotic mechanical component manipulators represented by 3 × 3 or 4 × 4 matrices.

Subspaces and Data Analysis

Subspaces address the issue of how to do efficient data analysis on a smaller subset S of a data set V. We assume the larger data set V is equipped with $\boxed{+}$ and $\boxed{\cdot}$ and has the 8-property toolkit: it is an abstract vector space by assumption.

Slot racer on a track. To illustrate the idea, consider a problem in planar kinematics and a laboratory data recorder that approximates the x, y, z location of an object in 3-dimensional space. The recorder puts the data set of the kinematics problem into fixed 3-vectors. After the recording, the data analysis begins.

From the beginning, the kinematics problem is planar, and we should have done the data recording using 2-vectors. However, the plane of action may not be nicely aligned with the axes set up by the data recorder, and this spin on the experiment causes the 3-dimensional recording.

The kinematics problem and its algebraic structure are exactly planar, but the geometry for the recorder data may be opaque. For instance, the experiment's acquisition plane might be given approximately by a **homogeneous restriction equation** like

$$x + 2y - 1000z = 0.$$

The **restriction equation** is preserved by operations $\boxed{+}$ and $\boxed{\cdot}$ (details postponed). Then data analysis on the smaller planar data set can proceed to use the toolkit at will, knowing that all calculations will be in the plane, hence physically relevant to the original kinematics problem.

Physical data in reality contains errors, preventing the data from exactly satisfying an ideal restriction equation like $x + 2y - 1000z = 0$. Methods like **least squares** can construct the idealized equations. The physical data is then converted by projection, making a new data set S that exactly satisfies the restriction equation $x + 2y - 1000z = 0$. It is this modified set S, the working data set of the application, that we call a subspace.

Applied scientists view subspaces as **working sets**, which are actively constructed and rarely discovered without mathematical effort. The construction is guided by the subspace criterion, Theorem 5.1, page 300.

Definition 5.2 (Subspace)

A subset S of an abstract vector space V is called a **subspace** if it is a nonempty vector space under the operations of addition and scalar multiplication inherited from V.

In applications, a subspace S of V is a smaller data set, recorded using the same data packages as V. The smaller set S contains at least the zero vector $\vec{\mathbf{0}}$. Required is that the algebraic operations of addition and scalar multiplication acting on S give answers back in S. Then the entire 8-property toolkit is available for calculations in the smaller data set S.

Theorem 5.1 (Subspace Criterion)
Assume abstract vector space V is equipped with addition $(+)$ and scalar multiplication $(\cdot)$. A subset S is a subspace of V provided these checkpoints hold:

Vector $\vec{\mathbf{0}}$ is in S (S is nonvoid).

For each pair $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ in S, the vector $\vec{\mathbf{v}}_1 + \vec{\mathbf{v}}_2$ is in S.

For each $\vec{\mathbf{v}}$ in S and constant c, the combination $c\vec{\mathbf{v}}$ belongs to S.

Actual use of the subspace criterion is rare, because most applications define a subspace S by a *restriction* on elements of V, normally realized as a set of linear homogeneous equations. Such systems can be re-written as a matrix equation $A\vec{\mathbf{u}} = \vec{\mathbf{0}}$. To illustrate, $x+y+z=0$ is re-written as a matrix equation as follows:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Theorem 5.2 (Subspaces of $\mathcal{R}^n$: The Kernel Theorem)
Let V be one of the vector spaces $\mathcal{R}^n$ and let A be an $m \times n$ matrix. Define the data set

$$S = \{\vec{\mathbf{v}} \ : \ \vec{\mathbf{v}} \text{ in } V \text{ and } A\vec{\mathbf{v}} = \vec{\mathbf{0}}\}.$$

Then S is a subspace of V, that is, operations of addition and scalar multiplication applied to data in S give data back in S and the 8-property toolkit applies to S-data.[2]
Proof on page 314.

When does Theorem 5.2 apply? Briefly, the kernel theorem hypothesis requires V to be a space of fixed vectors and S a subset defined by homogeneous restriction equations. A vector space of functions, used as data sets in differential equations, does not satisfy the hypothesis of Theorem 5.2, because V is not one of the spaces $\mathcal{R}^n$. This is why a subspace check for a function space uses the basic subspace criterion, and not Theorem 5.2.

Theorem 5.3 (Subspaces of $\mathcal{R}^n$: Restriction Equations)
Let V be one of the vector spaces $\mathcal{R}^n$ and let data set S be defined by a system of restriction equations. If the restriction equations are homogeneous linear algebraic equations, then S is a subspace of V.

How to apply Theorem 5.2 and Theorem 5.3. We illustrate with V the vector space $\mathcal{R}^4$ of all fixed 4-vectors with components x_1, x_2, x_3, x_4. Let S be the subset of V defined by the *restriction equation* $x_4 = 0$.

By Theorem 5.3, S is a subspace of V, with no further details required.

[2] This key theorem is named the **kernel theorem**, because solutions $\vec{x}$ of $A\vec{x} = \vec{0}$ define the **kernel** of A. It is also named the **Nullspace Theorem**.

To apply Theorem 5.2, the restriction equations have to be re-written as a homogeneous matrix equation $A\vec{\mathbf{x}} = \vec{\mathbf{0}}$:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Then Theorem 5.2 applies to conclude that S is a subspace of V.

When is S not a subspace? The following test enumerates three common conditions for which S fails to pass the subspace test. It is justified from the subspace criterion.

Theorem 5.4 (Test S not a Subspace)
Let V be an abstract vector space and assume S is a subset of V. Then S is not a subspace of V provided one of the following holds.

(1) The vector 0 is not in S.

(2) Some $\vec{\mathbf{x}}$ and $-\vec{\mathbf{x}}$ are not both in S.

(3) Vector $\vec{\mathbf{x}} + \vec{\mathbf{y}}$ is not in S for some $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ in S.

Linear Combinations and Closure

Definition 5.3 (Linear Combination)
A **linear combination** of vectors $\vec{\mathbf{v}}_1,\ldots,\vec{\mathbf{v}}_k$ is defined to be a sum

$$\vec{\mathbf{x}} = c_1\vec{\mathbf{v}}_1 + \cdots + c_k\vec{\mathbf{v}}_k,$$

where $c_1,\ldots,c_k$ are constants.

The **closure** property for a subspace S can be stated as *linear combinations of vectors in S are again in S*. Therefore, according to the subspace criterion, S is a subspace of V provided $\vec{\mathbf{0}}$ is in S and S is closed under the operations $+$ and $\cdot$ inherited from the larger data set V.

Definition 5.4 (Span)
Let vectors $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_k$ be given in a vector space V. The subset S of V consisting of all linear combinations $\vec{\mathbf{v}} = c_1\vec{\mathbf{v}}_1 + \cdots + c_k\vec{\mathbf{v}}_k$ is called the **span** of the vectors $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_k$ and written

$$S = \mathbf{span}(\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_k).$$

Important: The symbols $c_1, \ldots, c_n$ exhaust all possible choices of scalars: expect the **span** to contain infinitely many data packages (called *abstract* vectors) from data set V.

Theorem 5.5 (Span of Vectors is a Subspace)
Let V be an abstract vector space. A subset $S = \mathbf{span}(\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_k)$ is a subspace of V. Proof on page 314.

The Parking Lot Analogy

A useful visualization for *vector space* and *subspace* is a parking lot with valet parking. The large lot represents the **storage space** of the larger data set associated with a vector space V. The parking lot rules, such as *display your ticket*, *park between the lines*, correspond to the toolkit of 8 vector space rules. The valet parking lot S, which is a smaller roped-off area within the larger lot V, is also storage space, subject to the same rules as the larger lot. The smaller data set S corresponds to a subspace of V. Just as additional restrictions apply to the valet lot, a subspace S is generally defined by equations, relations or restrictions on the data items of V.

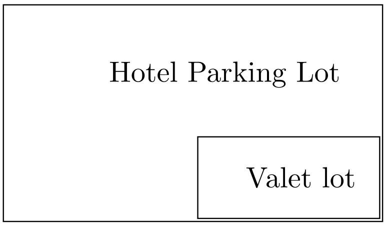


Figure 5. Parking lot analogy. An abstract vector space V and one of its subspaces S can be visualized through the analogy of a parking lot (V) containing a valet lot (S).

Vector Algebra

Definition 5.5 (Norm of a Fixed Vector)
The **norm** or **length** of a fixed vector $\vec{X}$ with components $x_1, \ldots, x_n$ is given by the formula

$$|\vec{X}| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

This measurement can be used to quantify the numerical error between two data sets stored in vectors $\vec{X}$ and $\vec{Y}$:

$$\textbf{norm-error} = |\vec{X} - \vec{Y}|.$$

Definition 5.6 (Dot Product or Scalar Product)
The **dot product** $\vec{X} \cdot \vec{Y}$ of two fixed vectors $\vec{X}$ and $\vec{Y}$ is defined by

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = x_1 y_1 + \cdots + x_n y_n.$$

Definition 5.7 (Angle Between Vectors)
Assume $|\vec{X}| > 0$ and $|\vec{Y}| > 0$. Define the angle θ, $0 \le \theta \le \pi$, between vectors $\vec{X}$ and $\vec{Y}$ by:

$$\cos\theta = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}||\vec{Y}|}.$$

Calculus vector geometry for $n = 3$ derives formula $|\vec{X}||\vec{Y}|\cos\theta = \vec{X} \cdot \vec{Y}$, which produces the above equation by solving for $\cos\theta$, motivation for the definition.

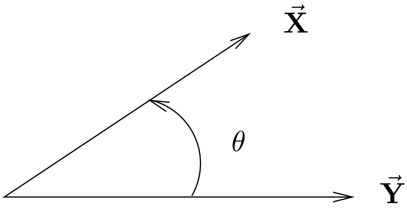


Figure 6. **Angle θ between two vectors $\vec{\mathbf{X}}$, $\vec{\mathbf{Y}}$.**

Definition 5.8 (Orthogonal Vectors)
Two n-vectors $\vec{X}$, $\vec{Y}$ are said to be **orthogonal** provided $\vec{X}\cdot\vec{Y}=0$.

If both vectors are nonzero, then $\cos(\theta)=\dfrac{\vec{X}\cdot\vec{Y}}{|\vec{X}||\vec{Y}|}=0$, which implies the angle between the vectors is $\theta=90^\circ$.

Definition 5.9 (Shadow Projection)
The **shadow projection** of vector $\vec{X}$ onto the direction of vector $\vec{Y}$ is the number d defined by

$$d=\frac{\vec{X}\cdot\vec{Y}}{|\vec{Y}|}.$$

The triangle determined by $\vec{X}$ and $(d/|\vec{Y}|)\vec{Y}$ is a right triangle.

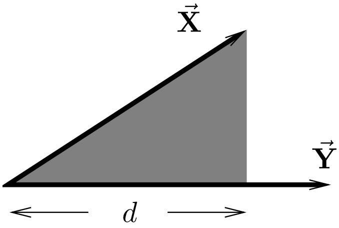


Figure 7. **Shadow projection d**
Distance d is the length of the shadow formed by vector $\vec{\mathbf{X}}$ onto the direction of vector $\vec{\mathbf{Y}}$.

Definition 5.10 (Vector Projection)
The **vector projection** of $\vec{X}$ onto the line L through the origin in the direction of $\vec{Y}$ is defined by

$$\mathbf{proj}_{\vec{Y}}(\vec{X})=d\frac{\vec{Y}}{|\vec{Y}|}=\frac{\vec{X}\cdot\vec{Y}}{\vec{Y}\cdot\vec{Y}}\vec{Y}.$$

Definition 5.11 (Vector Reflection)
The **vector reflection** of vector $\vec{X}$ in the line L through the origin having the direction of vector $\vec{Y}$ is defined to be the vector

$$\mathbf{refl}_{\vec{Y}}(\vec{X})=2\,\mathbf{proj}_{\vec{Y}}(\vec{X})-\vec{X}=2\frac{\vec{X}\cdot\vec{Y}}{\vec{Y}\cdot\vec{Y}}\vec{Y}-\vec{X}.$$

It is the formal analog of the complex conjugate map $a+ib\to a-ib$ with the x-axis replaced by line L.

Matrices are Vector Packages

A **matrix** A is a package of so many fixed vectors, considered together, and written as a 2-dimensional array

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

The packaging can be in terms of **column vectors** or **row vectors**:

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \cdots \\ a_{n1} \end{pmatrix} \cdots \begin{pmatrix} a_{1m} \\ a_{2m} \\ \cdots \\ a_{nm} \end{pmatrix} \quad \text{or} \quad \left\{ \begin{array}{c} (a_{11}, a_{12}, \ldots, a_{1n}) \\ (a_{21}, a_{22}, \ldots, a_{2n}) \\ \vdots \\ (a_{m1}, a_{m2}, \ldots, a_{mn}) \end{array} \right. .$$

Definition 5.12 (Equality of Matrices)
Two matrices A and B are said to be **equal** provided they have identical row and column dimensions and corresponding entries are equal. Equivalently, A and B are equal if they have identical columns, or identical rows.

Mailbox analogy. A matrix A can be visualized as a rectangular collection of so many mailboxes labeled (i, j) with contents a_{ij}, where the row index is i and the column index is j; see Table 2.

Table 2. The Mailbox Analogy for Matrices.
A matrix A is visualized as a block of mailboxes, each located by row index i and column index j. The box at (i, j) contains data a_{ij}.

a_{11}	a_{12}	$\cdots$	a_{1n}
a_{21}	a_{22}	$\cdots$	a_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{m1}	a_{m2}	$\cdots$	a_{mn}

Computer Storage

Computer programs might store matrices as a long single array. Array contents are fetched by computing the index into the long array followed by retrieval of the numeric content a_{ij}. From this computer viewpoint, vectors and matrices are the same objects.

For instance, a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ can be stored by stacking its rows into a column vector, the mathematical equivalent being the one-to-one and onto mapping

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \longleftrightarrow \quad \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

This mapping uniquely associates the 2×2 matrix A with a vector in $\mathcal{R}^4$. Similarly, a matrix of size $m \times n$ is associated with a column vector in $\mathcal{R}^k$, where $k = mn$.

Matrix Addition and Scalar Multiplication

Addition of two matrices is defined by applying fixed vector addition on corresponding columns. Similarly, an organization by rows leads to a second definition of matrix addition, which is exactly the same:

$$\begin{pmatrix} a_{11} \cdots a_{1n} \\ a_{21} \cdots a_{2n} \\ \vdots \\ a_{m1} \cdots a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} \cdots b_{1n} \\ b_{21} \cdots b_{2n} \\ \vdots \\ b_{m1} \cdots b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & \cdots & a_{2n} + b_{2n} \\ & \vdots & \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix}.$$

Scalar multiplication of matrices is defined by applying scalar multiplication to the columns or rows:

$$k \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ & \vdots & \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} ka_{11} & \cdots & ka_{1n} \\ ka_{21} & \cdots & ka_{2n} \\ & \vdots & \\ ka_{m1} & \cdots & ka_{mn} \end{pmatrix}.$$

Both operations on matrices are motivated by considering a matrix to be a long single array or *fixed vector*, to which the standard fixed vector definitions are applied. The operation of addition is properly defined exactly when the two matrices have the same row and column dimensions.

Digital Photographs

A digital camera stores image sensor data as a matrix A of numbers corresponding to the color and intensity of tiny sensor sites called **pixels** or **dots**. The pixel position in the print is given by row and column location in the matrix A.

A visualization of the image sensor is a checkerboard. Each square is stacked with a certain number of checkers, the count proportional to the number of electrons knocked loose by light falling on the photodiode site.

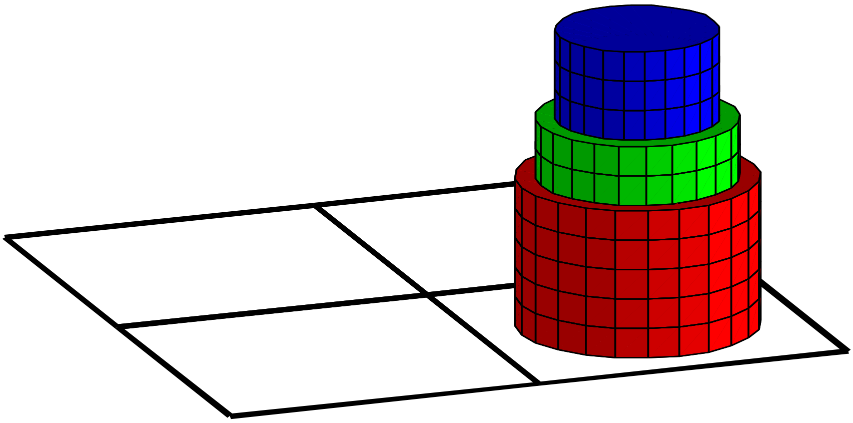


Figure 8. Checkerboard visualization.
Illustrated is a stack of checkers, representing one photodiode site on an image sensor inside a digital camera. There are 5 red, 2 green and 3 blue checkers stacked on one square, representing electron counts.

In 24-bit color, a pixel could be represented in matrix A by a coded integer $a = r + (2^8)g + (2^{16})b$. Symbols r, g, b are integers between 0 and 255 which represent the intensity of colors red, green and blue, respectively. For example, $r = g = b = 0$ is the color **black** while $r = g = b = 255$ is the color **white**.

A matrix of size $m \times n$ is visualized as a checkerboard with mn squares, each square stacked with red, green and blue checkers. Higher resolution image sensors store image data in huge matrices with richer color information, for instance 32-bit and 128-bit color.[3]

Visualization of Matrix Addition and Scalar Multiply

Matrix addition can be visualized through matrices representing color separations.[4] When three monochrome transparencies of colors red, green and blue (RGB) are projected simultaneously by a projector, the colors add to make a full color screen projection. The three transparencies can be associated with matrices R, G, B which contain pixel data for the monochrome images. Then the projected image is associated with the matrix sum $R + G + B$.

Matrix scalar multiplication has a similar visualization. The pixel information in a monochrome image (red, green or blue) is coded for intensity. The associated matrix A of pixel data when multiplied by a scalar k gives a new ma-

[3] A beginner's digital camera manufactured in the early days of digital photography made low resolution color photos using 24-bit color. The photo is constructed from 240 rows of dots with 320 dots per row. The associated storage matrix A is of size 240×320. The identical small format was used for video clips.

The storage format **BMP** stores data as bytes, in groups of three b, g, r, starting at the lower left corner of the photo. Therefore, 240×320 photos have $230,400$ data bytes. Storage format **JPEG** has replaced the early formats on phones.

[4] James Clerk Maxwell is credited with the idea of color separation.

trix kA of pixel data with the intensity of each pixel adjusted by factor k. The photographic effect is to adjust the range of intensities. In the checkerboard visualization of an image sensor, Figure 8 page 305, factor k increases or decreases the checker stack height at each square.

Color Separation Illustration

Consider the coded matrix

$$\vec{\mathbf{X}} = \begin{pmatrix} 514 & 3 \\ 131843 & 197125 \end{pmatrix}.$$

We will determine the monochromatic pixel data R, G, B in the equation $X = R + 2^8 G + 2^{16} B$.

First we decode the scalar equation $x = r + 2^8 g + 2^{16} b$ by these algebraic steps, which use the modulus function $\mathbf{mod}(x, m)$, defined to be the remainder after division of x by m. We assume r, g, b are integers between 0 and 255.

$y = \mathbf{mod}(x, 2^{16})$	The remainder should be $y = r + 2^8 g$.
$r = \mathbf{mod}(y, 2^8)$	Because $y = r + 2^8 g$, the remainder equals r.
$g = (y - r)/2^8$	Divide $y - r = 2^8 g$ by 2^8 to obtain g.
$b = (x - y)/2^{16}$	Because $x - y = x - r - 2^8 g$ has remainder b.
$r + 2^8 g + 2^{16} b$	Answer check. This should equal x.

Computer algebra systems can provide an answer for matrices R, G, B by duplicating the scalar steps. Below is a `maple` implementation that gives the answers

$$R = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}, \; G = \begin{pmatrix} 2 & 0 \\ 3 & 2 \end{pmatrix}, \; B = \begin{pmatrix} 0 & 0 \\ 2 & 3 \end{pmatrix}.$$

```
with(LinearAlgebra:-Modular):
X:=Matrix([[514,3],[131843,197125]]);
Y:=Mod(2^16,X,integer); # y=mod(x,65536)
R:=Mod(2^8,Y,integer);  # r=mod(y,256)
G:=(Y-R)/2^8;           # g=(y-r)/256
B:=(X-Y)/2^16;          # b=(x-y)/65536
X-(R+G*2^8+B*2^16);     # answer check
```

The result can be visualized through a checkerboard of 4 squares. The second square has 5 red, 2 green and 3 blue checkers stacked, representing the color $x = (5) + 2^8(2) + 2^{16}(3)$ - see Figure 8 page 305.

Matrix Multiply

College algebra texts cite the definition of matrix multiplication as *the product AB equals a matrix C given by the relations*

$$c_{ij} = a_{i1}b_{1j} + \cdots + a_{in}b_{nj}, \quad 1 \le i \le m,\ 1 \le j \le k.$$

Below, we motivate the definition of matrix multiplication from an applied point of view, based upon familiarity with the dot product.

Matrix multiplication as a dot product extension. To illustrate the basic idea by example, let

$$A = \begin{pmatrix} -1 & 2 & 1 \\ 3 & 0 & -3 \\ 4 & -2 & 5 \end{pmatrix}, \quad \vec X = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}.$$

The product equation $A\vec X$ is displayed as the *dotless juxtaposition*

$$\begin{pmatrix} -1 & 2 & 1 \\ 3 & 0 & -3 \\ 4 & -2 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix},$$

which represents an *unevaluated request* to **gang** the dot product operation onto the rows of the matrix on the left:

$$(-1\,2\,1)\cdot\begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = 3, \quad (3\,0\,-3)\cdot\begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = -3, \quad (4\,-2\,5)\cdot\begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = 21.$$

The *evaluated request* produces a column vector containing the dot product answers, called the **product of a matrix and a vector** (no mention of dot product), written as

$$\begin{pmatrix} -1 & 2 & 1 \\ 3 & 0 & -3 \\ 4 & -2 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ 21 \end{pmatrix}.$$

The general scheme which gangs the dot product operation onto the matrix rows can be written as

$$\begin{pmatrix} \cdots & \text{row 1} & \cdots \\ \cdots & \text{row 2} & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \text{row m} & \cdots \end{pmatrix} \vec X = \begin{pmatrix} (\text{row 1})\cdot\vec X \\ (\text{row 2})\cdot\vec X \\ \vdots \\ (\text{row m})\cdot\vec X \end{pmatrix}.$$

The product is properly defined only in case the number of matrix columns equals the number of entries in $\vec X$, so that the dot products on the right are defined.

Matrix multiply as a linear combination of columns. The identity

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} a \\ c \end{pmatrix} + x_2 \begin{pmatrix} b \\ d \end{pmatrix}$$

implies that $A\vec{\mathbf{x}}$ is a linear combination of the columns of A, where A is the 2×2 matrix on the left.

This result holds in general, a relation used so often that it deserves a formal statement.

Theorem 5.6 (Matrix Multiply as a Linear Combination of Columns)
Let matrix A have vector columns $\vec{\mathbf{v}}_1$,...,$\vec{\mathbf{v}}_n$ and let vector $\vec{X}$ have scalar components x_1, ..., x_n. Then the definition of matrix multiply implies

$$A\vec{X} = x_1\vec{\mathbf{v}}_1 + x_2\vec{\mathbf{v}}_2 + \cdots + x_n\vec{\mathbf{v}}_n.$$

General matrix product AB. The evaluation of matrix products $A\vec{Y}_1$, $A\vec{Y}_2$, ..., $A\vec{Y}_k$ is a list of k column vectors which can be packaged into a matrix C. Let B be the matrix which packages the columns $\vec{Y}_1$, ..., $\vec{Y}_k$. *Define* $C = AB$ by the dot product definition

$$c_{ij} = \mathbf{row}(A, i) \cdot \mathbf{col}(B, j).$$

This definition makes sense provided the column dimension of A matches the row dimension of B. It is consistent with the earlier definition from college algebra and the definition of $A\vec{Y}$, therefore it may be taken as *the basic definition for a matrix product.*

How to multiply matrices on paper. More arithmetic errors are made when computing dot products written in the form

$$\begin{pmatrix} -7 & 3 & 5 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 3 \\ -5 \end{pmatrix} = -9,$$

because alignment of corresponding entries must be done mentally. It is visually easier when the entries are aligned.

On paper, work can be arranged for a matrix times a vector as below, so that the entries align. The boldface transcription above the columns is temporary, erased after the dot product step.

$$\begin{array}{c} \begin{array}{rrr} \mathbf{-1} & \mathbf{3} & \mathbf{-5} \end{array} \\ \left(\begin{array}{rrr} -7 & 3 & 5 \\ -5 & -2 & 3 \\ 1 & -3 & -7 \end{array}\right) \end{array} \cdot \begin{pmatrix} -1 \\ 3 \\ -5 \end{pmatrix} = \begin{pmatrix} -9 \\ -16 \\ 25 \end{pmatrix}$$

Visualization of Matrix Multiply

Discussed here is a key example of how to interpret 2×2 matrix multiply as a geometric operation.

Let's begin by inspecting a 2×2 system $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ for its geometric meaning. Consider the system

$$\left|\begin{array}{rcl} y_1 &=& ax_1 + bx_2 \\ y_2 &=& cx_1 + dx_2 \end{array}\right| \quad \text{or} \quad \vec{\mathbf{y}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \vec{\mathbf{x}} \tag{2}$$

Geometric rotation and scaling of planar figures have equations of this form. Adopt below definitions of A, B:

$$\begin{array}{cc} \textbf{Rotation by angle } \theta & \textbf{Scale by factor } k \\ \hline A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} & B = \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix} \end{array} \tag{3}$$

The geometric effect of mapping points $\vec{\mathbf{x}}$ on an ellipse by the equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ is to rotate the ellipse. If we choose $\theta = \pi/2$, then it is a rotation by 90 degrees. The mapping $\vec{\mathbf{z}} = B\vec{\mathbf{y}}$ re-scales the axes by factor k. If we choose $k = 2$, then the geometric result is to double the dimensions of the rotated ellipse. The resulting geometric transformation of $\vec{\mathbf{x}}$ into $\vec{\mathbf{z}}$ has algebraic realization

$$\vec{\mathbf{z}} = B\vec{\mathbf{y}} = BA\vec{\mathbf{x}},$$

which means the composite transformation of rotation followed by scaling is represented by system (2), with coefficient matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = BA = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \cos\pi/2 & \sin\pi/2 \\ -\sin\pi/2 & \cos\pi/2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}.$$

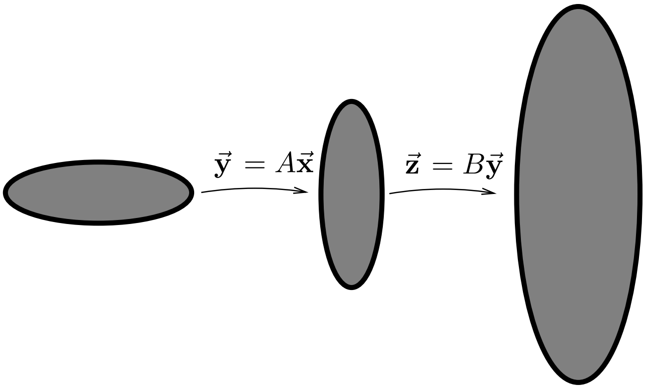

$\vec{\mathbf{y}} = A\vec{\mathbf{x}}$
$\vec{\mathbf{z}} = B\vec{\mathbf{y}}$

Figure 9. An ellipse is mapped into a rotated and re-scaled ellipse.
The rotation is $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, which is followed by re-scaling $\vec{\mathbf{z}} = B\vec{\mathbf{y}}$. The composite geometric transformation is $\vec{\mathbf{z}} = BA\vec{\mathbf{x}}$, which maps the ellipse into a rotated and re-scaled ellipse.

Special Matrices

The **zero matrix**, denoted **0**, is the $m \times n$ matrix all of whose entries are zero. The **identity matrix**, denoted I, is the $n \times n$ matrix with ones on the diagonal and zeros elsewhere: $a_{ij} = 1$ for $i = j$ and $a_{ij} = 0$ for $i \neq j$.

$$\mathbf{0} = \begin{pmatrix} 0\,0\cdots\,0 \\ 0\,0\cdots\,0 \\ \vdots \\ 0\,0\cdots\,0 \end{pmatrix}, \quad I = \begin{pmatrix} 1\,0\cdots\,0 \\ 0\,1\cdots\,0 \\ \vdots \\ 0\,0\cdots\,1 \end{pmatrix}.$$

The identity I is a package of column vectors called the **standard unit vectors** of size n. Literature may write the columns of I as $\vec{e}_1, \ldots, \vec{e}_n$ or as $\mathbf{col}(I, 1), \ldots, \mathbf{col}(I, n)$.

The **negative** of a matrix A is $(-1)A$, which multiplies each entry of A by the factor (-1):

$$-A = \begin{pmatrix} -a_{11} \cdots -a_{1n} \\ -a_{21} \cdots -a_{2n} \\ \vdots \\ -a_{m1} \cdots -a_{mn} \end{pmatrix}.$$

Square Matrices

An $n \times n$ matrix A is said to be **square**. The entries a_{kk}, $k = 1, \ldots, n$ of a square matrix make up its **diagonal**. A square matrix A is **lower triangular** if $a_{ij} = 0$ for $i > j$, and **upper triangular** if $a_{ij} = 0$ for $i < j$; it is **triangular** if it is either upper or lower triangular. Therefore, an upper triangular matrix has all zeros below the diagonal and a lower triangular matrix has all zeros above the diagonal. A square matrix A is a **diagonal matrix** if $a_{ij} = 0$ for $i \neq j$, that is, the off-diagonal elements are zero. A square matrix A is a **scalar matrix** if $A = cI$ for some constant c.

$$\text{upper triangular} = \begin{pmatrix} a_{11}\, a_{12} \cdots a_{1n} \\ 0 \quad a_{22} \cdots a_{2n} \\ \vdots \\ 0 \quad 0 \quad \cdots a_{nn} \end{pmatrix}, \quad \text{lower triangular} = \begin{pmatrix} a_{11}\, 0 \quad \cdots 0 \\ a_{21}\, a_{22} \cdots 0 \\ \vdots \\ a_{n1}\, a_{n2} \cdots a_{nn} \end{pmatrix},$$

$$\text{diagonal} = \begin{pmatrix} a_{11}\, 0 \quad \cdots 0 \\ 0 \quad a_{22} \cdots 0 \\ \vdots \\ 0 \quad 0 \quad \cdots a_{nn} \end{pmatrix}, \quad \text{scalar} = \begin{pmatrix} c\,0\cdots\,0 \\ 0\,c\cdots\,0 \\ \vdots \\ 0\,0\cdots\,c \end{pmatrix}.$$

Matrix Algebra

A matrix can be viewed as a single long array, or fixed vector, therefore the vector space toolkit page 297 for fixed vectors applies to matrices.

Let A, B, C be matrices of the same row and column dimensions and let k_1, k_2, k be constants. Then

Closure	The operations $A+B$ and kA are defined and result in a new matrix of the same dimensions.	
Addition rules	$A+B=B+A$	commutative
	$A+(B+C)=(A+B)+C$	associative
	Matrix $\mathbf{0}$ is defined and $\mathbf{0}+A=A$	zero
	Matrix $-A$ is defined and $A+(-A)=\mathbf{0}$	negative
Scalar multiply rules	$k(A+B)=kA+kB$	distributive I
	$(k_1+k_2)A=k_1A+k_2B$	distributive II
	$k_1(k_2A)=(k_1k_2)A$	distributive III
	$1\,A=A$	identity

These rules collectively establish that the set of all $m\times n$ matrices is an abstract vector space (page 298).

The operation of matrix multiplication gives rise to some new matrix rules, which are in common use, but do not qualify as vector space rules. The rules are proved by expansion of each side of the equation. Techniques are sketched in the exercises, which carry out the steps of each proof.

Associative	$A(BC)=(AB)C$, provided products BC and AB are defined.
Distributive	$A(B+C)=AB+AC$, provided products AB and AC are defined.
Right Identity	$AI=A$, provided AI is defined.
Left Identity	$IA=A$, provided IA is defined.

Transpose. Swapping rows and columns of a matrix A results in a new matrix B whose entries are given by $b_{ij}=a_{ji}$. The matrix B is denoted A^T (pronounced "*A-transpose*"). The transpose has the following properties. Exercises outline the proofs.

$(A^T)^T=A$	Identity
$(A+B)^T=A^T+B^T$	Sum
$(AB)^T=B^TA^T$	Product
$(kA)^T=kA^T$	Scalar

Inverse Matrix

Definition 5.13 (Inverse Matrix)
A square matrix B is said to be an **inverse** of a square matrix A provided $AB = BA = I$. The symbol I is the identity matrix of matching dimension.

To illustrate, $B = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ is an inverse of $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ because

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The zero matrix does not have an inverse. To justify, let $A = \mathbf{0}$ and assume square matrix B is a inverse of A. Then relation $\mathbf{0}B = B\mathbf{0} = I$ holds. The zero matrix times any matrix is the zero matrix, which leads to the contradiction $\mathbf{0} = I$.

A given matrix A may not have an inverse.

Definition 5.14 (Inverse Notation A^{-1})
If matrix A has an inverse B, then notation A^{-1} is used for B:

$$AA^{-1} = A^{-1}A = I$$

Theorem 5.7 (Inverses)
Let A, B, C denote square matrices. Then

(a) A matrix has at most one inverse, that is, if $AB = BA = I$ and $AC = CA = I$, then $B = C$.

(b) If A has an inverse, then so does A^{-1} and $(A^{-1})^{-1} = A$.

(c) If A has an inverse, then $(A^{-1})^T = (A^T)^{-1}$.

(d) If A and B have inverses , then $(AB)^{-1} = B^{-1}A^{-1}$.

Proofs on page 315.

Left to be discussed is how to find the inverse A^{-1}. For a 2×2 matrix, there is an easily justified formula.

Theorem 5.8 (Inverse of a 2×2)

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

The formula is commonly committed to memory, because of repeated use. In words, the theorem says:

> Swap the diagonal entries, change signs on the off-diagonal entries, then divide by the determinant $ad - bc$.

There is a generalization of this formula to $n \times n$ matrices, which is equivalent to the formulas in **Cramer's rule**. It will be derived during the study of determinants; the statement is paraphrased as follows:

$$A^{-1} = \frac{\text{adjugate matrix of } A}{\text{determinant of } A}.$$

A general and efficient method for computing inverses, based upon **rref** methods, will be presented in the next section. The method can be implemented on hand calculators, computer algebra systems and computer numerical laboratories.

Definition 5.15 (Symmetric Matrix)
A matrix A is said to be **symmetric** if $A^T = A$, which implies that the row and column dimensions of A are the same and $a_{ij} = a_{ji}$.

If A is symmetric and invertible, then its inverse is symmetric. If B is any matrix, not necessarily square, then $A = B^T B$ is symmetric. Proofs are in the exercises.

Proofs and Details

Proof of the Kernel Theorem 5.2: Zero is in S because $A\vec{\mathbf{0}} = \vec{\mathbf{0}}$ for any matrix A. To verify the subspace criterion, we verify that, for $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ in S, the vector $\vec{\mathbf{z}} = c_1\vec{\mathbf{x}} + c_2\vec{\mathbf{y}}$ also belongs to S. The details:

$$\begin{aligned} A\vec{\mathbf{z}} &= A(c_1\vec{\mathbf{x}} + c_2\vec{\mathbf{y}}) && \\ &= A(c_1\vec{\mathbf{x}}) + A(c_2\vec{\mathbf{y}}) && \\ &= c_1 A\vec{\mathbf{x}} + c_2 A\vec{\mathbf{y}} && \\ &= c_1\vec{\mathbf{0}} + c_2\vec{\mathbf{0}} && \text{Because } A\vec{\mathbf{x}} = A\vec{\mathbf{y}} = \vec{\mathbf{0}}\text{, due to } \vec{\mathbf{x}}, \vec{\mathbf{y}} \text{ in } S. \\ &= \vec{\mathbf{0}} && \text{Therefore, } A\vec{\mathbf{z}} = \vec{\mathbf{0}}\text{, and } \vec{\mathbf{z}} \text{ is in } S. \end{aligned}$$

■

Proof of the Span Theorem 5.5: Details will be supplied for $k = 3$, because the text of the proof can be easily edited to give the details for general k. The vector space V is an abstract vector space, and we do not assume that the vectors are fixed vectors. It is impossible, therefore, to **burst** the vectors into components! Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ be given vectors in V and let

$$S = \mathbf{span}(\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3) = \{\vec{\mathbf{v}} \ : \ \vec{\mathbf{v}} = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\}.$$

The subspace criterion will be applied to prove that S is a subspace of V.

(1) We show $\vec{\mathbf{0}}$ is in S. Choose $c_1 = c_2 = c_3 = 0$, then $\vec{\mathbf{v}} = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3 = \vec{\mathbf{0}}$. Therefore, $\vec{\mathbf{0}}$ is in S.

(2) Assume $\vec{\mathbf{v}} = a_1\vec{\mathbf{v}}_1 + a_2\vec{\mathbf{v}}_2 + a_3\vec{\mathbf{v}}_3$ and $\vec{\mathbf{w}} = b_1\vec{\mathbf{v}}_1 + b_2\vec{\mathbf{v}}_2 + b_3\vec{\mathbf{v}}_3$ are in S. We show that $\vec{\mathbf{v}} + \vec{\mathbf{w}}$ is in S, by adding the equations:

$$\begin{array}{rcl} \vec{\mathbf{v}} + \vec{\mathbf{w}} & = & a_1\vec{\mathbf{v}}_1 + a_2\vec{\mathbf{v}}_2 + a_3\vec{\mathbf{v}}_3 + b_1\vec{\mathbf{v}}_1 + b_2\vec{\mathbf{v}}_2 + b_3\vec{\mathbf{v}}_3 \\ & = & (a_1+b_1)\vec{\mathbf{v}}_1 + (a_2+b_2)\vec{\mathbf{v}}_2 + (a_3+b_3)\vec{\mathbf{v}}_3 \\ & = & c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3 \end{array}$$

where the constants are defined by $c_1 = a_1 + b_1$, $c_2 = a_2 + b_2$, $c_3 = a_3 + b_3$. Then $\vec{\mathbf{v}} + \vec{\mathbf{w}}$ is in S.

(3) Assume $\vec{\mathbf{v}} = a_1\vec{\mathbf{v}}_1 + a_2\vec{\mathbf{v}}_2 + a_3\vec{\mathbf{v}}_3$ and c is a constant. We show $c\vec{\mathbf{v}}$ is in S. Multiply the equation for $\vec{\mathbf{v}}$ by c to obtain

$$\begin{array}{rcl} c\vec{\mathbf{v}} & = & ca_1\vec{\mathbf{v}}_1 + ca_2\vec{\mathbf{v}}_2 + ca_3\vec{\mathbf{v}}_3 \\ & = & c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3 \end{array}$$

where the constants are defined by $c_1 = ca_1$, $c_2 = ca_2$, $c_3 = ca_3$. Then $c\vec{\mathbf{v}}$ is in S. ■

Proof of the Inverse Theorem 5.7:

(a) If $AB = BA = I$ and $AC = CA = I$, then $B = BI = BAC = IC = C$.
(b) Let $B = A^{-1}$. Given $AB = BA = I$, then by definition A is an inverse of B, but by (a) it is the only one, so $(A^{-1})^{-1} = B^{-1} = A$.
(c) Let $B = A^{-1}$. We show $B^T = (A^T)^{-1}$ or equivalently $C = B^T$ satisfies $A^TC = CA^T = I$. Start with $AB = BA = I$, take the transpose to get $B^TA^T = A^TB^T = I$. Substitute $C = B^T$, then $CA^T = A^TC = I$, which was to be proved.
(d) The formula is proved by showing that $C = B^{-1}A^{-1}$ satisfies $(AB)C = C(AB) = I$. The left side is $(AB)C = ABB^{-1}A^{-1} = I$ and the right side $C(AB) = B^{-1}A^{-1}AB = I$, proving LHS = RHS.

Exercises 5.1

Fixed vectors

Perform the indicated operation(s).

1. $\begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

2. $\begin{pmatrix} 2 \\ -2 \end{pmatrix} - \begin{pmatrix} 1 \\ -3 \end{pmatrix}$

3. $\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix}$

4. $\begin{pmatrix} 2 \\ -2 \\ 9 \end{pmatrix} - \begin{pmatrix} 1 \\ -3 \\ 7 \end{pmatrix}$

5. $2\begin{pmatrix} 1 \\ -1 \end{pmatrix} + 3\begin{pmatrix} -2 \\ 1 \end{pmatrix}$

6. $3\begin{pmatrix} 2 \\ -2 \end{pmatrix} - 2\begin{pmatrix} 1 \\ -3 \end{pmatrix}$

7. $5\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} + 3\begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix}$

8. $3\begin{pmatrix} 2 \\ -2 \\ 9 \end{pmatrix} - 5\begin{pmatrix} 1 \\ -3 \\ 7 \end{pmatrix}$

9. $\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}$

10. $\begin{pmatrix} 2 \\ -2 \\ 4 \end{pmatrix} - \begin{pmatrix} 1 \\ -3 \\ 5 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$

Parallelogram Rule

Determine the resultant vector in two ways: (a) the parallelogram rule, and (b) fixed vector addition.

11. $\begin{pmatrix} 2 \\ -2 \end{pmatrix} + \begin{pmatrix} 1 \\ -3 \end{pmatrix}$

12. $(2\vec{\imath} - 2\vec{\jmath}) + (\vec{\imath} - 3\vec{\jmath})$

13. $\begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix}$

14. $(2\vec{\imath} - 2\vec{\jmath} + 3\vec{k}) + (\vec{\imath} - 3\vec{\jmath} - \vec{k})$

Toolkit

Let V be the data set of all fixed 2-vectors, $V = \mathcal{R}^2$. Define addition and scalar multiplication componentwise. Verify the following toolkit rules by direct computation.

15. (Commutative)
$\vec{X} + \vec{Y} = \vec{Y} + \vec{X}$

16. (Associative)
$\vec{X} + (\vec{Y} + \vec{Z}) = (\vec{Y} + \vec{X}) + \vec{Z}$

17. (Zero)
Vector $\vec{0}$ is defined and $\vec{0} + \vec{X} = \vec{X}$

18. (Negative)
Vector $-\vec{X}$ is defined and
$\vec{X} + (-\vec{X}) = \vec{0}$

19. (Distributive I)
$k(\vec{X} + \vec{Y}) = k\vec{X} + k\vec{Y}$

20. (Distributive II)
$(k_1 + k_2)\vec{X} = k_1\vec{X} + k_2\vec{X}$

21. (Distributive III)
$k_1(k_2\vec{X}) = (k_1k_2)\vec{X}$

22. (Identity)
$1\vec{X} = \vec{X}$

Subspaces

Verify that the given restriction equation defines a subspace S of $V = \mathcal{R}^3$. Use Theorem 5.2, page 300.

23. $z = 0$

24. $y = 0$

25. $x + z = 0$

26. $2x + y + z = 0$

27. $x = 2y + 3z$

28. $x = 0,\ z = x$

29. $z = 0,\ x + y = 0$

30. $x = 3z - y,\ 2x = z$

31. $x + y + z = 0,\ x + y = 0$

32. $x + y - z = 0,\ x - z = y$

Test S Not a Subspace

Test the following restriction equations for $V = \mathcal{R}^3$ and show that the corresponding subset S is not a subspace of V. Use Theorem 5.4 page 301.

33. $x = 1$

34. $x + z = 1$

35. $xz = 2$

36. $xz + y = 1$

37. $xz + y = 0$

38. $xyz = 0$

39. $z \geq 0$

40. $x \geq 0$ and $y \geq 0$

41. Octant I

42. The interior of the unit sphere

Dot Product

Find the dot product of $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$.

43. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$.

44. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

45. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$.

46. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$.

47. $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are in $\mathcal{R}^{169}$, $\vec{\mathbf{a}}$ has all 169 components 1 and $\vec{\mathbf{b}}$ has all components -1, except four, which all equal 5.

48. $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are in $\mathcal{R}^{200}$, $\vec{\mathbf{a}}$ has all 200 components -1 and $\vec{\mathbf{b}}$ has all components -1 except three, which are zero.

Length of a Vector

Find the length of the vector $\vec{\mathbf{v}}$.

49. $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

50. $\vec{\mathbf{v}} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$.

51. $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$.

52. $\vec{\mathbf{v}} = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}$.

Shadow Projection

Find the shadow projection $d = \vec{\mathbf{a}} \cdot \vec{\mathbf{b}}/|\vec{\mathbf{b}}|$.

53. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$.

54. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

55. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$.

56. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$.

Projections and Reflections

Let L denote a line through the origin with unit direction $\vec{\mathbf{u}}$.

The **projection** of vector $\vec{\mathbf{x}}$ onto L is $P(\vec{\mathbf{x}}) = d\vec{\mathbf{u}}$, where $d = \vec{\mathbf{x}} \cdot \vec{\mathbf{u}}$ is the shadow projection.

The **reflection** of vector $\vec{\mathbf{x}}$ across L is $R(\vec{\mathbf{x}}) = 2d\vec{\mathbf{u}} - \vec{\mathbf{x}}$ (a generalized complex conjugate).

57. Let $\vec{\mathbf{u}}$ be the direction of the x-axis in the plane. Establish that $P(\vec{\mathbf{x}})$ and $R(\vec{\mathbf{x}})$ are sides of a right triangle and P duplicates the complex conjugate operation $z \to \overline{z}$. Include a figure.

58. Let $\vec{\mathbf{u}}$ be any direction in the plane. Establish that $P(\vec{\mathbf{x}})$ and $R(\vec{\mathbf{x}})$ are sides of a right triangle. Draw a suitable figure, which includes $\vec{\mathbf{x}}$.

59. Let $\vec{\mathbf{u}}$ be the direction of $2\vec{\imath} + \vec{\jmath}$. Define $\vec{\mathbf{x}} = 4\vec{\imath} + 3\vec{\jmath}$. Compute the vectors $P(\vec{\mathbf{x}})$ and $R(\vec{\mathbf{x}})$.

60. Let $\vec{\mathbf{u}}$ be the direction of $\vec{\imath} + 2\vec{\jmath}$. Define $\vec{\mathbf{x}} = 3\vec{\imath} + 5\vec{\jmath}$. Compute the vectors $P(\vec{\mathbf{x}})$ and $R(\vec{\mathbf{x}})$.

Angle

Find the angle θ between the given vectors.

61. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$.

62. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

63. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$.

64. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$.

65. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 0 \\ -2 \\ 1 \\ 1 \end{pmatrix}$.

66. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \\ 0 \\ 0 \end{pmatrix}$.

67. $\vec{\mathbf{a}} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$.

68. $\vec{\mathbf{a}} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$ and $\vec{\mathbf{b}} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$.

Matrix Multiply

Find the given matrix product or else explain why it does not exist.

69. $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix}$

70. $\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ -2 \end{pmatrix}$

71. $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$

72. $\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \end{pmatrix}$

73. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$

74. $\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$

75. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$

76. $\begin{pmatrix} 1 & 2 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$

77. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

78. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

79. $\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$

80. $\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}$

81. $\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

82. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}$

Matrix Classification

Classify as square, non-square, upper triangular, lower triangular, scalar, diagonal, symmetric, non-symmetric. Cite as many terms as apply.

83. $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

84. $\begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix}$

85. $\begin{pmatrix} 1 & 3 \\ 4 & 2 \end{pmatrix}$

86. $\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$

87. $\begin{pmatrix} 1 & 3 & 4 \\ 5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

88. $\begin{pmatrix} 1 & 0 & 4 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

89. $\begin{pmatrix} 1 & 3 & 4 \\ 3 & 2 & 0 \\ 4 & 0 & 3 \end{pmatrix}$

90. $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$

91. $\begin{pmatrix} i & 0 \\ 0 & 2i \end{pmatrix}$

92. $\begin{pmatrix} i & 3 \\ 3 & 2i \end{pmatrix}$

Digital Photographs

Assume integer 24-bit color encoding $x = r + (256)g + (65536)b$, which means r units **red**, g units **green** and b units **blue**. Given matrix $X = R + 256G + 65536B$, find the red, green and blue color separation matrices R, G, B. Computer assist expected.

93. $X = \begin{pmatrix} 514 & 3 \\ 131843 & 197125 \end{pmatrix}$

94. $X = \begin{pmatrix} 514 & 3 \\ 131331 & 66049 \end{pmatrix}$

95. $X = \begin{pmatrix} 513 & 7 \\ 131333 & 66057 \end{pmatrix}$

96. $X = \begin{pmatrix} 257 & 7 \\ 131101 & 66057 \end{pmatrix}$

97. $X = \begin{pmatrix} 257 & 17 \\ 131101 & 265 \end{pmatrix}$

98. $X = \begin{pmatrix} 65537 & 269 \\ 65829 & 261 \end{pmatrix}$

99. $X = \begin{pmatrix} 65538 & 65803 \\ 65833 & 7 \end{pmatrix}$

100. $X = \begin{pmatrix} 259 & 65805 \\ 299 & 5 \end{pmatrix}$

Matrix Properties

Verify the result.

101. Let C be an $m \times n$ matrix. Let $\vec{X}$ be column i of the $n \times n$ identity I. Define $\vec{Y} = C\vec{X}$. Verify that $\vec{Y}$ is column i of C.

102. Let A and C be an $m \times n$ matrices such that $AC = \mathbf{0}$. Verify that each column $\vec{Y}$ of C satisfies $A\vec{Y} = \vec{\mathbf{0}}$.

103. Let A be a 2×3 matrix and let $\vec{Y}_1$, $\vec{Y}_2$, $\vec{Y}_3$ be column vectors packaged into a 3×3 matrix C. Assume each column vector $\vec{Y}_i$ satisfies the equation $A\vec{Y}_i = \vec{\mathbf{0}}$, $1 \le i \le 3$. Show that $AC = \mathbf{0}$.

104. Let A be an $m \times n$ matrix and let $\vec{Y}_1$, $\ldots$, $\vec{Y}_n$ be column vectors packaged into an $n \times n$ matrix C. Assume each column vector $\vec{Y}_i$ satisfies the equation $A\vec{Y}_i = \vec{\mathbf{0}}$, $1 \le i \le n$. Show that $AC = \mathbf{0}$.

Triangular Matrices

Verify the result.

105. The product of two upper triangular 2×2 matrices is upper triangular.

106. The product of two upper triangular $n \times n$ matrices is upper triangular.

107. The product of two triangular 2×2 matrices is not necessarily triangular.

108. The product of two lower triangular $n \times n$ matrices is upper triangular.

109. The product of two lower triangular 2×2 matrices is lower triangular.

110. The only 3×3 matrices which are both upper and lower triangular are the 3×3 diagonal matrices.

Matrix Multiply Properties

Verify the result.

111. The associative law $A(BC) = (AB)C$ holds for matrix multiplication.
Sketch: Expand $L = A(BC)$ entry L_{ij} according to matrix multiply rules. Expand $R = (AB)C$ entry R_{ij} the same way. Show $L_{ij} = R_{ij}$.

112. The distributive law $A(B + C) = AB + AC$ holds for matrices.
Sketch: Expand $L = A(B + C)$ entry L_{ij} according to matrix multiply rules. Expand $R = AB + AC$ entry R_{ij} the same way. Show $L_{ij} = \sum_{k=1}^{n} a_{ik}(b_{kj} + c_{kj})$ and $R_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} + a_{ik}c_{kj}$. Then $L_{ij} = R_{ij}$.

113. For any matrix A the transpose formula $(A^T)^T = A$ holds.
Sketch: Expand $L = (A^T)^T$ entry L_{ij} according to matrix transpose rules. Then $L_{ij} = a_{ij}$.

114. For matrices A, B the transpose formula $(A + B)^T = A^T + B^T$ holds.
Sketch: Expand $L = (A + B)^T$ entry L_{ij} according to matrix transpose rules. Repeat for entry R_{ij} of $R = A^T + B^T$. Show $L_{ij} = R_{ij}$.

115. For matrices A, B the transpose formula $(AB)^T = B^T A^T$ holds.
Sketch: Expand $L = (AB)^T$ entry L_{ij} according to matrix multiply and transpose rules. Repeat for entry R_{ij} of $R = B^T A^T$. Show $L_{ij} = R_{ij}$.

116. For a matrix A and constant k, the transpose formula $(kA)^T = kA^T$ holds.

Invertible Matrices

Verify the result.

117. There are infinitely many 2×2 matrices A, B such that $AB = 0$

118. The zero matrix is not invertible.

119. The matrix $A = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$ is not invertible.

120. The matrix $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ is invertible.

121. The matrices $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ satisfy

$$AB = BA = (ad - bc)I.$$

122. If $AB = 0$, then one of A or B is not invertible.

Symmetric Matrices

Verify the result.

123. The product of two symmetric $n \times n$ matrices A, B such that $AB = BA$ is symmetric.

124. The product of two symmetric 2×2 matrices may not be symmetric.

125. If A is symmetric, then so is A^{-1}.
Sketch: Let $B = A^{-1}$. Compute B^T using transpose rules.

126. If B is an $m \times n$ matrix and $A = B^T B$, then A is $n \times n$ symmetric.
Sketch: Compute A^T using transpose rules.

5.2 Matrix Equations

Linear Algebraic Equations

An $m \times n$ system of linear equations

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2, \\ & \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & b_m, \end{array} \tag{1}$$

can be written as a matrix multiply equation $A\vec{X} = \vec{b}$. Let A be the matrix of coefficients a_{ij}, let $\vec{X}$ be the column vector of variable names $x_1, \ldots, x_n$ and let $\vec{b}$ be the column vector with components $b_1, \ldots, b_n$. Assume equations (1) hold. Then:

$$\begin{array}{rcl} A\vec{X} & = & \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ & = & \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix} \\ & = & \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \qquad \text{by equation (1)} \end{array}$$

Therefore, equations (1) imply $A\vec{X} = \vec{b}$. Conversely, assume matrix equation $A\vec{X} = \vec{b}$. Reversible steps above give the last vector equality. Vector equality page 304 implies system (1) is satisfied.

A system of linear equations can be represented by its **variable list** $x_1, x_2, \ldots, x_n$ and its **augmented matrix**.

Definition 5.16 (Augmented Matrix)
The augmented matrix of A and $\vec{\mathbf{b}}$ for system $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ is

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ & & \vdots & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_n \end{array}\right) \quad \text{or symbol} \quad \left\langle A \,|\, \vec{b} \right\rangle \tag{2}$$

Vertical Line Notation. The present text uses a vertical line in a matrix display to mean it is an augmented matrix. While symbol $\left\langle A \,|\, \vec{b}\right\rangle$ has a vertical bar, the matrix itself has no vertical line as in display (2). Given a matrix C, it certainly has no vertical line. It may be a coefficient matrix in some system $C\vec{\mathbf{x}} = \vec{\mathbf{d}}$, or C could be an augmented matrix for some system $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$. Computers do not display nor store the vertical line appearing in equation (2). References may not use a vertical line.

Convert Augmented Matrix to Linear Algebraic Equations. Given an augmented $n \times (n+1)$ matrix C and a variable list $x_1, \ldots, x_n$, the conversion back to a linear system of algebraic equations is made by expanding $C\vec{Y} = \mathbf{0}$, where $\vec{Y}$ has components $x_1, \ldots, x_n, -1$. Hand work might contain an exposition like this:

$$(3)\qquad \begin{array}{c} \begin{array}{ccccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n & \end{array} \\ \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ & & \vdots & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_n \end{array}\right) \end{array}$$

In (3), a dot product is applied to the first n elements of each row, using the variable list written above the columns. The symbolic answer is set equal to the rightmost column's entry, in order to recover the equations. An example:

$$\begin{array}{c} \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \end{array} \\ \left(\begin{array}{ccc|c} 1 & 5 & -2 & 7 \\ 2 & 0 & -1 & 10 \\ 3 & 2 & 4 & 12 \end{array}\right) \end{array} \longrightarrow \left\{ \begin{array}{lcr} x_1 + 5x_2 - 2x_3 & = & 7 \\ 2x_1 + 0x_2 - x_3 & = & 10 \\ 3x_1 + 2x_2 + 4x_3 & = & 12 \end{array} \right.$$

Homogeneous System Augmented Matrix. It is usual in homogeneous systems $A\vec{\mathbf{x}} = \vec{\mathbf{0}}$ to *omit* the column of zeros and deal directly with A instead of $\left\langle A \,|\, \vec{0}\right\rangle$. The convention is justified by arguing that the rightmost column of zeros is unchanged by swap, multiply and combination rules which are defined below. A negative is remembering to insert the column of zeros when using a computation. An example:

$$\left\{ \begin{array}{rcrcrcl} x_1 & + & 5x_2 & - & 2x_3 & = & 0 \\ 2x_1 & + & 0x_2 & - & x_3 & = & 0 \\ 3x_1 & + & 2x_2 & + & 4x_3 & = & 0 \end{array} \right.$$

$$\text{Use} \begin{pmatrix} 1 & 5 & -2 \\ 2 & 0 & -1 \\ 3 & 2 & 4 \end{pmatrix} \text{ instead of } \left(\begin{array}{ccc|c} 1 & 5 & -2 & 0 \\ 2 & 0 & -1 & 0 \\ 3 & 2 & 4 & 0 \end{array}\right)$$

Elementary Row Operations

The three operations on equations which produce equivalent systems can be translated directly to row operations on the augmented matrix for the system.

The rules produce **equivalent systems**, that is, the three rules neither create nor destroy solutions.

Swap	Two rows can be interchanged.
Multiply	A row can be multiplied by multiplier $m \neq 0$.
Combination	A multiple of one row can be added to a different row.

Documentation of Row Operations

Throughout the display below, symbol **s** stands for **source**, symbol **t** for **target**, symbol **m** for **multiplier** and symbol **c** for **constant**.

Swap	**swap(s,t)** $\equiv$ swap rows **s** and **t**.
Multiply	**mult(t,m)** $\equiv$ multiply row **t** by **m**$\neq 0$.
Combination	**combo(s,t,c)** $\equiv$ add **c** times row **s** to row **t** $\neq$ **s**.

The standard for documentation is to write the notation next to the target row, which is the row to be changed. For swap operations, the notation is written next to the first row that was swapped, and optionally next to both rows. The notation was developed from early `maple` notation for the corresponding operations `swaprow`, `mulrow` and `addrow`, appearing in the `maple` package `linalg`. For instance, `addrow(A,1,3,-5)` selects matrix A as the target of the combination rule, which is documented in written work as `combo(1,3,-5)`. In written work on paper, symbol A is omitted, because A is the matrix appearing on the previous line of the sequence of steps.

Maple Remarks. Versions of `maple` use packages to perform toolkit operations. A short conversion table appears below.

On paper	**Maple** `with(linalg)`	**Maple** `with(LinearAlgebra)`
`swap(s,t)`	`swaprow(A,s,t)`	`RowOperation(A,[t,s])`
`mult(t,c)`	`mulrow(A,t,c)`	`RowOperation(A,t,c)`
`combo(s,t,c)`	`addrow(A,s,t,c)`	`RowOperation(A,[t,s],c)`

Conversion between packages can be controlled by the following function definitions, which causes the maple code to be the same regardless of which linear algebra package is used.[5]

Maple linalg

```
combo:=(a,s,t,c)->addrow(a,s,t,c);
swap:=(a,s,t)->swaprow(a,s,t);
mult:=(a,t,c)->mulrow(a,t,c);
```

[5] The acronym ASTC is used for the signs of the trigonometric functions in quadrants I through IV. The argument lists for `combo, swap, mult` use the same order, ASTC, memorized in trigonometry as **A***ll* **S***tudents* **T***ake* **C***alculus*.

Maple LinearAlgebra

```
combo:=(a,s,t,c)->RowOperation(a,[t,s],c);
swap:=(a,s,t)->RowOperation(a,[t,s]);
mult:=(a,t,c)->RowOperation(a,t,c);
macro(matrix=Matrix);
```

RREF Test

A linear algebraic equation example of *RREF*:

$$\begin{array}{ccccccccc} x_1 & +2x_2 & & +3x_4 & +4x_5 & & +5x_7 & & = 6 \\ & & x_3 & +7x_4 & +8x_5 & & +9x_7 & & = 10 \\ & & & & & x_6 & +11x_7 & & = 12 \\ & & & & & & & x_8 & = 13 \end{array} \tag{4}$$

The corresponding vector-matrix augmented matrix, no vertical line:

$$\begin{pmatrix} 1 & 2 & 0 & 3 & 4 & 0 & 5 & 0 & 6 \\ 0 & 0 & 1 & 7 & 8 & 0 & 9 & 0 & 10 \\ 0 & 0 & 0 & 0 & 0 & 1 & 11 & 0 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

Definition 5.17 (Reduced Row-echelon Form or RREF)
The reduced row-echelon form of a matrix, or **rref**, is defined by:

1. Zero rows appear last. Each nonzero row has first element 1, called a **leading one**. The column in which the leading one appears, called a **pivot column**, has all other entries zero.

2. The pivot columns appear as consecutive initial columns of the identity matrix I. Trailing columns of I might be absent.

Matrix (5) is a typical **rref** which satisfies the preceding properties. The initial 4 columns of the 7×7 identity matrix I appear in natural order in matrix (5); the trailing 3 columns of I are absent.

If the **rref** of the augmented matrix has a leading one in the last column, then the corresponding system of equations then has an equation "$0 = 1$" displayed, which signals an **inconsistent** system. Important: the **rref** always exists, even if the corresponding linear algebraic equations are inconsistent.

Elimination Method

The elimination algorithm for equations page 197 has an implementation for matrices. A row is marked **processed** if either (1) the row is all zeros, or else (2) the row contains a leading one and all other entries in that column are zero. Otherwise, the row is called **unprocessed**.

1. Move each unprocessed row of zeros to the last row using **swap** and mark it *processed*.
2. Identify an unprocessed nonzero row having the least number of leading zeros. Apply the **swap** rule to make this row the very first unprocessed row. Apply the **multiply** rule to insure a leading one. Apply the **combination** rule to change to zero all other entries in that column. The number of leading ones (lead variables) has been increased by one and the current column is a column of the identity matrix. Mark the row as *processed*, e.g., box the leading one: $\boxed{1}$.
3. Repeat steps 1–2, until all rows have been processed. Then all leading ones have been defined and the resulting matrix is in reduced row-echelon form.

Computer algebra systems and computer numerical laboratories automate computation of the **reduced row-echelon form** of a matrix A.

Literature calls the algorithm **Gauss-Jordan elimination**. Two examples:

$\mathbf{rref}(\mathbf{0}) = \mathbf{0}$ In **step 2**, all rows of the zero matrix $\mathbf{0}$ are zero. No changes are made to the zero matrix.

$\mathbf{rref}(I) = I$ In **step 2**, each row has a leading one. No changes are made to the identity matrix I.

Visual RREF Test. The habit to mark pivots with a box leads to a visual test for a RREF. An illustration:

$$\left(\begin{array}{cccc|c} \boxed{1} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \boxed{1} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \boxed{1} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Each boxed leading one $\boxed{1}$ appears in a column of the identity matrix. The boxes trail downward, ordered by columns 1, 2, 3 of the identity. There is no 4th pivot, therefore trailing identity column 4 is not used.

Toolkit Sequence

A sequence of swap, multiply and combination steps applied to a system of equations is called a **toolkit sequence**. The viewpoint is that a camera is pointed over the shoulder of an expert who writes the mathematics, and after the completion of each toolkit step, a photo is taken. The ordered sequence of cropped photo frames is a **filmstrip** or a sequence of frames. The **First Frame** displays the original system and the **Last Frame** displays the reduced row echelon system.

The terminology applies to systems $A\vec{x} = \vec{b}$ represented by an augmented matrix $C = \left\langle A \,|\, \vec{b} \right\rangle$. The First Frame is C and the Last Frame is **rref**(C).

Documentation of toolkit sequence steps will use this textbook's notation, page 323:

```
swap(s,t), mult(t,m), combo(s,t,c),
```

each written next to the target row `t`. During the sequence, consecutive initial columns of the identity, called **pivot columns**, are created as steps toward the **rref**. The remaining consecutive columns of the identity might not appear. An illustration:

Frame 1: $$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 1 \\ 1 & 4 & -1 & 0 & 2 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ Original augmented matrix.

Frame 2: $$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 1 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ `combo(1,2,-1)`
Pivot column 1 completed.

Frame 3: $$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ `swap(2,3)`

Frame 4: $$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ `combo(2,3,-2)`

Frame 5: $$\left(\begin{array}{cccc|c} 1 & 0 & -3 & 0 & -1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ Pivot column 2 completed by operation `combo(2,1,-2)`. Back-substitution postpones this step.

Frame 6: $$\left(\begin{array}{cccc|c} 1 & 0 & -3 & 0 & -1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ All leading ones found.

`mult(3,-1/2)`

Frame 7: $$\left(\begin{array}{cccc|c} 1 & 0 & -3 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$ `combo(3,2,-1)`
Zero other column 3 entries.
Next, finish pivot column 3.

Last Frame: $\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$ `combo(3,1,3)` **rref** found. Column 4 of the identity does not appear! There is no 4th pivot column.

Avoiding fractions. A matrix A with only integer entries can often be put into reduced row-echelon form without introducing fractions. The **multiply** rule introduces fractions, so its use should be limited. It is advised that leading ones be introduced only when convenient, otherwise make the leading coefficient nonzero and positive. Divisions at the end of the computation will produce the **rref**.

Clever use of the **combination** rule can sometimes create a leading one without introducing fractions. Consider the two rows

$$\begin{array}{ccccc} 25 & 0 & 1 & 0 & 5 \\ 7 & 0 & 2 & 0 & 2 \end{array}$$

The second row multiplied by -4 and added to the first row effectively replaces the 25 by -3, whereupon adding the first row twice to the second gives a leading one in the second row. The resulting rows are fraction-free.

$$\begin{array}{ccccc} -3 & 0 & -7 & 0 & -3 \\ 1 & 0 & -12 & 0 & -4 \end{array}$$

Rank and Nullity. What does it mean, if the first column of a **rref** is the zero vector? It means that the corresponding variable x_1 is a **free variable**. In fact, every column that does not contain a leading one corresponds to a free variable in the standard general solution of the system of equations. Symmetrically, each leading one identifies a pivot column and corresponds to a **leading variable**.

The number of leading ones is the **rank** of the matrix, denoted **rank**(A). The rank cannot exceed the row dimension nor the column dimension. The column count less the number of leading ones is the **nullity** of the matrix, denoted **nullity**(A). It equals the number of free variables.

Regardless of how matrix B arises, augmented or not, we have the relation

$$\text{variable count} = \textbf{rank}(B) + \textbf{nullity}(B).$$

If $B = \left\langle A \,|\, \vec{b} \right\rangle$ for $A\vec{X} = \vec{b}$, then the variable count n comes from $\vec{X}$ and the column count of B is *one more*, or $n + 1$. Replacing the *variable count* by the *column count* can therefore lead to fundamental errors.

Back-substitution and efficiency. The algorithm implemented in the preceding toolkit sequence is *easy to learn*, because the actual work is organized by creating pivot columns, via swap, combination and multiply. The created pivot columns are initial columns of the identity. You are advised to learn the algorithm in this form, but *please change the algorithm* as you become more efficient at doing the steps. See the examples for illustrations.

Back Substitution. Computer implementations and also hand computation can be made more efficient by changing steps 2 and 3, then adding **step 4**, as outlined below.

1. Move each unprocessed row of zeros to the last row using **swap** and mark it *processed.*

2a. Identify an unprocessed nonzero row having the least number of leading zeros. Apply the **swap** rule to make this row the very first unprocessed row. Apply the **multiply** rule to insure a leading one. Apply the **combination** rule to change to zero all other entries in that column which are **below** the leading one.

3a. Repeat steps 1–2a, until all rows have been processed. The matrix has all leading ones identified, a triangular shape, but it is not generally a RREF.

4. **Back-Substitution**. Identify the last row with a leading one. Apply the **combination** rule to change to zero all other entries in that column which are **above** the leading one. Repeat until all rows have been processed. The resulting matrix is a RREF.

Literature refers to **step 4** as **back-substitution**, a process which is exactly the original elimination algorithm applied to the system created by step 3a with *reversed variable list.*

Inverse Matrix. An efficient method to find the inverse B of a square matrix A, should it happen to exist, is to form the augmented matrix $C = \left\langle A \,|\, I \right\rangle$ and then read off B as the package of the last n columns of $\mathbf{rref}(C)$. This method is based upon the equivalence

$$\mathbf{rref}(\left\langle A \,|\, I \right\rangle) = \left\langle I \,|\, B \right\rangle \quad \text{if and only if} \quad AB = I.$$

The next theorem aids not only in establishing this equivalence but also in the practical matter of testing a candidate solution for the inverse matrix.

Theorem 5.9 (Inverse Test for Matrices)
If A and B are square matrices such that $AB = I$, then also $BA = I$. Therefore, only one of the equalities $AB = I$ or $BA = I$ is required to check an inverse. Proof on page 338.

Theorem 5.10 (Matrix Inverse and the rref)
Let A and B denote square matrices. Then

(a) If $\mathbf{rref}\left(\left\langle A \,|\, I \right\rangle\right) = \left\langle I \,|\, B \right\rangle$, then $AB = BA = I$ and B is the inverse of A.

(b) If $AB = BA = I$, then $\mathbf{rref}\left(\left\langle A \,|\, I \right\rangle\right) = \left\langle I \,|\, B \right\rangle$.

(c) If $\mathbf{rref}\left(\left\langle A\,|\,I\right\rangle\right) = \left\langle C\,|\,B\right\rangle$, then $C = \mathbf{rref}(A)$. If $C \neq I$, then A is not invertible. If $C = I$, then B is the inverse of A.

(d) Identity $\mathbf{rref}(A) = I$ holds if and only if A has an inverse.

Proof on page 338.

Matrix Inverse: Find A^{-1}

The method will be illustrated for the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Define the first frame of the sequence to be $C_1 = \left\langle A\,|\,I\right\rangle$, then compute the toolkit sequence to $\mathbf{rref}(C_1)$ as follows.

$$C_1 = \left(\begin{array}{rrr|rrr} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \qquad \text{First Frame}$$

$$C_2 = \left(\begin{array}{rrr|rrr} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & -1 & 1 \end{array}\right) \qquad \texttt{combo(3,2,-1)}$$

$$C_3 = \left(\begin{array}{rrr|rrr} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1/2 & 1/2 \end{array}\right) \qquad \texttt{mult(3,1/2)}$$

$$C_4 = \left(\begin{array}{rrr|rrr} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & -1/2 & 1/2 \end{array}\right) \qquad \texttt{combo(3,2,1)}$$

$$C_5 = \left(\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1/2 & -1/2 \\ 0 & 1 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & -1/2 & 1/2 \end{array}\right) \qquad \begin{array}{l}\texttt{combo(3,1,-1)} \\ \\ \text{Last Frame}\end{array}$$

The theory implies that the inverse of A is the matrix in the right half of the last frame:

$$A^{-1} = \begin{pmatrix} 1 & 1/2 & -1/2 \\ 0 & 1/2 & 1/2 \\ 0 & -1/2 & 1/2 \end{pmatrix}$$

Answer Check. Let B equal the matrix of the last display, claimed to be A^{-1}. The **Inverse Test**, Theorem 5.9 page 328, says that only one of $AB = I$ or $BA = I$ needs to be checked. Details:

$$\begin{aligned} AB &= \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & -1/2 \\ 0 & 1/2 & 1/2 \\ 0 & -1/2 & 1/2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1/2-1/2 & -1/2+1/2 \\ 0 & 1/2+1/2 & 1/2-1/2 \\ 0 & 1/2-1/2 & 1/2+1/2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

Elementary Matrices

> Elementary matrices express toolkit operations of swap, combination and multiply as matrix multiply equations.

Typically, toolkit operations produce a finite sequence of k linear algebraic equations, the first is the original system and the last is the reduced row echelon form of the system. We are going to re-write a typical toolkit sequence as matrix multiply equations. Each step is obtained from the previous by left-multiplication by a square matrix E:

$$(6) \qquad \begin{array}{rcrl} A\vec{X} &=& \vec{b} & \text{Original system} \\ E_1A\vec{X} &=& E_1\vec{b} & \text{After one toolkit step} \\ E_2E_1A\vec{X} &=& E_2E_1\vec{b} & \text{After two toolkit steps} \\ E_3E_2E_1A\vec{X} &=& E_3E_2E_1\vec{b} & \text{After three toolkit steps} \end{array}$$

Definition 5.18 (Elementary Matrix)
An elementary matrix E is created from the identity matrix by applying a single toolkit operation, that is, exactly one of the operations combination, multiply or swap.

Elementary Combination Matrix. Create square matrix E by applying the operation `combo(s,t,c)` to the identity matrix. The result equals the identity matrix except for the zero in row t and column s which is replaced by c.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ Identity matrix.

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix}$$ Elementary combination matrix, `combo(2,3,c)`.

Elementary Multiply Matrix.
Create square matrix E by applying `mult(t,m)` to the identity matrix. The result equals the identity matrix except the one in row t is replaced by m.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Identity matrix.

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & m \end{pmatrix}$$

Elementary multiply matrix, `mult(3,m)`.

Elementary Swap Matrix. Create square matrix E by applying `swap(s,t)` to the identity matrix.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Identity matrix.

$$E = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Elementary swap matrix, `swap(1,3)`.

If square matrix E represents a combination, multiply or swap rule, then the definition of matrix multiply applied to matrix EB gives the same matrix as obtained by applying the toolkit rule directly to matrix B. The statement is justified by experiment. See the exercises and Theorem 5.11.

Elementary 3×3 matrices (C=Combination, M=Multiply, S=Swap) can be displayed in computer algebra system `maple` as follows.

On Paper	**Maple** `with(linalg)`	**Maple** `with(LinearAlgebra)`
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	`B:=diag(1,1,1);`	`B:=IdentityMatrix(3);`
`combo(2,3,c)`	`C:=addrow(B,2,3,c);`	`C:=RowOperation(B,[3,2],c);`
`mult(3,m)`	`M:=mulrow(B,3,m);`	`M:=RowOperation(B,3,m);`
`swap(1,3)`	`S:=swaprow(B,1,3);`	`S:=RowOperation(B,[3,1]);`

A helpful project is to write out several examples of elementary 5 matrices by hand or machine. Such experiments lead to the following observations and theorems, proofs delayed to page 339.

Constructing an Elementary Matrix E

Combination Change a zero in the identity matrix to symbol c.

Multiply Change a one in the identity matrix to symbol $m \neq 0$.

Swap Interchange two rows of the identity matrix.

Constructing E^{-1} from an Elementary Matrix E

Combination	Change multiplier c in E to $-c$.
Multiply	Change diagonal multiplier $m \neq 0$ in E to $1/m$.
Swap	The inverse of E is E itself.

Theorem 5.11 (Matrix Multiply by an Elementary Matrix)
Let B_1 be a given matrix of row dimension n. Select a toolkit operation combination, multiply or swap, then apply it to matrix B_1 to obtain matrix B_2. Apply the identical toolkit operation to the $n \times n$ identity I to obtain elementary matrix E. Then

$$B_2 = EB_1.$$

Theorem 5.12 (Toolkit Sequence Identity)
If C and D are any two frames in a sequence, then corresponding toolkit operations are represented by square elementary matrices E_1, E_2, ..., E_k and the two frames C, D satisfy the matrix multiply equation

$$D = E_k \cdots E_2 E_1 C.$$

Theorem 5.13 (The rref and Elementary Matrices)
Let A be a given matrix of row dimension n. Then there exist $n \times n$ elementary matrices E_1, E_2, ..., E_k representing certain toolkit operations such that

$$\mathbf{rref}(A) = E_k \cdots E_2 E_1 A.$$

Illustration

Consider the following 6-frame toolkit sequence.

$$A_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 0 \\ 3 & 6 & 3 \end{pmatrix}$$ Frame 1, original matrix.

$$A_2 = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & -6 \\ 3 & 6 & 3 \end{pmatrix}$$ Frame 2, combo(1,2,-2).

$$A_3 = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 3 & 6 & 3 \end{pmatrix}$$ Frame 3, mult(2,-1/6).

$$A_4 = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & -6 \end{pmatrix}$$ Frame 4, combo(1,3,-3).

$$A_5 = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Frame 5, combo(2,3,-6).

$$A_6 = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Frame 6, combo(2,1,-3). Found **rref**.

The corresponding 3×3 elementary matrices are

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Frame 2, combo(1,2,-2) applied to I.

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/6 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Frame 3, mult(2,-1/6) applied to I.

$$E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}$$

Frame 4, combo(1,3,-3) applied to I.

$$E_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -6 & 1 \end{pmatrix}$$

Frame 5, combo(2,3,-6) applied to I.

$$E_5 = \begin{pmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Frame 6, combo(2,1,-3) applied to I.

Because each frame of the sequence has the succinct form EB, where E is an elementary matrix and B is the previous frame, the complete toolkit sequence can be written as follows.

$A_2 = E_1 A_1$	Frame 2, E_1 equals combo(1,2,-2) on I.
$A_3 = E_2 A_2$	Frame 3, E_2 equals mult(2,-1/6) on I.
$A_4 = E_3 A_3$	Frame 4, E_3 equals combo(1,3,-3) on I.
$A_5 = E_4 A_4$	Frame 5, E_4 equals combo(2,3,-6) on I.
$A_6 = E_5 A_5$	Frame 6, E_5 equals combo(2,1,-3) on I.
$A_6 = E_5 E_4 E_3 E_2 E_1 A_1$	Summary, frames 1-6. This relation is $\mathbf{rref}(A_1) = E_5 E_4 E_3 E_2 E_1 A_1$, which is the result claimed in Theorem 5.13.

The summary is the equation

$$\mathbf{rref}(A_1) = \begin{pmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -6 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{6} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} A_1$$

The inverse relationship $A_1 = E_1^{-1}E_2^{-1}E_3^{-1}E_4^{-1}E_5^{-1}\,\mathbf{rref}(A_1)$ is formed by the rules for constructing E^{-1} from elementary matrix E, page 331, the result being

$$A_1 = \begin{pmatrix}1&0&0\\2&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&-6&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\3&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&6&1\end{pmatrix}\begin{pmatrix}1&3&0\\0&1&0\\0&0&1\end{pmatrix}\mathbf{rref}(A_1)$$

Examples and Methods

Example 5.1 (Identify a Reduced Row–Echelon Form)
Identify the matrices in reduced row–echelon form using the RREF Test page 324.

$$A = \begin{pmatrix}0&1&3&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}\quad B = \begin{pmatrix}1&1&3&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

$$C = \begin{pmatrix}2&1&1&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}\quad D = \begin{pmatrix}0&1&3&0\\0&0&0&1\\1&0&0&0\\0&0&0&0\end{pmatrix}$$

Solution:

Matrix A. There are two nonzero rows, each with a leading one. The pivot columns are 2, 4 and they are consecutive columns of the 4×4 identity matrix. Yes, it is a RREF.

Matrix B. Same as A but with pivot columns 1, 4. Yes, it is a RREF. Column 2 is not a pivot column. The example shows that a scan for columns of the identity is not enough.

Matrix C. Immediately not a RREF, because the leading nonzero entry in row 1 is not a one.

Matrix D. Not a RREF. Swapping row 3 twice to bring it to row 1 will make it a RREF. This example has pivots in columns 1, 4 but the pivot columns fail to be columns 1, 2 of the identity (they are columns 3, 2).

Visual RREF Test. More experience is needed to use the visual test for RREF, but the effort is rewarded. Details are very brief. The ability to use the visual test is learned by working examples that use the basic RREF test.

Leading ones are boxed:

$$A = \begin{pmatrix}0&\boxed{1}&3&0\\0&0&0&\boxed{1}\\0&0&0&0\\0&0&0&0\end{pmatrix}\quad B = \begin{pmatrix}\boxed{1}&1&3&0\\0&0&0&\boxed{1}\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

$$C = \begin{pmatrix}2&1&1&0\\0&0&0&\boxed{1}\\0&0&0&0\\0&0&0&0\end{pmatrix}\quad D = \begin{pmatrix}0&\boxed{1}&3&0\\0&0&0&\boxed{1}\\\boxed{1}&0&0&0\\0&0&0&0\end{pmatrix}$$

Matrices A, B pass the visual test. Matrices C, D fail the test. Visually, we look for a boxed one starting on row 1. Boxes occupy consecutive rows, marching down and right, to make a triangular diagram. Columns with boxed ones are expected to be consecutive initial columns of identity matrix I.

Example 5.2 (Reduced Row–Echelon Form)

Find the reduced row–echelon form of the coefficient matrix A using the **elimination method**, page 325. Then solve the system.

$$\begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & - & x_3 & + & x_4 & = & 0, \\ x_1 & + & 3x_2 & - & x_3 & + & 2x_4 & = & 0, \\ & & x_2 & & & + & x_4 & = & 0. \end{array}$$

Solution: The coefficient matrix A and its **rref** are given by (details below)

$$A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 1 & 3 & -1 & 2 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathbf{rref}(A) = \begin{pmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Using variable list x_1, x_2, x_2, x_4, the equivalent reduced echelon system is

$$\begin{array}{rcrcrcrcl} x_1 & & & - & x_3 & - & x_4 & = & 0, \\ & & x_2 & & & + & x_4 & = & 0, \\ & & & & & & 0 & = & 0. \end{array}$$

which has lead variables x_1, x_2 and free variables x_3, x_4.

The *last frame algorithm* applies to write the standard general solution. This algorithm assigns invented symbols t_1, t_2 to the free variables, then back-substitution is applied to the lead variables. The solution to the system is

$$\begin{array}{rcl} x_1 & = & t_1 + t_2, \\ x_2 & = & -t_2, \\ x_3 & = & t_1, \\ x_4 & = & t_2, \qquad -\infty < t_1, t_2 < \infty. \end{array}$$

Details of the Elimination Method.

$$\begin{pmatrix} 1^* & 2 & -1 & 1 \\ 1 & 3 & -1 & 2 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

The coefficient matrix A. Leading one identified and marked as 1^*.

$$\begin{pmatrix} \boxed{1} & 2 & -1 & 1 \\ 0 & 1^* & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Apply the **combination** rule to zero the other entries in column 1. Mark the row processed. Identify the next leading one, marked 1^*.

$$\begin{pmatrix} \boxed{1} & 0 & -1 & -1 \\ 0 & \boxed{1} & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Apply the **combination** rule to zero the other entries in column 2. Mark the row processed. The matrix passes the **Visual RREF Test**.

Example 5.3 (Back-Substitution)

Display a toolkit sequence which uses numerical efficiency ideas of back substitution, page 328, in order to find the RREF of the matrix

$$A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 1 & 3 & -1 & 2 \\ 0 & 1 & 0 & 1 \end{pmatrix},$$

Solution: The answer for the reduced row-echelon form of matrix A is

$$\mathbf{rref}(A) = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Back-substitution details appear below.

Meaning of the computation. Finding a RREF is part of solving the homogeneous system $A\vec{X} = \vec{0}$. The Last Frame Algorithm is used to write the general solution. The algorithm requires a toolkit sequence applied to the augmented matrix $\left\langle A \,|\, \vec{0} \right\rangle$, ending in the Last Frame, which is the RREF with an added column of zeros.

$$\begin{pmatrix} 1 & 2 & -1 & 1 \\ 1 & 3 & -1 & 2 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

The given matrix A. Identify row 1 for the first pivot.

$$\begin{pmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

`combo(1,2,-1)` applied to introduce zeros below the leading one in row 1.

$$\begin{pmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

`combo(2,3,-1)` applied to introduce zeros below the leading one in row 2. The RREF has not yet been found. The matrix is triangular.

$$\begin{pmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Begin back-substitution: `combo(2,1,-2)` applied to introduce zeros above the leading one in row 2.

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Continue back-substitution: `combo(3,2,-1)` and `combo(3,1,1)` applied to introduce zeros above the leading one in row 3.

$$\begin{pmatrix} \boxed{1} & 0 & -1 & 0 \\ 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & \boxed{1} \end{pmatrix}$$

RREF Visual Test passed.
This matrix is the answer.

Example 5.4 (Answer Check a Matrix Inverse)

Display the answer check details for the given matrix A and its proposed inverse B.

$$A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -3 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Solution:

Details. We apply the **Inverse Test**, Theorem 5.9, which requires one matrix multiply:

$$AB = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -3 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{Expect } AB = I.$$

$$= \begin{pmatrix} 1 & -3+2+1 & 1-2+1 & 1-1 \\ 0 & 1 & -1+1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1-1 & -1+1 & 1 \end{pmatrix} \quad \text{Multiply.}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Simplify. Then $AB = I$. Because of Theorem 5.9, we don't check $BA = I$.

Example 5.5 (Find the Inverse of a Matrix)
Compute the inverse matrix of

$$A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$

Solution: The answer:

$$A^{-1} = \begin{pmatrix} 1 & -3 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Details. Form the augmented matrix $C = \left\langle A \,|\, I \right\rangle$ and compute its reduced row-echelon form by toolkit steps.

$$\left(\begin{array}{cccc|cccc} 1 & 2 & -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

Augment I onto A.

$$\left(\begin{array}{cccc|cccc} 1 & 2 & -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

`swap(3,4)`.

$$\left(\begin{array}{cccc|cccc} 1 & 2 & -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

`combo(2,3,-1)`. Triangular matrix.

$$\left(\begin{array}{cccc|cccc} 1 & 2 & -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

Back-substitution: `combo(4,2,-1)`.

$$\left(\begin{array}{cccc|cccc} 1 & 2 & -1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

`combo(4,1,-1)`.

$$\left(\begin{array}{cccc|cccc} 1 & 0 & -1 & 0 & 1 & -2 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

`combo(2,1,-2)`.

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & -3 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$ `combo(3,1,1)`. Identity left, inverse right.

Details and Proofs

Proof of Theorem 5.9:

Assume $AB = I$. Let $C = BA - I$. We intend to show $C = \mathbf{0}$, then $BA = C + I = I$, as claimed.

Compute $AC = ABA - A = AI - A = \mathbf{0}$. It follows that the columns $\vec{y}$ of C are solutions of the homogeneous equation $A\vec{y} = \vec{0}$. To complete the proof, we show that the only solution of $A\vec{y} = \vec{0}$ is $\vec{y} = \vec{0}$, because then C has all zero columns, which means C is the zero matrix.

First, $B\vec{u} = \vec{0}$ implies $\vec{u} = I\vec{u} = AB\vec{u} = A\vec{0} = \vec{0}$, hence B has an inverse, and then $B\vec{x} = \vec{y}$ has a unique solution $\vec{x} = B^{-1}\vec{y}$.

Suppose $A\vec{y} = \vec{0}$. Write $\vec{y} = B\vec{x}$. Then $\vec{x} = I\vec{x} = AB\vec{x} = A\vec{y} = \vec{0}$. This implies $\vec{y} = B\vec{x} = B\vec{0} = \vec{0}$. ■

Proof of Theorem 5.10:

Details for (a). Let $C = \left\langle A \,|\, I \right\rangle$ and assume $\mathbf{rref}(C) = \left\langle I \,|\, B \right\rangle$. Solving the $n \times 2n$ system $C\vec{X} = \vec{\mathbf{0}}$ is equivalent to solving the system $A\vec{Y} + I\vec{Z} = \vec{\mathbf{0}}$ with n-vector unknowns $\vec{Y}$ and $\vec{Z}$. This system has exactly the same solutions as $I\vec{Y} + B\vec{Z} = \vec{\mathbf{0}}$, by the equation $\mathbf{rref}(C) = \left\langle I \,|\, B \right\rangle$. The latter is a reduced echelon system with lead variables equal to the components of $\vec{Y}$ and free variables equal to the components of $\vec{Z}$. Multiplying by A gives $A\vec{Y} + AB\vec{Z} = \vec{\mathbf{0}}$, hence $-\vec{Z} + AB\vec{Z} = \vec{\mathbf{0}}$, or equivalently $AB\vec{Z} = \vec{Z}$ for every vector $\vec{Z}$ (because its components are free variables). Letting $\vec{Z}$ be a column of I shows that $AB = I$. Then $AB = BA = I$ by Theorem 5.9, and B is the inverse of A.

Details for (b). Assume $AB = I$. We prove the identity $\mathbf{rref}(\left\langle A \,|\, I \right\rangle) = \left\langle I \,|\, B \right\rangle$. Let the system $A\vec{Y} + I\vec{Z} = \vec{\mathbf{0}}$ have a solution $\vec{Y}$, $\vec{Z}$. Multiply by B to obtain $BA\vec{Y} + B\vec{Z} = \vec{\mathbf{0}}$. Use $BA = I$ to give $\vec{Y} + B\vec{Z} = \vec{\mathbf{0}}$. The latter system therefore has $\vec{Y}$, $\vec{Z}$ as a solution. Conversely, a solution $\vec{Y}$, $\vec{Z}$ of $\vec{Y} + B\vec{Z} = \vec{\mathbf{0}}$ is a solution of the system $A\vec{Y} + I\vec{Z} = \vec{\mathbf{0}}$, because of multiplication by A. Therefore, $A\vec{Y} + I\vec{Z} = \vec{\mathbf{0}}$ and $\vec{Y} + B\vec{Z} = \vec{\mathbf{0}}$ are *equivalent* systems. The latter is in reduced row-echelon form, and therefore $\mathbf{rref}(\left\langle A \,|\, I \right\rangle) = \left\langle I \,|\, B \right\rangle$.

Details for (c). Toolkit steps that compute $\mathbf{rref}(\left\langle A \,|\, I \right\rangle)$ must also compute $\mathbf{rref}(A)$. This fact is learned first by working examples. Elementary matrix formulas can make the proof more transparent: see the Miscellany exercises. Conclusion: $\mathbf{rref}(\left\langle A \,|\, I \right\rangle) = \left\langle C \,|\, B \right\rangle$ implies $C = \mathbf{rref}(A)$.

Let's prove $C \neq I$ implies A is not invertible. Suppose not, then $C \neq I$ and A is invertible. Then (b) implies $\left\langle C \,|\, B \right\rangle = \mathbf{rref}(\left\langle A \,|\, I \right\rangle) = \left\langle I \,|\, B \right\rangle$. Comparing columns, this equation implies $C = I$, a contradiction.

To prove $C = I$ implies B is the inverse of A, apply (a).

Details for (d). Assume A is invertible. We are to prove $\mathbf{rref}(A) = I$. Part (b) says $F = \left\langle A \,|\, I \right\rangle$ satisfies $\mathbf{rref}(F) = \left\langle I \,|\, B \right\rangle$ where B is the inverse of A. Part (c) says $\mathbf{rref}(F) = \left\langle \mathbf{rref}(A) \,|\, \vec{b} \right\rangle$. Comparing matrix columns gives $\mathbf{rref}(A) = I$.

Converse: assume $\mathbf{rref}(A) = I$, to prove A invertible. Let $F = \left\langle A \,|\, I \right\rangle$, then $\mathbf{rref}(F) = \left\langle C \,|\, B \right\rangle$ for some C, B. Part (c) says $C = \mathbf{rref}(A) = I$. Part (a) says B is the inverse of A. This proves A is invertible and completes (d).

Proof of Theorem 5.11: It is possible to organize the proof into three cases, by considering the three possible toolkit operations. We don't do the tedious details. Instead, we refer to the *Elementary Matrix Multiply* exercises page 340, for suitable experiments that provide the intuition needed to develop formal proof details.

Proof of Theorem 5.12: The idea of the proof begins with writing Frame 1 as $C_1 = E_1C$, using Theorem 5.11. Repeat to write Frame 2 as $C_2 = E_2C_1 = E_2E_1C$. By induction, Frame k is $C_k = E_kC_{k-1} = E_k \cdots E_2E_1C$. But Frame k is matrix D in the sequence. ■

Proof of Theorem 5.13: The reduced row-echelon matrix $D = \mathbf{rref}(A)$ paired with $C = A$ imply by Theorem 5.12 that $\mathbf{rref}(A) = D = E_k \cdots E_2E_1C = E_k \cdots E_2E_1A$. ■

Exercises 5.2

Identify RREF

Mark the matrices which pass the RREF Test, page 324. Explain the failures.

1. $\begin{pmatrix} 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

2. $\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$

3. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$

4. $\begin{pmatrix} 1 & 1 & 4 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

Lead and Free Variables

For each matrix A, assume a homogeneous system $A\vec{X} = \vec{\mathbf{0}}$ with variable list $x_1, \ldots, x_n$. List the lead and free variables. Then report the rank and nullity of matrix A.

5. $\begin{pmatrix} 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

6. $\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$

7. $\begin{pmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

8. $\begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

9. $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

10. $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

11. $\begin{pmatrix} 1 & 1 & 3 & 5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

12. $\begin{pmatrix} 1 & 2 & 0 & 3 & 4 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

13. $\begin{pmatrix} 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

14. $\begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

15. $\begin{pmatrix} 0 & 1 & 0 & 5 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

16. $\begin{pmatrix} 1 & 0 & 3 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Elementary Matrices

Write the 3×3 elementary matrix E and its inverse E^{-1} for each of the following operations, defined on page 323.

17. `combo(1,3,-1)`

18. `combo(2,3,-5)`

19. `combo(3,2,4)`

20. `combo(2,1,4)`

21. `combo(1,2,-1)`

22. `combo(1,2,-`e^2`)`

23. `mult(1,5)`

24. `mult(1,-3)`

25. `mult(2,5)`

26. `mult(2,-2)`

27. `mult(3,4)`

28. `mult(3,5)`

29. `mult(2,-`π`)`

30. `mult(1,`e^2`)`

31. `swap(1,3)`

32. `swap(1,2)`

33. `swap(2,3)`

34. `swap(2,1)`

35. `swap(3,2)`

36. `swap(3,1)`

Elementary Matrix Multiply

For each given matrix B_1, perform the toolkit operation (`combo`, `swap`, `mult`) to obtain the result B_2. Then compute the elementary matrix E for the identical toolkit operation. Finally, verify the matrix multiply equation $B_2 = EB_1$.

37. $\begin{pmatrix} 1 & 1 \\ 0 & 3 \end{pmatrix}$, `mult(2,1/3)`.

38. $\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$, `mult(1,3)`.

39. $\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, `combo(3,2,-1)`.

40. $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$, `combo(2,1,-3)`.

41. $\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$, `swap(2,3)`.

42. $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$, `swap(1,2)`.

Inverse Row Operations

Given the final frame B of a sequence starting with matrix A, and the given operations, find matrix A. Do not use matrix multiply.

43. $B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,-3)`, `mult(1,-2)`, `swap(2,3)`.

44. $B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,2)`, `swap(3,2)`.

45. $B = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,3)`, `mult(1,4)`, `swap(1,3)`.

46. $B = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,4)`, `mult(1,3)`, `swap(3,2)`.

Elementary Matrix Products

Given the first frame B_1 of a sequence and elementary matrix operations E_1, E_2, E_3, find matrices $F = E_3E_2E_1$ and $B_4 = FB_1$. Hint: Compute $\left\langle B_4|F \right\rangle$ from toolkit operations on $\left\langle B_1|I \right\rangle$.

47. $B_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,-3)`, `mult(1,-2)`.

48. $B_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,3)`, `swap(3,2)`.

49. $B_1 = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `mult(1,4)`, `swap(1,3)`.

50. $B_1 = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}$, operations `combo(1,2,-1)`, `combo(2,3,4)`, `mult(1,3)`.

Miscellany

51. Justify with English sentences why all possible 2×2 matrices in reduced row-echelon form must look like

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & * \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where $*$ denotes an arbitrary number.

52. Display all possible 3×3 matrices in reduced row-echelon form. Besides the zero matrix and the identity matrix, please report five other forms, most containing symbol $*$ representing an arbitrary number.

53. Determine all possible 4×4 matrices in reduced row-echelon form.

54. Display a 6×6 matrix in reduced row-echelon form with rank 4 and only entries of zero and one.

55. Display a 5×5 matrix in reduced row-echelon form with nullity 2 having entries of zero, one and two, but no other entries.

56. Display the rank and nullity of any $n \times n$ elementary matrix.

57. Let $F = \left\langle C|D \right\rangle$ and let E be a square matrix with row dimension matching F. Display the details for the equality

$$EF = \left\langle EC|ED \right\rangle.$$

58. Let $F = \left\langle C|D \right\rangle$ and let E_1, E_2 be $n \times n$ matrices with n equal to the row dimension of F. Display the details for the equality

$$E_2E_1F = \left\langle E_2E_1C|E_2E_1D \right\rangle.$$

59. Assume matrix A is invertible. Display details explaining why $\mathbf{rref}(\left\langle A|I \right\rangle)$ equals the matrix $\left\langle R|E \right\rangle$, where matrix $R = \mathbf{rref}(A)$ and matrix $E = E_k \cdots E_1$. Symbols E_i are elementary matrices in toolkit steps taking matrix A into reduced row-echelon form. Suggestion: Use the preceding exercises.

60. Assume E_1, E_2 are elementary matrices in toolkit steps taking A into reduced row-echelon form. Prove that $A^{-1} = E_2E_1$. In words, A^{-1} is found by doing the same toolkit steps to the identity matrix.

61. Assume matrix A is invertible and $E_1, \ldots, E_k$ are elementary matrices in toolkit steps taking A into reduced row-echelon form. Prove that $A^{-1} = E_k \cdots E_1$.

62. Assume A, B are 2×2 matrices. Assume A is invertible and $\mathbf{rref}(\big\langle A|B\big\rangle) = \big\langle I|D\big\rangle$. Explain why the first column $\vec{x}$ of D is the unique solution of $A\vec{x} = \vec{b}$, where $\vec{b}$ is the first column of B.

63. Assume A, B are $n \times n$ matrices with A invertible. Explain how to solve the matrix equation $AX = B$ for matrix X using the augmented matrix of A, B.

5.3 Determinants and Cramer's Rule

Unique Solution of a 2×2 System

The 2×2 system

$$\begin{array}{rcrcl} ax & + & by & = & e, \\ cx & + & dy & = & f, \end{array} \tag{1}$$

has a unique solution provided $\Delta = ad - bc$ is nonzero, in which case the solution is given by

$$x = \frac{de - bf}{ad - bc}, \quad y = \frac{af - ce}{ad - bc}. \tag{2}$$

This result, called **Cramer's Rule** for 2×2 systems, is first learned in college algebra as a part of determinant theory.

Determinants of Order 2

College algebra introduces matrix notation and determinant notation:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad |A| \text{ or } \det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Evaluation of a 2×2 determinant is by **Sarrus' Rule**:

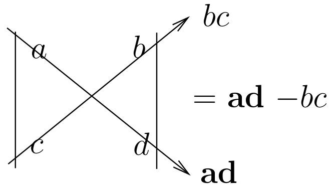


Figure 10. Sarrus' 2×2 rule.
A diagram for $|A| = (ad) - (bc)$.

The boldface product **ad** is the product of the main diagonal entries and the other product bc is from the anti-diagonal.

Cramer's 2×2 rule in determinant notation is

$$x = \frac{\begin{vmatrix} e & b \\ f & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} a & e \\ c & f \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}. \tag{3}$$

Unique Solution of an $n \times n$ System

Cramer's rule can be generalized to an $n \times n$ system of equations in matrix form $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ or in scalar form

$$\begin{array}{rcrcccrcl} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ \vdots & & \vdots & & \cdots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n. \end{array} \tag{4}$$

Determinants will be defined shortly; intuition from the 2 × 2 case and Sarrus' rule should suffice for the moment.

System (4) has a unique solution provided the **determinant of coefficients** $\Delta = \det(A)$ is nonzero, in which case the solution is given by

$$x_1 = \frac{\Delta_1}{\Delta},\ x_2 = \frac{\Delta_2}{\Delta},\ \ldots,\ x_n = \frac{\Delta_n}{\Delta}. \tag{5}$$

The determinant Δ_j equals $\det(B_j)$ where matrix B_j is matrix A modified to have column j equal to $\vec{\mathbf{b}} = (b_1, \ldots, b_n)$. Vector $\vec{\mathbf{b}}$ is the right side of system (4). The result is called **Cramer's Rule** for $n \times n$ systems.

Determinant Notation for Cramer's Rule

The **determinant of coefficients** for system $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ is denoted by

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}. \tag{6}$$

The other n determinants in Cramer's rule (5) are given by

$$\Delta_1 = \begin{vmatrix} b_1 & a_{12} & \cdots & a_{1n} \\ b_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ b_n & a_{n2} & \cdots & a_{nn} \end{vmatrix}, \ldots, \Delta_n = \begin{vmatrix} a_{11} & a_{12} & \cdots & b_1 \\ a_{21} & a_{22} & \cdots & b_2 \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & b_n \end{vmatrix}. \tag{7}$$

Determinant Notation Conflicts. The literature is filled with various notations for matrices, vectors and determinants. The expected notation **uses vertical bars only** for determinants and absolute values, e.g., $|A|$ makes sense for a matrix A or a constant A. For clarity, the notation $\det(A)$ may be preferred.

Value of a Determinant. Notation $|A|$ for $\det(A)$ implies that *a determinant is a number*, computed by $|A| = a_{11}a_{22} - a_{12}a_{21}$ when $n = 2$. For $n \geq 3$, $|A|$ is computed by similar but increasingly complicated formulas; see *Sarrus' Rule* page 345 and **Four Determinant Properties** *infra*.

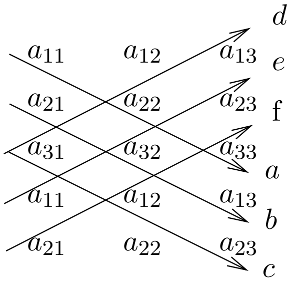
It is false that $|A| = A$ for a 1×1 matrix, because $|A|$ is a number and A is a matrix. The symbol $|c|$ for a *constant* c (not a matrix) is evaluated by algebra rules: $|c| = c$ for $c \geq 0$ and otherwise $|c| = -c$. Overloading of symbols causes equations like $|A| = -1$ for 1×1 matrix $A = (-1)$, whereas $|-1| = 1$ for constant -1.

Sarrus' Rule for 3 × 3 Matrices

College algebra supplies the following formula for the determinant of a 3×3 matrix A:

$$
\begin{aligned}
\det(A) &= \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \\
&= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\
&\quad -a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13}.
\end{aligned} \tag{8}
$$

The number $\det(A)$ can be computed by an algorithm similar to the one for 2×2 matrices, as in Figure 11. Important: no further generalizations are possible. *There is no Sarrus' rule for* 4×4 *or larger matrices*!



a_{11} a_{12} a_{13}
a_{21} a_{22} a_{23}
a_{31} a_{32} a_{33}
a_{11} a_{12} a_{13}
a_{21} a_{22} a_{23}
d
e
f
a
b
c

Figure 11. Sarrus' 3×3 rule.
The down arrow sum $a + b + c$ and the up arrow sum $d + e + f$ are subtracted:

$$\det(A) = (a + b + c) - (d + e + f).$$

College Algebra Definition of Determinant

The impractical definition is the formula

$$\det(A) = \sum_{\sigma \in S_n} (-1)^{\text{parity}(\sigma)}\, a_{1\sigma_1} \cdots a_{n\sigma_n}. \tag{9}$$

In formula (9), a_{ij} denotes the element in row i and column j of the matrix A. The symbol $\sigma = (\sigma_1, \ldots, \sigma_n)$ stands for a rearrangement of the subscripts 1, 2, …, n and S_n is the set of all possible rearrangements. The nonnegative integer parity(σ) is determined by counting the minimum number of pairwise interchanges required to assemble the list of integers σ_1, …, σ_n into natural order 1, …, n.

Formula (9) reproduces the definition for 3×3 matrices given in equation (8). We will have no computational use for (9). For computing the value of a determinant, see **four properties** and **cofactor expansion**, *infra.*

Four Determinant Properties

The definition of determinant (9) implies the following four properties:

Triangular The value of $\det(A)$ for either an upper triangular or a lower triangular matrix A is the product of the diagonal elements: $\det(A) = a_{11}a_{22}\cdots a_{nn}$.

Combination The value of $\det(A)$ is unchanged by adding a multiple of a row to a different row.

Multiply If one row of A is multiplied by constant $c \neq 0$ to create matrix B, then $\det(B) = c\det(A)$.

Swap If B results from A by swapping two rows, then $\det(A) = (-1)\det(B)$.

It is known that these four rules suffice to compute the value of any $n \times n$ determinant. The proof of the four properties is delayed until page 360.

Elementary Matrices and the Four Rules

The rules can be stated in terms of elementary matrices as follows.

Triangular The value of $\det(A)$ for either an upper triangular or a lower triangular matrix A is the product of the diagonal elements: $\det(A) = a_{11}a_{22}\cdots a_{nn}$. This is a one-arrow Sarrus' rule valid for dimension n.

Combination If E is an elementary matrix for a combination rule, then $\det(EA) = \det(A)$.

Multiply If E is an elementary matrix for a multiply rule with multiplier $m \neq 0$, then $\det(EA) = m\det(A)$.

Swap If E is an elementary matrix for a swap rule, then $\det(EA) = (-1)\det(A)$.

Since $\det(E) = 1$ for a combination rule, $\det(E) = -1$ for a swap rule and $\det(E) = c$ for a multiply rule with multiplier $c \neq 0$, it follows that for any elementary matrix E there is the determinant multiplication rule $\det(EA) = \det(E)\det(A)$.

Theorem 5.14 (Four Rules Compressed)
The Four rules to compute the value of any determinant can be written as two rules.

Triangular Rule The value of $|A|$ for a triangular matrix A is the product of the diagonal elements

Determinant Product Rule Let E be an elementary matrix, then $\det(EA) = \det(E)\det(A)$.

Additional Determinant Rules

The following rules make for efficient evaluation of certain special determinants. The results are stated for rows, but they also hold for columns, because of Theorem 5.15.

Zero row If one row of A is zero, then $\det(A) = 0$.

Duplicate rows If two rows of A are identical, then $\det(A) = 0$.

Dependent rows If a row of A is a linear combination of the other rows, then $\det(A) = 0$.

RREF $\ne I$ If $\mathbf{rref}(A) \ne I$, then $\det(A) = 0$.

Common factor The relation $\det(A) = c\det(B)$ holds, provided A and B differ only in one row, say row j, for which $\mathbf{row}(A, j) = c\,\mathbf{row}(B, j)$.

Row linearity The relation $\det(A) = \det(B) + \det(C)$ holds, provided A, B and C differ only in one row, say row j, for which $\mathbf{row}(A, j) = \mathbf{row}(B, j) + \mathbf{row}(C, j)$.

The proofs of these properties are delayed until page 360.

Determinant of a Transpose

A consequence of (9) is the relation $|A| = \left|A^T\right|$ where A^T means the transpose of A, obtained by swapping rows and columns.

Theorem 5.15 (Determinant of the Transpose)
The relation

$$\det\left(A^T\right) = \det(A) \quad \text{or} \quad \left|A^T\right| = |A|$$

implies that all determinant theory results for rows also apply to columns.

Cofactor Expansion

The special subject of cofactor expansions is used to justify Cramer's rule and to provide an alternative method for computation of determinants. There is no claim that cofactor expansion is efficient, only that it is possible, and different than Sarrus' rule or the use of the four properties.

Background from College Algebra

The cofactor expansion theory is most easily understood from the college algebra topic in dimension 3. Cofactor row expansion computes $|A|$ by one of three possible formulas, recorded below. The **pattern**:

$$|A| = \Sigma(\text{row element} \times \text{checkerboard sign} \times \text{cross-out determinant}).$$

$$\begin{aligned}
|A| &= \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \\
|A| &= a_{11}(+1)\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{12}(-1)\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}(+1)\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\
|A| &= a_{21}(-1)\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{22}(+1)\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + a_{23}(-1)\begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \\
|A| &= a_{31}(+1)\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} + a_{32}(-1)\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} + a_{33}(+1)\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}
\end{aligned}$$

The formulas expand a 3×3 determinant in terms of 2×2 determinants, along a row of A. The attached signs ± 1 are called the **checkerboard signs**, to be defined shortly. The 2×2 **cross-out determinants** are officially called **minors** of the 3×3 determinant $|A|$. The checkerboard sign multiplied against a minor is called a **cofactor**.

These formulas are generally used when a row has one or two zeros, making it unnecessary to evaluate one or two of the 2×2 determinants in the expansion. To illustrate, row 1 expansion gives

$$\begin{vmatrix} 3 & 0 & 0 \\ 2 & 1 & 7 \\ 5 & 4 & 8 \end{vmatrix} = 3(+1)\begin{vmatrix} 1 & 7 \\ 4 & 8 \end{vmatrix} + 0(-1)\begin{vmatrix} 2 & 7 \\ 5 & 8 \end{vmatrix} + 0(+1)\begin{vmatrix} 2 & 1 \\ 5 & 4 \end{vmatrix} = -60.$$

A clever time–saving choice is always a row which has the most zeros, although success does not depend upon cleverness. What has been said for rows also applies to columns, due to the transpose formula $|A| = |A^T|$.

Minors and Cofactors

The $(n-1) \times (n-1)$ determinant obtained from $\det(A)$ by crossing-out row i and column j is called the (i,j)–minor of A and denoted $\textsf{\textbf{minor}}(A,i,j)$ (M_{ij} is common in literature). The (i,j)–cofactor of $|A|$ is $\textsf{\textbf{cof}}(A,i,j) = (-1)^{i+j}\,\textsf{\textbf{minor}}(A,i,j)$. Multiplicative factor $(-1)^{i+j}$ is called the **checkerboard sign**, because its value can be determined by counting *plus*, *minus*, *plus*, etc., from location $(1,1)$ to location (i,j) in any checkerboard fashion.

To illustrate how to create the smaller cross-out determinant, denoted by the symbol $\textsf{\textbf{minor}}(A,i,j)$, consider this example:

$$\textsf{\textbf{minor}}\left(\begin{pmatrix} 3 & 0 & 0 \\ 2 & 1 & 7 \\ 5 & 4 & 8 \end{pmatrix}, 2, 3\right) = \begin{vmatrix} 3 & 0 & \cancel{0} \\ \cancel{2} & \cancel{1} & \cancel{7} \\ 5 & 4 & \cancel{8} \end{vmatrix} = \begin{vmatrix} 3 & 0 \\ 5 & 4 \end{vmatrix}$$

cross-out row=2 and column=3, red strikeouts removed

Expansion of Determinants by Cofactors

The formulas are

$$\text{(10)} \qquad \det(A) = \sum_{j=1}^{n} a_{kj}\,\mathbf{cof}(A,k,j), \quad \det(A) = \sum_{i=1}^{n} a_{i\ell}\,\mathbf{cof}(A,i,\ell),$$

where $1 \le k \le n$, $1 \le \ell \le n$. The first expansion in (10) is called a **cofactor row expansion** and the second is called a **cofactor column expansion**. The value $\mathbf{cof}(A,i,j)$ is the cofactor of element a_{ij} in $\det(A)$, that is, the checkerboard sign times the minor of a_{ij}. The proof of expansion (10) is delayed until page 361.

The Adjugate Matrix

The **adjugate** of an $n \times n$ matrix A, denoted $\mathbf{adj}(A)$, is the transpose of the matrix of cofactors:

$$\mathbf{adj}(A) = \begin{pmatrix} \mathbf{cof}(A,1,1) & \mathbf{cof}(A,1,2) & \cdots & \mathbf{cof}(A,1,n) \\ \mathbf{cof}(A,2,1) & \mathbf{cof}(A,2,2) & \cdots & \mathbf{cof}(A,2,n) \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{cof}(A,n,1) & \mathbf{cof}(A,n,2) & \cdots & \mathbf{cof}(A,n,n) \end{pmatrix}^T.$$

A cofactor $\mathbf{cof}(A,i,j)$ is the checkerboard sign $(-1)^{i+j}$ times the corresponding cross-out determinant $\mathbf{minor}(A,i,j)$. In the 2×2 case,

$$\mathbf{adj}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

In words: *swap the diagonal elements and change the sign of the off–diagonal elements.*

The Inverse Matrix

The adjugate appears in the inverse matrix formula for a 2×2 matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

This formula is verified by direct matrix multiplication:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = (ad-bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The $n \times n$ matrix identity $A \cdot \mathbf{adj}(A) = \det(A)\, I$ implies

$$A^{-1} = \frac{1}{\det(A)}\begin{pmatrix} \mathbf{cof}(A,1,1) & \mathbf{cof}(A,1,2) & \cdots & \mathbf{cof}(A,1,n) \\ \mathbf{cof}(A,2,1) & \mathbf{cof}(A,2,2) & \cdots & \mathbf{cof}(A,2,n) \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{cof}(A,n,1) & \mathbf{cof}(A,n,2) & \cdots & \mathbf{cof}(A,n,n) \end{pmatrix}^T$$

Theorem 5.16 (Fundamental Adjugate Identity)

$$A \cdot \textbf{adj}(A) = \textbf{adj}(A) \cdot A = \det(A)\, I$$

The proof is delayed until page 362.

Determinants of Elementary Matrices

An elementary matrix E is the result of applying a combination, multiply or swap rule to the identity matrix. This definition implies that an elementary matrix is the identity matrix with a minor change applied, to wit:

Combination	Change an off-diagonal zero of I to c.
Multiply	Change a diagonal one of I to multiplier $m \neq 0$.
Swap	Swap two rows of I.

Theorem 5.17 (Determinants and Elementary Matrices)
Let E be an $n \times n$ elementary matrix. Then

Combination	$\det(E) = 1$
Multiply	$\det(E) = m$ for multiplier m.
Swap	$\det(E) = -1$
Product	$\det(EX) = \det(E)\det(X)$ for all $n \times n$ matrices X.

Theorem 5.18 (Determinants and Invertible Matrices)
Let A be a given invertible matrix. Then

$$\det(A) = \frac{(-1)^s}{m_1 m_2 \cdots m_r}$$

where s is the number of swap rules applied and m_1, m_2, ..., m_r are the nonzero multipliers used in multiply rules when A is reduced to $\mathbf{rref}(A)$.

Determinant Product Rule

The determinant rules of combination, multiply and swap imply that $\det(EX) = \det(E)\det(X)$ for elementary matrices E and square matrices X. We show that a more general relationship holds.

Theorem 5.19 (Product Rule for Determinants)
Let A and B be given $n \times n$ matrices. Then

$$\det(AB) = \det(A)\det(B).$$

Proof:

Used in the proof is the equivalence of invertibility of a square matrix C with $\det(C) \neq 0$ and $\mathbf{rref}(C) = I$.

Assume one of A or B has zero determinant. Then $\det(A)\det(B) = 0$. If $\det(B) = 0$, then $B\vec{\mathbf{x}} = \vec{\mathbf{0}}$ has infinitely many solutions, in particular a nonzero solution $\vec{\mathbf{x}}$. Multiply $B\vec{\mathbf{x}} = \vec{\mathbf{0}}$ by A, then $AB\vec{\mathbf{x}} = \vec{\mathbf{0}}$ which implies AB is not invertible. Then the identity $\det(AB) = \det(A)\det(B)$ holds, because both sides are zero. If $\det(B) \neq 0$ but $\det(A) = 0$, then there is a nonzero $\vec{\mathbf{y}}$ with $A\vec{\mathbf{y}} = \vec{\mathbf{0}}$. Because B has an inverse, then $\vec{\mathbf{x}} = B^{-1}\vec{\mathbf{y}}$ is defined and nonzero. Then $AB\vec{\mathbf{x}} = A\vec{\mathbf{y}} = \vec{\mathbf{0}}$, with $\vec{\mathbf{x}} \neq \vec{\mathbf{0}}$, which implies $\mathbf{rref}(AB) \neq I$ and $|AB| = 0$. Therefore, both sides of $\det(AB) = \det(A)\det(B)$ are zero and the identity holds.

Assume A, B are invertible. Then $C = AB$ is invertible. In particular, $\mathbf{rref}(A^{-1}) = \mathbf{rref}(B^{-1}) = I$. Write $I = \mathbf{rref}(A^{-1}) = E_1E_2\cdots E_kA^{-1}$ and $I = \mathbf{rref}(B^{-1}) = F_1F_2\cdots F_mB^{-1}$ for elementary matrices E_i, F_j. Then

$$AB = E_1E_2\cdots E_kF_1F_2\cdots F_m. \tag{11}$$

The theorem follows from repeated application of identity $\det(EX) = \det(E)\det(X)$ to relation (11), because

$$\det(A) = \det(E_1)\cdots\det(E_k), \quad \det(B) = \det(F_1)\cdots\det(F_m).$$

Cramer's Rule and the Determinant Product Formula

The equation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ in the 3×3 case is used routinely to produce the three matrix multiply equations

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 & 0 & 0 \\ x_2 & 1 & 0 \\ x_3 & 0 & 1 \end{pmatrix} = \begin{pmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{pmatrix},$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 & x_1 & 0 \\ 0 & x_2 & 0 \\ 0 & x_3 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{pmatrix},$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & x_1 \\ 0 & 1 & x_2 \\ 0 & 0 & x_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{pmatrix}.$$

The determinant of the second matrix on the left in the first equation evaluates to x_1. Similarly, in the other equations, the determinant of the second matrix evaluates to x_2, x_3, respectively. Therefore, **the determinant product theorem** applied to these three equations, followed by dividing by $\det(A)$, derives Cramer's Rule:

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{|A|}, x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}}{|A|}, x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}}{|A|}.$$

Examples

Example 5.6 (Four Properties)

Apply the four properties of a determinant to justify the formula

$$\begin{vmatrix} 12 & 6 & 0 \\ 11 & 5 & 1 \\ 10 & 2 & 2 \end{vmatrix} = 24.$$

Solution: The details:

$$\begin{vmatrix} 12 & 6 & 0 \\ 11 & 5 & 1 \\ 10 & 2 & 2 \end{vmatrix}$$

Given.

$$= \begin{vmatrix} 12 & 6 & 0 \\ -1 & -1 & 1 \\ -2 & -4 & 2 \end{vmatrix}$$

Combination rule twice: `combo(1,2,-1)`, `combo(1,3,-1)`.

$$= 6\begin{vmatrix} 2 & 1 & 0 \\ -1 & -1 & 1 \\ -2 & -4 & 2 \end{vmatrix}$$

Multiply rule: factor out 6 from row 1.

$$= 6\begin{vmatrix} 0 & -1 & 2 \\ -1 & -1 & 1 \\ 0 & -3 & 2 \end{vmatrix}$$

Combination rule twice: `combo(1,3,1)`, `combo(2,1,2)`.

$$= 6(-1)\begin{vmatrix} -1 & -1 & 1 \\ 0 & -1 & 2 \\ 0 & -3 & 2 \end{vmatrix}$$

Swap rule: `swap(1,2)`.

$$= 6(-1)^2\begin{vmatrix} 1 & 1 & -1 \\ 0 & -1 & 2 \\ 0 & -3 & 2 \end{vmatrix}$$

Multiply rule: factor out (-1) from row 1.

$$= 6\begin{vmatrix} 1 & 1 & -1 \\ 0 & -1 & 2 \\ 0 & 0 & -4 \end{vmatrix}$$

Combination rule: `combo(2,3,-3)`.

$$= 6(1)(-1)(-4)$$

Triangular rule.

$$= 24$$

Formula verified.

Example 5.7 (Determinant of an Elementary Matrix)

Compute the determinants of the following elementary matrices.

$$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}, \quad \begin{vmatrix} 1 & 0 & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}, \quad \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}.$$

Solution: The matrices correspond to toolkit operations:

```
swap(1,2), combo(3,1,c), mult(3,10).
```

Therefore, the determinant values are $-1, 1, 10$, by Theorem 5.17.

Example 5.8 (Additional Determinant Rules)

Compute the determinants by applying the additional determinant rules, page 347.

$$\begin{vmatrix} 0 & 0 \\ 1 & 0 \end{vmatrix}, \quad \begin{vmatrix} 1 & 0 & 10 \\ 0 & 1 & 0 \\ 1 & 1 & 10 \end{vmatrix}, \quad \begin{vmatrix} 1 & 3 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 6 & 4 & 2 \end{vmatrix}.$$

Solution: Answer: $0, 0, 0$. A row of zeros implies determinant zero, for the 2×2. Row 3 equal to the sum of rows 1 and 2 implies determinant zero, for the 3×3. Row 4 equals twice row 1 implies determinant zero, for the 4×4.

Example 5.9 (Adjugate and Inverse)

Compute the adjugate matrix $\mathbf{adj}(A)$ and the inverse matrix $B = \frac{\mathbf{adj}(A)}{|A|}$, given

$$A = \begin{pmatrix} 1 & 3 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Solution: The adjugate matrix is the transpose of the matrix of cofactors. **A common mistake** is to compute instead the transpose matrix, a tragic over-simplification, considering the effort required: the matrix of cofactors requires the evaluation of 16 determinants of size 3×3.

For example, the effort for one 3×3 cofactor (=(checkerboard sign)(3×3 minor determinant)) is about 30 seconds:

$$\mathbf{cof}(A, 1, 2) = (-1)^{1+2}\,\mathbf{minor}(A, 1, 2) = - \begin{vmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 4 & 2 \end{vmatrix} = 0.$$

Reported here is the answer for the adjugate matrix, an effort on paper of about 8 minutes.

$$\begin{aligned} \mathbf{adj}(A) &= \text{transpose of} \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 2 \\ 0 & 0 & -1 & 2 \\ -1 & 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ -1 & 2 & 2 & 1 \end{pmatrix} \end{aligned}$$

The determinant of A is already known, because of the formula $A\,\mathbf{adj}(A) = |A|I$. For instance, the $(1, 1)$-position in matrix $|A|I$ has value $|A|$, which from the left side of $A\,\mathbf{adj}(A) = |A|I$ equals the dot product of row 1 of A and column 1 of $\mathbf{adj}(A)$. Then $|A| = -1$.

The inverse matrix B is the adjugate matrix $\textbf{adj}(A)$ divided by the determinant $|A| = -1$:

$$B = \frac{\textbf{adj}(A)}{|A|} = \begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & -2 & -1 \end{pmatrix}.$$

Answer Check. The inverse answer can be checked by matrix multiply, using the equation $A\,\textbf{adj}(A) = |A|I$, or the equation $AB = I$. For example,

$$AB = \begin{pmatrix} 1 & 3 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & -2 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Alternate solution without determinants. Define $C = \left\langle A|I \right\rangle$ and compute with toolkit steps $\mathbf{rref}(C) = \left\langle I|B \right\rangle$. Toolkit steps can evaluate $|A|$, and B is the inverse of A. Report $\textbf{adj}(A) = |A|B$.

Example 5.10 (Cofactor Expansion Method)

Justify by cofactor expansion the identity

$$\begin{vmatrix} 10 & 5 & 0 & 0 \\ 11 & 5 & a & 0 \\ 10 & 2 & b & 0 \\ 15 & 8 & 4 & 2 \end{vmatrix} = 10(6a - b).$$

Solution: The plan is to choose the row or column with most zeros, then expand by cofactors. The greatest advantage is column 4, effectively reducing the determinant to 3×3. The resulting 3×3 is treated by a hybrid method in the next example. Here, we will expand it by cofactors, again choosing a column or row with most zeros. The details:

$$\begin{vmatrix} 10 & 5 & 0 & 0 \\ 11 & 5 & a & 0 \\ 10 & 2 & b & 0 \\ 15 & 8 & 4 & 2 \end{vmatrix}$$

Given 4×4 determinant with symbols a, b.

$$= 2(-1)^{4+4} \begin{vmatrix} 10 & 5 & 0 \\ 11 & 5 & a \\ 10 & 2 & b \end{vmatrix}$$

Cofactor expansion on column 4. Three zero terms are not written. See $\boxed{1}$ below.

$$= \begin{array}{l} 2\left(a(-1)^{2+3} \begin{vmatrix} 10 & 5 \\ 10 & 2 \end{vmatrix}\right) + \\ 2\left(b(-1)^{3+3} \begin{vmatrix} 10 & 5 \\ 11 & 5 \end{vmatrix}\right) \end{array}$$

Expand by cofactors on column 3. The zero term is not written. See $\boxed{2}$ below.

$$= \begin{array}{l} 2\left(a(-1)^{2+3}(-30)\right) + \\ 2\left(b(-1)^{3+3}(-5)\right) \end{array}$$

Expand 2×2 determinants by Sarrus' rule.

$$= 60a - 10b$$

Final answer with symbols a, b.

$\boxed{1}$ The factor 2 is from element $4,4$. The factor $(-1)^{4+4}$ is the checkerboard sign of element $4,4$. The 3×3 determinant is the **minor** obtained by cross-out of row 4, column 4.

$\boxed{2}$ For example, $2\left(a(-1)^{2+3}\begin{vmatrix} 10 & 5 \\ 10 & 2 \end{vmatrix}\right)$ is decoded as follows. Factor 2 is from the 4×4 cofactor expansion. Inside the parentheses, factor a is from the 3×3 determinant element in row 2, column 3. Factor $(-1)^{2+3}$ is the checkerboard sign of that row and column. Factor $\begin{vmatrix} 10 & 5 \\ 10 & 2 \end{vmatrix}$ is the minor determinant obtained by cross–out of row 2 and column 3.

Example 5.11 (Hybrid Method)

Justify by cofactor expansion and the four properties the identity

$$\begin{vmatrix} 10 & 5 & 0 \\ 11 & 5 & a \\ 10 & 2 & b \end{vmatrix} = 5(6a - b).$$

Solution: The details:

$$\begin{aligned}
&\begin{vmatrix} 10 & 5 & 0 \\ 11 & 5 & a \\ 10 & 2 & b \end{vmatrix} && \text{Given.} \\
&= \begin{vmatrix} 10 & 5 & 0 \\ 1 & 0 & a \\ 0 & -3 & b \end{vmatrix} && \text{Combination: subtract row 1 from the other rows.} \\
&= \begin{vmatrix} 0 & 5 & -10a \\ 1 & 0 & a \\ 0 & -3 & b \end{vmatrix} && \text{Combination: add } -10 \text{ times row 2 to row 1.} \\
&= (1)(-1)\begin{vmatrix} 5 & -10a \\ -3 & b \end{vmatrix} && \text{Cofactor expansion on column 1.} \\
&= (1)(-1)(5b - 30a) && \text{Sarrus' rule for } n = 2. \\
&= 5(6a - b). && \text{Formula verified.}
\end{aligned}$$

Example 5.12 (Determinant Product Rule)

Let A, B be 4×4 matrices. Let E_1, E_2, E_3 be elementary matrices of the same size corresponding to toolkit operations

```
combo(1,3,-2), mult(3,-5), swap(2,4).
```

Find $|A|$, given $|B| = 3$ and the equation

$$A^3 B^2 = E_3 E_2 E_1 B.$$

Solution: The idea is to use the determinant product rule $|CD| = |C||D|$ repeatedly, on the given equation, to obtain the scalar equation

$$|A|^3 |B|^2 = |E_3||E_2||E_1||B|.$$

Determinant values for elementary matrices are completely determined by the given toolkit operation: $|E_1| = 1, |E_2| = -5, |E_3| = -1$. Then the scalar equation above reduces, because of $|B| = 3$, to the algebraic equation

$$|A|^3(3)^2 = (-1)(-5)(1)(3).$$

Solving for symbol $|A|$ gives the answer $|A| = \sqrt[3]{15/9} = 1.1856$.

Example 5.13 (Cramer's Rule)

Solve by Cramer's rule the system of equations

$$\begin{array}{rcrcrcrcr} 2x_1 & + & 3x_2 & + & x_3 & - & x_4 & = & 1, \\ x_1 & + & x_2 & - & & & x_4 & = & -1, \\ & & 3x_2 & + & x_3 & + & x_4 & = & 3, \\ x_1 & + & & & x_3 & - & x_4 & = & 0, \end{array}$$

verifying $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 2$.

Solution: Form the four determinants $\Delta_1, \ldots, \Delta_4$ from the determinant of coefficients Δ as follows:

$$\Delta = \begin{vmatrix} 2 & 3 & 1 & -1 \\ 1 & 1 & 0 & -1 \\ 0 & 3 & 1 & 1 \\ 1 & 0 & 1 & -1 \end{vmatrix},$$

$$\Delta_1 = \begin{vmatrix} \mathbf{1} & 3 & 1 & -1 \\ \mathbf{-1} & 1 & 0 & -1 \\ \mathbf{3} & 3 & 1 & 1 \\ \mathbf{0} & 0 & 1 & -1 \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} 2 & \mathbf{1} & 1 & -1 \\ 1 & \mathbf{-1} & 0 & -1 \\ 0 & \mathbf{3} & 1 & 1 \\ 1 & \mathbf{0} & 1 & -1 \end{vmatrix},$$

$$\Delta_3 = \begin{vmatrix} 2 & 3 & \mathbf{1} & -1 \\ 1 & 1 & \mathbf{-1} & -1 \\ 0 & 3 & \mathbf{3} & 1 \\ 1 & 0 & \mathbf{0} & -1 \end{vmatrix}, \quad \Delta_4 = \begin{vmatrix} 2 & 3 & 1 & \mathbf{1} \\ 1 & 1 & 0 & \mathbf{-1} \\ 0 & 3 & 1 & \mathbf{3} \\ 1 & 0 & 1 & \mathbf{0} \end{vmatrix}.$$

Five repetitions of the methods used in the previous examples give the answers $\Delta = -2$, $\Delta_1 = -2$, $\Delta_2 = 0$, $\Delta_3 = -2$, $\Delta_4 = -4$, therefore Cramer's rule implies the solution $x_i = \Delta_i/\Delta$, $1 \le i \le 4$. Then $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 2$.

Answer Check. The details of the computation above can be checked in computer algebra system `maple` as follows.

```
A:=Matrix([[2,  3,  1, -1], [1,  1,  0, -1],
           [0,  3,  1,  1], [1,  0,  1, -1]]);
B1:=Matrix([[ 1,  3,  1, -1], [-1,  1,  0, -1],
            [ 3,  3,  1,  1], [ 0,  0,  1, -1]]);
Delta:= linalg[det](A); Delta1:=linalg[det](B1);
x[1]:=Delta1/Delta;
```

The Cayley-Hamilton Theorem

Presented here is an adjoint formula $F^{-1} = \mathbf{adj}(F)/\det(F)$ derivation for the celebrated Cayley-Hamilton formula

$$(-A)^n + p_{n-1}(-A)^{n-1} + \cdots + p_0 I = 0.$$

The $n \times n$ matrix A is given and I is the identity matrix. The coefficients p_k in (14) are determined by the **characteristic polynomial** of matrix A, which is defined by the determinant expansion formula

$$(12) \qquad |A - \lambda I| = (-\lambda)^n + p_{n-1}(-\lambda)^{n-1} + \cdots + p_0(-\lambda)^0.$$

The **characteristic equation of** A is $|A - \lambda I| = 0$, explicitly

$$(13) \qquad (-\lambda)^n + p_{n-1}(-\lambda)^{n-1} + \cdots + p_0(-\lambda)^0 = 0.$$

Theorem 5.20 (Cayley-Hamilton)
A square matrix A satisfies its own characteristic equation. In detail, given characteristic equation $(-\lambda)^n + p_{n-1}(-\lambda)^{n-1} + \cdots + p_0(-\lambda)^0 = 0$, then replace λ on the left by A and zero on the right side by the zero matrix $\mathbf{0}$ to obtain

$$(14) \qquad (-A)^n + p_{n-1}(-A)^{n-1} + \cdots + p_0 I = \mathbf{0}.$$

Proof of (14): Define $x = -\lambda$, $F = A + xI$ and $G = \mathbf{adj}(F)$. A cofactor of $\det(F)$ is a polynomial in x of degree at most $n-1$. Therefore, there are $n \times n$ constant matrices $C_0, \ldots, C_{n-1}$ such that

$$\mathbf{adj}(F) = x^{n-1}C_{n-1} + \cdots + xC_1 + C_0.$$

The adjugate identity $\det(F)I = \mathbf{adj}(F)\,F$ is valid for any square matrix F, even if $\det(F)$ is zero. Relation (13) implies $\det(F) = x^n + p_{n-1}x^{n-1} + \cdots + p_0$. Expand the matrix product $\mathbf{adj}(F)F$ in powers of x as follows:

$$\begin{aligned} \mathbf{adj}(F)F &= \left(\sum_{j=0}^{n-1} x^j C_j\right)(A + xI) \\ &= C_0 A + \sum_{i=1}^{n-1} x^i (C_i A + C_{i-1}) + x^n C_{n-1}. \end{aligned}$$

Match coefficients of powers of x on each side of $\det(F)I = \mathbf{adj}(F)F$ to give the relations

$$(15) \qquad \begin{cases} p_0 I &= C_0 A, \\ p_1 I &= C_1 A + C_0, \\ p_2 I &= C_2 A + C_1, \\ & \vdots \\ I &= C_{n-1}. \end{cases}$$

To complete the proof of the Cayley-Hamilton identity (14), multiply the equations in (15) by I, $(-A)$, $(-A)^2$, $(-A)^3$, $\ldots$, $(-A)^n$, respectively. Then add all the equations. The left side matches the left side of (14). The right side is a telescoping sum which adds to the zero matrix. ■

An Applied Definition of Determinant

To be developed here is another way to look at formula (9), which emphasizes the column and row structure of a determinant. The definition, which agrees with (9), leads to a short proof of the four properties, which are used to find the value of any determinant.

Permutation Matrices

A matrix P obtained from the identity matrix I by swapping rows is called a **permutation matrix**. There are $n!$ permutation matrices. To illustrate, the 3×3 permutation matrices are

$$\begin{pmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{pmatrix}, \begin{pmatrix} 1\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \end{pmatrix}, \begin{pmatrix} 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \end{pmatrix}, \begin{pmatrix} 0\,1\,0 \\ 0\,0\,1 \\ 1\,0\,0 \end{pmatrix}, \begin{pmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{pmatrix}, \begin{pmatrix} 0\,0\,1 \\ 0\,1\,0 \\ 1\,0\,0 \end{pmatrix}.$$

Define for a permutation matrix P the determinant by

$$\det(P) = (-1)^k$$

where k is the least number of row swaps required to convert P to the identity. The number k satisfies $r = k+2m$, where r is *any* count of row swaps that changes P to the identity, and m is some integer. Therefore, $\det(P) = (-1)^k = (-1)^r$. In the illustration, the corresponding determinants are 1, -1, -1, 1, 1, -1, as computed from $\det(P) = (-1)^r$, where r row swaps change P into I.

It can be verified that $\det(P)$ agrees with the value reported by formula (9). Each σ in (9) corresponds to a permutation matrix P with rows arranged in the order specified by σ. The summation in (9) for $A = P$ has exactly one nonzero term.

Sampled Product

Let A be an $n \times n$ matrix and P an $n \times n$ permutation matrix. The matrix P has ones in exactly n locations. Sampled product $A.P$ multiplies entries from the matrix A, selected by the location of the ones in P.

Definition 5.19 (Sampled Product $A.P$)
Let $\vec{A}_1, \ldots, \vec{A}_n$ be the rows of A and let $\vec{P}_1, \ldots, \vec{P}_n$ be the rows of P. Let the rows of P be rows $\sigma_1,\ldots,\sigma_n$ of identity matrix I. Define via the normal dot product $(\cdot)$ the **sampled product**

$$(16)\qquad \begin{aligned} A.P &= (A_1 \cdot P_1)(A_2 \cdot P_2) \cdots (A_n \cdot P_n) \\ &= a_{1\sigma_1} \cdots a_{n\sigma_n}. \end{aligned}$$

Equation (16) implies that $A.P$ is a linear function of the rows of A. Replace rows by columns and repeat definition (16) to show $A.P$ is a linear function of the columns of A with value $a_{\sigma_1 1} \cdots a_{\sigma_n n}$.

Sampled-Product Determinant Formula

An alternative definition of determinant is

$$\det(A) = \textstyle\sum_P \det(P)\, A.P, \tag{17}$$

where the summation extends over all possible permutation matrices P. The definition emphasizes the explicit linear dependence of the determinant upon the rows of A (or the columns of A). A tedious but otherwise routine justification shows that the college algebra definition of determinant (9) and the sampled product definition of determinant (17) give the same value.

Three Properties that Define a Determinant

Write the determinant $\det(A)$ in terms of the rows $A_1, \ldots, A_n$ of the matrix A as follows:

$$D_1(A_1, \ldots, A_n) = \sum_P \det(P)\, A.P.$$

Already known is that $D_1(A_1, \ldots, A_n)$ is a function D that satisfies the following **three properties**:

Linearity	D is linear in each argument $A_1, \ldots, A_n$.
Swap	D changes sign if two arguments are swapped. Equivalently, $D = 0$ if two arguments are equal.
Identity	$D = 1$ when $A = I$.

The equivalence reported in **swap** is obtained by expansion, e.g., for $n = 2$, $A_1 = A_2$ implies $D(A_1, A_2) = -D(A_1, A_2)$ and hence $D = 0$. Similarly, $D(A_1 + A_2, A_1 + A_2) = 0$ implies by linearity that $D(A_1, A_2) = -D(A_2, A_1)$, which is the swap property for $n = 2$.

It is less obvious that *the three properties uniquely define the determinant*:

Theorem 5.21 (Uniqueness)
If $D(A_1, \ldots, A_n)$ satisfies the properties of **linearity**, **swap** and **identity**, then $D(A_1, \ldots, A_n) = \det(A)$.

Proof: The rows of the identity matrix I are denoted $E_1, \ldots, E_n$, so that for $1 \le j \le n$ we may write the expansion

$$A_j = a_{j1}E_1 + a_{j2}E_2 + \cdots + a_{jn}E_n. \tag{18}$$

We illustrate the proof for the case $n = 2$:

$$\begin{aligned} D(A_1, A_2) &= D(a_{11}E_1 + a_{12}E_2, A_2) && \text{By (18).} \\ &= a_{11}D(E_1, A_2) + a_{12}D(E_2, A_2) && \text{By linearity.} \\ &= a_{11}a_{22}D(E_1, E_2) + a_{11}a_{21}D(E_1, E_1) && \text{Repeat for } A_2. \\ &\quad + a_{12}a_{21}D(E_2, E_1) + a_{12}a_{22}D(E_2, E_2) \end{aligned}$$

The swap and identity properties give $D(E_1, E_1) = D(E_2, E_2) = 0$ and $1 = D(E_1, E_2) = -D(E_2, E_1)$. Therefore, $D(A_1, A_2) = a_{11}a_{22} - a_{12}a_{21}$ and this implies that $D(A_1, A_2) = \det(A)$.

The proof for general n depends upon the identity

$$\begin{aligned} D(E_{\sigma_1}, \ldots, E_{\sigma_n}) &= (-1)^{\text{parity}(\sigma)} D(E_1, \ldots, E_n) \\ &= (-1)^{\text{parity}(\sigma)} \end{aligned}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a rearrangement of the integers 1, …, n. This identity is implied by the **swap** and **identity** properties. Then, as in the case $n = 2$, **linearity** implies that

$$\begin{aligned} D(A_1, \ldots, A_n) &= \textstyle\sum_\sigma a_{1\sigma_1} \cdots a_{n\sigma_n} D(E_{\sigma_1}, \ldots, E_{\sigma_n}) \\ &= \textstyle\sum_\sigma (-1)^{\text{parity}(\sigma)} a_{1\sigma_1} \cdots a_{n\sigma_n} \\ &= \det(A). \end{aligned}$$

Proofs and Details

Verification of the Four Properties:

The details will use the sampled product $A.P$ defined on page 358 and the sampled product determinant formula (17) page 359. This is done only for clarity of proof, because it is possible to use the clumsier college algebra definition of determinant (9) page 345.

Triangular. If A is $n \times n$ triangular, then in (17) appears only one nonzero term, due to zero factors in the product $A.P$. The term that appears corresponds to P=identity, therefore $A.P$ is the product of the diagonal elements of A. Since $\det(P) = \det(I) = 1$, the result follows. A similar proof can be constructed from college algebra determinant definition (9), using intuition from Sarrus' rule.

Swap. Let elementary swap matrix Q be obtained from I by swapping rows i and j. Let $B = QA$, then B equals matrix A with rows i and j swapped. To be shown: $\det(A) = -\det(B)$. By definition, $B.P = QA.P$. With effort, it is possible to show that $QA.P = P.QA = PQ.A = A.PQ$ and $\det(PQ) = -\det(P)$. Matrices PQ over all possible P duplicates the list of all permutation matrices. Then definition (17) implies the result.

Combination. Let matrix B be obtained from matrix A by adding to row j the row vector k times row i ($i \neq j$). Then $\mathbf{row}(B, j) = \mathbf{row}(A, j) + k\,\mathbf{row}(A, i)$ and $B.P = (B_1 \cdot P) \cdots (B_n \cdot P) = A.P + k\,C.P$, where C is the matrix obtained from A by *replacing* $\mathbf{row}(A, j)$ with $\mathbf{row}(A, i)$.

Matrix C has equal rows $\mathbf{row}(C, i) = \mathbf{row}(C, j) = \mathbf{row}(A, i)$. By the swap rule applied to rows i and j, $|C| = -|C|$, or $|C| = 0$. Add on P across $B.P = A.P + k\,C.P$ to obtain $|B| = |A| + k|C|$. Then $|B| = |A|$.

Multiply. Let matrices A and B have the same rows, except $\mathbf{row}(B, i) = c\,\mathbf{row}(A, i)$ for some index i. Then $B.P = c\,A.P$. Add on P across this equation to obtain $|B| = c|A|$.

Verification of the Additional Rules:

Zero row. Apply the common factor rule with $c = 2$, possible since the row has all zero entries. Then $|A| = 2|A|$, which implies $|A| = 0$.

Duplicate rows. The swap rule applies to the two duplicate rows to give $|A| = -|A|$, which implies $|A| = 0$.

Dependent rows. The determinant is unchanged by adding a linear combination of rows of A to a different row, the result a matrix B. Then $|A| = |B|$. Select the combination to create a row of zeros in B. Then $|B| = 0$ from **zero row**, implying $|A| = 0$.

RREF $\neq I$. Each step in a toolkit sequence to the RREF gives $|A| = |EB| = |E||B|$ where E is an elementary matrix and B is one frame closer to $rref(A)$. At some point $B = \mathbf{rref}(A)$, then $B \neq I$ means B has a row of zeros. Therefore, $|B| = 0$, which implies $|A| = |E||B| = 0$.

Common factor and row linearity. The sampled product $A.P$ is a linear function of each row, therefore the same is true of $|A|$ by the sampled product determinant formula (17) page 359.

Derivation of cofactor expansion (10): The column expansion formula is derived from the row expansion formula applied to the transpose. We consider only the derivation of the row expansion formula (10) for $k = 1$, because the case for general k is the same except for notation. The plan is to establish equality of the two sides of (10) for $k = 1$, which in terms of $\textsf{minor}(A, 1, j) = (-1)^{1+j}\,\textsf{cof}(A, 1, j)$ is the equality

$$\det(A) = \sum_{j=1}^{n} a_{1j}(-1)^{1+j}\,\textsf{minor}(A, 1, j). \tag{19}$$

The details require expansion of $\textsf{minor}(A, 1, j)$ in (19) via the definition of determinant $\det(A) = \sum_\sigma (-1)^{\text{parity}(\sigma)} a_{1\sigma_1} \cdots a_{n\sigma_n}$. A typical term on the right in (19) after expansion looks like

$$a_{1j}\,(-1)^{1+j}(-1)^{\text{parity}(\alpha)} a_{2\alpha_2} \cdots a_{n\alpha_n}.$$

Here, α is a rearrangement of the set of $n-1$ elements consisting of $1, \ldots, j-1, j+1, \ldots, n$. Define $\sigma = (j, \alpha_2, \ldots, \alpha_n)$, which is a rearrangement of symbols $1, \ldots, n$. After parity(α) interchanges, α is changed into $(1, \ldots, j-1, j+1, \ldots, n)$ and therefore these same interchanges transform σ into $(j, 1, \ldots, j-1, j+1, \ldots, n)$. An additional $j-1$ interchanges will transform σ into natural order $(1, \ldots, n)$. This establishes, because of $(-1)^{j-1} = (-1)^{j+1}$, the identity

$$\begin{aligned}(-1)^{\text{parity}(\sigma)} &= (-1)^{j-1+\text{parity}(\alpha)} \\ &= (-1)^{j+1+\text{parity}(\alpha)}.\end{aligned}$$

Collecting formulas gives

$$(-1)^{\text{parity}(\sigma)} a_{1\sigma_1} \cdots a_{n\sigma_n} = a_{1j}\,(-1)^{1+j}(-1)^{\text{parity}(\alpha)} a_{2\alpha_2} \cdots a_{n\alpha_n}.$$

Adding across this formula over all α and j gives a sum on the right which matches the right side of (19). Some additional thought reveals that the terms on the left add exactly to $\det(A)$, hence (19) is proved.

Derivation of Cramer's Rule: The cofactor column expansion theory implies that the Cramer's rule solution of $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ is given by

$$x_j = \frac{\Delta_j}{\Delta} = \frac{1}{\Delta}\sum_{k=1}^{n} b_k\,\textsf{cof}(A, k, j). \tag{20}$$

We will verify that $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$. Let $E_1, \ldots, E_n$ be the rows of the identity matrix. The question reduces to showing that $E_p A\vec{\mathbf{x}} = b_p$. The details will use the fact

$$\sum_{j=1}^{n} a_{pj}\,\textbf{cof}(A,k,j) = \begin{cases} \det(A) & \text{for } k = p, \\ 0 & \text{for } k \ne p, \end{cases} \tag{21}$$

Equation (21) follows by cofactor row expansion, because the sum on the left is $\det(B)$ where B is matrix A with row k replaced by row p. If B has two equal rows, then $\det(B) = 0$; otherwise, $B = A$ and $\det(B) = \det(A)$.

$$\begin{aligned} E_p A\vec{\mathbf{x}} &= \sum_{j=1}^{n} a_{pj} x_j \\ &= \frac{1}{\Delta}\sum_{j=1}^{n} a_{pj} \sum_{k=1}^{n} b_k\,\textbf{cof}(A,k,j) && \text{Apply formula (20).} \\ &= \frac{1}{\Delta}\sum_{k=1}^{n} b_k \left(\sum_{j=1}^{n} a_{pj}\,\textbf{cof}(A,k,j)\right) && \text{Switch order of summation.} \\ &= b_p && \text{Apply (21).} \end{aligned}$$

Derivation of $A \cdot$ adj$(A) = \det(A)I$: The proof uses formula (21). Consider column k of **adj**(A), denoted $\vec{X}$, multiplied against matrix A, which gives

$$A\vec{X} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}\,\textbf{cof}(A,k,j) \\ \sum_{j=1}^{n} a_{2j}\,\textbf{cof}(A,k,j) \\ \vdots \\ \sum_{j=1}^{n} a_{nj}\,\textbf{cof}(A,k,j) \end{pmatrix}.$$

By formula (21),

$$\sum_{j=1}^{n} a_{ij}\,\textbf{cof}(A,k,j) = \begin{cases} \det(A) & i = k, \\ 0 & i \ne k. \end{cases}$$

Therefore, $A\vec{X}$ is $\det(A)$ times column k of the identity I. ■

Exercises 5.3

Determinant Notation

Write formulae for x and y as quotients of 2×2 determinants. Do not evaluate the determinants!

1. $\begin{pmatrix} 1 & -1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -10 \\ 3 \end{pmatrix}$

2. $\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 10 \\ -6 \end{pmatrix}$

3. $\begin{pmatrix} 0 & -1 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 10 \end{pmatrix}$

4. $\begin{pmatrix} 0 & -3 \\ 3 & 10 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$

Sarrus' 2×2 rule

Evaluate $\det(A)$.

5. $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

6. $A = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}$

7. $A = \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix}$

8. $A = \begin{pmatrix} 5a & 1 \\ -1 & 2a \end{pmatrix}$

Sarrus' rule 3×3

Evaluate $\det(A)$.

9. $A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$

10. $A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

11. $A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

12. $A = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 2 & -1 \\ 1 & 1 & -1 \end{pmatrix}$

Inverse of a 2×2 Matrix

Define matrix A and its adjugate C:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad C = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

13. Verify $AC = |A| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

14. Display the details of the argument that $|A| \ne 0$ implies A^{-1} exists and $A^{-1} = \dfrac{C}{|A|}$.

15. Show that A^{-1} exists implies $|A| \ne 0$. Suggestion: Assume not, then $AB = BA = I$ for some matrix B and also $|A| = 0$. Find a contradiction using $AC = |A|I$ from Exercise 13.

16. Calculate the inverse of $\begin{pmatrix} 1 & 2 \\ -2 & 3 \end{pmatrix}$ using the formula developed in these exercises.

Unique Solution of a 2×2 System

Solve $A\vec{\mathbf{X}} = \vec{\mathbf{b}}$ for $\vec{\mathbf{X}}$ using Cramer's rule for 2×2 systems.

17. $A = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$, $\vec{\mathbf{b}} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

18. $A = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$, $\vec{\mathbf{b}} = \begin{pmatrix} 5 \\ -5 \end{pmatrix}$

19. $A = \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix}$, $\vec{\mathbf{b}} = \begin{pmatrix} -4 \\ 4 \end{pmatrix}$

20. $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, $\vec{\mathbf{b}} = \begin{pmatrix} -10 \\ 10 \end{pmatrix}$

Definition of Determinant

21. Let A be 3×3 with zero first row. Use the college algebra definition of determinant to show that $\det(A) = 0$.

22. Let A be 3×3 with equal first and second row. Use the college algebra definition of determinant to show that $\det(A) = 0$.

23. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Use the college algebra definition of determinant to verify that $|A| = ad - bc$.

24. Let $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$. Use the college algebra definition of determinant to verify that the determinant of A equals

$$\begin{aligned} & a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} \\ & + a_{31}a_{12}a_{23} - a_{11}a_{32}a_{23} \\ & - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13} \end{aligned}$$

Four Properties

Evaluate $\det(A)$ using the four properties for determinants, page 345.

25. $A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

26. $A = \begin{pmatrix} 0 & 0 & 1 \\ 3 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

27. $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

28. $A = \begin{pmatrix} 2 & 4 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

29. $A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 \end{pmatrix}$

30. $A = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

31. $A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$

32. $A = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}$

Elementary Matrices and the Four Rules

Find $\det(A)$.

33. A is 3×3 and obtained from the identity matrix I by three row swaps.

34. A is 7×7, obtained from I by swapping rows 1 and 2, then rows 4 and 1, then rows 1 and 3.

35. A is obtained from the matrix $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ by swapping rows 1 and 3, then two row combinations.

36. A is obtained from the matrix $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ by two row combinations, then multiply row 2 by -5.

More Determinant Rules

Cite the determinant rule that verifies $\det(A) = 0$. **Never** expand $\det(A)$! See page 347.

37. $A = \begin{pmatrix} -1 & 5 & 1 \\ 2 & -4 & -4 \\ 1 & 1 & -3 \end{pmatrix}$

38. $A = \begin{pmatrix} 0 & 0 & 0 \\ 2 & -4 & -4 \\ 1 & 1 & -3 \end{pmatrix}$

39. $A = \begin{pmatrix} 4 & -8 & -8 \\ 2 & -4 & -4 \\ 1 & 1 & -3 \end{pmatrix}$

40. $A = \begin{pmatrix} -1 & 5 & 0 \\ 2 & -4 & 0 \\ 1 & 1 & 0 \end{pmatrix}$

41. $A = \begin{pmatrix} -1 & 5 & 3 \\ 2 & -4 & 0 \\ 1 & 1 & 3 \end{pmatrix}$

42. $A = \begin{pmatrix} -1 & 5 & 4 \\ 2 & -4 & -2 \\ 1 & 1 & 2 \end{pmatrix}$

Cofactor Expansion and College Algebra

Evaluate the determinant with an efficient cofactor expansion.

43. $\begin{vmatrix} 2 & 5 & 1 \\ 2 & 0 & -4 \\ 1 & 0 & 0 \end{vmatrix}$

44. $\begin{vmatrix} 2 & 5 & 1 \\ 2 & 0 & -4 \\ 1 & 0 & 1 \end{vmatrix}$

45. $\begin{vmatrix} 2 & 5 & 0 & 0 \\ 2 & 1 & 4 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{vmatrix}$

46. $\begin{vmatrix} 0 & 2 & 0 & 1 \\ 2 & 3 & 2 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \end{vmatrix}$

47. $\begin{vmatrix} 2 & 5 & 1 & -1 & 1 \\ 0 & -1 & -4 & 1 & -1 \\ 1 & 2 & 3 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 1 & 2 & 0 & 0 & 1 \end{vmatrix}$

48. $\begin{vmatrix} 2 & 0 & 1 & -1 & 1 \\ 0 & -1 & -4 & 1 & -1 \\ 1 & 2 & 3 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 1 & 2 & 0 & 1 & 1 \end{vmatrix}$

Minors and Cofactors

Write out and then evaluate the minor and cofactor of each element cited for the matrix $A = \begin{pmatrix} 2 & 5 & y \\ x & -1 & -4 \\ 1 & 2 & z \end{pmatrix}$

49. Row 1 and column 3.

50. Row 2 and column 1.

51. Row 3 and column 2.

52. Row 3 and column 1.

Cofactor Expansion

Use cofactors to evaluate the determinant.

53. $\begin{vmatrix} 2 & 7 & 1 \\ -1 & 0 & -4 \\ 1 & 0 & 3 \end{vmatrix}$

54. $\begin{vmatrix} 2 & 7 & 7 \\ -1 & 1 & 0 \\ 1 & 2 & 0 \end{vmatrix}$

55. $\begin{vmatrix} 0 & 2 & 7 & 7 \\ 0 & -1 & 1 & 0 \\ 3 & 1 & 2 & 0 \\ 0 & -1 & 1 & 0 \end{vmatrix}$

56. $\begin{vmatrix} 0 & 2 & 7 & 7 \\ 0 & -1 & y & 0 \\ x & 1 & 2 & 0 \\ 0 & -1 & 1 & 0 \end{vmatrix}$

57. $\begin{vmatrix} 0 & 2 & 7 & 7 & 3 \\ 0 & -1 & 0 & 0 & 1 \\ x & 1 & 2 & 0 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 \end{vmatrix}$

58. $\begin{vmatrix} 0 & 2 & 7 & 7 & 3 \\ 0 & -1 & 2 & 0 & 1 \\ x & 1 & 2 & 0 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 \end{vmatrix}$

Adjugate and Inverse Matrix

Find the adjugate of A and the inverse B of A. Check the answers via the formulas $A\,\mathbf{adj}(A) = \det(A)I$ and $AB = I$.

59. $A = \begin{pmatrix} 2 & 7 \\ -1 & 0 \end{pmatrix}$

60. $A = \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix}$

61. $A = \begin{pmatrix} 5 & 1 & 1 \\ 0 & 0 & 2 \\ 1 & 0 & 3 \end{pmatrix}$

62. $A = \begin{pmatrix} 5 & 1 & 2 \\ 2 & 0 & 0 \\ 1 & 0 & 3 \end{pmatrix}$

63. $A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 2 \end{pmatrix}$

64. $A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 2 & 1 \end{pmatrix}$

Transpose and Inverse

65. Verify that $A = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ satisfies $A^T = A^{-1}$.

66. Find all 2×2 matrices $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that $\det(A) = 1$ and $A^T = A^{-1}$.

67. Find all 3×3 diagonal matrices A such that $A^T = A^{-1}$.

68. Find all 3×3 upper triangular matrices A such that $A^T = A^{-1}$.

69. Find all $n \times n$ diagonal matrices A such that $A^T = A^{-1}$.

70. Determine the $n \times n$ triangular matrices A such that $\det(A) = 1$ and $A^T = \mathbf{adj}(A)$.

Elementary Matrices

Find the determinant of A from the given equation.

71. Let $A = 5E_2E_1$ be 3×3, where E_1 multiplies row 3 of the identity by -7 and E_2 swaps rows 3 and 1 of the identity. Hint: $A = (5I)E_2E_1$.

72. Let $A = 2E_2E_1$ be 5×5, where E_1 multiplies row 3 of the identity by -2 and E_2 swaps rows 3 and 5 of the identity.

73. Let $A = E_2E_1B$ be 4×4, where E_1 multiplies row 2 of the identity by 3 and E_2 is a combination. Find $|A|$ in terms of $|B|$.

74. Let $A = 3E_2E_1B$ be 3×3, where E_1 multiplies row 2 of the identity by 3 and E_2 is a combination. Find $|A|$ in terms of $|B|$.

75. Let $A = 4E_2E_1B$ be 3×3, where E_1 multiplies row 1 of the identity by 2, E_2 is a combination and $|B| = -1$.

76. Let $A = 2E_3E_2E_1B^3$ be 3×3, where E_1 multiplies row 2 of the identity by -1, E_2 and E_3 are swaps and $|B| = -2$.

Determinants and the Toolkit

Display the toolkit steps for $\mathbf{rref}(A)$. Using only the steps, report:

- The determinant of the elementary matrix E for each step.
- The determinant of A.

77. $A = \begin{pmatrix} 2 & 3 & 1 \\ 0 & 0 & 2 \\ 1 & 0 & 4 \end{pmatrix}$

78. $A = \begin{pmatrix} 2 & 3 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$

79. $A = \begin{pmatrix} 2 & 3 & 1 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 3 & 0 & 2 \\ 1 & 0 & 2 & 1 \end{pmatrix}$

80. $A = \begin{pmatrix} 2 & 3 & 1 & 2 \\ 0 & 3 & 0 & 0 \\ 2 & 6 & 1 & 2 \\ 1 & 0 & 2 & 1 \end{pmatrix}$

Determinant Product Rule

Apply the product rule $\det(AB) = \det(A)\det(B)$.

81. Let $\det(A) = 5$ and $\det(B) = -2$. Find $\det(A^2B^3)$.

82. Let $\det(A) = 4$ and $A(B - 2A) = 0$. Find $\det(B)$.

83. Let $A = E_1E_2E_3$ where E_1, E_2 are elementary swap matrices and E_3 is an elementary combination matrix. Find $\det(A)$.

84. Assume $\det(AB + A) = 0$ and $\det(A) \neq 0$. Show that $\det(B + I) = 0$.

Cramer's 2×2 Rule

Assume

$$\begin{pmatrix} a\,b \\ c\,d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix}.$$

85. Derive the formula

$$\begin{pmatrix} a\,b \\ c\,d \end{pmatrix}\begin{pmatrix} x\,0 \\ y\,1 \end{pmatrix} = \begin{pmatrix} e\,b \\ f\,d \end{pmatrix}.$$

86. Derive the formula

$$\begin{pmatrix} a\,b \\ c\,d \end{pmatrix}\begin{pmatrix} 1\,x \\ 0\,y \end{pmatrix} = \begin{pmatrix} a\,e \\ c\,f \end{pmatrix}.$$

87. Use the determinant product rule to derive the Cramer's Rule formula

$$x = \frac{\begin{vmatrix} e & b \\ f & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}.$$

88. Derive, using the determinant product rule, the Cramer's Rule formula

$$y = \frac{\begin{vmatrix} a & e \\ c & f \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}.$$

Cramer's 3×3 Rule

Let A be the coefficient matrix in the equation

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

89. Derive the formula

$$A \begin{pmatrix} x_1 & 0 & 0 \\ x_2 & 1 & 0 \\ x_3 & 0 & 1 \end{pmatrix} = \begin{pmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{pmatrix}$$

90. Derive the formula

$$A \begin{pmatrix} 1 & 0 & x_1 \\ 0 & 1 & x_2 \\ 0 & 0 & x_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{pmatrix}$$

91. Derive, using the determinant product rule, the Cramer's Rule formula

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}.$$

92. Use the determinant product rule to derive the Cramer's Rule formula

$$x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}.$$

Cayley-Hamilton Theorem

93. Let $A = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$. Expand $|A - rI|$ to compute the characteristic polynomial of A. Answer: $r^2 - 4r + 5$.

94. Let $A = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$. Apply the Cayley-Hamiltion theorem to justify the equation

$$A^2 - 4A + 5 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

95. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Expand $|A - rI|$ by Sarrus' Rule to obtain $r^2 - (a+b)r + (ad - bc)$.

96. The result of the previous exercise is often written as $(-r)^2 + \textbf{trace}(A)(-r) + |A|$ where $\textbf{trace}(A) = a + d =$ sum of the diagonal elements. Display the details.

97. Let $\lambda^2 - 2\lambda + 1 = 0$ be the characteristic equation of a matrix A. Find a formula for A^2 in terms of A and I.

98. Let A be an $n \times n$ triangular matrix with all diagonal entries zero. Prove that $A^n = 0$.

99. Find all 2×2 matrices A such that $A^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, discovered from values of $\textbf{trace}(A)$ and $|A|$.

100. Find four 2×2 matrices A such that $A^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Applied Definition of Determinant

Miscellany for permutation matrices and the sampled product page 358

$$\begin{aligned} A.P &= (A_1 \cdot P_1)(A_2 \cdot P_2) \cdots (A_n \cdot P_n) \\ &= a_{1\sigma_1} \cdots a_{n\sigma_n}. \end{aligned}$$

101. Compute the sampled product of $\begin{pmatrix} 5 & 3 & 1 \\ 0 & 5 & 7 \\ 1 & 9 & 4 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

102. Compute the sampled product of $\begin{pmatrix} 5 & 3 & 3 \\ 0 & 2 & 7 \\ 1 & 9 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

103. Determine the permutation matrices P required to evaluate $\det(A)$ when A is 2×2.

104. Determine the permutation matrices P required to evaluate $\det(A)$ when A is 4×4.

Three Properties

Reference: Page 359, three properties that define a determinant

105. Assume $n = 3$. Prove that the three properties imply $D = 0$ when two rows are identical.

106. Assume $n = 3$. Prove that the three properties imply $D = 0$ when a row is zero.

5.4 Vector Spaces, Independence, Basis

The technical topics of independence, dependence and span apply to the study of Euclidean spaces $\mathcal{R}^2$, $\mathcal{R}^3$, ..., $\mathcal{R}^n$ and also to the continuous function space $C(E)$, the space of differentiable functions $C^1(E)$ and its generalization $C^n(E)$, and to general abstract vector spaces.

Basis and General Solution: Algebraic Equations

The term **basis** was introduced on page 207 for systems of linear algebraic equations. To review, a basis is obtained from the vector general solution $\vec{x}$ of matrix equation $A\vec{x} = \vec{0}$ by computing the partial derivatives ∂_{t_1}, ∂_{t_2}, ...of $\vec{x}$, where t_1, t_2, ...is the list of invented symbols assigned to the free variables identified in $\mathbf{rref}(A)$. The partial derivatives are **Strang's special solutions**[6] to the homogeneous equation $A\vec{x} = \vec{0}$. Solution $\vec{v}_i$ is also found by letting $t_i = 1$ with all other invented symbols zero, $1 \le i \le k$. Knowing the special solutions enables reconstruction of the general solution: multiply by constants and add.

> The general solution of $A\vec{\mathbf{x}} = \vec{\mathbf{0}}$ is the sum of constants times Strang's special solutions (they are a **basis**).

Deeper properties have been isolated for the list of Strang's special solutions, the partial derivatives $\partial_{t_1}\vec{x}$, $\partial_{t_2}\vec{x}$, ... The most important properties are **span** and **independence**.

Span, Independence and Basis

Definition 5.20 (Span of a Set of Vectors)
A list of vectors $\vec{v}_1$, ..., $\vec{v}_k$ is said to **span** an abstract vector space V (page 301), written

$$V = \mathbf{span}(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k),$$

provided V consists of exactly the set of all linear combinations

$$\vec{v} = c_1\vec{v}_1 + \cdots + c_k\vec{v}_k,$$

for all choices of constants $c_1, \ldots, c_k$.

The notion originates with the general solution $\vec{v}$ of a homogeneous matrix system $A\vec{v} = \vec{0}$, where the invented symbols t_1, ..., t_k are the constants c_1, ..., c_k and the vector partial derivative list $\partial_{t_1}\vec{v}$, ..., $\partial_{t_k}\vec{v}$ is the list of vectors $\vec{v}_1$, ..., $\vec{v}_k$.

[6] The nomenclature is due to Gilbert Strang [Strang], with **Strang's special solutions** an appropriate reference.

Definition 5.21 (Independence of Vectors)
Vectors $\vec{v}_1$, ..., $\vec{v}_k$ in an abstract vector space V are said to be **Independent** or **Linearly independent** provided each linear combination $\vec{v} = c_1\vec{v}_1 + \cdots + c_k\vec{v}_k$ is represented by a unique set of constants c_1, ..., c_k. The unique constants are called the **weights** of vector $\vec{v}$ relative to $\vec{v}_1$, ..., $\vec{v}_k$.

See pages 377 and 382 for independence tests.

Unique representation of linear combinations has an algebraic equivalent:

Linear Independence of Vectors $\vec{v}_1$, ..., $\vec{v}_k$

If two linear combinations are equal,

$$a_1\vec{v}_1 + \cdots + a_k\vec{v}_k = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k,$$

then the coefficients match

$$a_1 = b_1,\ a_2 = b_2, \ldots, a_k = b_k.$$

Definition 5.22 (Basis)
A **basis** of an abstract vector space V is defined to be a list of independent vectors $\vec{v}_1$, ..., $\vec{v}_k$ which spans V. A basis is tested by two checkpoints:

1. The list of vectors $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_k$ is independent.
2. The vectors span V, written $V = \mathbf{span}(\vec{v}_1, \ldots, \vec{v}_k)$.

A basis expresses the **general solution** of a linear problem with *the fewest possible terms.*

Theorem 5.22 (Independence of Strang's Special Solutions)
Assume matrix equation $A\vec{x} = \vec{0}$ with scalar general solution $x_1, x_2, \ldots, x_n$ using invented symbols $t_1, t_2, \ldots, t_k$. Define k **special solutions** by partial differentiation:

$$\vec{v}_1 = \partial_{t_1}\vec{x}, \quad \vec{v}_2 = \partial_{t_2}\vec{x}, \quad \ldots, \quad \vec{v}_k = \partial_{t_k}\vec{x}$$

Then:

1. Each solution $\vec{x}$ of $A\vec{x} = \vec{0}$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$.
2. The vectors $\vec{v}_1, \ldots, \vec{v}_k$ are independent.

Briefly: Strang's special solutions are independent and they form a basis for the set of solutions to $A\vec{x} = \vec{0}$. See also the Kernel Theorem 5.2 page 300.

Proof on page 393

Vector Space $\mathcal{R}^n$

The vector space $\mathcal{R}^n$ of n-element fixed column vectors (or row vectors) is from the view of applications a *storage system for organization of numerical data sets* that is equipped with an algebraic toolkit. The scheme induces a *data structure* onto the numerical data set. In particular, whether needed or not, there are pre-defined operations of addition $(+)$ and scalar multiplication $(\cdot)$ which apply to fixed vectors. The two operations on fixed vectors satisfy the *closure law* and in addition obey the *eight algebraic vector space properties.* The vector space $V = \mathcal{R}^n$ is viewed as a **data set** consisting of data item packages.

Algebraic Toolkit

The **toolkit** for an abstract vector space V is the following set of eight algebraic properties. Set V is a data set. Elements of V are data packages called **vectors**, denoted $\vec{\mathbf{X}}$ and $\vec{\mathbf{Y}}$ in the toolkit.

Closure	The operations $\vec{X}+\vec{Y}$ and $k\vec{X}$ are defined and result in a new vector which is also in the set V.	
Addition	$\vec{X}+\vec{Y}=\vec{Y}+\vec{X}$	commutative
	$\vec{X}+(\vec{Y}+\vec{Z})=(\vec{X}+\vec{Y})+\vec{Z}$	associative
	Vector $\vec{0}$ is defined and $\vec{0}+\vec{X}=\vec{X}$	zero
	Vector $-\vec{X}$ is defined and $\vec{X}+(-\vec{X})=\vec{0}$	negative
Scalar multiply	$k(\vec{X}+\vec{Y})=k\vec{X}+k\vec{Y}$	distributive I
	$(k_1+k_2)\vec{X}=k_1\vec{X}+k_2\vec{X}$	distributive II
	$k_1(k_2\vec{X})=(k_1k_2)\vec{X}$	distributive III
	$1\vec{X}=\vec{X}$	identity

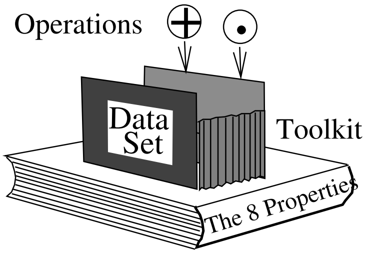


Figure 12. A Data Storage System. A vector space is a data set of data item packages plus a storage system which organizes the data. A toolkit is provided consisting of operations + and · plus 8 algebraic vector space properties.

Fixed Vectors and the Toolkit

Scalar multiplication of fixed vectors is commonly used for re-scaling, especially to unit systems fps, cgs and mks. For instance, a numerical data set of lengths recorded in meters (mks) is re-scaled to centimeters (cgs) using scale factor $k = 100$.

Addition and subtraction of fixed vectors is used in a variety of calculations, which includes averages, difference quotients and calculus operations like integration.

Planar Plot Vector Toolkit Example

The data set for a plot problem consists of plot points in $\mathcal{R}^2$ which are the **dots** for the connect-the-dots graphic. Assume the function $y(x)$ to be plotted comes from differential equation $y' = f(x, y)$. Euler's numerical method applies to compute the sequence of dots in the graphic. In this algorithm, the next dot is represented as $\vec{v}_2 = \vec{v}_1 + \vec{E}(\vec{v}_1)$ where symbol $\vec{v}_1$ is the previous dot. Symbol $\vec{E}(\vec{v}_1)$ is the Euler increment. Definitions:

$$\begin{aligned} \vec{v}_1 &= \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \quad \vec{E}(\vec{v}_1) = h \begin{pmatrix} 1 \\ f(x_0, y_0) \end{pmatrix}, \\ \vec{v}_2 &= \vec{v}_1 + \vec{E}(\vec{v}_1) = \begin{pmatrix} x_0 + h \\ y_0 + hf(x_0, y_0) \end{pmatrix}. \end{aligned}$$

Step size $h = 0.05$ is a common instance. The Euler increment $\vec{E}(\vec{v}_1)$ is defined as scalar multiplication by h against an $\mathcal{R}^2$-vector which contains an evaluation of f at the previous dot $\vec{v}_1$.

Summary. The **dots** for the graphic of $y(x)$ form a data set in the vector space $\mathcal{R}^2$. The dots are obtained by algorithm rules, which are easily expressed by vector addition $(+)$ and scalar multiplication $(\cdot)$. The 8 properties of the toolkit were used in a limited way.

Digital Photographs

A digital photo has many **pixels** arranged in a two dimensional array. Structure can be assigned to the photo by storing the pixel digital color data in a matrix A of size $n \times m$. Each entry of A is an integer which encodes the color information at a specific pixel location.

The set V of all $n \times m$ matrices is a vector space under the usual rules for matrix addition and scalar multiplication. Initially, V is just a storage system for photos. However, the algebraic toolkit for V (page 371) is a convenient way to express operations on photos. An illustration: reconstruction of a photo from RGB (Red, Green, Blue) separation photos.

Let $A = (a_{ij})$ be an $n \times m$ matrix of color data for a photo. One way to encode each entry of A is to define $a_{ij} = r_{ij} + g_{ij}x + b_{ij}x^2$ where x is some convenient base. The integers r_{ij}, g_{ij}, b_{ij} represent the amount of red, green and blue present in the pixel with data a_{ij}. Then $A = R + Gx + Bx^2$ where $R = [r_{ij}]$, $G = [g_{ij}]$, $B = [b_{ij}]$ are $n \times m$ matrices that represent the color separation photos. Construction of matrices R, G, B can be done from A by decoding integer a_{ij} into respective matrix entries. It is done with modular arithmetic. Matrices R, xG and x^2B correspond to three monochromatic photos, which can be realized as color transparencies. The transparencies placed on a standard overhead projector will reconstruct the original photograph.

Printing machinery from many years ago employed separation negatives and multiple printing runs in primary ink colors to make book photos. The advent of digital printers and simpler inexpensive technologies has made the separation process nearly obsolete. To document the historical events, we quote Sam Wang[7]:

> I encountered many difficulties when I first began making gum prints: it was not clear which paper to use; my exposing light (a sun lamp) was highly inadequate; plus a myriad of other problems. I was also using panchromatic film, making in–camera separations, holding RGB filters in front of the camera lens for three exposures onto 3 separate pieces of black and white film. I also made color separation negatives from color transparencies by enlarging in the darkroom. Both of these methods were not only tedious but often produced negatives very difficult to print — densities and contrasts that were hard to control and working in the dark with panchromatic film was definitely not fun. The fact that I got a few halfway decent prints is something of a small miracle, and represents hundreds of hours of frustrating work! Digital negatives by comparison greatly simplify the process. Nowadays (2004) I use color images from digital cameras as well as scans from slides, and the negatives print much more predictably.

Function Spaces

The default storage system used for applications involving ordinary or partial differential equations is a *function space*. The data item packages for differential equations are their solutions, which are *functions*, or in an applied context, a graphic defined on a certain graph window. They are **not** column vectors of numbers.

Functions and Column Vectors

An alternative view, adopted by researchers in numerical solutions of differential equations, is that a solution is a table of numbers, consisting of pairs of x and y values.

It is possible to think of the function as being a fixed vector. The viewpoint is that a function is a **graph** and a graph is determined by so many **dots**, which are practically obtained by **sampling** the function $y(x)$ at a reasonably dense set of x-values. The approximation is

$$y \approx \begin{pmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_n) \end{pmatrix}$$

where $x_1, \ldots, x_n$ are the **samples** and $y(x_1), \ldots, y(x_n)$ are the **sampled values** of function y.

The trouble with the approximation is that two different functions may need different sampling rates to properly represent their graphic. The result is that

[7] Sam Wang lectured on photography and art with computer at Clemson University in South Carolina. **Reference**: *A Gallery of Tri-Color Prints*, by Sam Wang

the two functions might need data storage of different dimensions, e.g., f needs its sampled values in $\mathcal{R}^{200}$ and g needs its sampled values in $\mathcal{R}^{400}$. The absence of a universal fixed vector storage system for sampled functions explains the appeal of a system like the set of all functions.

Infinitely Long Column Vectors

Is there a way around the lack of a universal numerical data storage system for sampled functions? Is it possible to *develop a theory of column vectors with infinitely many components*? It may help you to think of any function f as an infinitely long column vector, with one entry $f(x)$ for each possible sample x, e.g.,

$$\vec{f} = \begin{pmatrix} \vdots \\ f(x) \\ \vdots \end{pmatrix} \qquad \text{level } x$$

It is not clear how to order or address the entries of such a column vector: at algebraic stages it hinders. Can computers store infinitely long column vectors? The safest path through the algebra is to deal exactly with functions and function notation. Still, there is something attractive about the change from sampled approximations to a single column vector with infinite extent:

$$\vec{f} \approx \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \to \begin{pmatrix} \vdots \\ f(x) \\ \vdots \end{pmatrix} \qquad \text{level } x$$

The thinking behind the *level* x annotation is that x stands for one of the infinite possibilities for an invented sample. Alternatively, with a rich set of invented samples $x_1, \ldots, x_n$, value $f(x)$ equals approximately $f(x_j)$, where x is closest to some sample x_j.

The Vector Space V of all Functions on a Set E

The rules for function addition and scalar multiplication come from college algebra and pre-calculus backgrounds:

$$(f+g)(x) = f(x) + g(x), \quad (cf)(x) = c \cdot f(x).$$

These rules can be motivated and remembered by the notation of infinitely long column vectors, where level x is an arbitrary **sample**:

$$c_1\vec{f} + c_2\vec{g} = c_1 \begin{pmatrix} \vdots \\ f(x) \\ \vdots \end{pmatrix} + c_2 \begin{pmatrix} \vdots \\ g(x) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ c_1 f(x) + c_2 g(x) \\ \vdots \end{pmatrix}$$

The rules define **addition** and **scalar multiplication** of functions. The closure law for a vector space holds. Routine tedious justifications show that V, under the above rules for addition and scalar multiplication, has the required 8-property toolkit to make it a vector space:

Closure	The operations $f + g$ and kf are defined and result in a new function which is also in the set V of all functions on the set E.	
Addition	$f + g = g + f$	commutative
	$f + (g + h) = (f + g) + h$	associative
	The zero function 0 is defined and $0 + f = f$	zero
	The function $-f$ is defined and $f + (-f) = 0$	negative
Scalar multiply	$k(f + g) = kf + kg$	distributive I
	$(k_1 + k_2)f = k_1 f + k_2 f$	distributive II
	$k_1(k_2 f) = (k_1 k_2)f$	distributive III
	$1f = f$	identity

Important subspaces of the vector space V of all functions appear in applied literature as the storage systems for solutions to differential equations and solutions of related models.

Vector Space $C(E)$

Let $E = \{x \;:\; a < x < b\}$ be an open interval on the real line, a, b possibly infinite. The set $C(E)$ is defined to be the subset S of the set V of all functions on E obtained by restricting the function to be continuous. Because sums and scalar multiples of continuous functions are continuous, then $S = C(E)$ is a subspace of V and a vector space in its own right. The definition applies to any nonvoid subset E of $\mathcal{R}^1$.

Vector Space $C^1(E)$

The set $C^1(E)$ is the subset of the vector space $C(E)$ of all continuous functions on open interval E obtained by restricting the function to be continuously differentiable. Because sums and scalar multiples of continuously differentiable functions are continuously differentiable, then $C^1(E)$ is a subspace of $C(E)$ and a vector space in its own right.

Vector Space $C^k(E)$

The set $C^k(E)$ is the subset of the vector space $C(E)$ of all continuous functions on open interval E obtained by restricting the function to be k times continuously differentiable. Because sums and scalar multiples of k times continuously differentiable functions are k times continuously differentiable, then $C^k(E)$ is a subspace of $C(E)$ and a vector space in its own right.

Solution Space of a Differential Equation

The differential equation $y'' - y = 0$ has general solution $y = c_1 e^x + c_2 e^{-x}$, which means that the set S of all solutions of the differential equation consists of all possible linear combinations of the two functions e^x and e^{-x}. Briefly,

$$S = \textbf{span}\left(e^x, e^{-x}\right).$$

The functions e^x, e^{-x} are in $C^2(E)$ for any interval E on the x-axis. Therefore, S is a subspace of $C^2(E)$ and a vector space in its own right.

More generally, every homogeneous linear differential equation, of any order, has a solution set S which is a vector space in its own right.

Invented Vector Spaces

The number of different vector spaces used as data storage systems in scientific literature is finite, but growing with new discoveries. There is really no limit to the number of different vector spaces possible, because creative individuals are able to invent new ones.

Here is an example of how creation begets new vector spaces. Consider the problem $y' = 2y + f(x)$ and the task of storing data for the plotting of an initial value problem with initial condition $y(x_0) = y_0$. The data set V suitable for plotting consists of column vectors

$$\vec{v} = \begin{pmatrix} x_0 \\ y_0 \\ f \end{pmatrix}.$$

A plot command takes such a data item, computes the solution

$$y(x) = y_0 e^{2x} + e^{2x} \int_0^x e^{-2t} f(t)dt$$

and then plots it in a window of fixed size with center at (x_0, y_0). The column vectors are not numerical vectors in $\mathcal{R}^3$, but some **hybrid** of vectors in $\mathcal{R}^2$ and the space of continuous functions $C(E)$ where E is the real line.

It is relatively easy to come up with definitions of vector addition and scalar multiplication on V. The closure law holds and the eight vector space properties can be routinely verified. Therefore, V is an abstract vector space, unlike any found in this text. To reiterate:

> An abstract vector space is a set V and two operations of $\boxed{+}$ and $\boxed{\cdot}$ such that the closure law holds and the eight algebraic vector space properties are satisfied.

The paycheck for having recognized a vector space setting in an application is clarity of exposition and economy of effort in details. Algebraic details in $\mathcal{R}^2$ often transfer unchanged to an abstract vector space setting, line for line, to obtain the details in the more abstract setting.

Independence and Dependence

Independence is defined in Definition 5.21 page 370:

Vectors $\vec{v}_1, \ldots, \vec{v}_k$ are called **independent** provided each linear combination $\vec{v} = c_1\vec{v}_1 + \cdots + c_k\vec{v}_k$ is represented by a **unique** set of constants $c_1, \ldots, c_k$.

Independence means **unique representation of linear combinations** of $\vec{v}_1$, $\ldots$, $\vec{v}_k$, which is the statement

$$a_1\vec{v}_1 + \cdots + a_k\vec{v}_k = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k$$

implies the coefficients match:

$$\begin{cases} a_1 & = & b_1 \\ a_2 & = & b_2 \\ & \vdots & \\ a_k & = & b_k \end{cases}$$

The subject of *independence* applies to coordinate spaces $\mathcal{R}^n$, function spaces and in particular solution spaces of differential equations, digital photos, sequences of Fourier coefficients or Taylor coefficients, and general abstract vector spaces. Introduced here are definitions for low dimensions, the geometrical meaning of independence, geometric tests for independence and basic algebraic tests for independence.

The motivation for the study of independence is the theory of general solutions, which are expressions representing *all possible solutions* of a linear problem. Independence is a central issue for discovery of *the shortest possible expression* for a general solution.

Definition 5.23 (Dependence)
Vectors $\vec{v}_1, \ldots, \vec{v}_k$ are called **dependent** provided they are not independent. This means that some linear combination $\vec{v} = a_1\vec{v}_1 + \cdots + a_k\vec{v}_k$ can be represented in a second way as $\vec{v} = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k$ where for at least one index j, $a_j \neq b_j$.

Publications and proofs routinely use a brief abstract definition of independence which is a consequence of Theorem 5.23 below. See Definition 5.24 page 381 for the abstract definition normally used in mathematical proofs and technical publications.

Theorem 5.23 (Unique Representation of the Zero Vector)
Vectors $\vec{v}_1, \ldots, \vec{v}_k$ are independent in vector space V if and only if the system of equations

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}$$

has unique solution $c_1 = \cdots = c_k = 0$. Proof on page 394.

Theorem 5.24 (Subsets of Independent Sets)

Any nonvoid subset of an independent set is also independent.

Subsets of dependent sets can be either independent or dependent.

Proof on page 394.

Independence of $1, x^2, x^4$ is decided by Theorem 5.24, because it is known that powers $1, x, x^2, x^3, x^4$ form an independent set.

Independence Test: Abstract Vector Space

Theorem 5.23 provides a simple independence / dependence test.[8]

Form the system of equations

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}.$$

Solve for the constants $c_1, \ldots, c_k$.

Independence is proved if $c_1, \ldots, c_k$ are all zero.

Dependence is proved if a **nonzero** solution $c_1, \ldots, c_k$ exists. This means $c_j \neq 0$ for at least one index j.

Example 5.14 (Independence of Fixed Vectors in $\mathcal{R}^2$)

Test $\mathcal{R}^2$ vectors $\vec{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, $\vec{v}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ for independence.

Details:

The two column vectors are tested for independence by forming the system of equations $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$ and solving for the weights c_1, c_2. Then:

$$c_1 \begin{pmatrix} -1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Write the vector equation as a homogeneous system $A\vec{c} = \vec{0}$:

$$\begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The system has $\mathbf{rref}(A) = I$, details omitted. Then $c_1 = c_2 = 0$, which verifies independence of the two vectors.

Theorem 5.29 page 382 provides a shorter independence test for two vectors: $\vec{v}_1 \neq (\text{constant})\vec{v}_2$.

[8] The *test* is used in publications and mathematical proofs, often without citing the definition of independence. See Definition 5.24 page 381.

Example 5.15 (Independence of Fixed Vectors in $\mathcal{R}^3$)

Test $\mathcal{R}^3$ vectors $\vec{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$, $\vec{v}_2 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ for independence.

Details: The two column vectors are tested for independence by forming the system of equations $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$ and solving for the weights c_1, c_2:

$$c_1 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Write the vector equation as a homogeneous system $A\vec{c} = \vec{0}$:

$$\begin{pmatrix} -1 & 2 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The 3×2 coefficient matrix A has reduced row echelon form

$$\mathbf{rref}(A) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

The original homogeneous system is then equivalent to $c_1 = 0$, $c_2 = 0$. This proves the two vectors are **independent** by the independence test page 378.

See the Rank Test page 383 and the Determinant Test page 383 for additional column vector independence tests. Determinants are defined only for square matrices, therefore it is an error to use the Determinant Test on non-square Example 5.15. Determinant shortcuts for non-square problems exist [EP], but they are not discussed here.

Geometric Independence and Dependence for Two Vectors

Two vectors $\vec{v}_1$, $\vec{v}_2$ in $\mathcal{R}^2$ or $\mathcal{R}^3$ are defined to be **geometrically independent** provided neither is the zero vector and one is not a scalar multiple of the other. Graphically, this means $\vec{v}_1$ and $\vec{v}_2$ form the edges of a non-degenerate parallelogram: Figure 13. Free vector arguments use the parallelogram rule for adding and subtracting vectors: Figure 14.

> Two vectors in $\mathcal{R}^2$ or $\mathcal{R}^3$ are geometrically independent if and only if they form the edges of a parallelogram of positive area.

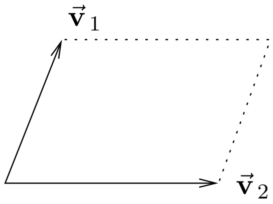


Figure 13. Geometric Independence.
Two nonzero nonparallel vectors $\vec{v}_1$, $\vec{v}_2$ form the edges of a parallelogram. A vector $\vec{v} = c_1\vec{v}_1 + c_2\vec{v}_2$ lies interior to the parallelogram if and only if the scaling constants satisfy $0 < c_1 < 1$, $0 < c_2 < 1$.

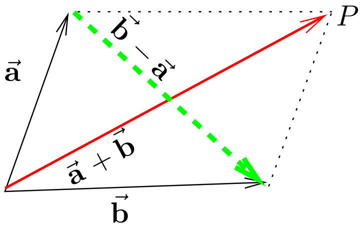

P
$\vec{\mathbf{a}}$
$\vec{\mathbf{b}}-\vec{\mathbf{a}}$
$\vec{\mathbf{a}}+\vec{\mathbf{b}}$
$\vec{\mathbf{b}}$

Figure 14. Parallelogram Rule. Given nonzero vectors $\vec{a}$, $\vec{b}$. Red sum vector $\vec{a}+\vec{b}$ has head at vertex P and tail at the joined tails of $\vec{a}$, $\vec{b}$. Green difference vector $\vec{b}-\vec{a}$ connects the head of $\vec{a}$ to the head of $\vec{b}$, according to the **head minus tail rule** on page 297.

Geometric Dependence of Two Vectors

Vectors $\vec{v}_1$, $\vec{v}_2$ in $\mathcal{R}^2$ or $\mathcal{R}^3$ are defined to be **geometrically dependent** provided they are **not geometrically independent**. This means *the two vectors do not form a parallelogram of positive area*: one of $\vec{v}_1$, $\vec{v}_2$ is the zero vector or else $\vec{v}_1$ and $\vec{v}_2$ lie along the same line.

> Two vectors in $\mathcal{R}^2$ or $\mathcal{R}^3$ are geometrically dependent if and only if one is the zero vector or else they are parallel vectors.

Geometric Independence for Three Fixed Vectors

Three vectors in $\mathcal{R}^3$ are said to be **geometrically independent** provided none of them are the zero vector and they form the edges of a non-degenerate parallelepiped of positive volume. Such vectors are called a **triad**. In the special case of all pairs orthogonal (the vectors are 90° apart) they are called an **orthogonal triad**.

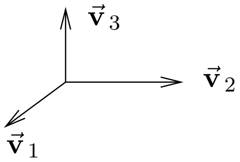


Figure 15. Geometric independence of three vectors. Vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ form the edges of a non-degenerate parallelepiped. A vector $\vec{v}=c_1\vec{v}_1+c_2\vec{v}_2+c_3\vec{v}_3$ is located interior to the parallelepiped provided $0<c_1,c_2,c_3<1$.

> Three vectors in $\mathcal{R}^3$ are **geometrically independent** if and only they form the edges of a parallelepiped of positive volume.

Geometric Dependence of Three Fixed Vectors

Given vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$, they are **dependent** if and only if they are **not independent**. The three subcases that occur can be analyzed geometrically using Theorem 5.24 page 378:

> A nonvoid subset of an independent set is independent.

1. There is a dependent subset of one vector. This vector is the zero vector.

2. There is a dependent subset of two nonzero vectors. Then two of them lie along the same line.

2. There is a dependent subset of three nonzero vectors. Then one of them is in the plane of the other two, because the three cannot form a parallelepiped of positive volume.

Three vectors in $\mathcal{R}^3$ are **geometrically dependent** if and only if one of them is in the span of the other two. The span is geometrically a point, line or plane.

Theorem 5.25 (Geometric Independence $\equiv$ Algebraic Independence)
The definitions of geometric independence and algebraic independence are equivalent. Proof on page 395.

Independence in an Abstract Vector Space

Linear algebra literature uses a purely algebraic definition of independence, which is equivalent to the **independence test** page 378. The definition and its consequences are recorded here for reference.

Definition 5.24 (Independence in an Abstract Vector Space)
Let $\vec{v}_1$, ..., $\vec{v}_k$ be a finite set of vectors in an abstract vector space V. The set is called **independent** if and only if the vector equation

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}$$

has unique solution $c_1 = \cdots = c_k = 0$.

The set of vectors is called **dependent** if and only if the set is not independent. This means that the equation in unknowns c_1, ..., c_k has a solution with at least one constant c_j nonzero.

Theorem 5.26 (Unique Representation)
Let $\vec{v}_1$, ..., $\vec{v}_k$ be independent vectors in an abstract vector space V. If scalars a_1, ..., a_k and b_1, ..., b_k satisfy the relation

$$a_1\vec{v}_1 + \cdots + a_k\vec{v}_k = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k$$

then the coefficients must match:

$$\begin{cases} a_1 & = & b_1, \\ a_2 & = & b_2, \\ \vdots & & \\ a_k & = & b_k. \end{cases}$$

Proof on page 395.
The result is often used to derive scalar equations from vector equations, e.g., the *Method of Undetermined Coefficients* in differential equations, page 560.

Theorem 5.27 (Zero Vector and Dependent Sets)
An independent set in an abstract vector space V cannot contain the zero vector. Equivalently, a set containing the zero vector is dependent. Proof on page 395

Theorem 5.28 (Linear Combination and Independence)
Let $\vec{v}_1, \ldots, \vec{v}_k$ be given vectors in abstract vector space V. Then:

1. Assume $\vec{v}_1, \ldots, \vec{v}_k$ is an independent set. Suppose $\vec{v}$ from V is not a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$. Then $\vec{v}_1, \ldots, \vec{v}_k, \vec{v}$ is an independent set.

2. If vector $\vec{v}$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$, then $\vec{v}_1, \ldots, \vec{v}_k, \vec{v}$ is a dependent set.

Proof on page 395

Theorem 5.29 (Independence of Two Vectors)
Two vectors in an abstract vector space V are independent if and only if neither is the zero vector and one is not a constant multiple of the other. Proof on page 396.

Independence and Dependence Tests for Fixed Vectors

Recorded here are a number of useful algebraic tests to determine independence or dependence of a finite list of fixed vectors.

Rank Test

In the vector space $\mathcal{R}^n$, the key to detection of independence is **zero free variables**, or nullity zero, or equivalently, maximal rank. The test is justified from the formula $\mathbf{nullity}(A) + \mathbf{rank}(A) = k$, where k is the column dimension of A.

Theorem 5.30 (Rank-Nullity Test for Three Vectors)
Let $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ be 3 column vectors in $\mathcal{R}^n$ and let their $n \times 3$ augmented matrix be

$$A = \left\langle \vec{v}_1 | \vec{v}_2 | \vec{v}_3 \right\rangle.$$

The vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ are independent if $\mathbf{rank}(A) = 3$ and dependent if $\mathbf{rank}(A) < 3$. The conditions are equivalent to $\mathbf{nullity}(A) = 0$ and $\mathbf{nullity}(A) > 0$, respectively. Proof on page 396.

Theorem 5.31 (Rank-Nullity Test)
Let $\vec{v}_1, \ldots, \vec{v}_k$ be k column vectors in $\mathcal{R}^n$ and let A be their $n \times k$ augmented matrix. The vectors are independent if $\mathbf{rank}(A) = k$ and dependent if $\mathbf{rank}(A) < k$. The conditions are equivalent to $\mathbf{nullity}(A) = 0$ and $\mathbf{nullity}(A) > 0$, respectively. Proof on page 396.

Determinant Test

In the unusual case when system $A\vec{c} = \vec{0}$ arising in the independence test is square (A is $n \times n$), then $\det(A) = 0$ detects dependence, and $\det(A) \neq 0$ detects independence. The reasoning applies formula $A^{-1} = \textbf{adj}(A)/\det(A)$, valid exactly when $\det(A) \neq 0$.

Theorem 5.32 (Determinant Test)
Let $\vec{v}_1, \ldots, \vec{v}_n$ be n column vectors in $\mathcal{R}^n$ and let A be the $n \times n$ augmented matrix of these vectors. The vectors are independent if $\det(A) \neq 0$ and dependent if $\det(A) = 0$. Proof on page 396.

Orthogonal Vector Test

In some applications the vectors being tested are known to satisfy **orthogonality conditions**. The dot product conditions for three vectors:

$$(1) \qquad \begin{array}{lll} \vec{v}_1 \cdot \vec{v}_1 > 0, & \vec{v}_2 \cdot \vec{v}_2 > 0, & \vec{v}_3 \cdot \vec{v}_3 > 0, \\ \vec{v}_1 \cdot \vec{v}_2 = 0, & \vec{v}_2 \cdot \vec{v}_3 = 0, & \vec{v}_3 \cdot \vec{v}_1 = 0. \end{array}$$

The conditions mean that the vectors are nonzero and pairwise 90° apart. The set of vectors is said to be **pairwise orthogonal**, or briefly, **orthogonal**. The orthogonality conditions for a list of k vectors are written

$$(2) \qquad \vec{v}_i \cdot \vec{v}_i > 0, \quad \vec{v}_i \cdot \vec{v}_j = 0, \quad 1 \leq i, j \leq k, \quad i \neq j.$$

Theorem 5.33 (Orthogonal Vector Test)
A set of nonzero pairwise orthogonal vectors $\vec{v}_1, \ldots, \vec{v}_k$ is linearly independent. Proof on page 397.

Independence Tests for Functions

It is not obvious how to solve for $c_1, \ldots, c_k$ in the algebraic independence test page 378, when the vectors $\vec{v}_1, \ldots, \vec{v}_k$ are not fixed vectors. If V is a set of functions, then the methods from linear algebraic equations do not directly apply. This algebraic problem causes development of special tools just for functions, called the **sampling test** and **Wronskian test**. Neither test is an equivalence. Such tests only apply to conclude independence. No results here are equipped to test dependence of a list of functions.

Sampling Test for Functions

Let f_1, f_2, f_3 be three functions defined on a domain D. Let V be the vector space of all functions $\vec{f}$ on D with the usual scalar multiplication and addition rules learned in college algebra.[9] Addressed here is the question of how to test independence and dependence of $\vec{f_1}$, $\vec{f_2}$, $\vec{f_3}$ in V. The vector relation

$$c_1\vec{f_1} + c_2\vec{f_2} + c_3\vec{f_3} = \vec{0}$$

means

$$c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) = 0, \quad x \text{ in } D.$$

An idea how to solve for c_1, c_2, c_3 arises by **sampling**, which means 3 relations are obtained by **inventing** 3 values for x, say x_1, x_2, x_3. The equations arising are

$$\begin{array}{ccccccc} c_1 f_1(x_1) & + & c_2 f_2(x_1) & + & c_3 f_3(x_1) & = & 0, \\ c_1 f_1(x_2) & + & c_2 f_2(x_2) & + & c_3 f_3(x_2) & = & 0, \\ c_1 f_1(x_3) & + & c_2 f_2(x_3) & + & c_3 f_3(x_3) & = & 0. \end{array}$$

This system of 3 equations in 3 unknowns can be written in matrix form $A\vec{c} = \vec{0}$, where the coefficient matrix A and vector $\vec{c}$ of unknowns c_1, c_2, c_3 are defined by

$$A = \begin{pmatrix} f_1(x_1) & f_2(x_1) & f_3(x_1) \\ f_1(x_2) & f_2(x_2) & f_3(x_2) \\ f_1(x_3) & f_2(x_3) & f_3(x_3) \end{pmatrix}, \quad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}.$$

The matrix A is called the **sampling matrix** for f_1, f_2, f_3 with **samples** x_1, x_2, x_3. Important: you must invent the values for the samples.

The system $A\vec{c} = \vec{0}$ has unique solution $\vec{c} = \vec{0}$, proving $\vec{f_1}$, $\vec{f_2}$, $\vec{f_3}$ independent, provided $\det(A) \neq 0$.

Definition 5.25 (Sampling Matrix)

Let functions $f_1, \ldots, f_k$ be given. Let k samples $x_1, \ldots, x_k$ be given. The **Sampling Matrix** A is defined by:

$$A = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_k(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_k(x_2) \\ \vdots & \vdots & \cdots & \vdots \\ f_1(x_k) & f_2(x_k) & \cdots & f_k(x_k) \end{pmatrix}.$$

Theorem 5.34 (Sampling Test for Functions)

The functions $f_1, \ldots, f_k$ are linearly independent on an x-set D provided there is a sampling matrix A constructed from invented samples $x_1, \ldots, x_k$ in D such that $\det(A) \neq 0$.

The converse is false. An independent list of functions may have $\det(A) = 0$ for a given sampling matrix. Proof on page 397.

[9] Symbol $\vec{f}$ is the vector package for function f. Symbol $f(x)$ is a number, a function value. Symbol f is a graph, equivalently the domain D plus equation $y = f(x)$. Vector $\vec{f}$ is the package of equation $y = f(x)$ and the domain.

Wronskian Test for Functions

The test will be explained first for two functions f_1, f_2. Independence of f_1, f_2, as in the sampling test, is decided by solving for constants c_1, c_2 in the equation

$$c_1 f_1(x) + c_2 f_2(x) = 0, \quad \text{for all } x.$$

J. M. Wronski[10] suggested to solve for the constants by differentiation of this equation, obtaining a pair of equations

$$\begin{array}{rcrcl} c_1 f_1(x) & + & c_2 f_2(x) & = & 0, \\ c_1 f_1'(x) & + & c_2 f_2'(x) & = & 0, \quad \text{for all } x. \end{array}$$

This is a system of equations $A\vec{c} = \vec{0}$ with coefficient matrix A and variable list vector $\vec{c}$ given by

$$A = \begin{pmatrix} f_1(x) & f_2(x) \\ f_1'(x) & f_2'(x) \end{pmatrix}, \quad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

The **Wronskian Test** is simply $\det(A) \neq 0$ implies $\vec{c} = \vec{0}$, similar to the sampling test:

$$\begin{vmatrix} f_1(x) & f_2(x) \\ f_1'(x) & f_2'(x) \end{vmatrix} \neq 0 \quad \text{for some } x \text{ implies } f_1,\ f_2 \text{ independent.}$$

Interesting about Wronski's idea is that it requires the invention of just one sample x such that the determinant is non-vanishing, in order to establish independence of the two functions.

Definition 5.26 (Wronskian Matrix)
Given functions f_1, ..., f_n each differentiable $n-1$ times on an interval $a < x < b$, the **Wronskian determinant** is defined by the relation

$$W(f_1, \ldots, f_n)(x) = \begin{vmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \cdots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}.$$

Theorem 5.35 (Wronskian Test)
Let functions f_1, ..., f_n be differentiable $n-1$ times on interval $a < x < b$. Assume the Wronskian determinant $W(f_1, \ldots, f_n)(x_0)$ is nonzero for some x_0 in (a, b). Then f_1, ..., f_n are independent functions in the vector space V of all functions on (a, b).

The converse is false. Independent functions may have Wronskian determinant identically zero on (a, b).

Proof on page 397.

[10] J. M. Wronski (1776-1853). Born Józef Maria Hoëné in Poland, he resided his final 40 years in France using the name Wronski.

Euler Solution Atom Test

The test originates in linear differential equations. It applies in a variety of situations outside that scope, providing basic intuition about independence of functions.

Definition 5.27 (Euler Solution Atom)
The infinite set of Euler solution atoms is a set of functions on $-\infty < x < \infty$ indexed by three variables a, b, n:

> **Index set**: real a, real $b > 0$, integer $n = 0, 1, 2, \ldots$
> **Distinct functions**: $x^n e^{ax}$, $x^n e^{ax}\cos(bx)$, $x^n e^{ax}\sin(bx)$

A **base atom** is one of $e^{ax}, e^{ax}\cos(bx), e^{ax}\sin(bx)$. An **Euler solution atom** is a base atom times x^n, index set as above.

Theorem 5.36 (Independence of Euler Solution Atoms)
A finite list of distinct Euler solution atoms is independent on any interval E in $-\infty < x < \infty$.

Outline of the proof on page 398. See also Example 5.22, page 391.

Application: Vandermonde Determinant

Choosing the functions in the *sampling test* to be 1, x, x^2 with invented samples x_1, x_2, x_3 gives the sampling matrix

$$V(x_1, x_2, x_3) = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix}.$$

The sampling matrix is called a **Vandermonde matrix**. Using the polynomial basis $f_1(x) = 1$, $f_2(x) = x$, $\ldots$, $f_k(x) = x^{k-1}$ and invented samples x_1, $\ldots$, x_k gives the $k \times k$ Vandermonde matrix

$$V(x_1, \ldots, x_k) = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{k-1} \\ 1 & x_2 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_k & \cdots & x_k^{k-1} \end{pmatrix}.$$

The most often used Vandermonde determinant identities are

$$\begin{vmatrix} 1 & a \\ 1 & b \end{vmatrix} = b - a,$$

$$\begin{vmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{vmatrix} = (c-b)(c-a)(b-a),$$

$$\begin{vmatrix} 1 & a & a^2 & a^3 \\ 1 & b & b^2 & b^3 \\ 1 & c & c^2 & c^3 \\ 1 & d & d^2 & d^3 \end{vmatrix} = (d-c)(d-b)(d-a)(c-b)(c-a)(b-a).$$

Theorem 5.37 (Vandermonde Determinant Identity)
The Vandermonde matrix has a nonzero determinant for distinct samples:

$$\det(V(x_1, \ldots, x_k)) = \prod_{i<j}(x_j - x_i).$$

Proof on page 398.

Examples

Example 5.16 (Vector General Solution)
Find the vector general solution $\vec{\mathbf{u}}$ of $A\vec{\mathbf{u}} = \vec{\mathbf{0}}$, given matrix

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Solution: The solution divides into two distinct sections: $\boxed{\mathbf{1}}$ and $\boxed{\mathbf{2}}$.

$\boxed{\mathbf{1}}$: Find the scalar general solution of the system $A\vec{x} = \vec{0}$.

The toolkit: combination, swap and multiply. Then we use the last frame algorithm. The usual shortcut applies to compute **rref**(A). We skip the augmented matrix $\left\langle A|\vec{0}\right\rangle$, knowing that the last column of zeros is unchanged by the toolkit. The details:

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

First frame.

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

`combo(1,2,-2)`.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

`combo(2,1,-2)`. Last frame, this is **rref**(A).

$$\left|\begin{array}{rcl} x_1 & = & 0, \\ x_2 & = & 0, \\ 0 & = & 0. \end{array}\right|$$

Translate to scalar equations.

$$\left|\begin{array}{lcl} x_1 &=& 0, \\ x_2 &=& 0, \\ x_3 &=& t_1. \end{array}\right|$$

Scalar general solution, obtained from the last frame algorithm: x_1, x_2=lead, x_3=free.

$\boxed{\mathbf{2}}$: Find the vector general solution of the system $A\vec{x} = \vec{0}$.

The plan is to use the answer from $\boxed{\mathbf{1}}$ and partial differentiation to display the vector general solution $\vec{x}$.

$$\left|\begin{array}{lcl} x_1 &=& 0, \\ x_2 &=& 0, \\ x_3 &=& t_1. \end{array}\right|$$

Scalar general solution, from $\boxed{\mathbf{1}}$.

$$\partial_{t_1}\vec{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Strang's special solution is the partial derivative on symbol t_1. Only one, because of only one invented symbol.

$$\vec{x} = t_1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

The vector general solution. It is the sum of terms, an invented symbol times the corresponding **special solution** (partial on that symbol). See also Example 5.19.

Example 5.17 (Independence)

Assume $\vec{v}_1, \vec{v}_2$ are independent vectors in abstract vector space V. Display the details which verify the independence of the vectors $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$.

Solution: The algebraic independence test page 378 will be applied. Form the equation

$$c_1(\vec{v}_1 + 3\vec{v}_2) + c_2(\vec{v}_1 - 2\vec{v}_2) = \vec{0}$$

and somehow solve for c_1, c_2. The plan is to re-write this equation in terms of $\vec{v}_1, \vec{v}_2$, then use the algebraic independence page 378 on vectors $\vec{v}_1, \vec{v}_2$ to obtain scalar equations for c_1, c_2. The equation re-arrangement:

$$(c_1 + c_2)\vec{v}_1 + (3c_1 - 2c_2)\vec{v}_2 = \vec{0}.$$

The independence test applied to a relation $a\vec{v}_1 + b\vec{v}_2 = \vec{0}$ implies scalar equations $a = 0$, $b = 0$. The re-arranged equation has $a = c_1 + c_2$, $b = 3c_1 - 2c_2$. Therefore, independence strips away the vectors from the re-arranged equation, leaving a system of scalar equations in symbols c_1, c_2:

$$\begin{array}{rcrcll} c_1 &+& c_2 &=& 0, & \text{The equation } a = 0, \\ 3c_1 &-& 2c_2 &=& = 0, & \text{The equation } b = 0. \end{array}$$

These equations have only the zero solution $c_1 = c_2 = 0$, because the coefficient matrix $\begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix}$ is invertible (nonzero determinant). The vectors $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$ are independent by the independence test page 378.

Example 5.18 (Span)

Let $\vec{v}_1, \vec{v}_2$ be two vectors in an abstract vector space V. Define two subspaces

$$S_1 = \mathbf{span}(\vec{v}_1, \vec{v}_2), \quad S_2 = \mathbf{span}(\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2).$$

(a) Display the technical details which show that the two subspaces are equal: $S_1 = S_2$.

(b) Use the result of (a) to prove that independence of $\vec{v}_1, \vec{v}_2$ implies independence of $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$.

Solution:

Details for (a). Sets S_1, S_2 are known to be subspaces of V by the **span theorem** page 301. To show $S_1 = S_2$, we will show each set is a subset of the other, that is, $S_2 \subset S_1$ and $S_1 \subset S_2$.

Show $S_2 \subset S_1$. By definition of **span** page 301, both vectors $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$ belong to the set S_1. Therefore, the span of these two vectors is also in subspace S_1, hence $S_2 \subset S_1$.

Show $S_1 \subset S_2$. Write $\vec{v}_1$ as a linear combination of $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$ in $\boxed{\mathbf{1}}$, $\boxed{\mathbf{2}}$ steps below. This will prove $\vec{v}_1$ belongs to S_2.

$\boxed{\mathbf{1}}$ $5\vec{v}_1 = 2(\vec{v}_1 + 3\vec{v}_2) + 3(\vec{v}_1 - 2\vec{v}_2)$. Eliminate $\vec{v}_2$ with a combination.

$\boxed{\mathbf{2}}$ $\vec{v}_1 = \frac{2}{5}(\vec{v}_1 + 3\vec{v}_2) + \frac{3}{5}(\vec{v}_1 - 2\vec{v}_2)$. Divide by 5.

Similarly, $\vec{v}_2$ belongs to S_2. Therefore, the span of $\vec{v}_1, \vec{v}_2$ belongs to S_2, or $S_1 \subset S_2$, as claimed.

Details for (b). Independence of $\vec{v}_1, \vec{v}_2$ implies $\dim(S_1) = 2$. Therefore, $\dim(S_2) = 2$. If $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$ fail to be independent, then they are dependent and span S_2. Then $\dim(S_2) \le 1$, a contradiction to $\dim(S_2) = 2$. This proves that $\vec{v}_1 + 3\vec{v}_2, \vec{v}_1 - 2\vec{v}_2$ are independent.

Example 5.19 (Independence, Span and Basis)

A 5×5 linear system $A\vec{x} = \vec{0}$ has scalar general solution

$$\begin{aligned} x_1 &= t_1 + 2t_2, \\ x_2 &= t_1, \\ x_3 &= t_2, \\ x_4 &= 4t_2 + t_3, \\ x_5 &= t_3. \end{aligned}$$

Find a basis for the solution space.

Solution: The answer is the set of **Strang's special solutions** obtained by taking partial derivatives on the symbols t_1, t_2, t_3. Details below.

$$\vec{X}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{X}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 4 \\ 0 \end{pmatrix}, \quad \vec{X}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Span. The vector general solution is expressed as the sum $\vec{x} = t_1\vec{X}_1 + t_2\vec{X}_2 + t_3\vec{X}_3$, which implies that the solution space is **span**$(\vec{X}_1, \vec{X}_2, \vec{X}_3)$.

Independence follows from Theorem 5.22, proof on page 393. Let's repeat the proof for the three special solutions $\vec{X}_1, \vec{X}_2, \vec{X}_3$, using the independence test in Theorem 5.23, which is the basis for Definition 5.24, page 381. Form the equation $c_1\vec{X}_1+c_2\vec{X}_2+c_3\vec{X}_3 = \vec{0}$ and solve for c_1, c_2, c_3. The left side of the equation is a vector solution $\vec{x}$ with invented symbols replaced by $t_1 = c_1, t_2 = c_2, t_3 = c_3$. The equation says that $\vec{x} = \vec{0}$, which in scalar form means $x_1 = x_2 = x_3 = x_4 = x_5 = 0$. The scalar general solution has lead variables x_1, x_4 and free variables x_2, x_3, x_5. The free variable equations are:

$$\begin{aligned} x_2 &= t_1, \\ x_3 &= t_2, \\ x_5 &= t_3. \end{aligned}$$

Because $x_2 = x_3 = x_5 = 0$, then $t_1 = t_2 = t_3 = 0$, which implies $c_1 = c_2 = c_3 = 0$. This proves independence of $\vec{X}_1, \vec{X}_2, \vec{X}_3$.

Special Solution Details. Take the partial derivative of the scalar general solution on symbol t_1 to create special solution $\vec{X}_1$. The others are found the same way, by partial derivatives on t_2, t_3. For symbol t_1:

$$\vec{X}_1 = \partial_{t_1}\vec{x} = \begin{pmatrix} \partial_{t_1}x_1 \\ \partial_{t_1}x_2 \\ \partial_{t_1}x_3 \\ \partial_{t_1}x_4 \\ \partial_{t_1}x_5 \end{pmatrix} = \begin{pmatrix} \partial_{t_1}(t_1+2t_2) \\ \partial_{t_1}(t_1) \\ \partial_{t_1}(t_2) \\ \partial_{t_1}(4t_2+t_3) \\ \partial_{t_1}(t_3) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Example 5.20 (Rank Test and Determinant Test)

Apply both the rank test and the determinant test to decide independence or dependence of the vectors

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_4 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 2 \end{pmatrix}.$$

Solution: Answer: The vectors are dependent.

Details for the Rank Test. Form the augmented matrix A of the four vectors and then compute the rank of A. If the rank is 4, then the rank test implies they are independent, otherwise dependent.

$$\begin{aligned} A &= \left\langle \vec{v}_1 | \vec{v}_2 | \vec{v}_3 | \vec{v}_4 \right\rangle \\ &= \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \end{pmatrix}. \end{aligned}$$

How to determine that the rank is not 4? Use rank of A equals the rank of A^T. Equivalently, the row rank equals the column rank. Then a row of zeros implies a dependent set of rows, which implies the row rank is not 4 (the rank is actually 2). Also, columns one and two of A are identical, they are dependent columns, therefore the column rank is not 4.

Details for the Determinant Test. The test uses the square matrix A defined above. The question of independence reduces to testing $|A|$ nonzero. If nonzero, then the columns of A are independent, which implies the four given vectors are independent. Otherwise, $|A| = 0$, which implies the columns of A are dependent, so the given four vectors are dependent.

All depends upon A being square: there is no determinant theory for non-square matrices.

Immediately$|A| = 0$, because A has a row of zeros. Alternatively, $|A| = 0$ because A has duplicate columns. Then the columns of A are dependent, which means dependence of the given four vectors.

Example 5.21 (Sampling Test and Wronskian Test)

Let $V = C(-\infty, \infty)$ and define vectors $\vec{v}_1 = x^2$, $\vec{v}_2 = x^{7/3}$, $\vec{v}_3 = x^5$.[11] Apply the sampling test and the Wronskian test to establish independence of the three vectors in V.

Solution: The vectors are not fixed vectors (column vectors in some $\mathcal{R}^n$), therefore the rank test and determinant test cannot apply. The Euler solution atom test does not apply: the functions are not atoms.

Sampling Test Details. A bad sample choice is $x = 0$, because it will produce a row of zeros, hence a zero determinant, leading to no test. Choose samples $x = 1, 2, 3$ for lack of insight, and then see if it works. The **sample matrix**:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 4 & (\sqrt[3]{2})^7 & 32 \\ 9 & (\sqrt[3]{3})^7 & 243 \end{pmatrix}.$$

Because $|A| \approx 132$ is nonzero, then the given vectors are independent by the sampling test.

Wronskian Test Details. Choose the sample x after finding the Wronskian matrix $W(x)$ for all x. Start with row vector $\begin{pmatrix} x^2, x^{7/3}, x^5 \end{pmatrix}$ and differentiate twice to compute the rows of the Wronskian matrix:

$$W(x) = \begin{pmatrix} x^2 & x^{7/3} & x^5 \\ 2x & \frac{7}{3}x^{4/3} & 5x^4 \\ 2 & \frac{28}{9}x^{1/3} & 20x^3 \end{pmatrix}.$$

The sample $x = 0$ won't work, because $|W(0)|$ has a row of zeros. Choose $x = 1$, then

$$W(1) = \begin{pmatrix} 1 & 1 & 1 \\ 2 & \frac{7}{3} & 5 \\ 2 & \frac{28}{9} & 20 \end{pmatrix}.$$

The determinant $|W(1)| = 8/3$ is nonzero, which implies the three functions are independent by the Wronskian test.

Example 5.22 (Solution Space of a Differential Equation)

A fifth order linear differential equation has general solution

$$y(x) = c_1 + c_2 x + c_3 e^x + c_4 e^{-x} + c_5 e^{2x}.$$

Write the solution space S in vector space $C^5(-\infty, \infty)$ as the span of basis vectors.

[11] Equation $\vec{v}_1 = x^2$ is an abuse of notation which defines vector package $\vec{v}_1$ in V with domain $(-\infty, \infty)$ and equation $y = x^2$. It is used without apology.

Solution: The answer is

$$S = \textbf{span}\left(1, x, e^x, e^{-x}, e^{2x}\right).$$

Details. A general solution is an expression for all solutions (no solutions skipped) in terms of arbitrary constants, in this case, the constants c_1 to c_5. We think of the constants as the invented symbols $t_1, t_2, \ldots$ in a matrix equation general solution. Then the expected basis vectors should be the partial derivatives on the symbols:

$$\begin{array}{rcl} \partial_{c_1} y(x) & = & 1, \\ \partial_{c_2} y(x) & = & x, \\ \partial_{c_3} y(x) & = & e^x, \\ \partial_{c_4} y(x) & = & e^{-x}, \\ \partial_{c_5} y(x) & = & e^{2x}. \end{array}$$

The five vectors so obtained already span the space S. All that remains is to prove they are independent. The easiest method to apply in this case is the Wronskian test.

Independence Details. Let $W(x)$ be the Wronskian of the five solutions above. Then row one is the list $1, x, e^x, e^{-x}, e^{2x}$ and the other four rows are successive derivatives of the first row.

$$W(x) = \begin{vmatrix} 1 & x & e^x & e^{-x} & e^{2x} \\ 0 & 1 & e^x & -e^{-x} & 2e^{2x} \\ 0 & 0 & e^x & e^{-x} & 4e^{2x} \\ 0 & 0 & e^x & -e^{-x} & 8e^{2x} \\ 0 & 0 & e^x & e^{-x} & 16e^{2x} \end{vmatrix}.$$

The cofactor rule applied twice in succession to column 1 gives

$$W(x) = \begin{vmatrix} e^x & e^{-x} & 4e^{2x} \\ e^x & -e^{-x} & 8e^{2x} \\ e^x & e^{-x} & 16e^{2x} \end{vmatrix}.$$

Choose sample $x = 0$ to simplify the work:

$$W(0) = \begin{vmatrix} 1 & 1 & 4 \\ 1 & -1 & 8 \\ 1 & 1 & 16 \end{vmatrix} = -24.$$

Then the determinant $|W(0)| = -24$ is nonzero, which implies independence of the functions in row one of $W(x)$, by the Wronskian test.

A Faster Independence Test. Generally, the Wronskian test is not used. Instead, apply the Euler solution atom test Theorem 5.36 page 386, which establishes independence without proof details.[12]

The details of the Euler solution atom test are brief: (1) The list $1, x, e^x, e^{-x}, e^{2x}$ is a finite set of distinct Euler solution atoms. (2) The test concludes that the set $1, x, e^x, e^{-x}, e^{2x}$ is independent.

Example 5.23 (Extracting a Basis from a List)

Let V be the vector space of all polynomials. Define subspace

$$S = \textbf{span}(x + 1, 2x - 1, 3x + 4, x^2).$$

Find a basis for S selected from the list $x + 1, 2x - 1, 3x + 4, x^2$.

[12] The proof of the Euler solution atom test, only outlined but not proved in this textbook, involves determinant evaluations similar to this example. An essential result used in the proof is *subsets of independent sets are independent.*

Solution: One possible answer: $x+1, 2x-1, x^2$.

The vectors $x+1, 2x-1$ are independent, because one is not a scalar multiple of the other (they are lines with slopes $1, 2$); see Theorem 5.29.

The list $x+1, 2x-1, 3x+4$ of three vectors is dependent. In detail, using Theorem 5.28, we first will show $\mathbf{span}(x+1, 2x-1) = \mathbf{span}(1, x)$, using these two stages:

$\boxed{1}$ $3x = (x+1) + (2x-1)$

$\boxed{2}$ $-3 = -2(x+1) + (2x-1)$

Divide $\boxed{1}$ by 3 and $\boxed{2}$ by -3 to show $\mathbf{span}(x+1, 2x-1) = \mathbf{span}(1, x)$. Then $3x+4$ is in $\mathbf{span}(1, x) = \mathbf{span}(x+1, 2x-1)$. Therefore, the list $x+1, 2x-1, 3x+4$ of three vectors is dependent. Skip $3x+4$ and go on to add x^2 to the list. Vector x^2 is not in $\mathbf{span}(x+1, 2x-1) = \mathbf{span}(1, x)$, because Euler solution atoms $1, x, x^2$ are independent, Theorem 5.36 page 386. The final independent set is $x+1, 2x-1, x^2$, and this is a basis for S. **Important**: a basis is not unique, for instance $1, x, x^2$ is also a basis for S. To extract a basis from the list means the expected answer is the list $x+1, 2x-1, 3x+4, x^2$ with dependent vectors removed. Many correct answers are possible.

Details and Proofs

Proof of Theorem 5.22, Independence of Special Solutions:

1. To prove: each solution $\vec{x}$ is a linear combination of $\vec{v_1}, \ldots, \vec{v}_k$. The general solution of $A\vec{x} = \vec{0}$ is written in scalar form by the last frame algorithm page 189, using invented symbols $t_1, \ldots, t_k$. Special solution $\vec{v}_i = \partial_{t_i}\vec{x}$ $(1 \le i \le k)$ can also be defined as the vector obtained from the scalar general solution with $t_i = 1$ and all other $t_1, \ldots, t_k$ set to zero. The vector general solution is a re-write of the scalar equations in vector form

$$\vec{x} = t_1\vec{v}_1 + \cdots + t_k\vec{v}_k \tag{3}$$

Therefore, each solution is a linear combination of the special solutions.

2. To prove: the vectors $\vec{v_1}, \ldots, \vec{v}_k$ are independent. Suppose a given solution $\vec{x}$ can be written in two ways as a linear combination of the special solutions:

$$\vec{x} = a_1\vec{v}_1 + \cdots + a_k\vec{v}_k, \quad \vec{x} = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k$$

Subtract the two equations and collect on $\vec{v_1}, \ldots, \vec{v}_k$:

$$(a_1 - b_1)\vec{v}_1 + \cdots + (a_k - b_k)\vec{v}_k = \vec{0}$$

Define $c_i = a_i - b_i$, $1 \le i \le k$, then rewrite the preceding equation as

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}$$

The left side of this equation is a solution of $A\vec{x} = \vec{0}$ in the form (3) produced by the last frame algorithm. Values $c_1, \ldots, c_k$ are values assigned to the invented symbols $t_1, \ldots, t_k$. Because this solution equals $\vec{0}$, then the corresponding scalar solution $x_1, \ldots, x_n$ of $A\vec{x} = \vec{0}$ is zero: $x_i = 0$ for $1 \le i \le n$. Variables x_i are divided into free variables and lead variables. The free variables in the last frame algorithm are set equal to $t_1, \ldots, t_k$. The lead variables are determined in terms of the free variables. Because all $x_i = 0$, then **all the free variables are zero**: $t_1 = \cdots = t_k = 0$, equivalently $c_1 = \cdots = c_k = 0$.

Equation $c_i = a_i - b_i$ and $c_i = 0$ implies $a_i = b_i$ for $1 \le i \le k$. This proves that a given solution cannot be represented in two different ways: vectors $\vec{v}_1, \ldots, \vec{v}_k$ are independent. ■

Proof of Theorem 5.23, Unique Representation of the Zero Vector: The proof will be given for the characteristic case $k = 3$, because details for general k can be written from this proof, by minor editing of the text.

Assume vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ are independent and $c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$. Then $a_1\vec{v}_1 + x_2\vec{v}_2 + a_3\vec{v}_3 = b_1\vec{v}_1 + b_2\vec{v}_2 + b_3\vec{v}_3$ where we define $a_1 = c_1$, $a_2 = c_2$, $a_3 = c_3$ and $b_1 = b_2 = b_3 = 0$. By independence, the coefficients match. By the definition of the symbols, this implies the equations $c_1 = a_1 = b_1 = 0$, $c_2 = a_2 = b_2 = 0$, $c_3 = a_3 = b_3 = 0$. Then $c_1 = c_2 = c_3 = 0$.

Conversely, assume $c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$ implies $c_1 = c_2 = c_3 = 0$. If

$$a_1\vec{v}_1 + a_2\vec{v}_2 + a_3\vec{v}_3 = b_1\vec{v}_1 + b_2\vec{v}_2 + b_3\vec{v}_3,$$

then subtract the right side from the left to obtain

$$(a_1 - b_1)\vec{v}_1 + (a_2 - b_2)\vec{v}_2 + (a_3 - b_3)\vec{v}_3 = \vec{0}.$$

This equation is equivalent to

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$$

where the symbols c_1, c_2, c_3 are defined by $c_1 = a_1 - b_1$, $c_2 = a_2 - b_2$, $c_3 = a_3 - b_3$. The theorem's condition implies that $c_1 = c_2 = c_3 = 0$, which in turn implies $a_1 = b_1$, $a_2 = b_2$, $a_3 = b_3$. ■

Proof of Theorem 5.24, Subsets of Independent Sets are Independent: The idea will be communicated for a set of three independent vectors $\vec{v}_1, \vec{v}_2, \vec{v}_3$. Let the subset to be tested consist of the two vectors $\vec{v}_1, \vec{v}_2$. To be applied: the algebraic independence test page 378. Form the vector equation

$$c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$$

and solve for the constants c_1, c_2. If $c_1 = c_2 = 0$ is the only solution, then $\vec{v}_1, \vec{v}_2$ is a an independent set.

Define $c_3 = 0$. Because $c_3\vec{v}_3 = \vec{0}$, the term $c_3\vec{v}_3$ can be added into the previous vector equation to obtain the new vector equation

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}.$$

Independence of the three vectors implies $c_1 = c_2 = c_3 = 0$, which in turn implies $c_1 = c_2 = 0$, completing the proof that $\vec{v}_1, \vec{v}_2$ are independent.

The proof for an arbitrary independent set $\vec{v}_1, \ldots, \vec{v}_k$ is similar. By renumbering, we can assume the subset to be tested for independence is $\vec{v}_1, \ldots, \vec{v}_m$ for some index $m \le k$. The proof amounts to adapting the proof for $k = 3$ and $m = 2$, given above. The details are omitted.

Because a single nonzero vector is an independent subset of any list of vectors, then a subset of a dependent set can be independent. If the subset of the dependent set is the whole set, then the subset is dependent. In conclusion, subsets of dependent sets can be either independent or dependent.

Proof of Theorem 5.25: The ideas below for $\mathcal{R}^2$ can be applied to supply details for $\mathcal{R}^3$, the $n = 3$ case omitted.

Assume vectors $\vec{v}_1$, $\vec{v}_2$ are geometrically independent: they are nonzero and nonparallel. To apply the independence test page 378, let's solve for c_1, c_2 in the equation

$$c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}.$$

Suppose $c_1 \neq 0$. Divide by c_1 to obtain $\vec{v}_1 = -(c_2/c_1)\vec{v}_2$. This equality says $\vec{v}_1$, $\vec{v}_2$ are parallel, so we conclude $c_1 = 0$. Replace $c_1 = 0$, then $0\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$, which implies $c_2\vec{v}_2 = \vec{0}$. Because $\vec{v}_2 \neq \vec{0}$, then $c_2 = 0$. This proves weights $c_1 = c_2 = 0$. By the independence test page 378, vectors $\vec{v}_1, \vec{v}_2$ are algebraically independent.

Assume vectors $\vec{v}_1, \vec{v}_2$ are algebraically independent. To show they are geometrically independent requires: (1) they are nonzero, (2) they are not parallel. If (1) fails, then one of the vectors is zero, say $\vec{v}_1$. The independence test page 378 detects dependence, because $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$ holds with $c_1 = 1, c_2 = 0$ (not both weights are zero). Similarly if $\vec{v}_2$ is zero. If (1) holds but (2) fails, then the vectors are nonzero and parallel, meaning $\vec{v}_1 = c\vec{v}_2$ for some scalar c. Let $c_1 = 1, c_2 = -c$ in the independence test page 378 to conclude dependence instead of independence. Therefore, (1) and (2) hold, meaning the vectors are geometrically independent. ■

Proof of Theorem 5.26, Unique Representation Abstract Space: Assume independence of $\vec{v}_1, \ldots, \vec{v}_k$. Suppose there are two equal linear combinations

$$a_1\vec{v}_1 + \cdots + a_k\vec{v}_k = b_1\vec{v}_1 + \cdots + b_k\vec{v}_k$$

Subtract:

$$(a_1 - b_1)\vec{v}_1 + \cdots + (a_k - b_k)\vec{v}_k = \vec{0}$$

Definition 5.24 page 381 says all the weights are zero: $a_j - b_j = 0$ for $1 \le j \le k$. Therefore, the coefficients must match: $a_j = b_j$ for $1 \le j \le k$. ■

Proof of Theorem 5.27, Zero Vector Abstract Space: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an independent set in abstract vector space V. Suppose $\vec{0}$ is in the set. Assume $\vec{v}_1 = \vec{0}$ by renumbering the list. Then:

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}$$

holds with $c_1 = 1$ and all other weights zero. Applying the independence test page 378 proves the set is *dependent*. ■

Proof of Theorem 5.28, Linear Combination and Independence:
1. Let $\vec{v}_1, \ldots, \vec{v}_k$ be a set of independent vectors in abstract vector space V. Assume $\vec{v}$ is not a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$. Independence test page 378 will be applied to set $\vec{v}_1, \ldots, \vec{v}_k, \vec{v}$. Form the equation

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k + c_{k+1}\vec{v} = \vec{0}$$

and solve for the coefficients. If $c_{k+1} \neq 0$, then divide by it and solve for vector $\vec{v}$ as a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$, a contradiction. Therefore, $c_{k+1} = 0$. Term $c_{k+1}\vec{v}$ is the zero vector, therefore the equation becomes

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0}$$

Independence implies $c_1 = \cdots = c_k = 0$. Then all weights are zero, proving independence of $\vec{v}_1, \ldots, \vec{v}_k, \vec{v}$ by the test page 378.

2. Suppose vector $\vec{v}$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$. Then for some constants $c_1, \ldots, c_k$:

$$\vec{v} = c_1\vec{v}_1 + \cdots + c_k\vec{v}_k$$

Define $c_{k+1} = -1$. Then

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k + c_{k+1}\vec{v} = \vec{0}$$

holds for weights $c_1, \ldots, c_{k+1}$ not all zero. Apply the independence test page 378 to prove the set is *dependent*. ■

Proof of Theorem 5.29, Independence Two Vectors Abstract Space: Let $\vec{v}_1, \vec{v}_2$ be two vectors in abstract vector space V.

If they are independent, then Theorem 5.27 implies neither can be the zero vector. If a vector is be a multiple of the other, then $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$ holds with either $c_1 = 1$ or $c_2 = 1$ (not both weights zero). Applying the independence test page 378 proves the set is *dependent*, a contradiction. Conclude that neither is a constant multiple of the other.

Assume neither is the zero vector and one is not a constant multiple of the other. Let's apply the independence test page 378. Form the system of equations

$$c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$$

and solve for c_1, c_2. If $c_1 = 0$, then $c_2\vec{v}_2 = \vec{0}$, which implies $c_2 = 0$ because $\vec{v}_2 \neq \vec{0}$. Then $c_1 = c_2 = 0$ and independence is proved by the test on page 378. Otherwise, $c_1 \neq 0$ and division results in

$$\vec{v}_1 = -\frac{c_2}{c_1}\vec{v}_2$$

which implies one vector is a constant multiple of the other, a contradiction. Conclusion: $c_1 = c_2 = 0$ and the two vectors are proved independent by the independence test page 378. ■

Proofs of Theorems 5.30, 5.31, Rank-Nullity Test: The proof will be given for $k = 3$, because a small change in the text of this proof is a proof for general k.

Suppose $\textbf{rank}(A) = 3$. Then there are 3 leading ones in $\textbf{rref}(A)$ and zero free variables. Therefore, $A\vec{c} = \vec{0}$ has unique solution $\vec{c} = \vec{0}$.

To be applied: the algebraic independence test page 378. Form the vector equation

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$$

and solve for the constants c_1, c_2, c_3. The vector equation says that a linear combination of the columns of matrix A is the zero vector, or equivalently, $A\vec{c} = \vec{0}$ where $\vec{c}$ has components c_1, c_2, c_3. Therefore, $\textbf{rank}(A) = 3$ implies $\vec{c} = \vec{0}$, or equivalently, $c_1 = c_2 = c_3 = 0$. This proves that the 3 vectors are linearly independent by the test page 378.

If $\textbf{rank}(A) < 3$, then there exists at least one free variable. Then the equation $A\vec{c} = \vec{0}$ has at least one nonzero solution $\vec{c}$. This proves that the vectors are dependent by the test page 378. ■

Proof of Theorem 5.32, Determinant Test: The proof details will be done for $n = 3$, because minor edits to this text will give the details for general n.

The algebraic independence test page 378 for vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ in $\mathcal{R}^3$ requires solving the system of linear algebraic equations

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$$

for constants c_1, c_2, c_3. The left side of the equation is a linear combination of the columns of the augmented matrix $A = \left\langle \vec{v}_1 | \vec{v}_2 | \vec{v}_3 \right\rangle$, and therefore the system can be represented as the matrix equation $A\vec{c} = \vec{0}$. If $\det(A) \ne 0$, then A^{-1} exists. Multiply the equation $A\vec{c} = \vec{0}$ by the inverse matrix to give

$$\begin{aligned} A\vec{c} &= \vec{0} \\ A^{-1}A\vec{c} &= A^{-1}\vec{0} \\ I\vec{c} &= A^{-1}\vec{0} \\ \vec{c} &= \vec{0}. \end{aligned}$$

Then $\vec{c} = \vec{0}$, or equivalently, $c_1 = c_2 = c_3 = 0$. The vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ are proved independent by the independence test page 378.

Conversely, if the vectors are independent and $A = \left\langle \vec{v}_1 | \vec{v}_2 | \vec{v}_3 \right\rangle$ is the augmented matrix of these vectors, then the system $A\vec{c} = \vec{0}$ has unique solution $\vec{c} = \vec{0}$ by the independence test page 378. The unique solution case for a homogeneous system $A\vec{c} = \vec{0}$ means no free variables or $\mathbf{rref}(A) = I$. Then A has a inverse. Because A^{-1} exists, then $\det(A) \ne 0$. ■

Proof of Theorem 5.33, Orthogonal Vector Test: The proof will be given for $k = 3$, because the details are easily supplied for k vectors, by editing the text in the proof. To be applied: the algebraic independence test page 378. Form the system of equations

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$$

and solve for the constants c_1, c_2, c_3. Constant c_1 is isolated by taking the dot product of the above equation with vector $\vec{v}_1$, to obtain the scalar equation

$$c_1\vec{v}_1 \cdot \vec{v}_1 + c_2\vec{v}_1 \cdot \vec{v}_2 + c_3\vec{v}_1 \cdot \vec{v}_3 = \vec{v}_1 \cdot \vec{0}.$$

The orthogonality relations $\vec{v}_1 \cdot \vec{v}_2 = 0$, $\vec{v}_2 \cdot \vec{v}_3 = 0$, $\vec{v}_3 \cdot \vec{v}_1 = 0$ reduce the scalar equation to

$$c_1\vec{v}_1 \cdot \vec{v}_1 + c_2(0) + c_3(0) = 0.$$

Because $\vec{v}_1 \cdot \vec{v}_1 > 0$, then $c_1 = 0$. Symmetrically, vector $\vec{v}_2$ replacing $\vec{v}_1$, the scalar equation becomes

$$c_1(0) + c_2\vec{v}_2 \cdot \vec{v}_2 + c_3(0) = 0.$$

Again, $c_2 = 0$. The argument for $c_3 = 0$ is similar. The conclusion: $c_1 = c_2 = c_3 = 0$. The three vectors are proved independent. ■

Proof of Theorem 5.34, Sampling Test: Let A be the sampling matrix of Definition 5.25. Let vector $\vec{c}$ have components $c_1, \ldots, c_k$. The algebraic independence test page 378 will be applied. Form the vector equation

$$c_1\vec{v}_1 + \cdots + c_k\vec{v}_k = \vec{0},$$

to be solved for $c_1, \ldots, c_k$. Substitute samples $x_1, \ldots, x_k$ into the vector equation and re-write as $A\vec{c} = \vec{0}$. Because $\det(A) \ne 0$, then equation $A\vec{c} = \vec{0}$ has unique solution $\vec{c} = \vec{0}$. Then all the weights are zero, proving that vectors $\vec{v}_1, \ldots, \vec{v}_k$ are independent. ■

Proof of Theorem 5.35, Wronskian Test: To be applied: the algebraic independence test page 378. Form the equation

$$c_1 f_1(x) + c_2 f_2(x) + \cdots + c_n f_n(x) = 0, \quad \text{for all } x,$$

and solve for the constants $c_1, \ldots, c_n$. The functions are proved independent provided all the constants are zero. The idea of the proof, attributed to Wronski, is to differentiate the above equation $n-1$ times, then substitute $x = x_0$ to obtain a homogeneous $n \times n$ system $A\vec{c} = \vec{0}$ for the components $c_1, \ldots, c_n$ of the vector $\vec{c}$. Because $|A| = W(f_1, \ldots, f_n)(x_0) \neq 0$, the inverse matrix $A^{-1} = \mathbf{adj}(A)/|A|$ exists. Multiply $A\vec{c} = \vec{0}$ on the left by A^{-1} to obtain $\vec{c} = \vec{0}$, completing the proof.

Proof of Theorem 5.36, Euler Solution Atom Test: An outline of the proof will be given, the excuse being that the details are long and uninteresting.[13] Unpleasantness includes complex numbers, real and imaginary parts of functions and the use of several support theorems.

$\boxed{1}$ The powers $1, x, \ldots, x^k$ are independent: Wronskian test Theorem 5.35.

$\boxed{2}$ Exponential e^x is independent of the powers $1, x, \ldots, x^k$. An easy argument uses Maclaurin series for the exponential. The same is true for e^{ax} with $a \neq 0$. Value a can be complex.

$\boxed{3}$ A list of distinct exponentials $e^{a_i x}$, $i = 1, \ldots, k$ with nonzero exponents is linearly independent. Details use the Wronskian test Theorem 5.35, Vandermonde matrices and determinants Theorem 5.37. Values a_i are allowed complex.

$\boxed{4}$ Powers $1, x, \ldots, x^k$ times e^{ax} $(a \neq 0)$ are independent. The result uses the algebraic independence test page 378 and $\boxed{1}$. Symbol a is allowed complex.

$\boxed{5}$ Powers $1, x, \ldots, x^p$ times a list of distinct complex exponentials $e^{a_i x}$, $i = 1, \ldots, q$ makes a list of pq distinct functions. This list of functions is independent. The details use the algebraic independence test page 378, double mathematical induction on p, q and $\boxed{1}$-$\boxed{4}$.

$\boxed{6}$ Restrict the values a_i in $\boxed{5}$ to be of the form $A + iB$ with $B > 0$. The real and imaginary parts of the list of functions in $\boxed{5}$ makes a set of $2pq$ distinct functions, all of which are Euler solution atoms. The set is independent.

The proof concludes by arguing that any finite set of distinct Euler solution atoms is a subset of an independent set described in $\boxed{6}$. Because *subsets of independent sets are independent*, Theorem 5.24, the proof ends. ■

Proof of Theorem 5.37, Vandermonde Determinant Identity: Let's prove the identity for the case $k = 3$, which simplifies notation. Assume distinct samples x_1, x_2, x_3. To be proved:

$$\det(V(x_1, x_2, x_3)) = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1).$$

The proof uses a recursion:

$$\det(V(x_1, x_2, x_3)) = \det(V(x_2, x_3))(x_3 - x_1)(x_2 - x_1).$$

Expansion of $\det(V(x_2, x_3)) = \begin{vmatrix} 1 & x_2 \\ 1 & x_3 \end{vmatrix} = x_3 - x_2$ by Sarrus' Rule gives the claimed $n = 3$ identity:

$$\det(V(x_1, x_2, x_3)) = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1).$$

Recursion proof. Define matrix $A = V(x, x_2, x_3)$ (x_1 replaced by x). Cofactor expansion along row one of $\det(A)$ gives a quadratic in variable x:

$$\det(A) = (1)\,\mathbf{cof}(A, 1, 1) + (x)\,\mathbf{cof}(A, 1, 2) + (x^2)\,\mathbf{cof}(A, 1, 3).$$

[13] Writing details for this is not preferred to eating shattered glass.

Because a determinant with duplicate rows has zero value, then quadratic equation $\det(A) = 0$ has roots $x = x_2$ and $x = x_3$. The factor and root theorems of college algebra apply: for some constant c,

$$\det(A) = c(x_3 - x)(x_2 - x).$$

Constant c is the coefficient of x^2 in $\det(A)$, therefore

$$c = \textbf{cof}(A, 1, 3) = (-1)^{1+3}\,\textbf{minor}(A, 1, 3) = \det(V(x_2, x_3)).$$

Then

$$\det(A) = \det(V(x_2, x_3))(x_3 - x)(x_2 - x).$$

Upon substitution of $x = x_1$, this equation becomes the claimed recursion

$$\det(V(x_1, x_2, x_3)) = \det(V(x_2, x_3))(x_3 - x_1)(x_2 - x_1).$$

Mathematical Induction. The $k \times k$ case first proves by cofactor expansion the recursion

$$\det(V(x_1, x_2, \ldots, x_k)) = \det(V(x_2, \ldots, x_k)\prod_{j=2}^{k}(x_j - x_1). \tag{4}$$

Identity (4) provides the induction step used to prove Theorem 5.37 by induction. To understand the derivation of identity (4), which also requires mathematical induction, experiment with special case $k = 4$:

$$\det(V(x_1, x_2, x_3, x_4)) = \det(V(x_2, x_3, x_4))\prod_{j=2}^{4}(x_j - x_1).$$

Exercises 5.4

Scalar and Vector General Solution

Given the scalar general solution of $A\vec{x} = \vec{0}$, find the vector general solution

$$\vec{x} = t_1\vec{u}_1 + t_2\vec{u}_2 + \cdots$$

where symbols t_1, t_2, ... denote arbitrary constants and $\vec{u}_1$, $\vec{u}_2$, ... are fixed vectors.

1. $x_1 = 2t_1$, $x_2 = t_1 - t_2$, $x_3 = t_2$

2. $x_1 = t_1 + 3t_2$, $x_2 = t_1$, $x_3 = 4t_2$, $x_4 = t_2$

3. $x_1 = t_1$, $x_2 = t_2$, $x_3 = 2t_1 + 3t_2$

4. $x_1 = 2t_1 + 3t_2 + t_3$, $x_2 = t_1$, $x_3 = t_2$, $x_4 = t_3$

Vector General Solution

Find the vector general solution $\vec{x}$ of $A\vec{x} = \vec{0}$.

5. $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$

6. $A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

7. $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

8. $A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

9. $A = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 2 & -1 & 0 \end{pmatrix}$

10. $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 2 \end{pmatrix}$

Dimension

11. Give four examples in $\mathcal{R}^3$ of $S = \textbf{span}(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ (3 vectors required) which have respectively dimensions $0, 1, 2, 3$.

12. Give an example in $\mathcal{R}^3$ of 2-dimensional subspaces S_1, S_2 with only the zero vector in common.

13. Let $S = \textbf{span}(\vec{v}_1, \vec{v}_2)$ in abstract vector space V. Explain why $\dim(S) \le 2$.

14. Let $S = \textbf{span}(\vec{v}_1, \ldots, \vec{v}_k)$ in abstract vector space V. Explain why $\dim(S) \le k$.

15. Let S be a subspace of $\mathcal{R}^3$ with basis $\vec{v}_1, \vec{v}_2$. Define $\vec{v}_3$ to be the **cross product** of $\vec{v}_1, \vec{v}_2$. What is $\dim(\textbf{span}(\vec{v}_2, \vec{v}_3))$?

16. Let S_1, S_2 be subspaces of $\mathcal{R}^4$ such that $\dim(S_1) = \dim(S_2) = 2$. Assume S_1, S_2 have only the zero vector in common. Prove or give a counter-example: the span of the union of S_1, S_2 equals $\mathcal{R}^4$.

Independence in Abstract Spaces

17. Assume linear combinations of vectors $\vec{v}_1$, $\vec{v}_2$ are uniquely determined, that is, $a_1\vec{v}_1 + a_2\vec{v}_2 = b_1\vec{v}_1 + b_2\vec{v}_2$ implies $a_1 = b_1$, $a_2 = b_2$. **Prove** this result: If $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$, then $c_1 = c_2 = 0$.

18. Assume the zero linear combination of vectors $\vec{v}_1$, $\vec{v}_2$ is uniquely determined, that is, $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$ implies $c_1 = c_2 = 0$. **Prove** this result: If $a_1\vec{v}_1 + a_2\vec{v}_2 = b_1\vec{v}_1 + b_2\vec{v}_2$, then $a_1 = b_1$, $a_2 = b_2$.

19. Prove that two **nonzero** vectors $\vec{v}_1$, $\vec{v}_2$ in an abstract vector space V are independent if and only if each of $\vec{v}_1$, $\vec{v}_2$ is not a constant multiple of the other.

20. Let $\vec{v}_1$ be a vector in an abstract vector space V. Prove that the one-element set $\vec{v}_1$ is independent if and only if $\vec{v}_1$ is not the zero vector.

21. Let V be an abstract vector space and assume $\vec{v}_1$, $\vec{v}_2$ are independent vectors in V. Define $\vec{u}_1 = \vec{v}_1 + \vec{v}_2$, $\vec{u}_2 = \vec{v}_1 + 2\vec{v}_2$. Prove that $\vec{u}_1$, $\vec{u}_2$ are independent in V. **Advice**: Fixed vectors not assumed! Bursting the vector packages is impossible, there are no components.

22. Let V be an abstract vector space and assume $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ are independent vectors in V. Define $\vec{u}_1 = \vec{v}_1 + \vec{v}_2$, $\vec{u}_2 = \vec{v}_1 + 4\vec{v}_2$, $\vec{u}_3 = \vec{v}_3 - \vec{v}_1$. Prove that $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$ are independent in V.

23. Let S be a finite set of independent vectors in an abstract vector space V. Prove that none of the vectors can be the zero vector.

24. Let S be a finite set of independent vectors in an abstract vector space V. Prove that no vector in the list can be a linear combination of the other vectors.

The Spaces $\mathcal{R}^n$

25. **(Scalar Multiply)** Let $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ have components measured in centimeters. Report constants c_1, c_2, c_3 for re-scaled data $c_1\vec{x}$, $c_2\vec{x}$, $c_3\vec{x}$ in units of kilometers, meters and millimeters.

26. **(Matrix Multiply)** Let $\vec{u} = \left(x_1, x_2, x_3, p_1, p_2, p_3 \right)^T$ have position x-units in kilometers and momentum p-units in kilogram-centimeters per millisecond. Determine a matrix M such that the vector $\vec{y} = M\vec{u}$ has SI units of meters and kilogram-meters per second.

27. Let $\vec{v}_1$, $\vec{v}_2$ be two independent vectors in $\mathcal{R}^n$. Assume $c_1\vec{v}_1 + c_2\vec{v}_2$ lies strictly interior to the parallelogram determined by $\vec{v}_1$, $\vec{v}_2$. Give geometric details explaining why $0 < c_1 < 1$ and $0 < c_2 < 1$.

28. Prove the 4 scalar multiply toolkit properties for fixed vectors in $\mathcal{R}^3$.

29. Define

$$\vec{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, -\vec{v} = \begin{pmatrix} -v_1 \\ -v_2 \\ -v_3 \end{pmatrix}.$$

Prove the 4 addition toolkit properties for fixed vectors in $\mathcal{R}^3$.

30. Use the 8 property toolkit in $\mathcal{R}^3$ to prove that zero times a vector is the zero vector.

31. Let A be an invertible 3×3 matrix. Let $\vec{v}_1, \vec{v}_2, \vec{v}_3$ be a basis for $\mathcal{R}^3$. Prove that $A\vec{v}_1, A\vec{v}_2, A\vec{v}_3$ is a basis for $\mathcal{R}^3$.

32. Let A be an invertible 3×3 matrix. Let $\vec{v}_1, \vec{v}_2, \vec{v}_3$ be dependent in $\mathcal{R}^3$. Prove that $A\vec{v}_1, A\vec{v}_2, A\vec{v}_3$ is a dependent set in $\mathcal{R}^3$.

Digital Photographs

Let V be the vector space of all 2×3 matrices. A matrix in V is a 6-pixel digital photo, a sub-section of a larger photo.

Let $B_1 = \begin{pmatrix} 1\,0\,0 \\ 0\,0\,0 \end{pmatrix}$, …, $B_6 = \begin{pmatrix} 0\,0\,0 \\ 0\,0\,1 \end{pmatrix}$. Each B_j **lights up** one pixel in the 2×3 sub-photo.

33. Prove that $B_1, \ldots, B_6$ are independent and span V: they are a **basis** for V.

34. Let $A = 2\begin{pmatrix} 1\,0\,0 \\ 0\,0\,0 \end{pmatrix} + 4\begin{pmatrix} 0\,0\,0 \\ 0\,1\,0 \end{pmatrix}$. Assume a black and white image and 0 means black. Describe photo A, from the checkerboard analogy.

Digital RGB Photos

Define red, green and blue monochrome matrices R, G, B by

$$\begin{pmatrix} 2\,0\,0 \\ 0\,1\,1 \\ 5\,8\,1 \end{pmatrix}, \begin{pmatrix} 3\,0\,0 \\ 0\,4\,0 \\ 0\,1\,0 \end{pmatrix}, \begin{pmatrix} 5\,0\,0 \\ 0\,3\,0 \\ 1\,0\,5 \end{pmatrix}.$$

35. Define base $x = 16$. Compute $A = R + xG + x^2B$.

36. Define base $x = 32$. Compute $A = R + xG + x^2B$.

Polynomial Spaces

Let V be the vector space of all cubic or less polynomials $p(x) = c_0 + c_1x + c_2x^2 + c_3x^3$.

37. Find a subspace S of V, $\dim(S) = 2$, which contains the vector $1 + x$.

38. Let S be the subset of V spanned by x, x^2 and x^3. Prove that S is a subspace of V which does not contain the polynomial $1 + x$.

39. Define set S by the conditions $p(0) = 0, p(1) = 0$. Find a basis for S.

40. Define set S by the condition $p(0) = \int_0^1 p(x)dx$. Find a basis for S.

The Space $C(E)$

Define $\vec{f}$ to be the vector package with domain $E = \{x \;:\; -2 \le x \le 2\}$ and equation $y = |x|$. Similarly, $\vec{g}$ is defined by equation $y = x$.

41. Show independence of $\vec{f}, \vec{g}$.

42. Find the dimension of $\mathbf{span}(\vec{f}, \vec{g})$.

43. Let $h(x) = 0$ on $-1 \le x \le 0$, $h(x) = -x$ on $0 \le x \le 1$. Show that $\vec{h}$ is in $C(E)$.

44. Let $h(x) = -1$ on $-2 \le x \le 0$, $h(x) = 1$ on $0 \le x \le 2$. Show that $\vec{h}$ is not in $C(E)$.

45. Let $h(x) = 0$ on $-2 \le x \le 0$, $h(x) = -x$ on $0 \le x \le 2$. Show that $\vec{h}$ is in $\mathbf{span}(\vec{f}, \vec{g})$.

46. Let $h(x) = \tan(\pi x/2)$ on $-2 < x < 2$, $h(2) = h(-2) = 0$. Explain why $\vec{h}$ is not in $C(E)$

The Space $C^1(E)$

Define $\vec{f}$ to be the vector package with domain $E = \{x \;:\; -1 \le x \le 1\}$ and equation $y = x|x|$. Similarly, $\vec{g}$ is defined by equation $y = x^2$.

47. Verify that $\vec{f}$ is in $C^1(E)$, but its derivative is not.

48. Show that $\vec{f}, \vec{g}$ are independent in $C^1(E)$.

The Space $C^k(E)$

49. Compute the first three derivatives of $y(x) = e^{-x^2}$ at $x = 0$.

50. Justify that $y(x) = e^{-x^2}$ belongs to $C^k(0, 1)$ for all $k \ge 1$.

51. Prove that the span of a finite list of distinct Euler solution atoms (page 386) is a subspace of $C^k(E)$ for any interval E.

52. Prove that $y(x) = |x|$ is in $C^k(0, 1)$ but not in $C^1(-1, 1)$.

Solution Space

A differential equations solver finds general solution $y = c_1 + c_2x + c_3e^x + c_4e^{-x}$. Use vector space $V = C^4(E)$ where E is the whole real line.

53. Write the solution set S as the span of four vectors in V.

54. Find a basis for the solution space S of the differential equation. Verify independence using the sampling test or Wronskian test.

55. Find a differential equation $y'' + a_1y' + a_0y = 0$ which has solution $y = c_1 + c_2x$.

56. Find a differential equation $y'''' + a_3y''' + a_2y'' + a_1y' + a_0y = 0$ which has solution $y = c_1 + c_2x + c_3e^x + c_4e^{-x}$.

Algebraic Independence Test for Two Vectors

Solve for c_1, c_2 in the independence test for two vectors, showing all details.

57. $\vec{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

58. $\vec{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

Dependence of two vectors

Solve for c_1, c_2 not both zero in the independence test for two vectors, showing all details for dependency of the two vectors.

59. $\vec{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$

60. $\vec{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}$

Independence Test for Three Vectors

Solve for the constants c_1, c_2, c_3 in the relation $c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 = \vec{0}$. Report dependent of independent vectors. If dependent, then display a dependency relation.

61. $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$

62. $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$

Independence in an Abstract Vector Space

In vector space V, report independence or a dependency relation for the given vectors.

63. Space $V = C(-\infty, \infty)$, $\vec{v}_1 = 1 + x$, $\vec{v}_2 = 2 + x$, $\vec{v}_3 = 3 + x^2$.

64. Space $V = C(-\infty, \infty)$, $\vec{v}_1 = x^{3/5}$, $\vec{v}_2 = x^2$, $\vec{v}_3 = 2x^2 + 3x^{3/5}$

65. Space V is all 3×3 matrices. Let $\vec{v}_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, $\vec{v}_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, $\vec{v}_3 = \begin{pmatrix} 2 & 5 & 0 \\ 0 & 2 & 5 \\ 0 & 3 & 5 \end{pmatrix}$.

66. Space V is all 2×2 matrices. Let $\vec{v}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\vec{v}_2 = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$, $\vec{v}_3 = \begin{pmatrix} 0 & 2 \\ 1 & 2 \end{pmatrix}$.

Rank Test

Compute the rank of the augmented matrix to determine independence or dependence of the given vectors.

67. $\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \end{pmatrix}$

68. $\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$

Determinant Test

Evaluate the determinant of the augmented matrix to determine independence or dependence of the given vectors.

69. $\begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 5 \\ 0 \end{pmatrix}$

70. $\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

Sampling Test for Functions

Invent samples to verify independence.

71. $\cosh(x), \sinh(x)$

72. $x^{7/3}, x\sin(x)$

73. $1, x, \sin(x)$

74. $1, \cos^2(x), \sin(x)$

Sampling Test and Dependence

For three functions f_1, f_2, f_3 to be dependent, constants c_1, c_2, c_3 must be found such that

$$c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) = 0.$$

The trick is that c_1, c_2, c_3 are not all zero and the relation holds **for all** x. The sampling test method can discover the constants, but it is **unable to prove dependence**!

75. Functions $1, x, 1+x$ are dependent. Insert $x = 1, 2, -1$ and solve for c_1, c_2, c_3, to discover a dependency relation.

76. Functions $1, \cos^2(x), \sin^2(x)$ are dependent. Cleverly choose 3 values of x, insert them, then solve for c_1, c_2, c_3, to discover a dependency relation.

Vandermonde Determinant

77. Let $V = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}$. Verify by direct computation the formula

$$|V| = x_2 - x_1.$$

78. Let $V = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix}$. Verify by direct computation the formula

$$|V| = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1).$$

Wronskian Test for Functions

Apply the Wronskian Test to verify independence.

79. $\cos(x), \sin(x)$.

80. $\cos(x), \sin(x), \sin(2x)$.

81. $x, x^{5/3}$.

82. $\cosh(x), \sinh(x)$.

Wronskian Test: Theory

83. The functions x^2 and $x|x|$ are continuously differentiable and have zero Wronskian. Verify that they **fail to be dependent** on $-1 < x < 1$.

84. The Wronskian Test can verify the independence of the powers $1, x, \ldots, x^k$. Show the determinant details.

Extracting a Basis

Given a list of vectors in space $V = \mathcal{R}^4$, extract a largest independent subset.

85. $\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$

86. $\begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$

Extracting a Basis

Given a list of vectors in space $V = C(-\infty, \infty)$, extract a largest independent subset.

87. $x, x\cos^2(x), x\sin^2(x), e^x, x + e^x$

88. $1, 2 + x, \frac{x}{1+x^2}, \frac{x^2}{1+x^2}$

Euler Solution Atom

Identify the Euler solution atoms in the given list. Strictly apply the definition: e^x is an atom but $2e^x$ is not.

89. $1, 2 + x, e^{2.15x}, e^{x^2}, \frac{x}{1+x^2}$

90. $2, x^3, e^{x/\pi}, e^{2x+1}, \ln|1 + x|$

Euler Solution Atom Test

Establish independence of set S_1.
Suggestion: First establish an identity $\mathbf{span}(S_1) = \mathbf{span}(S_2)$, where S_2 is an invented list of distinct atoms. The Test implies S_2 is independent. Extract a largest independent subset of S_1, using independence of S_2.

91. Set S_1 is the list $2, 1 + x^2, 4 + 5e^x, \pi e^{2x+\pi}, 10x\cos(x)$.

92. Set S_1 is the list $1 + x^2, 1 - x^2, 2\cos(3x), \cos(3x) + \sin(3x)$.

5.5 Basis, Dimension and Rank

The topics of basis, dimension and rank apply to the study of Euclidean spaces, continuous function spaces, spaces of differentiable functions and general abstract vector spaces.

Definition 5.28 (Basis)
A **basis** for a vector space V is defined to be an independent set of vectors such that each vector in V is a linear combination of finitely many vectors in the basis. The independent vectors are said to **span** V, with notation

$$V = \mathbf{span}(\text{the set of basis vectors}).$$

If the set of independent vectors is finite, then V is called **finite dimensional**. An important example is $\mathcal{R}^n$. Otherwise, V is said to be **infinite dimensional**. A Fourier series example: the space V spanned by $\sin(nx)$ on $-\pi \le x \le \pi$, $n = 1, 2, 3, \ldots$ is infinite dimensional.

Theorem 5.38 (Size of a Basis)
If vector space V has two bases $\vec{v}_1, \ldots, \vec{v}_p$ and $\vec{u}_1, \ldots, \vec{u}_q$, then $p = q$. Proof on page 422.

Definition 5.29 (Dimension)
The **dimension** of a finite-dimensional vector space V is defined to be the number of vectors in a basis.
Because of Theorem 5.38, the term *dimension* is well-defined.

Theorem 5.39 (Basis of a Finite-Dimensional Vector Space)
Let V be an n-dimensional vector space and $L = \{\vec{v}_1, \ldots, \vec{v}_p\}$ a list of vectors in V, not assumed linearly independent. Then:

1. If $p = n$ and L is an independent set, then L is a basis for V.

2. If $p = n$ and $\mathbf{span}(L) = V$, then L is a basis for V.

3. Always V has a basis containing a given independent subset of L.

4. If $\mathbf{span}(L) = V$, then L contains a basis for V.

Proof on page 422.

Euclidean Spaces

The space $\mathcal{R}^n$ has a **standard basis** consisting of the columns of the $n \times n$ identity matrix:

$$\begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}, \quad \cdots \quad, \begin{pmatrix}0\\0\\0\\\vdots\\1\end{pmatrix}.$$

The determinant test implies they are independent. They span $\mathcal{R}^n$ due to the formula

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \quad \cdots \quad + c_n \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Definition 5.29 implies the columns of the identity matrix form a basis of $\mathcal{R}^n$ of dimension n.

Theorem 5.40 (Basis and Dimension in $\mathcal{R}^n$)
Any basis of $\mathcal{R}^n$ has exactly n independent vectors. Further, any list of $n+1$ or more vectors in $\mathcal{R}^n$ is dependent.
Proof on page 423.

Polynomial Spaces

The vector space of all polynomials $p(x) = p_0 + p_1x + p_2x^2$ has dimension 3, justified by producing a basis 1, x, x^2. Formally, the basis elements are obtained from the expression $p(x)$ by partial differentiation on the symbols p_0, p_1, p_2.

Illustration. The subspace $S = \textbf{span}(1-x, 1+x, x)$ is the set of combinations $c_1(1-x) + c_2(1+x) + c_3x$. Partial differentiation on symbols c_1, c_2, c_3 produces the list of vectors $1-x, 1+x, x$. While they span S, they fail to be independent. Extract a **largest independent subset** from this list to find a basis for S, for example $1-x, 1+x$. Basis size 2 verifies that S has **dimension** 2: see Theorem 5.38 and Definition 5.29.

Differential Equations

The equation $y'' + y = 0$ has general solution $y = c_1 \cos x + c_2 \sin x$. Therefore, the formal partial derivatives ∂_{c_1}, ∂_{c_2} applied to the general solution y give a basis $\cos x$, $\sin x$. The solution space of $y'' + y = 0$ has dimension 2.

Similarly, $y''' = 0$ has a solution basis 1, x, x^2 and therefore its solution space has dimension 3. Generally, an nth order linear homogeneous scalar differential equation has solution space V of dimension n, and an $n \times n$ linear homogeneous system $\frac{d}{dx}\vec{y} = A\vec{y}$ has solution space V of dimension n. There is a general procedure for finding a basis for a differential equation:

> Let a linear differential equation have general solution expressed in terms of arbitrary constants c_1, c_2, ..., then a basis is found by taking the partial derivatives ∂_{c_1}, ∂_{c_2}, ...

Largest Subset of Independent Fixed Vectors

Let vectors $\vec{v}_1, \ldots, \vec{v}_k$ be given in $\mathcal{R}^n$. Then the subset

$$S = \mathbf{span}(\vec{v}_1, \ldots, \vec{v}_k)$$

of $\mathcal{R}^n$ consisting of all linear combinations $\vec{v} = c_1\vec{v}_1 + \cdots + c_k\vec{v}_k$ is a subspace of $\mathcal{R}^n$ by Theorem 5.5. The subset S is identical to the set of all linear combinations of the columns of the augmented matrix A of $\vec{v}_1, \ldots, \vec{v}_k$.

Because matrix multiply is a linear combination of columns, that is,

$$A \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = c_1\vec{v}_1 + \quad \cdots \quad + c_k\vec{v}_k,$$

then S is also equals the **image** of the matrix, $S = \mathbf{Image}(A)$.

Definition 5.30 (Image of a Matrix)

$$\mathbf{Image}(A) = \{A\vec{c} \ : \ \text{vector } \vec{c} \text{ arbitrary}\}.$$

Discussed here are efficient methods for finding a basis for any subspace S given as the span of a finite list L of vectors: $S = \mathbf{span}(L)$. The methods apply in particular when the list L consists of the columns of a matrix. Equivalently, the methods find a **largest subset of independent vectors** L_1 from the vectors in set L. This largest subset L_1 is independent and spans S, which makes it a basis for S.

Iterative Method for a Largest Independent Subset

A largest independent subset of vectors $\vec{v}_1, \ldots, \vec{v}_k$ in an abstract vector space V is identified as $\vec{v}_{i_1}, \ldots, \vec{v}_{i_p}$ for some distinct subscripts $i_1 < \cdots < i_p$. Described here is how to find such subscripts. A set containing only the zero vector is dependent, therefore let's assume at least one nonzero vector is listed. Let i_1 be the first subscript such that $\vec{v}_{i_1} \neq \vec{0}$. Define i_2 to be the *first* subscript greater than i_1 such that $\vec{v}_{i_2}$ is not a scalar multiple of $\vec{v}_{i_1}$. The process terminates if there is no such $i_2 > i_1$. Otherwise, proceed in a similar way to define i_3, i_4, ..., i_p. At each stage q we let $S = \{\vec{v}_{i_1}, \ldots, \vec{v}_{i_q}\}$ and select another vector $\vec{v}_{i_{q+1}}$ from $\vec{v}_1, \ldots, \vec{v}_k$ which is not in $\mathbf{span}(S)$. Then

$$\dim(\mathbf{span}(S)) < \dim(\mathbf{span}(S \cup \{\vec{v}_{i_{q+1}}\})).$$

Why does it work? Because each vector added which increases the dimension cannot be a linear combination of the preceding vectors, in short, the list of vectors at each stage is independent. See Example 5.24.

Pivot Theorem Method

Definition 5.31 (Pivot Column of Matrix A)
A column j of A is called a **pivot column** provided $\mathbf{rref}(A)$ has a leading one in column j. The leading ones in $\mathbf{rref}(A)$ belong to consecutive initial columns of the identity matrix I; the **matching columns in** A are the pivot columns of A.

Theorem 5.41 (Pivot Theorem: Independent Columns of A)

1. The pivot columns of a matrix A are linearly independent.

2. A non-pivot column is a linear combination of the pivot columns.

Proof on page 423.

Example 5.24 (Largest Independent Subset)
Find a largest independent subset from the five vectors

$$\vec{v}_1 = \begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \vec{v}_2 = \begin{pmatrix}1\\1\\0\\0\end{pmatrix}, \vec{v}_3 = \begin{pmatrix}1\\1\\0\\1\end{pmatrix}, \vec{v}_4 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \vec{v}_5 = \begin{pmatrix}1\\1\\0\\2\end{pmatrix}.$$

Solution:

The **Iterative Method** applies. A visual inspection shows that we should skip the zero vector $\vec{v}_1$ and add $\vec{v}_2, \vec{v}_3$ to the proposed largest independent set. Here, we use the fact that two nonzero vectors are independent if one is not a scalar multiple of the other. Because $\vec{v}_4 = \vec{v}_2 + \vec{v}_3$, we also skip $\vec{v}_4$. Formally, this dependence relation can be computed from toolkit steps on augmented matrix $B = \left\langle \vec{v}_2 | \vec{v}_3 | \vec{v}_4 \right\rangle$. Similarly, $\vec{v}_5 = 2\vec{v}_4 + \vec{v}_2$, causing a skip of $\vec{v}_5$. A largest independent subset is $\vec{v}_2, \vec{v}_3$.

The **Pivot Theorem** applies. This method has a computer implementation. Form the augmented matrix A of the five vectors and then compute $\mathbf{rref}(A)$.

$$A = \begin{pmatrix} 0\,1\,0\,0 & -2 \\ 0\,0\,1\,0 & 3 \\ 0\,0\,0\,1 & -1 \\ 0\,0\,0\,0 & 0 \end{pmatrix}, \quad \mathbf{rref}(A) = \begin{pmatrix} 0\,1\,0\,0 & -2 \\ 0\,0\,1\,0 & 3 \\ 0\,0\,0\,1 & -1 \\ 0\,0\,0\,0 & 0 \end{pmatrix}.$$

Then columns 2, 3 of matrix A are the pivot columns of A. We report $\vec{v}_2, \vec{v}_3$ as a largest independent subset, namely the pivot columns of A. **Beware**: The wrong answer is column 2, 3 of $\mathbf{rref}(A)$, because $\mathbf{rref}(A)$ columns are not in the original list of vectors! Example 5.24 is complete.

The Pivot Theorem can be restated as a method, called the **pivot method**, for finding a largest independent subset.

Theorem 5.42 (Pivot Method)
Let A be the augmented matrix of fixed vectors $\vec{v}_1$, ..., $\vec{v}_k$. Let the leading ones in $\mathbf{rref}(A)$ occur in columns i_1, ..., i_p. Then a largest independent subset of the k vectors $\vec{v}_1$, ..., $\vec{v}_k$ is the set of pivot columns of A, that is, the vectors

$$\vec{v}_{i_1}, \vec{v}_{i_2}, \ldots, \vec{v}_{i_p}.$$

Proof on page 424.

Rank and Nullity

Definition 5.32 (Rank of a Matrix)
The **rank** of an $m \times n$ matrix A, symbol $\mathbf{rank}(A)$, equals the number of leading ones in $\mathbf{rref}(A)$. Alternatively, the rank is the number of pivot columns of A.

Definition 5.33 (Nullity of a Matrix)
The **nullity** of an $m \times n$ matrix A is the number of free variables in the system $\mathbf{rref}(A)\vec{u} = \vec{0}$, or equivalently, the number of columns of A less the rank of A. The nullity equals the number of non-pivot columns of A in the Pivot Theorem.

The variable count in $\vec{u}$ equals the column dimension of A, which leads to the main result for rank and nullity.

Theorem 5.43 (Rank-Nullity Theorem)

$$\mathbf{rank}(A) + \mathbf{nullity}(A) = \text{column dimension of } A.$$

Proof on page 424.

In terms of homogeneous system $A\vec{u} = \vec{0}$, the rank of A is the number of leading variables and the nullity of A is the number of free variables, reliably computed from the system $\mathbf{rref}(A)\vec{x} = \vec{0}$.

Theorem 5.44 (Basis for $Ax = 0$)
Assume

$$k = \mathbf{nullity}(A) = \dim\left\{\vec{x} : A\vec{x} = \vec{0}\right\} > 0.$$

Then the solution set of $A\vec{x} = \vec{0}$ can be expressed as

$$(1) \qquad \vec{x} = t_1\vec{X}_1 + \cdots + t_k\vec{X}_k$$

where $\vec{X}_1$, ..., $\vec{X}_k$ are special linearly independent solutions of $A\vec{x} = \vec{0}$ and t_1, ..., t_k are arbitrary scalars (free variable invented symbols).
Proof on page 424.

Theorem 5.45 (Row Rank Equals Column Rank)
The number of independent rows of a matrix A equals the number of independent columns of A. Equivalently, $\mathbf{rank}(A) = \mathbf{rank}(A^T)$.
Proof on page 424.

Nullspace, Column Space and Row Space

Definition 5.34 (Kernel and Nullspace)
The **kernel** or **nullspace** of an $m \times n$ matrix A is the vector space of all solutions $\vec{x}$ to the homogeneous system $A\vec{x} = \vec{0}$. In symbols,

$$\mathbf{kernel}(A) \quad = \quad \mathbf{nullspace}(A) \quad = \quad \{\vec{x} \; : \; A\vec{x} = \vec{0}\}.$$

Definition 5.35 (Column Space)
The **column space** of $m \times n$ matrix A is the vector space consisting of all vectors $\vec{y} = A\vec{x}$, where $\vec{x}$ is arbitrary in $\mathcal{R}^n$.

In literature, the column space is also called the **image** of A, or the **range** of A, or the span of the columns of A. Because $A\vec{x}$ can be written as a linear combination of the columns $\vec{v}_1, \ldots, \vec{v}_n$ of A, the column space is the set of all linear combinations

$$\vec{y} = x_1\vec{v}_1 + \cdots + x_n\vec{v}_n.$$

In symbols,

$$\begin{aligned} \mathbf{colspace}(A) &= \{\vec{y} \; : \; \vec{y} = A\vec{x} \text{ for some } \vec{x}\} \\ &= \mathbf{Image}(A) \\ &= \mathbf{Range}(A) \\ &= \mathbf{span}(\vec{v}_1, \ldots, \vec{v}_n). \end{aligned}$$

Definition 5.36 (Row Space)
The **row space** of $m \times n$ matrix A is the vector space consisting of vectors $\vec{w} = A^T\vec{y}$, where $\vec{y}$ is arbitrary in $\mathcal{R}^m$. Technically, the row space of A is the column space of A^T. This vector space is viewed as the set of all linear combinations of rows of A. In symbols,

$$\begin{aligned} \mathbf{rowspace}(A) &= \mathbf{colspace}\left(A^T\right) \\ &= \{\vec{w} \; : \; \vec{w} = A^T\vec{y} \text{ for some } \vec{y}\} \\ &= \mathbf{Image}\left(A^T\right) \\ &= \mathbf{Range}\left(A^T\right). \end{aligned}$$

The row space of A and the null space of A live in $\mathcal{R}^n$, but the column space of A lives in $\mathcal{R}^m$, a different dimension. The correct bases are obtained as follows. If an alternative basis for **rowspace**(A) is suitable (rows of A not reported), then bases for **rowspace**(A), **colspace**(A), **nullspace**(A) can all be found by calculating just **rref**(A).

Null Space. Compute **rref**(A). Write out the general solution $\vec{x}$ to $A\vec{x} = \vec{0}$, where the free variables are assigned invented symbols $t_1, \ldots, t_k$. Report the basis for **nullspace**(A) as the list of partial derivatives $\partial_{t_1}\vec{x}, \ldots, \partial_{t_k}\vec{x}$, which are **special solutions** of $A\vec{x} = \vec{0}$.

Column Space. Compute $\mathbf{rref}(A)$. Identify the lead variable columns $i_1, \ldots, i_k$. Report the basis for $\mathbf{colspace}(A)$ as the list of columns $i_1, \ldots, i_k$ of A. These are the **pivot columns** of A.

Row Space. Compute $\mathbf{rref}\left(A^T\right)$. Identify the lead variable columns $i_1, \ldots, i_k$. Report the basis for $\mathbf{rowspace}(A)$ as the list of rows $i_1, \ldots, i_k$ of A.

Alternatively, compute $\mathbf{rref}(A)$, then $\mathbf{rowspace}(A)$ has a basis consisting of the list of nonzero rows of $\mathbf{rref}(A)$. The two bases obtained by these methods are different, but equivalent.

Due to the identity $\mathbf{nullity}(A) + \mathbf{rank}(A) = n$, where n is the column dimension of A, the following results hold. Notation: $\dim(V)$ is the dimension of vector space V, which equals the number of elements in a basis for V. Subspaces $\mathbf{nullspace}(A) = \mathbf{kernel}(A)$ and $\mathbf{colspace}(A) = \mathbf{Image}(A)$ have dual naming conventions in the literature.

Theorem 5.46 (Dimension Identities)

(a) $\dim(\mathbf{nullspace}(A)) = \dim(\mathbf{kernel}(A)) = \mathbf{nullity}(A)$

(b) $\dim(\mathbf{colspace}(A)) = \dim(\mathbf{Image}(A)) = \mathbf{rank}(A)$

(c) $\dim(\mathbf{rowspace}(A)) = \dim(\mathbf{Image}\left(A^T\right) = \mathbf{rank}(A)$

(d) $\dim(\mathbf{kernel}(A)) + \dim(\mathbf{Image}(A)) =$ column dimension of A

(e) $\dim(\mathbf{kernel}(A)) + \dim(\mathbf{kernel}\left(A^T\right)) =$ column dimension of A

Proof on page 425.

Equivalent Bases

Assume $\vec{v}_1, \ldots, \vec{v}_k$ are independent vectors in an abstract vector space V and let $S = \mathbf{span}(\vec{v}_1, \ldots, \vec{v}_n)$. Let $\vec{u}_1, \ldots, \vec{u}_\ell$ be another set of independent vectors in V.

Studied here is the question of whether or not $\vec{u}_1, \ldots, \vec{u}_\ell$ is a basis for S. First of all, all the vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ have to be in S, otherwise this set cannot possibly span S. Secondly, to be a basis, the vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ must be independent. Two bases for S must have the same number of elements, by Theorem 5.38. Therefore, $k = \ell$ is necessary for a possible second basis of S. These remarks establish:

Theorem 5.47 (Equivalent Bases of a Subspace S)

Let $\vec{v}_1, \ldots, \vec{v}_k$ be independent vectors in an abstract vector space V. Let S be the subspace of V consisting of all linear combinations of $\vec{v}_1, \ldots, \vec{v}_k$.

A set of vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ in V is an equivalent basis for S if and only

(1) Each of $\vec{u}_1, \ldots, \vec{u}_\ell$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$.

(2) The set $\vec{u}_1, \ldots, \vec{u}_\ell$ is independent.

(3) The sets are the same size, $k = \ell$.

Proof on page 425.

Equivalent Basis Test in $\mathcal{R}^n$

Assume given two sets of fixed vectors $\vec{v}_1, \ldots, \vec{v}_k$ and $\vec{u}_1, \ldots, \vec{u}_\ell$, in the same space $\mathcal{R}^n$. A test is developed for equivalence of bases, in a form suited for use in computer algebra systems and numerical laboratories.

Theorem 5.48 (Equivalence Test for Bases in $\mathcal{R}^n$)
Define augmented matrices

$$B = \left\langle \vec{v}_1 | \cdots | \vec{v}_k \right\rangle, \quad C = \left\langle \vec{u}_1 | \cdots | \vec{u}_\ell \right\rangle, \quad W = \left\langle B | C \right\rangle.$$

The relation

$$k = \ell = \mathbf{rank}(B) = \mathbf{rank}(C) = \mathbf{rank}(W)$$

implies

1. $\vec{v}_1, \ldots, \vec{v}_k$ is an independent set.
2. $\vec{u}_1, \ldots, \vec{u}_\ell$ is an independent set.
3. $\mathbf{span}\{\vec{v}_1, \ldots, \vec{v}_k\} = \mathbf{span}\{\vec{u}_1, \ldots, \vec{u}_\ell\}$

In particular, $\mathbf{colspace}(B) = \mathbf{colspace}(C)$ and each set of vectors is an equivalent basis for this vector space.

Proof on page 426.

Examples

Example 5.25 (Basis and Dimension)
Let S be the solution space in $V = \mathcal{R}^4$ of the system of equations

$$\begin{array}{rcrcrcl} x_1 & + & 2x_2 & & & = & 0, \\ 2x_1 & + & 5x_2 & & & = & 0, \\ & & & & x_4 & = & 0, \\ & & & & 0 & = & 0. \end{array} \tag{2}$$

Find a basis for S, then report the dimension of S.

Solution: The solution divides into two distinct sections: $\boxed{\mathbf{1}}$ and $\boxed{\mathbf{2}}$.

$\boxed{\mathbf{1}}$: Find the scalar general solution of system (2).

The toolkit: matrix combination, swap and multiply on the coefficient matrix. The last frame algorithm finds the general solution. The details:

$$\begin{pmatrix} 1&2&0&0\\ 2&5&0&0\\ 0&0&0&1\\ 0&0&0&0 \end{pmatrix}$$

First frame.

$$\begin{pmatrix} 1&2&0&0\\ 0&1&0&0\\ 0&0&0&1\\ 0&0&0&0 \end{pmatrix}$$

`combo(1,2,-2)`.

$$\begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&0&1\\ 0&0&0&0 \end{pmatrix}$$

`combo(2,1,-2)`. Last frame, this is the **rref**.

$$\left|\begin{array}{rcl} x_1 &=& 0,\\ x_2 &=& 0,\\ x_4 &=& 0,\\ 0 &=& 0. \end{array}\right|$$

Translate to scalar equations.

$$\left|\begin{array}{rcl} x_1 &=& 0,\\ x_2 &=& 0,\\ x_3 &=& t_1,\\ x_4 &=& 0. \end{array}\right|$$

Scalar general solution, obtained from the last frame algorithm: x_1, x_2, x_4=lead, x_3=free.

$\boxed{\mathbf{2}}$: Find the vector general solution of the system (2).

The plan is to use the answer from $\boxed{\mathbf{1}}$ and partial differentiation to display the vector general solution $\vec{x}$.

$$\partial_{t_1}\vec{x} = \begin{pmatrix} 0\\0\\1\\0 \end{pmatrix}$$

The **special solution** is the partial on symbol t_1. Only one, because there is only one invented symbol.

$$\vec{x} = t_1 \begin{pmatrix} 0\\0\\1\\0 \end{pmatrix}$$

The vector general solution.

Therefore, solution space $S = \mathbf{span}(\vec{X}_1)$, where $\vec{X}_1$ is the special solution obtained above. Because the spanning set is independent with one element, then $\dim(S) = 1$.

Example 5.26 (Euclidean Spaces)

Let A be an $m \times n$ matrix with columns $\vec{v}_1, \ldots, \vec{v}_n$ and let $\vec{b}$ be a vector in $\mathcal{R}^m$. Write a mathematical proof for each of the following facts.

$\boxed{\mathbf{1}}$. If the equation $A\vec{x} = \vec{b}$ has a solution $\vec{x}$, then $\vec{b}$ belongs to the span of the columns of A.

$\boxed{\mathbf{2}}$. If $\vec{b}$ belongs to the span of the columns of A, then the equation $A\vec{x} = \vec{b}$ has a solution $\vec{x}$.

$\boxed{\mathbf{3}}$. If $A\vec{x} = \vec{b}$ has a solution $\vec{x}$, then $\vec{b}, \vec{v}_1, \ldots, \vec{v}_n$ is a dependent set.

Solution:
$\boxed{\mathbf{1}}$: Let equation $A\vec{x} = \vec{b}$ have a solution $\vec{x}$. Write the equation backwards, then express the matrix product as a linear combination of the columns of A:

$$\vec{b} = A\vec{x} = x_1\vec{v}_1 + \cdots + x_n\vec{v}_n.$$

This proves $\vec{b}$ is a linear combination of the columns of A.

$\boxed{\mathbf{1}}$: Let $\vec{b}$ be a linear combination of the columns of A. We show $A\vec{x} = \vec{b}$ has a solution $\vec{x}$. By hypothesis, there are constants $x_1, \ldots, x_n$ such that

$$\vec{b} = x_1\vec{v}_1 + \cdots + x_n\vec{v}_n.$$

Let $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Because $A\vec{x}$ can be written as a linear combination of the columns of A, then $A\vec{x} = x_1\vec{v}_1 + \cdots + x_n\vec{v}_n = \vec{b}$, which proves that $A\vec{x} = \vec{b}$ has a solution $\vec{x}$.

$\boxed{\mathbf{3}}$: Assume $A\vec{x} = \vec{b}$ has a solution $\vec{x}$. Write the equation backwards, then express the matrix product as a linear combination of the columns of A:

$$\vec{b} = A\vec{x} = x_1\vec{v}_1 + \cdots + x_n\vec{v}_n.$$

Define $c_0 = -1$, $c_1 = x_1, \ldots, c_n = x_n$. Then

$$c_0\vec{b} + c_1\vec{v}_1 + \cdots + c_n\vec{v}_n = \vec{0}.$$

The definition of dependence implies that vectors $\vec{b}, \vec{v}_1, \ldots, \vec{v}_n$ are dependent.

The details for $\boxed{\mathbf{1}}$, $\boxed{\mathbf{2}}$, $\boxed{\mathbf{3}}$ are complete.

Example 5.27 (Sequence Spaces)

Let V be the vector space of all real sequences $\{x_n\}_{n=1}^{\infty}$ with componentwise addition and scalar multiplication. Let S be the subset of V defined by the equation $x_1 = 0$. Show that S is an infinite-dimensional subspace of V.

Solution: The space V is the abstraction of addition and scalar multiplication of Taylor series

$$f(t) = \sum_{n=1}^{\infty} x_n t^{n-1}.$$

The subspace S corresponds to all Taylor series which satisfy $f(0) = 0$. We assume it is known, or easily verified, that the larger set V is a vector space.

The **subspace criterion** applies to prove that S is a subspace of V. The omitted details are constructed from a similar set of details for the $\mathcal{R}^3$ subspace defined by the linear algebraic restriction equation $x_1 = 0$.

The remainder of the proof establishes $\dim(S) = \infty$. These details produce a list L with $\mathbf{span}(L) \subset S$. It is false that $S = \mathbf{span}(L)$, even though $\mathbf{span}(L)$ is a subspace by the **span theorem**. Further details are delayed to after $\dim(S) = \infty$ is established.

A standard method to find a basis L for S computes the partial derivatives on the symbols used to define S. The symbols are $x_2, x_3, \ldots$. We abuse notation and think of the sequences as column vectors with infinitely many components:

$$\{x_n\}_{n=1}^{\infty} \longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix}.$$

Then S is the subset of all infinitely long column vectors with $x_1 = 0$. Take partial derivatives on $x_2, x_3, \ldots$ to obtain the candidate basis vectors:

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}, \ldots$$

The list is infinite. Any finite subset of this list is independent. The intuition:

$$c_1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} + c_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

implies

$$\begin{pmatrix} 0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

and therefore $c_1 = c_2 = c_3 = 0$.

The list L of infinite sequences is correspondingly

$$\begin{array}{c} 0, 1, 0, 0, 0, \ldots \\ 0, 0, 1, 0, 0, \ldots \\ 0, 0, 0, 1, 0, \ldots \\ \vdots \end{array}$$

and there are infinitely many.

The details are finished by the method of contradiction. Suppose a true hypothesis and false conclusion. Then S has finite dimension n. Let Z be the span of a list L_1 of $n+1$ vectors from the above list. A proof can be constructed, based upon the above ideas, for independence of L_1, and then $\dim(Z) = n+1$. Because Z a subset of S, then $\dim(Z) \le \dim(S) = n$, a contradiction to $\dim(Z) = n+1$. Therefore, S cannot have finite dimension.

Conclusion: S is an infinite dimensional subspace of V.

Complaints. The preceding details do not prove Z is an independent set. The notation with infinitely many components is certainly not standard notation, therefore the reader is advised not to use it to present proof details. But it is excellent for intuition, and that is why you see it presented here, instead of more abstract details.

Is L a basis of S? The answer is **NO**.
Subspace $W = \mathbf{span}(L)$ is contained in subspace S. Then L is a basis for W. But L is not a basis for S. For example, the Taylor series for $f(t) = e^t - 1$ corresponds to a sequence in S with $x_k > 0$ for $k > 1$, and this sequence cannot be written as a finite linear combination of vectors selected from L.

Example 5.28 (Polynomial Spaces)

Let V be the vector space of all polynomials $p(x)$. Find a basis and hence the dimension of the subspace S defined by these conditions:

1. Polynomial $p(x)$ has degree no larger than two.

2. The equation $p(0) = \int_0^1 xp(x)dx$ is satisfied.

Solution: The answer is a list of independent polynomials in S: $\frac{2}{3} + x$, $\frac{1}{2} + x^2$. Then $\dim(S) = 2 =$ number of basis elements.

Details. Start by inventing symbols for the coefficients of $p(x)$, for example $p(x) = a_1 + a_2x + a_3x^2$ because of requirement **1**. Insert the $p(x)$ expression into requirement **2**, in order to find a relation for the three symbols a_1, a_2, a_3.

$p(0) = \int_0^1 xp(x)dx$ — Requirement **2**.

$a_1 = \int_0^1 (a_1x + a_2x^2 + a_3x^3)dx$ — Insert for $p(x)$ the expression $a_1 + a_2x + a_3x^2$. Then evaluate $p(0) = a_1$.

$a_1 = \dfrac{a_1}{2} + \dfrac{a_2}{3} + \dfrac{a_3}{4}$ — Evaluate integral.

Rearrangement of the last equation gives the linear equation $a_1 - \dfrac{2}{3}a_2 - \dfrac{1}{2}a_3 = 0$ in unknowns a_1, a_2, a_3. This linear system is in reduced echelon form. It has general solution

$$\begin{aligned} a_1 &= \tfrac{2}{3}t_1 + \tfrac{1}{2}t_2, \\ a_2 &= t_1, \\ a_3 &= t_2, \end{aligned} \tag{3}$$

with basis of solutions

$$\vec{v}_1 = \begin{pmatrix} \frac{2}{3} \\ 1 \\ 0 \end{pmatrix}, \quad \vec{v}_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ 1 \end{pmatrix}.$$

Translation to the corresponding polynomials, via the correspondence

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \longrightarrow p(x) = a_1 + a_2x + a_3x^2$$

gives the two polynomials

$$p_1(x) = \frac{2}{3} + x, \quad p_2(x) = \frac{1}{2} + x^2.$$

Why are these polynomials a basis for S?

A sophisticated answer is that the correspondence used to find the two polynomials is a one-to-one linear map from $W = \mathbf{span}(\vec{v}_1, \vec{v}_2)$ onto S, mapping $\vec{v}_1 \to p_1$ and $\vec{v}_2 \to p_2$.

A computational method will justify independence and span for the polynomials p_1, p_2. Start with $p(x) = a_1 + a_2x + a_3x^2$ in S. Equation $a_1 = \frac{2}{3}a_2 + \frac{1}{2}a_3$ holds because p belongs to S. Define $t_1 = a_2$, $t_2 = a_3$ (idea from equation (3)). Then all three equations in (3) are satisfied. Expand:

$$\begin{aligned} t_1p_1 + t_2p_2 &= a_2\left(\frac{2}{3} + x\right) + a_3\left(\frac{1}{2} + x^2\right) \\ &= \frac{2a_2}{3} + \frac{a_3}{2} + a_2x + a_3x^2 \\ &= a_1 + a_2x + a_3x^2. \end{aligned}$$

This computation proves that each polynomial in S is also in $\mathbf{span}(p_1, p_2)$, written as $S \subset \mathbf{span}(p_1, p_2)$. Because p_1, p_2 are already in S, then $\mathbf{span}(p_1, p_2) \subset S$. Then $S = \mathbf{span}(p_1, p_2)$, proving list p_1, p_2 spans S. The two polynomials p_1, p_2 are independent, because one is not a scalar multiple of the other. Then p_1, p_2 is independent and spans, which proves that p_1, p_2 is a basis for S and $\dim(S) = 2$.

Example 5.29 (Differential Equations)

A given homogeneous 5th order linear differential equation has general solution $y(x) = c_1 + c_2x + c_3x^2 + c_4\cos x + c_5\sin x$. Find a basis for the solution space S, a subspace of vector space $V = C^5(-\infty, \infty)$.

Solution: The answer is a list of 5 independent solutions in S: 1, x, x^2, $\cos x$, $\sin x$.

Details. The general solution expression implies S is the span of the reported list. We explain how to find the list. In the case of linear algebraic equations, we would take partial derivatives on the invented symbols to determine the list of special solutions, which is the basis. Here, we imagine c_1 to c_5 to be the invented symbols and take partial derivatives to determine a list of special vectors which span S. Let y abbreviate $y(x) = c_1 + c_2x + c_3x^2 + c_4\cos x + c_5\sin x$.

$$\partial_{c_1}y = 1, \quad \partial_{c_2}y = x, \quad \partial_{c_3}y = x^2, \quad \partial_{c_4}y = \cos x, \quad \partial_{c_5}y = \sin x.$$

The five answers are Euler solution atoms (defined on page 386). They are independent by Theorem 5.36, page 386. The general solution expression implies are solutions and they span S. They are a basis for S, dimension five.

Alternative Independence Test. The Wronskian test applies with sample $x = 0$. The Wronskian matrix is formed by rows which are successive derivatives of the list in row 1:

$$W(x) = \begin{pmatrix} 1 & x & x^2 & \cos x & \sin x \\ 0 & 1 & 2x & -\sin x & \cos x \\ 0 & 0 & 2 & -\cos x & -\sin x \\ 0 & 0 & 0 & \sin x & -\cos x \\ 0 & 0 & 0 & \cos x & \sin x \end{pmatrix}.$$

The determinant of $W(x)$ for sample $x = 0$ is $|W(0)| = 2$. The Wronskian test page 385 implies the list in row 1 of $W(x)$ is independent.

Example 5.30 (Largest Independent Subset)

Let $V = C(-\infty, \infty)$ and consider this list of vectors in V:

$$1,\ x + x^2,\ 2 + x,\ 1 + x^2,\ e^x,\ x + e^x.$$

Find a largest independent subset of this list.

Solution: One answer of the many possible answers is the list

$$1, x + x^2, 2 + x, e^x.$$

Details. Start with the nonzero vectors $1, x + x^2$. They are not scalar multiples of each other, hence they are independent. The initial independent subset is $1, x + x^2$. Vector $1 + x$ cannot be expressed as a combination of 1 and $x + x^2$, because such a relation

$$2 + x = c_1(1) + c_2(x + x^2)$$

requires both c_1 and c_2 nonzero, in which case we reach the impossibility that a linear polynomial equals a quadratic polynomial. The vector is added to the list to extend the initial independent subset to $1, x + x^2, 2 + x$. The different growth rate at $x = \infty$ of exponential term e^x explains why the independent subset is extended to $1, x+x^2, 2+x, e^x$.

Why is $x + e^x$ eliminated from the list? First, assemble two facts:

1. Vector x belongs to $\mathbf{span}(1, x + x^2, 2 + x)$.
2. Vectors x and e^x belong to $\mathbf{span}(1, x + x^2, 2 + x, e^x)$.

Then $x + e^x$ is in the span of the preceding vectors in the independent subset $1, x + x^2, 1 + x, e^x$. The final independent subset has been found.

Remark on the method. The pivot theorem does not directly apply to this example, because the vector space V is not a space $\mathcal{R}^n$ of fixed vectors. The pivot theorem can be used by reducing the original problem to an equivalent problem in some $\mathcal{R}^n$. This method is explored later, keyword **isomorphism**.

Example 5.31 (Pivot Theorem Method)

Extract a largest independent subset from the columns of the matrix

$$A = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Solution: The answer is columns 2,3.

Details. The quickest solution is to observe that column 5 equals column 3 minus column 2, but columns 2,3 are nonzero and not scalar multiples of one another, therefore they are independent. Zero columns do not add to an independent subset of columns, therefore a largest independent subset of columns is obtained from columns 2,3.

A solution with computer implementation computes the pivot columns of A to be columns 2,3, and then we report a largest independent set of columns of A to be the pivot columns $2, 3$.

The pivot columns of A are computed from the $\mathbf{rref}(A)$, which is found on paper using the toolkit *combo, swap, multiply.* It is a one-step process with computer assist: enter the matrix A and then write a command line for $\mathbf{rref}(A)$. The answer:

$$\mathbf{rref}(A) = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then the pivot columns are columns $2, 3$ of matrix A. This is a largest independent subset of the columns of A.

Sample Code, Computer Algebra System `maple`:

```
A:=Matrix([[0,1,2,0,1],[0,1,1,0,0],[0,2,1,0,-1],
           [0,0,1,0,1],[0,0,1,0,1]]);
LinearAlgebra[ReducedRowEchelonForm](A);
```

Example 5.32 (Nullspace, Row Space, Column Space)

Compute the nullspace, column space and row space of the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

Solution: The answers appear below.

Details. The first computation is $\mathbf{rref}(A)$, which provides one answer for each of the three subspaces. The steps:

$\begin{pmatrix} 0&1&0&1 \\ 1&1&1&0 \\ 0&0&0&0 \\ 1&1&1&0 \end{pmatrix}$ Given matrix A.

$\begin{pmatrix} 1&1&1&0 \\ 0&1&0&1 \\ 0&0&0&0 \\ 1&1&1&0 \end{pmatrix}$ `swap(1,2)`

$\begin{pmatrix} 1&1&1&0 \\ 0&1&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}$ `combo(1,4,-1)`

$\begin{pmatrix} 1&0&1&-1 \\ 0&1&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}$ Begin back-substitution: `combo(2,1,-1)`. Found $\mathbf{rref}(A)$.

The last frame algorithm is applied to find the general solution of $A\vec{x} = \vec{0}$, using the scalar form of the last frame:

$$\begin{array}{rcrcrcrcl} x_1 & & & + & x_3 & - & x_4 & = & 0, \\ & & x_2 & & & + & x_4 & = & 0, \\ & & & & & & 0 & = & 0, \\ & & & & & & 0 & = & 0. \end{array}$$

The lead variables are x_1, x_2 and the free variables x_3, x_4. Using invented symbols t_1, t_2 gives the general solution

$$\begin{aligned} x_1 &= -t_1 + t_2, \\ x_2 &= -t_2, \\ x_3 &= t_1, \\ x_4 &= t_2. \end{aligned}$$

Nullspace. The partial derivatives on the invented symbols, the *special solutions*, form a basis for the nullspace of A:

$$\textbf{nullspace}(A) = \textbf{kernel}(A) = \textbf{span}\left(\begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}\right).$$

Column Space. The column space of A is the span of the pivot columns of A, which according to the computed **rref** are columns 1, 2 of A. Then

$$\textbf{colspace}(A) = \textbf{span}(\text{pivot columns of } A) = \textbf{span}\left(\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}\right).$$

Row space. One answer is the set of nonzero rows of **rref**(A). This gives the first answer

$$\textbf{rowspace}(A) = \textbf{span}\left(\begin{pmatrix} 1 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}\right).$$

A second answer is the set of pivot columns of A^T, columns 1,2 of A^T, found from

$$A^T = \begin{pmatrix} 0\,1\,0\,1 \\ 1\,1\,0\,1 \\ 0\,1\,0\,1 \\ 1\,0\,0\,0 \end{pmatrix}, \quad \textbf{rref}(A^T) = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{pmatrix}.$$

Then the second answer for the row space is

$$\textbf{rowspace}(A) = \textbf{span}\left(\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}\right).$$

Example 5.33 (Fundamental Subspaces)

Compute the nullspace and column space for both A and A^T, given

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

The 4 computed subspaces are known as *Gilbert Strang's Four Fundamental Subspaces.*

Solution: Let $N_1 = \textbf{nullspace}(A) = \textbf{span}(\text{Strang's special solutions for } A)$ and $C_1 = \textbf{colspace}(A) = \textbf{span}(\text{pivot columns of } A)$. Both N_1 and C_1 were computed in the previous example:

$$N_1 = \textbf{span}\left(\begin{pmatrix}-1\\0\\1\\0\end{pmatrix}, \begin{pmatrix}1\\-1\\0\\1\end{pmatrix}\right), \quad C_1 = \textbf{span}\left(\begin{pmatrix}0\\1\\0\\1\end{pmatrix}, \begin{pmatrix}1\\1\\0\\1\end{pmatrix}\right).$$

Define

$$\begin{aligned} N_2 &= \textbf{nullspace}(A^T) \\ &= \textbf{span}(\text{Strang's special solutions of } A^T), \\ C_2 &= \textbf{colspace}(A^T) \\ &= \textbf{span}(\text{pivot columns of } A^T). \end{aligned}$$

The computation of C_2 was completed in the previous example, which also computed

$$A^T = \begin{pmatrix}0\,1\,0\,1\\1\,1\,0\,1\\0\,1\,0\,1\\1\,0\,0\,0\end{pmatrix}, \quad \textbf{rref}(A^T) = \begin{pmatrix}1\,0\,0\,0\\0\,1\,0\,1\\0\,0\,0\,0\\0\,0\,0\,0\end{pmatrix}.$$

Then the general solution for $A^T\vec{x} = \vec{0}$ is

$$x_1 = 0, x_2 = -t_2, x_3 = t_1, x_4 = t_2, \quad \text{Strang's special solutions} = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\-1\\0\\1\end{pmatrix}.$$

The newly found answer for N_2 plus the transcribed answer for C_2, taken from the previous example, give the equations

$$N_2 = \textbf{span}\left(\begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\-1\\0\\1\end{pmatrix}\right), \quad C_2 = \textbf{span}\left(\begin{pmatrix}0\\1\\0\\1\end{pmatrix}, \begin{pmatrix}0\\1\\0\\0\end{pmatrix}\right).$$

Example 5.34 (Equivalent Bases)

Let

$$\vec{v}_1 = \begin{pmatrix}0\\1\\\frac{3}{2}\end{pmatrix}, \vec{v}_2 = \begin{pmatrix}1\\0\\-\frac{1}{2}\end{pmatrix}, \vec{u}_1 = \begin{pmatrix}1\\3\\4\end{pmatrix}, \vec{u}_2 = \begin{pmatrix}3\\1\\0\end{pmatrix}.$$

Verify that $\{\vec{v}_1, \vec{v}_2\}$ and $\{\vec{u}_1, \vec{u}_2\}$ are equivalent bases for a subspace S.

Solution:

Define $B = \begin{pmatrix}0 & 1\\1 & 0\\\frac{3}{2} & -\frac{1}{2}\end{pmatrix}$, $C = \begin{pmatrix}1\,3\\3\,1\\4\,0\end{pmatrix}$, $W = \begin{pmatrix}0 & 1\,1\,3\\1 & 0\,3\,1\\\frac{3}{2} & -\frac{1}{2}\,4\,0\end{pmatrix}$. Compute the **rank** of each matrix to be 2. Apply the theorem.

Maple Illustration.

```
v1:=<0,1,3/2>;v2:=<1,0,-1/2>; # Basis v1,v2
u1:=<1,3,4>;u2:=<3,1,0>;
B:=<v1|v2>; C:=<u1|u2>; W:=<B|C>;
# Test: ranks of B, C, W must equal 2
linalg[rank](B),linalg[rank](C),linalg[rank](W);
```

Example 5.35 (Equivalent Bases: False Test)
Does $\mathbf{rref}(B) = \mathbf{rref}(C)$ imply that each column of C is a linear combination of the columns of B? The answer is **no**. Supply a counter-example.

Solution: Define $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$, $C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Then $\mathbf{rref}(B) = \mathbf{rref}(C) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$, but column 2 of C is not a linear combination of the columns of B. This means $S_1 = \mathbf{colspace}(B)$ is not equal to $S_2 = \mathbf{colspace}(C)$. Geometrically, S_1 and S_2 are planes in $\mathcal{R}^3$ which intersect only along the line L through the two points $(0,0,0)$ and $(1,0,1)$.

What went wrong? The culprit is the toolkit operation `swap`.

Details and Proofs

Proof of Theorem 5.38, Size of a Basis: The proof proceeds by the formal method of contradiction. Assume the hypotheses are true and the conclusion is false. Then $p \ne q$. Without loss of generality, let the larger basis be listed first, $p > q$.

Because $\vec{u}_1, \ldots, \vec{u}_q$ is a basis of the vector space V, then there are coefficients $\{a_{ij}\}$ such that

$$\begin{array}{rcccccc} \vec{v}_1 &=& a_{11}\vec{u}_1 &+& \cdots &+& a_{1q}\vec{u}_q, \\ \vec{v}_2 &=& a_{21}\vec{u}_1 &+& \cdots &+& a_{2q}\vec{u}_q, \\ && \vdots &&&& \\ \vec{v}_p &=& a_{p1}\vec{u}_1 &+& \cdots &+& a_{pq}\vec{u}_q. \end{array}$$

Let $A = [a_{ij}]$ be the $p \times q$ matrix of coefficients. Because $p > q$, then $\mathbf{rref}(A^T)$ has at most q leading variables and at least $p - q > 0$ free variables.

Then the $q \times p$ homogeneous system $A^T\vec{x} = \vec{0}$ has infinitely many solutions. Let $\vec{x}$ be a nonzero solution of $A^T\vec{x} = \vec{0}$.

The equation $A^T\vec{x} = \vec{0}$ means $\sum_{i=1}^p a_{ij}x_i = 0$ for $1 \le j \le p$, giving the dependence relation

$$\begin{array}{rcl} \sum_{i=1}^p x_i\vec{v}_i &=& \sum_{i=1}^p x_i \sum_{j=1}^q a_{ij}\vec{u}_j \\ &=& \sum_{j=1}^q \sum_{i=1}^p a_{ij}x_i\vec{u}_j \\ &=& \sum_{j=1}^q (0)\vec{u}_j \\ &=& \vec{0} \end{array}$$

The independence of $\vec{v}_1, \ldots, \vec{v}_p$ is contradicted. Arrival of the contradiction implies $p = q$. ■

Proof of Theorem 5.39, Basis of a finite dimensional vector space:

$\boxed{1}$ Let $S = \textbf{span}(L)$, a subspace of V. By independence, $\dim(S) = n$. By hypothesis, $\dim(V) = n$. Suppose $\vec{v}$ is in V. Let L equal the list of $n+1$ elements $\vec{v}_1, \ldots, \vec{v}_n, \vec{v}$. Then L is contained in V. Space V has dimension n, which means that no independent subset exists of size larger than n. So list L is not an independent set, which implies that $\vec{v}$ is in $S = \textbf{span}(L)$. Therefore $S = V$ and L is a basis for V.

$\boxed{2}$ It suffices to prove under the given hypotheses that L is an independent set. If not, then $V = \textbf{span}(L)$ is spanned by less than n independent vectors. This implies the dimension of V is less than n. A contradiction is reached, therefore L is an independent set.

$\boxed{3}$ Choose any independent subset of L, call it $\vec{w}_1, \ldots, \vec{w}_q$. If $q = n$, then we are done, by $\boxed{1}$. Otherwise, the span of these q vectors is a subspace of V not equal to V. Choose a vector $\vec{v}_{q+1}$ not in the subspace. Then $\vec{w}_1, \ldots, \vec{w}_{q+1}$ is an independent set in V. Repeat the construction until the number of constructed vectors equals n. Then the constructed list is a basis for V.

$\boxed{4}$ Assume $S = \textbf{span}(L) = V$. If L contains fewer than n independent vectors, then V would have a basis of fewer than n elements, a violation of $\dim(V) = n$. Therefore, L contains n independent vectors. It cannot have more than n, without violating $\dim(V) = n$. Therefore, L contains exactly n independent vectors, which form a basis for V.

Proof of Theorem 5.40, Basis and Dimension in $\mathcal{R}^n$: The first result is due to the fact that all bases contain the same identical number of vectors. Because the columns of the $n \times n$ identity are independent and span $\mathcal{R}^n$, then all bases must contain n vectors, exactly.

A list of $n+1$ vectors $\vec{v}_1, \ldots, \vec{v}_{n+1}$ generates a subspace $S = \textbf{span}(\vec{v}_1, \ldots, \vec{v}_{n+1})$. Because S is contained in $\mathcal{R}^n$, then S has a basis of n elements or less. Therefore, the list of $n+1$ vectors is dependent. ■

Proof of Theorem 5.41, The Pivot Theorem:

$\boxed{1}$: To prove: the pivot columns of A are independent. Let $\vec{v}_1$, …, $\vec{v}_k$ be the vectors columns of matrix A. Let i_1, …, i_p be the pivot columns of A.

To apply the independence test, form the system of equations

$$c_1\vec{v}_{i_1} + \cdots + c_p\vec{v}_{i_p} = \vec{0}$$

and solve for the constants c_1, …, c_p, independence confirmed if they are all zero. The tool used to solve for the constants is the elementary matrix formula

$$A = M\,\textbf{rref}(A), \quad M = E_1E_2\cdots E_r,$$

where E_1, …, E_r denote certain elementary matrices. Each elementary matrix is the inverse of a swap, multiply or combination operation applied to A, in order to reduce A to $\textbf{rref}(A)$. Because elementary matrices are invertible, then M is invertible. The equation $A = \left\langle \vec{v}_1|\cdots|\vec{v}_k \right\rangle$ implies the pivot columns of A satisfy the equation

$$\vec{v}_{i_q} = M\vec{e}_q, \quad q = 1, \ldots, p,$$

where $\vec{e}_1 = \textbf{col}(I,1)$, …, $\vec{e}_p = \textbf{col}(I,p)$ are the consecutive columns of the identity matrix which occupy the columns of the leading ones in $\textbf{rref}(A)$. Then

$$\begin{aligned} \vec{0} &= c_1\vec{v}_{i_1} + \cdots + c_p\vec{v}_{i_p} \\ &= M(c_1\vec{e}_1 + \cdots + c_p\vec{e}_p) \end{aligned}$$

implies by invertibility of M that

$$c_1\vec{e}_1 + \cdots + c_p\vec{e}_p = \vec{0}.$$

Distinct columns of the identity matrix are independent (subsets of independent sets are independent), therefore $c_1 = \cdots = c_p = 0$. The independence of the pivot columns of A is established. The proof of $\boxed{\mathbf{1}}$ is complete.

$\boxed{\mathbf{2}}$: To prove: a non-pivot column of A is a linear combination of the pivot columns of A. Let column j of A be non-pivot. Let's express this column as a linear combination of the pivot columns of A.

Consider the homogeneous system $A\vec{x} = \vec{0}$ and its equivalent system $\mathbf{rref}(A)\vec{x} = \vec{0}$. The pivot column subscripts determine the leading variables and the remaining column subscripts determine the free variables. Then column j matches a free variable x_j. Define $x_j = 1$. Define all other free variables to be zero. The lead variables are now determined and the resulting nonzero vector $\vec{x}$ satisfies the homogeneous equation $\mathbf{rref}(A)\vec{x} = \vec{0}$, and hence also $A\vec{x} = \vec{0}$. Translating this equation into a linear combination of columns implies

$$\left(\sum_{\text{pivot subscripts } i} x_i\vec{v}_i \right) + \vec{v}_j = \vec{0}$$

which in turn implies that column j of A is a linear combination of the pivot columns of A. The proof of $\boxed{\mathbf{2}}$ is complete.

Proof of Theorem 5.42, The Pivot Method: According to the Pivot Theorem 5.41, the fixed vectors are independent. An attempt to add another column of A to these chosen columns results in a non-pivot column being added. The Pivot Theorem applies: the column added is dependent on the pivot columns. Therefore, the set of pivot columns of A forms a largest independent subset of the columns of A. ■

Proof of Theorem 5.43, The Rank-Nullity Theorem: The rank of A is the number of leading ones in $\mathbf{rref}(A)$. The nullity of A is the number of non-pivot columns in A. The sum of the rank and nullity is the number of variables, which is the column dimension n of A. Then the rank + nullity = n, as claimed. ■

Proof of Theorem 5.44, Basis for $A\vec{x} = \vec{0}$: The system $\mathbf{rref}(A)\vec{x} = \vec{0}$ has exactly the same solution set as $A\vec{x} = \vec{0}$. This system has a standard general solution $\vec{x}$ expressed in terms of invented symbols $t_1, \ldots, t_k$. Define $\vec{X}_j = \partial_{t_j}\vec{x}$, $j = 1, \ldots, k$. Then (1) holds. It remains to prove independence, which means we are to solve for $c_1, \ldots, c_k$ in the system

$$c_1\vec{X}_1 + \cdots + c_k\vec{X}_k = \vec{0}.$$

The left side is a solution $\vec{x}$ of $A\vec{x} = \vec{0}$ in which the invented symbols have been assigned values $c_1, \ldots, c_k$. The right side implies each component of $\vec{x}$ is zero. Because the standard general solution assigns invented symbols to free variables, the relation above implies that each free variable is zero. But free variables have already been assigned values $c_1, \ldots, c_k$. Therefore, $c_1 = \cdots = c_k = 0$. ■

Proof Theorem 5.45, Row Rank equals Column Rank: Let S be the set of all linear combinations of columns of A. Then $S = \mathbf{span}(\text{columns of } A) = \mathbf{Image}(A)$. The non-pivot columns of A are linear combinations of pivot columns of A. Therefore, any linear combination of columns of A is a linear combination of the $p = \mathbf{rank}(A)$ linearly independent pivot columns. By definition, the pivot columns form a **basis** for the vector space S, and $p = \mathbf{rank}(A) = \dim(S)$.

The **span** R **of the rows of** A is defined to be the set of all linear combinations of the columns of A^T.

Let $q = \mathbf{rank}(A^T) = \dim(R)$. It will be shown that $p = q$, which proves the theorem.

Let $\mathbf{rref}(A) = E_1 \cdots E_k A$ where $E_1, \ldots, E_k$ are elementary swap, multiply and combination matrices. The invertible matrix $M = E_1 \cdots E_k$ satisfies the equation $\mathbf{rref}(A) = MA$. Then:

$$\mathbf{rref}(A)^T = A^T M^T$$

Matrix $\mathbf{rref}(A)^T$ has its first p columns independent and its remaining columns are zero. Each nonzero column of $\mathbf{rref}(A)^T$ is expressed on the right as a linear combination of the columns of A^T. Therefore, R contains p independent vectors. The number $q = \dim(R)$ is the vector count in any basis for R. This implies $p \le q$.

The preceding display can be solved for A^T, because M^T is invertible, giving

$$A^T = \mathbf{rref}(A)^T (M^T)^{-1}.$$

Then every column of A^T is a linear combination of the p nonzero columns of $\mathbf{rref}(A)^T$. This implies a basis for R contains at most p elements, i.e., $q \le p$.

Combining $p \le q$ with $q \le p$ proves $p = q$. ■

Proof of Theorem 5.46, Dimension Identities:

(a) $\dim(\mathbf{nullspace}(A)) = \dim(\mathbf{kernel}(A)) = \mathbf{nullity}(A)$

The nullspace is the kernel, defined as the set of solutions to $A\vec{\mathbf{x}} = \vec{\mathbf{0}}$. This set has basis *Strang's Special Solutions*, the number of which matches the number of free variables. That number is the nullity of A.

(b) $\dim(\mathbf{colspace}(A)) = \dim(\mathbf{Image}(A)) = \mathbf{rank}(A)$

The column space has as a basis the pivot columns of A. The number of pivot columns is the rank of A.

(c) $\dim(\mathbf{rowspace}(A)) = \dim(\mathbf{Image}\left(A^T\right) = \mathbf{rank}(A)$

The row space has a basis given by the pivot columns of A^T. The number of columns is the number of independent rows of A, or the row rank of A, which by Theorem 5.45 equals the rank of A.

(d) $\dim(\mathbf{kernel}(A)) + \dim(\mathbf{Image}(A)) =$ column dimension of A

This identity restates the Rank-Nullity Theorem 5.43.

(e) $\dim(\mathbf{kernel}(A)) + \dim(\mathbf{kernel}\left(A^T\right)) =$ column dimension of A

Apply part (d) to A^T. If $\dim(\mathbf{kernel}(A)) = \dim(\mathbf{Image}(A^T))$ then identity (e) follows. Let $r = \dim(\mathbf{kernel}(A))$ and $s = \dim(\mathbf{Image}(A^T))$. We must show $r = s$. Already known is $r = \mathbf{nullity}(A)$, which equals the number of Strang's Special Solutions. Number s is the number of independent columns in A^T, which equals the row rank of A. Theorem 5.45 applies: s equals the row rank of A, which is the rank of A, which is r. Then $r = s$, as claimed. ■

Proof of Theorem 5.47, Equivalent Bases: Vectors $\vec{w}_1, \ldots, \vec{w}_k$ are a basis for S provided they are independent and span S. The three items from the theorem:

(1) Each of $\vec{u}_1, \ldots, \vec{u}_\ell$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_k$.

(2) The set $\vec{u}_1, \ldots, \vec{u}_\ell$ is independent.

(3) The sets are the same size, $k = \ell$.

Sufficiency. Assume given vectors $\vec{v}_1, \ldots, \vec{v}_k$ which form a basis for S. Assume vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ are also a basis for S. Then these vectors are independent and span S. The spanning condition $S = \mathbf{span}(\vec{u}_1, \ldots, \vec{u}_k)$ implies (1). Independence implies (2). Theorem 5.38 applies: the two bases have the same size: $k = \ell$, which proves (3) holds.

Necessity. Assume that vectors $\vec{v}_1, \ldots, \vec{v}_k$ form a basis for S. Assume given vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ in S satisfying (1), (2), (3). We prove $\vec{u}_1, \ldots, \vec{u}_\ell$ is a basis for S. Item (2) implies the vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ are independent and (1) implies they span S, because $\vec{v}_1, \ldots, \vec{v}_k$ span S. The definition of basis applies: vectors $\vec{u}_1, \ldots, \vec{u}_\ell$ form a basis for S. ■

Proof of Theorem 5.48, Equivalence test for bases in $\mathcal{R}^n$:
Because $\mathbf{rank}(B) = k$, then the first k columns of W are independent. If some column of C is independent of the columns of B, then W would have $k+1$ independent columns, which violates $k = \mathbf{rank}(W)$. Therefore, the columns of C are linear combinations of the columns of B. The vector space $\mathcal{U} = \mathbf{colspace}(C)$ is therefore a subspace of the vector space $\mathcal{V} = \mathbf{colspace}(B)$. Because each vector space has dimension k, then $\mathcal{U} = \mathcal{V}$. ■

Exercises 5.5

Basis and Dimension

Compute a basis and the report the dimension of the subspace S.

1. In $\mathcal{R}^3$, S is the solution space of

$$\left|\begin{array}{rcrcrcl} x_1 & & & + & x_3 & = & 0, \\ & & x_2 & + & x_3 & = & 0. \end{array}\right|$$

2. In $\mathcal{R}^4$, S is the solution space of

$$\left|\begin{array}{lrcl} x_1 + 2x_2 + x_3 & & = & 0, \\ & x_4 & = & 0. \end{array}\right|$$

3. In $\mathcal{R}^2$, $S = \mathbf{span}(\vec{v}_1, \vec{v}_2)$. Vectors $\vec{v}_1, \vec{v}_2$ are columns of an invertible matrix.

4. Set $S = \mathbf{span}(\vec{v}_1, \vec{v}_2)$, in $\mathcal{R}^4$. The vectors are columns in a 4×4 invertible matrix.

5. Set $S = \mathbf{span}(\sin^2 x, \cos^2 x, 1)$, in the vector space V of continuous functions.

6. Set $S = \mathbf{span}(x, x-1, x+2)$, in the vector space V of all polynomials.

7. Set $S = \mathbf{span}(\sin x, \cos x)$, the solution space of $y'' + y = 0$.

8. Set $S = \mathbf{span}\left(e^{2x}, e^{3x}\right)$, the solution space of $y'' - 5y' + 6y = 0$.

Euclidean Spaces

9. Let A be 3×2. Why is it impossible for the columns of A to be a basis for $\mathcal{R}^3$?

10. Let A be $m \times n$. What condition on indices m, n implies it is impossible for the columns of A to be a basis for $\mathcal{R}^m$?

11. Find a pairwise orthogonal basis for $\mathcal{R}^3$ which contains $\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$.

12. Display a basis for $\mathcal{R}^4$ which contains the independent columns of $\begin{pmatrix} 0 & 1 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$.

13. Let S be a subspace of $\mathcal{R}^{10}$ of dimension 5. Insert a basis for S into an $m \times n$ augmented matrix A. What are m and n?

14. Suppose A and B are 3×3 matrices and let $C = AB$. Assume the columns of A are not a basis for $\mathcal{R}^3$. Is there a matrix B so that the columns of C form a basis for $\mathcal{R}^3$?

15. The term **Hyperplane** is used for an equation like $x_4 = 0$, which in $\mathcal{R}^4$ defines a subspace S of dimension 3. Find a basis for S.

16. Find a 3-dimensional subspace S of $\mathcal{R}^4$ which has no basis consisting of columns of the identity matrix.

Polynomial Spaces

Symbol V is the vector space of all polynomials $p(x)$. Given subspace S of V, find a basis for S and $\dim(S)$.

17. The subset S of $\mathbf{span}(1, x, x^2)$ is defined by $\frac{dp}{dx}(1) = 0$.

18. The subset S of $\mathbf{span}(1, x, x^2, x^3)$ is defined by $p(0) = \frac{dp}{dx}(1) = 0$.

19. The subset S of $\mathbf{span}(1, x, x^2)$ is defined by $\int_0^1 p(x)dx = 0$.

20. The subset S of $\mathbf{span}(1, x, x^2, x^3)$ is defined by $\int_0^1 xp(x)dx = 0$.

Differential Equations

Find a basis for solution subspace S. Assume the general solution of the 4th order linear differential equation is

$$y(x) = c_1 + c_2 x + c_3 e^x + c_4 e^{-x}.$$

21. Subspace S_1 is defined by $y(0) = \frac{dy}{dx}(0) = 0$.

22. Subspace S_2 is defined by $y(1) = 0$.

23. Subspace S_3 is defined by $y(0) = \int_0^1 y(x)dx$.

24. Subspace S_4 is defined by $y(1) = 0, \int_0^1 y(x)dx = 0$.

Largest Subset of Independent Vectors

Find a largest independent subset of the given vectors.

25. The columns of $\begin{pmatrix} 0&0&1&1\\ 0&0&1&1\\ 0&1&1&0\\ 0&1&2&1 \end{pmatrix}$.

26. The columns of $\begin{pmatrix} 3&1&2&0&5\\ 2&1&1&0&4\\ 3&2&1&0&7\\ 1&0&1&0&1\\ 3&2&1&0&7 \end{pmatrix}$.

27. The polynomials $x, 1+x, 1-x, x^2$.

28. The continuous functions x, e^x, $x+e^x$, e^{2x}.

Pivot Theorem Method

Extract a largest independent set from the columns of the given matrix A. The answer is a list of independent columns of A, called the pivot columns of A.

29. $\begin{pmatrix} 1&2&1\\ 1&1&0\\ 2&1&0 \end{pmatrix}$

30. $\begin{pmatrix} 0&1&2&1\\ 0&1&1&0\\ 0&2&1&0\\ 0&0&1&1 \end{pmatrix}$

31. $\begin{pmatrix} 0&2&1&0&1\\ 1&5&2&0&3\\ 1&3&1&0&2\\ 0&2&1&0&3\\ 0&2&1&0&1 \end{pmatrix}$

32. $\begin{pmatrix} 0&0&2&1&0&1\\ 0&1&5&2&0&3\\ 0&1&3&1&0&2\\ 0&2&4&1&0&3\\ 0&0&2&1&0&1\\ 0&2&4&1&0&3 \end{pmatrix}$

Row and Column Rank

Justify by direct computation that $\mathbf{rank}(A) = \mathbf{rank}\left(A^T\right)$, which means that the row rank equals the column rank.

33. $A = \begin{pmatrix} 1&0&1\\ 0&1&1\\ 0&0&0 \end{pmatrix}$

34. $A = \begin{pmatrix} 1&0&1&0\\ 0&1&1&1\\ 0&0&0&0\\ 0&1&1&1 \end{pmatrix}$

Nullspace or Kernel

Find a basis for the nullspace of A, which is also called the kernel of A.

35. $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

36. $A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}$

Row Space

Find a basis for the row space of A. There are two possible answers: (1) The nonzero rows of **rref**(A), (2) The pivot columns of A^T. Answers (1) and (2) can differ wildly.

37. $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

38. $A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}$

Column Space

Find a basis for the column space of A, in terms of the columns of A. Normally, we report the pivot columns of A.

39. $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

40. $A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}$

Dimension Identities

Let A be an $m \times n$ matrix of rank r. Prove the following dimension identities in Theorem 5.46.

41. $\dim(\textbf{nullspace}(A)) = n - r$

42. $\dim(\textbf{colspace}(A)) = r$

43. $\dim(\textbf{rowspace}(A)) = r$

44. The dimensions of **nullspace**(A) and **colspace**(A) add to n.

Orthogonal Complement $S^\perp$

Let S be a subspace of vector space $V = \mathcal{R}^n$. Define the **Orthogonal complement** by

$$(4) \qquad S^\perp = \{\vec{x} \ : \ \vec{x}^T\vec{y} = 0, \ \vec{y} \text{ in } S\}.$$

45. Let $V = \mathcal{R}^3$ and let S be the xy-plane. Compute $S^\perp$. Answer: The z-axis.

46. Prove that $S^\perp$ is a subspace, using the **Subspace Criterion**.

47. Prove that the orthogonal complement of $S^\perp$ is S. In symbols, $\left(S^\perp\right)^\perp = S$.

48. Prove that

$$V = \{\vec{x} + \vec{y} \ : \ \vec{x} \in S, \vec{y} \in S^\perp\}.$$

This relation is called the **Direct Sum** of S and $S^\perp$.

Fundamental Theorem of Linear Algebra

Let A be an $m \times n$ matrix.

49. Write a short proof:
Lemma. Any solution of $A\vec{x} = \vec{0}$ is orthogonal to every row of A.

50. Find the dimension of the kernel and image for both A and A^T. The four answers use symbols $m, n, \textbf{rank}(A)$. The main tool is the rank-nullity theorem.

51. Prove
kernel$(A) =$ **Image** $\left(A^T\right)^\perp$. Use Exercise 49.

52. Prove
kernel $\left(A^T\right) =$ **Image** $(A)^\perp$.

Fundamental Subspaces

The kernel and image of both A and A^T are called *The Four Fundamental Subspaces* by Gilbert Strang. Let A denote an $n \times m$ matrix.

53. Prove using Exercise 51:
kernel$(A) =$ **rowspace**$(A)^\perp$

54. Establish these four identities.
kernel$(A) =$ **Image** $\left(A^T\right)^\perp$
kernel $\left(A^T\right) =$ **Image** $(A)^\perp$
Image $(A) =$ **kernel**$(A^T)^\perp$
Image $\left(A^T\right) =$ **kernel**$(A)^\perp$

Notation. *kernel* is null space, *image* is column space, symbol $\perp$ is orthogonal complement: see equation (4).

Equivalent Bases

Test the given subspaces for equality.

55. $S_1 = \textbf{span}\left(\begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}\right)$,

$S_2 = \textbf{span}\left(\begin{pmatrix}3\\3\\-1\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}\right)$

56. $S_3 = \textbf{span}\left(\begin{pmatrix}1\\0\\1\end{pmatrix}, \begin{pmatrix}1\\2\\1\end{pmatrix}\right)$,

$S_4 = \textbf{span}\left(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right)$

57. $S_5 = \textbf{span}\left(\begin{pmatrix}1\\0\\1\\1\end{pmatrix}, \begin{pmatrix}1\\2\\1\\1\end{pmatrix}\right)$,

$S_6 = \textbf{span}\left(\begin{pmatrix}1\\0\\1\\1\end{pmatrix}, \begin{pmatrix}0\\1\\0\\1\end{pmatrix}\right)$

58. $S_7 = \textbf{span}\left(\begin{pmatrix}2\\1\\1\\1\end{pmatrix}, \begin{pmatrix}1\\2\\1\\1\end{pmatrix}\right)$,

$S_8 = \textbf{span}\left(\begin{pmatrix}1\\-1\\0\\0\end{pmatrix}, \begin{pmatrix}3\\3\\2\\2\end{pmatrix}\right)$

Chapter 6

Scalar Linear Differential Equations

Contents

Studied here are linear differential equations of the second order

$$a(x)y'' + b(x)y' + c(x)y = f(x) \tag{1}$$

and corresponding nth order models. Important to the theory is continuity of the **coefficients** $a(x)$, $b(x)$, $c(x)$ and the **non-homogeneous term** $f(x)$, which is also called the **forcing term** or the **input**.

6.1 Linear 2nd Order Constant

Studied is the homogeneous 2nd order equation

$$Ay'' + By' + Cy = 0$$

where $A \neq 0$, B and C are constants. An explicit formula for the general solution is developed. Prerequisites are the quadratic formula, complex numbers, Cramer's rule for 2×2 linear algebraic equations and first order linear differential equations.

Theorem 6.1 (How to Solve Second Order Constant Equations)
In the differential equation $Ay''+By'+Cy=0$, let $A \neq 0$, B and C be real constants. Let r_1, r_2 denote the two roots of the quadratic equation $Ar^2+Br+C=0$. If the roots are complex, then let $r_1 = a+ib$ with $b>0$, and $r_2 = \overline{r_1} = a-ib$. Define solutions $y_1(x)$, $y_2(x)$ of $Ay''+By'+Cy=0$ according to the following three cases, which are organized by the sign of the college algebra discriminant $\mathcal{D} = B^2-4AC$:

Case 1. $\mathcal{D}>0$ (Real distinct) $y_1(x) = e^{r_1x}$, $y_2(x) = e^{r_2x}$.

Case 2. $\mathcal{D}=0$ (Real equal) $y_1(x) = e^{r_1x}$, $y_2(x) = xe^{r_1x}$.

Case 3. $\mathcal{D}<0$ (Conjugate roots) $y_1(x){=}e^{ax}\cos(bx), y_2(x){=}e^{ax}\sin(bx)$.

Then each solution of $Ay''+By'+Cy=0$ is obtained, for some specialization of the constants c_1, c_2, from the expression

$$y(x) = c_1y_1(x)+c_2y_2(x).$$

Proof on page 437. Examples 6.1–6.3, page 434, consider the three cases.

A **general solution** is an expression that represents all solutions of the differential equation. Theorem 6.1 gives an expression of the form

$$y(x) = c_1y_1(x)+c_2y_2(x)$$

where c_1 and c_2 are *symbols* representing constants and y_1, y_2 are special solutions of the differential equation, determined by the roots of the **characteristic equation** $Ar^2+Br+C=0$ as in Theorem 6.1.

The **initial value problem** for $Ay''+By'+Cy=0$ selects the constants c_1, c_2 in the general solution $y = c_1y_2+c_2y_2$ from **initial conditions** of the form

$$y(x_0) = g_1, \quad y'(x_0) = g_2.$$

In these conditions, x_0 is a given point in $-\infty < x < \infty$ and g_1, g_2 are two real numbers, e.g., $g_1 =$ position, $g_2 =$ velocity at $x = x_0$.

Theorem 6.2 (Picard-Lindelöf Existence-Uniqueness)
Let $A \neq 0$, B, C, x_0, g_1 and g_2 be constants. Then the initial value problem $Ay''+By'+Cy=0$, $y(x_0)=g_1$, $y'(x_0)=g_2$ has one and only one solution, found from the general solution $y = c_1y_1+c_2y_2$ by applying Cramer's rule or the method of elimination. The solution is defined on $-\infty < x < \infty$.

Proof on page 437. Cramer's rule details are in Example 6.4, page 435.

Working Rule to solve $Ay''+By'+Cy=0$.

Find the roots of the characteristic equation $Ar^2+Br+C=0$. Apply Theorem 6.2 to write down y_1, y_2. The general solution is then $y = c_1y_1+c_2y_2$. If initial conditions are given, then determine c_1, c_2 explicitly, otherwise c_1, c_2 remain symbols.

Theorem 6.3 (Superposition)
In differential equation $Ay'' + By' + Cy = 0$, let $A \neq 0$, B and C be constants. Assume y_1, y_2 are solutions and c_1, c_2 are constants. Then $y = c_1y_1 + c_2y_2$ is a solution of $Ay'' + By' + Cy = 0$.

A proof appears on page 438. The result is implicitly used in Theorem 6.1, in order to show that a general solution satisfies the differential equation.

Structure of Solutions

The special solutions y_1, y_2 constructed in Theorem 6.1 have the form

$$e^{ax}, xe^{ax}, e^{ax}\cos bx, e^{ax}\sin bx.$$

These functions will be called **Euler solution atoms** or briefly **Atoms**.

Definition 6.1 (Euler Solution Atoms)
Define an **Euler base atom** to be one of the functions

$$e^{ax}, e^{ax}\cos bx, e^{ax}\sin bx,$$

where a, $b > 0$ are real constants with $b > 0$. Define

$$\textbf{Euler solution atom} = x^n(\textbf{base atom}), \quad n = 0, 1, 2, \ldots.$$

L. Euler (1707-1783) discovered these special solutions by substitution of $y = e^{rx}$ into the differential equation $Ay'' + By' + Cy = 0$, which results in the equations

$Ar^2e^{rx} + Bre^{rx} + Ce^{rx} = 0$	**Euler's Substitution** $y = e^{rx}$.
$Ar^2 + Br + C = 0$	**Characteristic equation**, found by canceling e^{rx}.

The same equations can also be found for the substitution $y = xe^{rx}$, called **Euler's substitution**. Together, the equations imply:

Theorem 6.4 (Euler's Exponential Substitution)

Euler atom $y = e^{rx}$ is a solution of $Ay'' + By' + Cy = 0$ if and only if r is a root of characteristic equation $Ar^2 + Br + C = 0$.

Euler atom $y = xe^{rx}$ is a solution of $Ay'' + By' + Cy = 0$ if and only if r is a double root of characteristic equation $Ar^2 + Br + C = 0$.

Euler atoms $y = e^{ax}\cos bx$ and $y = e^{ax}\sin bx$ are real solutions of $Ay'' + By' + Cy = 0$ if and only if $r = a + ib$ and $\overline{r} = a - ib$ are complex roots of characteristic equation $Ar^2 + Br + C = 0$.

Proof on page 439.

Theorem 6.1 may be succinctly summarized as follows.

The general solution y of a second order linear homogeneous constant-coefficient differential equation is a sum of constants times Euler solution atoms. The atoms are found from Euler's Theorem.

Speed

The time taken to write out the general solution varies among individuals and according to the algebraic complexity of the characteristic equation. Judge your understanding of the *Theorem* by these statistics: most persons can write out the general solution in under 60 seconds. Especially simple equations like $y'' = 0$, $y'' + y = 0$, $y'' - y = 0$, $y'' + 2y' + y = 0$, $y'' + 3y' + 2y = 0$ are finished in less than 30 seconds.

Graphics

Computer programs can produce plots for initial value problems. Computers cannot plot **symbolic solutions** containing unevaluated symbols c_1, c_2 that appear in the general solution.

Errors

Recorded below in Table 1 are some common but fatal errors made in displaying the general solution.

Table 1. Errors in Applying Theorem 6.1.

Bad equation	For $y'' - y = 0$, the correct characteristic equation is $r^2 - 1 = 0$. A common error is to write $r^2 - r = 0$.
Sign reversal	For factored equation $(r + 1)(r + 2) = 0$, the roots are $r = -1$, $r = -2$. A common error is to claim $r = 1$ and/or $r = 2$ is a root.
Miscopy signs	The equation $r^2+2r+2 = 0$ has complex conjugate roots $a \pm bi$, where $a = -1$ and $b = 1$ ($b > 0$ is required). A common error is to miscopy signs on a and/or b.
Copying $\pm i$	The equation $r^2 + 2r + 5 = 0$ has roots $a \pm ib$ where $a = -1$ and $b = 2$. A common mistake is to display $e^{-x}\cos(\pm 2ix)$ and $e^{-x}\sin(\pm 2ix)$. These expressions are not real solutions: neither ± nor the complex unit i should be copied.

Examples

Example 6.1 (Case 1)

Solve $y'' + y' - 2y = 0$.

Solution: The general solution is $y = c_1 e^x + c_2 e^{-2x}$. Ordering is not important; an equivalent answer is $y = c_1 e^{-2x} + c_2 e^x$. The answer will be justified below, by finding the two solutions y_1, y_2 in Theorem 6.1.

The characteristic equation $r^2 + r - 2 = 0$ is found formally by replacements $y'' \to r^2$, $y' \to r$ and $y \to 1$ in the differential equation $y'' + y' - 2y = 0$.[1]

A college algebra method[2] called **inverse-FOIL** applies to factor $r^2 + r - 2 = 0$ into $(r-1)(r+2) = 0$. The roots are $r = 1$, $r = -2$. Used implicitly here are the college algebra **factor theorem** and **root theorem**[3].

Applying case $\mathcal{D} > 0$ of Theorem 6.1 gives solutions $y_1 = e^x$ and $y_2 = e^{-2x}$.

Example 6.2 (Case 2)

Solve $4y'' + 4y' + y = 0$.

Solution: The general solution is $y = c_1 e^{-x/2} + c_2 x e^{-x/2}$. To justify this formula, find the characteristic equation $4r^2 + 4r + 1 = 0$ and factor it by the **inverse-FOIL** method or **square completion** to obtain $(2r+1)^2 = 0$. The roots are both $-1/2$.

Case $\mathcal{D} = 0$ of Theorem 6.1 gives $y_1 = e^{-x/2}$, $y_2 = xe^{-x/2}$. Then the general solution is $y = c_1 y_1 + c_2 y_2$, which completes the verification.

Example 6.3 (Case 3)

Solve $4y'' + 2y' + y = 0$.

Solution: The solution is $y = c_1 e^{-x/4}\cos(\sqrt{3}x/4) + c_2 e^{-x/4}\sin(\sqrt{3}x/4)$. This formula is justified below, by showing that the solutions y_1, y_2 of Theorem 6.1 are given by $y_1 = e^{-x/4}\cos(\sqrt{3}x/4)$ and $y_2 = e^{-x/4}\sin(\sqrt{3}x/4)$.

The characteristic equation is $4r^2 + 2r + 1 = 0$. The roots by the *quadratic formula* are

$$r = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$ College algebra formula for the roots of the quadratic $Ar^2 + Br + C = 0$.

$$= \frac{-2 \pm \sqrt{2^2 - (4)(4)(1)}}{(2)(4)}$$ Substitute $A = 4$, $B = 2$, $C = 1$.

$$= -\frac{1}{4} \pm \frac{\sqrt{-1}\sqrt{12}}{8}$$ Simplify. Used $\sqrt{(-1)(12)} = \sqrt{-1}\sqrt{12}$.

$$= -\frac{1}{4} \pm i\frac{\sqrt{3}}{4}$$ Convert to complex form, $i = \sqrt{-1}$.

[1] Some history. Euler's formal substitution $y = e^{rx}$ into the differential equation $y'' + y' - 2y = 0$ produces $r^2 + r - 2 = 0$ directly. Formal replacement $y'' \to r^2$, $y' \to r$ and $y \to 1$ gives the same characteristic equation $r^2 + r - 2 = 0$, with a reduction in errors. We prefer the shortcut, to increase the speed.

[2] **FOIL** is an abbreviation for **F**irst=AC, **O**utside=AD, **I**nside=BC, **L**ast=BD in the expansion of the algebraic product $(A+B)(C+D)$.

[3] **Theorem**. $r = r_0$ is a root of $p(r) = 0$ if and only if $(r - r_0)$ is a factor of $p(r)$.

The real part of the root is labeled $a = -1/4$. The two imaginary parts are $\sqrt{3}/4$ and $-\sqrt{3}/4$. Only the positive one is labeled, the other being discarded: $b = \sqrt{3}/4$.

Theorem 6.1 applies in the discriminant case $\mathcal{D} < 0$ to give solutions $y_1 = e^{ax}\cos(bx)$ and $y_2 = e^{ax}\sin(bx)$. Substitution of $a = -1/4$ and $b = \sqrt{3}/4$ results in the formulas $y_1 = e^{-x/4}\cos(\sqrt{3}x/4)$, $y_2 = e^{-x/4}\sin(\sqrt{3}x/4)$. The verification is complete.

The substitutions of a, b are remembered from the following diagram.

$$\begin{array}{cc} \boxed{-1/4} & \boxed{\sqrt{3}/4} \\ \downarrow & \downarrow \\ e^{\boxed{\mathbf{a}}x} & \cos\left(\boxed{\mathbf{b}}x\right) \\ \multicolumn{2}{c}{e^{-x/4}\cos\left(\sqrt{3}4x\right),} \end{array} \qquad \begin{array}{cc} \boxed{-1/4} & \boxed{\sqrt{3}/4} \\ \downarrow & \downarrow \\ e^{\boxed{\mathbf{a}}x} & \sin\left(\boxed{\mathbf{b}}x\right) \\ \multicolumn{2}{c}{e^{-x/4}\sin\left(\sqrt{3}4x\right)} \end{array}$$

It is recommended to perform the a, b substitution to find the first atom, which is $e^{-x/4}\cos\left(\sqrt{3}4x\right)$. Then replace $\boxed{\cos}$ by $\boxed{\sin}$ in that expression to obtain the second atom $e^{-x/4}\sin\left(\sqrt{3}4x\right)$.

Example 6.4 (Initial Value Problem)

Solve $y'' + y' - 2y = 0$, $y(0) = 1$, $y'(0) = -2$ and graph the solution on $0 \le x \le 2$.

Solution: The solution to the initial value problem is $y = e^{-2x}$. The graph appears in Figure 1.

Details. The general solution is $y = c_1e^x + c_2e^{-2x}$, from Example 6.1. The problem of finding c_1, c_2 uses the two equations $y(0) = 1$, $y'(0) = -2$ and the general solution to obtain expanded equations for c_1, c_2. For instance, $y(0) = 1$ expands to $\left.(c_1e^x + c_2e^{-2x})\right|_{x=0} = 1$, which is an equation for symbols c_1, c_2. The second equation $y'(0) = -2$ expands similarly, to give the two equations

$$\begin{array}{rcrcr} e^0c_1 & + & e^0c_2 & = & 1, \\ e^0c_1 & - & 2e^0c_2 & = & -2. \end{array}$$

The equations will be solved by the method of elimination. Because $e^0 = 1$, the equations simplify. Subtracting them eliminates the variable c_1 to give $3c_2 = 3$. Therefore, $c_2 = 1$ and back-substitution finds $c_1 = 0$. Then $y = c_1e^x + c_2e^{-2x}$ reduces, after substitution of $c_1 = 0$, $c_2 = 1$, to the equation $y = e^{-2x}$.

Graph. The solution $y = e^{-2x}$ is graphed by a routine application of curve library methods, which appear in the appendices, page 1015. No hand-graphing methods will be discussed here. To produce a computer graphic of the solution, the following code is offered. Calculator plots are similar.

```
plot(exp(-2*x),x=0..2);         Maple
plot2d(exp(-2*x),[x,0,2]);      Maxima
Plot[{exp(-2 x)},{x,0,2}];      Mathematica
plot [0:2] exp(-2*x)            Gnuplot
x=0:0.05:2; plot(x,exp(-2*x)) Matlab and Scilab
```

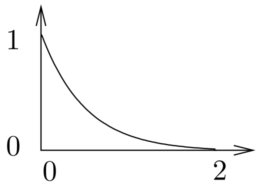


Figure 1. Exponential solution $y = e^{-2x}$. The graph decreases to zero at $x = \infty$.

Example 6.5 (Euler Solution Atoms)
Consider the list

$$1,\ x^2,\ 2,\ 3x+4x^2,\ x^3e^{x/\pi},\ 2x+3\cos x,\ \frac{x}{1+x^2}.$$

Box each entry that is precisely an atom and identify its base atom. Double-box the non-atom list entries that are a sum of constants times atoms.

Solution:
The answers and explanations:

$\boxed{1}$	An atom. Base atom $=1$.
$\boxed{x^2}$	An atom. Base atom $=1$.
$\boxed{\boxed{2}}$	$\boxed{\mathbf{X}}$ Not an atom. Constant 2 times the atom 1, which is a linear combination of atoms.
$\boxed{\boxed{3x+4x^2}}$	$\boxed{\mathbf{X}}$ **Not** an atom. Linear combination of atoms x, x^2.
$\boxed{e^{x/\pi}}$	An atom. Base atom $=e^{ax}$ where $a=1/\pi$.
$\boxed{\boxed{2x+3\cos x}}$	$\boxed{\mathbf{X}}$ **Not** an atom. Linear combination of atoms $x, \cos x$.
$\dfrac{x}{1+x^2}$	$\boxed{\mathbf{X}}$ **Not** an atom. Not a linear combination of atoms.

Example 6.6 (Inverse Problem)
Consider a 2nd order differential equation $Ay''+By'+Cy=0$, the coefficients A,B,C initially unknown. Find a set of coefficients for each of the following three examples, given the supplied information about the differential equation.

(a) The characteristic equation is $r^2+2r+5=0$.

(b) The characteristic equation has roots $r=-1,2$.

(c) Two solutions are e^x and xe^x.

Solution:
(a) The characteristic equation of $Ay''+By'+Cy=0$ is $Ar^2+Br+C=0$. Comparing terms to $r^2+2r+5=0$ implies a differential equation is $y''+2y'+5y=0$. The substitutions $y\to 1, y'\to r, y''\to r^2$ are used here in reverse.

(b) The characteristic polynomial Ar^2+Br+C factors into $A(r-r_1)(r-r_2)$ where r_1, r_2 are the two roots of the quadratic equation. Given $r_1=-1$ and $r_2=2$, then the characteristic equation has to be $A(r-(-1))(r-2)=0$ for some number $A\neq 0$. Assume $A=1$ to find one equation. Multiply out the product $(r+1)(r-2)$ to give characteristic equation $r^2-r-2=0$. This reduces the problem to methods in part (a). Then a differential equation is $y''-y'-2y=0$.

(c) The two given solutions are Euler solution atoms created from root $r=1$. Consulting Theorem 6.4, these two atoms are solutions of a second order equation with characteristic equation roots $r=1,1$ (a double root). The method in (b) is then applied: multiply out the product $(r-1)(r-1)$ to get characteristic equation $r^2-2r+1=0$. Then apply the method of (a). A differential equation is $y''-2y'+y=0$.

Proofs and Details

Proof of Theorem 6.1: To show that y_1 and y_2 are solutions is left to the exercises. For the remainder of the proof, assume y is a solution of $Ay'' + By' + Cy = 0$. It has to be shown that $y = c_1y_1 + c_2y_2$ for some real constants c_1, c_2.

Algebra background. In college algebra it is shown that the polynomial $Ar^2 + Br + C$ can be written in terms of its roots r_1, r_2 as $A(r - r_1)(r - r_2)$. In particular, the sum and product of the roots satisfy the relations $B/A = -r_1 - r_2$ and $C/A = r_1r_2$.

Case $\mathcal{D} > 0$. The equation $Ay'' + By' + Cy = 0$ can be re-written in the form $y'' - (r_1 + r_2)y' + r_1r_2y = 0$ due to the college algebra relations for the sum and product of the roots of a quadratic equation. The equation *factors* into $(y' - r_2y)' - r_1(y' - r_2y) = 0$, which suggests the substitution $u = y' - r_2y$. Then $Ay'' + By' + Cy = 0$ is equivalent to the first order system

$$\begin{aligned} u' - r_1u &= 0, \\ y' - r_2y &= u. \end{aligned}$$

Growth-decay theory, page 3, applied to the first equation gives $u = u_0e^{r_1x}$. The second equation $y' - r_2y = u$ is then solved by the integrating factor method, as in Example 2.14, page 99. This gives $y = y_0e^{r_2x} + u_0e^{r_1x}/(r_1 - r_2)$. Therefore, any possible solution y has the form $c_1e^{r_1x} + c_2e^{r_2x}$ for some c_1, c_2. This completes the proof of the case $\mathcal{D} > 0$.

Case $\mathcal{D} = 0$. The details follow the case $\mathcal{D} > 0$, except that $y' - r_2y = u$ has a different solution, $y = y_0e^{r_1x} + u_0xe^{r_1x}$ (exponential factors e^{r_1x} and e^{r_2x} cancel because $r_1 = r_2$). Therefore, any possible solution y has the form $c_1e^{r_1x} + c_2xe^{r_1x}$ for some c_1, c_2. This completes the proof of the case $\mathcal{D} = 0$.

Case $\mathcal{D} < 0$. The equation $Ay'' + By' + Cy = 0$ can be re-written in the form $y'' - (r_1 + r_2)y' + r_1r_2y = 0$ as in the case $\mathcal{D} > 0$, even though y is real and the roots are complex. The substitution $u = y' - r_2y$ gives the same equivalent system as in the case $\mathcal{D} > 0$. The solutions are symbolically the same, $u = u_0e^{r_1x}$ and $y = y_0e^{r_1x} + u_0e^{r_1x}/(r_1 - r_2)$. Therefore, any possible real solution y has the form $C_1e^{r_1x} + C_2e^{r_2x}$ for some possibly complex C_1, C_2.

Taking the real part of both sides of this equation gives $y = c_1e^{ax}\cos(bx) + c_2e^{ax}\sin(bx)$ for some real constants c_1, c_2. Details follow.

$$\begin{aligned}
y &= \mathcal{R}\mathrm{e}(y) && \text{Because } y \text{ is real.} \\
&= \mathcal{R}\mathrm{e}(C_1e^{r_1x} + C_2e^{r_2x}) && \text{Substitute } y = C_1e^{r_1x} + C_2e^{r_2x}. \\
&= e^{ax}\,\mathcal{R}\mathrm{e}(C_1e^{ibx} + C_2e^{-ibx}) && \text{Use } e^{u+iv} = e^ue^{iv}. \\
&= e^{ax}\,\mathcal{R}\mathrm{e}\,(C_1\cos bx + iC_1\sin bx && \text{Use } e^{i\theta} = \cos\theta + i\sin\theta. \\
&\quad + C_2\cos bx - iC_2\sin bx) \\
&= e^{ax}\,\mathcal{R}\mathrm{e}(C_1 + C_2))\cos bx && \text{Collect on trigonometric factors.} \\
&\quad + e^{ax}\,\mathcal{R}\mathrm{e}(iC_1 - iC_2)\sin bx \\
&= c_1e^{ax}\cos(bx) + c_2e^{ax}\sin(bx) && \text{Where } c_1 = \mathcal{R}\mathrm{e}(C_1{+}C_2) \text{ and } c_2 = \mathcal{I}\mathrm{m}(C_2{-}C_1) \\
&&& \text{are real.}
\end{aligned}$$

This completes the proof of the case $\mathcal{D} < 0$.

Proof of Theorem 6.2: The left sides of the two requirements $y(x_0) = g_1$, $y'(x_0) = g_2$ are expanded using the relation $y = c_1y_1 + c_2y_2$ to obtain the following system of equations for the unknowns c_1, c_2:

$$\begin{array}{rcrcl} y_1(x_0)c_1 & + & y_2(x_0)c_2 & = & g_1, \\ y_1'(x_0)c_1 & + & y_2'(x_0)c_2 & = & g_2. \end{array}$$

If the determinant of coefficients

$$\Delta = y_1(x_0)y_2'(x_0) - y_1'(x_0)y_2(x_0)$$

is nonzero, then Cramer's rule says that the solutions c_1, c_2 are given as quotients

$$c_1 = \frac{g_1 y_2'(x_0) - g_2 y_2(x_0)}{\Delta}, \quad c_2 = \frac{y_1(x_0)g_2 - y_1'(x_0)g_1}{\Delta}.$$

The organization of the proof is made from the three cases of Theorem 6.1, using x instead of x_0, to simplify notation. The issue of a unique solution has now been reduced to verification of $\Delta \neq 0$, in the three cases.

Case $\mathcal{D} > 0$. Then

$$\begin{aligned} \Delta &= e^{r_1 x} r_2 e^{r_2 x} - r_1 e^{r_1 x} e^{r_2 x} && \text{Substitute for } y_1,\ y_2. \\ &= (r_2 - r_1)e^{r_1 x + r_2 x} && \text{Simplify.} \\ &\neq 0 && \text{Because } r_1 \neq r_2. \end{aligned}$$

Case $\mathcal{D} = 0$. Then

$$\begin{aligned} \Delta &= e^{r_1 x}(e^{r_1 x} + r_1 x e^{r_1 x}) - r_1 e^{r_1 x} x e^{r_1 x} && \text{Substitute for } y_1,\ y_2. \\ &= e^{2r_1 x} && \text{Simplify.} \\ &\neq 0 \end{aligned}$$

Case $\mathcal{D} < 0$. Then $r_1 = \overline{r_2} = a + ib$ and

$$\begin{aligned} \Delta &= be^{2ax}(\cos^2 bx + \sin^2 bx) && \text{Two terms cancel.} \\ &= be^{2ax} && \text{Use } \cos^2\theta + \sin^2\theta = 1. \\ &\neq 0 && \text{Because } b > 0. \end{aligned}$$

In applications, the method of elimination is sometimes used to find c_1, c_2. In some references, it is called *Gaussian elimination.*

Proof of Superposition Theorem 6.3: The three terms of the differential equation are computed using the expression $y = c_1 y_1 + c_2 y_2$:

$$\begin{aligned} &\text{Term 1:} && cy = cc_1 y_1 + cc_2 y_2 \\ &\text{Term 2:} && by' = b(c_1 y_1 + c_2 y_2)' \\ &&& \quad\;\; = bc_1 y_1' + bc_2 y_2' \\ &\text{Term 3:} && ay'' = a(c_1 y_1 + c_2 y_2)'' \\ &&& \quad\;\;\; = ac_1 y_1'' + ac_2 y_2'' \end{aligned}$$

The left side of the differential equation, denoted LHS, is the sum of the three terms. It is simplified as follows:

$$\begin{aligned} \text{LHS} &= c_1[ay_1'' + by_1' + cy_1] && \text{Add terms 1,2 and 3,} \\ &\quad + c_2[ay_2'' + by_2' + cy_2] && \text{then collect on } c_1,\ c_2. \\ &= c_1[0] + c_2[0] && \text{Both } y_1,\ y_2 \text{ satisfy } ay'' + by' + cy = 0. \\ &= \text{RHS} && \text{The left and right sides match.} \end{aligned}$$

Proof of Euler's Theorem 6.4 The substitution $y = e^{rx}$ requires the derivative formulas $y' = re^{rx}$, $y'' = r^2e^{rx}$, which then imply from $Ay'' + By' + Cy = 0$ the relation

$$Ar^2e^{rx} + Bre^{rx} + Ce^{rx} = 0. \tag{2}$$

Assume that $y = e^{rx}$ is a solution of the differential equation. Then relation (2) holds. Cancel e^{rx} to obtain $Ar^2 + Br + C = 0$, then r is a root of the characteristic equation.

Conversely, if r is a root of the characteristic equation, then multiply $Ar^2 + Br + C = 0$ by e^{rx} to give relation (2). Then $y = e^{rx}$ is a solution of the differential equation.

This completes the proof of the first statement in Euler's theorem, in the special case for r real. Examination of the details reveals it is also valid for complex $r = a + ib$, with $y = e^{rx}$ a complex solution.

We go on to prove the third statement in Euler's theorem. A complex exponential solution $y = e^{rx}$, with $r = a + ib$, can be expanded as $y = e^{rx} = e^{ax+ibx} = e^{ax}\cos bx + ie^{ax}\sin bx$, because of Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$. Write $u = e^{ax}\cos bx$ and $v = e^{ax}\sin bx$, then $y = u + iv$ with u, v real. Expand the differential equation $Ay'' + By' + Cy = 0$ using $y = u + iv$ as

$$(Au'' + Bu' + Cu) + i(Av'' + Bv' + Cv) = 0 + 0i.$$

Then A, B, C, u, v all real implies, by equality of complex numbers, the two equations

$$\begin{aligned} Au'' + Bu' + Cu &= 0, \\ Av'' + Bv' + Cv &= 0. \end{aligned}$$

Together, these equations imply that $u = e^{ax}\cos bx$ and $v = e^{ax}\sin bx$ are solutions of the differential equation. Conversely, if both u and v are solutions, then the steps can be reversed to show $y = e^{rx}$ is a solution, which in turn implies $r = a + ib$ is a root of the characteristic equation. Finally, if $a + ib$ is a root and A, B, C are real, then college algebra implies $a - ib$ is a root. This completes the proof of the last statement of Euler's theorem.

The second statement of Euler's theorem will be proved. Substitute $y = xe^{rx}$ into the differential equation using the formulas $y' = e^{rx} + rxe^{rx}$, $y'' = 2re^{rx} + r^2xe^{rx}$ to obtain the relation

$$(Ar^2 + Br + C)xe^{rx} + (2Ar + B)e^{rx} = 0. \tag{3}$$

If $y = xe^{rx}$ is a solution of the differential equation, then relation (3) holds for all x. Cancel e^{rx} to get the polynomial relation

$$(Ar^2 + Br + C)x + (2Ar + B) = 0, \quad \text{for all } x.$$

Substitute $x = 0$ and then $x = 1$ to obtain $2Ar + B = 0$ and $Ar^2 + Br + C = 0$. These equations say that r is a double root of the characteristic equation, because the polynomial $p(t) = At^2 + Bt + C$ then satisfies $p(r) = p'(r) = 0$.

Conversely, suppose that r is a double root of $Ar^2+Br+C = 0$. Then $p(t) = At^2+Bt+C$ must satisfy the relations $p(r) = p'(r) = 0$, which imply $Ar^2+Br+C = 0$ and $2Ar+B = 0$. Then for all x, relation (3) holds, which in turn implies that $y = xe^{rx}$ is a solution. ■

Exercises 6.1

General Solution 2nd Order

Solve the constant equation using Theorem 6.1, page 431. Report the general solution using symbols c_1, c_2. Model the solution after Examples 6.1–6.3, page 434.

1. $y'' = 0$
Ans: $y = c_1 + c_2 x$

2. $3y'' = 0$

3. $y'' + y' = 0$

4. $3y'' + y' = 0$

5. $y'' + 3y' + 2y = 0$

6. $y'' - 3y' + 2y = 0$

7. $y'' - y' - 2y = 0$

8. $y'' - 2y' - 3y = 0$

9. $y'' + y = 0$

10. $y'' + 4y = 0$

11. $y'' + 16y = 0$

12. $y'' + 8y = 0$

13. $y'' + y' + y = 0$

14. $y'' + y' + 2y = 0$

15. $y'' + 2y' + y = 0$

16. $y'' + 4y' + 4y = 0$

17. $3y'' + y' + y = 0$

18. $9y'' + y' + y = 0$

19. $5y'' + 25y' = 0$

20. $25y'' + y' = 0$

21. $2y'' + y' - y = 0$

22. $2y'' - 3y' - 2y = 0$

23. $2y'' + 7y' + 3y = 0$

24. $4y'' + 8y' + 3y = 0$

25. $6y'' + 7y' + 2y = 0$

26. $6y'' + y' - 2y = 0$

27. $y'' + 4y' + 8y = 0$

28. $y'' - 2y' + 4y = 0$

29. $y'' + 2y' + 4y = 0$

30. $y'' + 4y' + 5y = 0$

31. $4y'' - 4y' + y = 0$

32. $4y'' + 4y' + y = 0$

33. $9y'' - 6y' + y = 0$

34. $9y'' + 6y' + y = 0$

35. $4y'' + 12y' + 9y = 0$

36. $4y'' - 12y' + 9y = 0$

Initial Value Problem 2nd Order

Solve the given problem, modeling the solution after Example 6.4.

37. $6y''+7y'+2y = 0$, $y(0) = 0$, $y'(0) = -1$

38. $2y''+7y'+3y = 0$, $y(0) = 5$, $y'(0) = -5$

39. $y'' - 2y' + 4y = 0$, $y(0) = 1$, $y'(0) = 1$

40. $y'' + 4y' + 5y = 0$, $y(0) = 1$, $y'(0) = 1$

41. $9y'' - 6y' + y = 0$, $y(0) = 3$, $y'(0) = 1$

42. $4y''+12y'+9y = 0$, $y(0) = 2$, $y'(0) = 1$

Detecting Euler Solution Atoms

A **Euler solution atom** is defined in Definition 6.1 page 432. Box each list entry that is precisely an atom. Double-box non-atom list entries that are a sum of constants times atoms. Follow Example 6.5 page 436.

43. 1, $e^{x/5}$, -1, $e^{1.1x}$, $2e^x$

44. $-x\cos\pi x$, $x^2\sin 2x$, x^3, $2x^3$

45. e^{2x}, $e^{-x^2/2}$, $\cos^2 2x$, $\sin 1.57x$

46. $x^7 e^x \cos 3x$, $x^{10} e^x \sin 4x$

47. $x^7 e^x \cosh 3x$, $x^{10} e^{-x} \sinh 5x$

48. $\cosh^2 x$, $x(1+x)$, $x^{1.5}$, $\sqrt{x}e^{-x}$

49. $x^{1/2}e^{x/2}$, $\dfrac{1}{x}e^x$, $e^x(1+x^2)$

50. $\dfrac{x}{1+x}$, $\dfrac{1}{x}(1+x^2)$, $\ln|x|$

Euler Base Atom

An **Euler base atom** is defined in Definition 6.1 page 432. Find the base atom for each Euler solution atom in the given list.

51. $x\cos\pi x$, x^3, $x^{10}e^{-x}\sin 5x$

52. x^6, x^4e^{2x}, $x^2e^{-x/\pi}$, $x^7e^x\cos 1.1x$

Inverse Problems

Find the homogeneous 2nd order differential equation, given the supplied information. Follow Example 6.6.

53. $e^{-x/5}$ and 1 are solutions.
Ans: $5y'' + y' = 0$.

54. e^{-x} and 1 are solutions.

55. $e^x + e^{-x}$ and $e^x - e^{-x}$ are solutions.

56. $e^{2x} + xe^{2x}$ and xe^{2x} are solutions.

57. x and $2+x$ are solutions.

58. $4e^x$ and $3e^{2x}$ are solutions.

59. The characteristic equation is $r^2+2r+1=0$.

60. The characteristic equation is $4r^2+4r+1=0$.

61. The characteristic equation has roots $r=-2,3$.

62. The characteristic equation has roots $r=2/3,3/5$.

63. The characteristic equation has roots $r=0,0$.

64. The characteristic equation has roots $r=-4,-4$.

65. The characteristic equation has complex roots $r=1\pm 2i$.

66. The characteristic equation has complex roots $r=-2\pm 3i$.

Details of proofs

67. **(Theorem 6.1, Background)** Expand the relation $Ar^2+Br+C = A(r-r_1)(r-r_2)$ and compare coefficients to obtain the sum and product of roots relations

$$\frac{B}{A} = -(r_1+r_2), \qquad \frac{C}{A} = r_1r_2.$$

68. **(Theorem 6.1, Background)**
Let r_1, r_2 be the two roots of $Ar^2+Br+C=0$. The discriminant is $\mathcal{D} = B^2-4AC$. Use the quadratic formula to derive these relations for $\mathcal{D}>0$, $\mathcal{D}=0$, $\mathcal{D}<0$, respectively:

$$\begin{aligned}
&r_1 = \tfrac{-B+\sqrt{\mathcal{D}}}{2A}, r_2 = \tfrac{-B-\sqrt{\mathcal{D}}}{2A},\\
&r_1 = r_2 = \tfrac{\sqrt{\mathcal{D}}}{2A}.\\
&r_1 = \tfrac{-B+i\sqrt{-\mathcal{D}}}{2A}, r_2 = \tfrac{-B-i\sqrt{-\mathcal{D}}}{2A}.
\end{aligned}$$

69. **(Theorem 6.1, Case 1)**
Let $y_1 = e^{r_1x}$, $y_2 = e^{r_2x}$. Assume $Ar^2+Br+C = A(r-r_1)(r-r_2)$. Show that y_1, y_2 are solutions of $Ay''+By'+Cy=0$.

70. **(Theorem 6.1, Case 2)**
Let $y_1 = e^{r_1x}$, $y_2 = x\,e^{r_1x}$. Assume $Ar^2+Br+C{=}A(r-r_1)(r-r_1)$. Show that y_1, y_2 are solutions of $Ay''+By'+Cy=0$.

71. **(Theorem 6.1, Case 3)**
Let a, b be real, $b>0$. Let $y_1 = e^{ax}\cos bx$, $y_2 = e^{ax}\sin bx$. Assume factorization
$Ar^2{+}Br{+}C{=}A(r{-}a{-}ib)(r{-}a{+}ib)$
then show that y_1, y_2 are solutions of $Ay''+By'+Cy=0$.

6.2 Continuous Coefficient Theory

The existence, uniqueness and structure of solutions for the equation

$$(1) \qquad a(x)y'' + b(x)y' + c(x)y = f(x)$$

will be studied, guided in part by the first order theory.

Continuous–Coefficient Equations

The **homogeneous equation** is $a(x)y'' + b(x)y' + c(x)y = 0$ while the **non-homogeneous equation** is $a(x)y'' + b(x)y' + c(x)y = f(x)$. An equation is said to have **constant coefficients** if a, b and c are scalars.

A **linear combination** of two functions y_1, y_2 is $c_1y_1(x) + c_2y_2(x)$, where c_1 and c_2 are constants. The natural domain is the common domain of y_1 and y_2.

The **general solution** of $a(x)y'' + b(x)y' + c(x)y = f(x)$ is an expression which describes all possible solutions of the equation. Exactly how to write such an expression is revealed in the theorems below.

An **initial value problem** is the problem of solving $a(x)y'' + b(x)y' + c(x)y = f(x)$ subject to **initial conditions** $y(x_0) = g_1$, $y'(x_0) = g_2$. It is assumed that x_0 is in the common domain of continuity of the coefficients and that g_1, g_2 are prescribed numbers.

Theorem 6.5 (Superposition)
The homogeneous equation $a(x)y'' + b(x)y' + c(x)y = 0$ has the *superposition property*:

> If y_1, y_2 are solutions and c_1, c_2 are constants, then the linear combination $y(x) = c_1y_1(x) + c_2y_2(x)$ is a solution.

Proof on page 445.

Theorem 6.6 (Picard-Lindelöf Existence-Uniqueness)
Let the coefficients $a(x)$, $b(x)$, $c(x)$, $f(x)$ be continuous on an interval J containing $x = x_0$. Assume $a(x) \neq 0$ on J. Let g_1 and g_2 be constants. Then the initial value problem

$$a(x)y'' + b(x)y' + c(x)y = f(x), \quad y(x_0) = g_1, \quad y'(x_0) = g_2$$

has a unique solution $y(x)$ defined on J.

Proof on page 446.

Theorem 6.7 (Homogeneous Structure)
The homogeneous equation $a(x)y'' + b(x)y' + c(x)y = 0$ has a general solution of the form $y_h(x) = c_1y_1(x) + c_2y_2(x)$, where c_1, c_2 are arbitrary constants and $y_1(x)$, $y_2(x)$ are solutions.

Proof on page 447.

Theorem 6.8 (Non-Homogeneous Structure)

The non-homogeneous equation $a(x)y'' + b(x)y' + c(x)y = f(x)$ has general solution $y = y_h + y_p$, where $y_h(x)$ is the general solution of the homogeneous equation $a(x)y''+b(x)y'+c(x)y = 0$ and $y_p(x)$ is a particular solution of the non-homogeneous equation $a(x)y'' + b(x)y' + c(x)y = f(x)$.

Proof on page 447.

Theorem 6.9 (Reduction of Order)

Let $y_1(x)$ be a solution of $a(x)y'' + b(x)y' + c(x)y = 0$ on an interval J. Assume $a(x) \neq 0$, $y_1(x) \neq 0$ on J. Let all coefficients be continuous on J. Select x_0 in J. Then the general solution has the form $y_h(x) = c_1y_1(x) + c_2y_2(x)$ where c_1, c_2 are constants and

$$y_2(x) = y_1(x) \int_{x_0}^{x} \frac{e^{-\int_{x_0}^{t}(b(r)/a(r))dr}}{y_1^2(t)} dt.$$

Proof on page 448.

Theorem 6.10 (Equilibrium Method)

A non-homogeneous equation

$$ay'' + by' + cy = f$$

has an easily-found particular solution $y_p(x)$ in the special case when all coefficients a, b, c, f are constant. The solution can be found by the *equilibrium method*. The answers:

$$\begin{array}{lll} c \neq 0 & y_p(x) & = \dfrac{f}{c}, \\ c = 0, b \neq 0 & y_p(x) & = \displaystyle\int \frac{f}{b} dx = \frac{f}{b}x, \\ c = b = 0, a \neq 0 & y_p(x) & = \displaystyle\int \left(\int \frac{f}{a} dt \right) dx = \frac{f}{a}\frac{x^2}{2}. \end{array}$$

See Example 6.11 page 445.

Equilibrium Method. The method applies to non-homogeneous equations with constant coefficients $ay'' + by' + cy = f$. The method:

Truncate the LHS of the differential equation to just the lowest order term, then solve the resulting equation by the method of quadrature.

Examples and Methods

Example 6.7 (Superposition)

Verify that $y = c_1y_1 + c_2y_2$ is a solution, given equation $y'' + 4y' + 4y = 0$ and solutions $y_1(x) = e^{2x}$, $y_2(x) = xe^{2x}$.

Solution: The answer check details can be simplified as follows.

LHS $= y'' + 4y' + 4y$	Given differential equation LHS.
$\text{LHS} = c_1y_1'' + c_2y_2''+$ $4(c_1y_1' + c_2y_2')+$ $4(c_1y_1 + c_2y_2)$	Substitute $y = c_1y_1 + c_2y_2$.
$\text{LHS} = c_1(y_1'' + 4y_1' + 4y_1)+$ $c_2(y_2'' + 4y_2' + 4y_2)$	Collect on c_1, c_2.
$\text{LHS} = c_1(0)+$ $c_2(0)$	Because y_1, y_2 are solutions of the equation $y''+$ $4y' + 4y = 0$.

Then $y = c_1y_1 + c_2y_2$ satisfies $y'' + 4y' + 4y = 0$, as claimed.

Example 6.8 (Continuous Coefficients)

Determine all intervals J of existence of $y(x)$, according to Picard's theorem, for the differential equation $y'' + \frac{1}{1+x}y' + \frac{x}{2+x}y = 0$.

Solution: The challenge is describe the open intervals J where $1+x \neq 0$ and $2+x \neq 0$, because the coefficients are continuous whenever both inequalities hold. The real line is divided by the exceptions $x = -1$, $x = -2$. Then $-\infty < x < -2$, $-2 < x < -1$, $-1 < x < \infty$ are the possible intervals J in Picard's theorem.

Example 6.9 (Recognizing y_h)

Consider $y'' + 4y = x$. Extract from the solution $y = 2\cos 2x + 3\sin 2x + x/4$ a particular solution y_p with fewest terms.

Solution: The homogeneous equation $y'' + 4y = 0$ has characteristic equation $r^2 + 4 = 0$ with complex roots $\pm 2i$ and Euler solution atoms $\cos 2x, \sin 2x$. Then $2\cos 2x + 3\sin 2x$ is a solution y_h of the homogeneous equation and $y = y_h + x/4$. Subtract the homogeneous solution to obtain a particular solution $x/4$. By Theorem 6.8, this is a particular solution y_p. It has the fewest possible terms.

Example 6.10 (Reduction of Order)

Given solution $y_1 = 1$, find an independent solution y_2 of $y'' + 4y' = 0$ by reduction of order.

Solution: The answer is $y_2 = \frac{1}{4}\left(1 - e^{-4x}\right)$. The method is Theorem 6.9.

We apply the theorem by inserting the formula $y_1 = 1$ into

$$y_2(x) = y_1(x) \int_{x_0}^{x} \frac{e^{-\int_{x_0}^{t}(b(r)/a(r))dr}}{y_1^2(t)} dt.$$

Then, using $x_0 = 0$, $a(x) = 1$, $b(x) = 4$, $c(x) = 0$ gives

$$\begin{aligned} y_2(x) &= (1)\int_0^x \frac{e^{-\int_0^t (4/1)dr}}{(1)^2} dt \\ &= (1)\int_0^x \frac{e^{-4t}}{(1)^2} dt \\ &= \frac{e^{-4x} - 1}{-4} \end{aligned}$$

Example 6.11 (Equilibrium Method)

Apply the equilibrium method to find y_p, then find the general solution $y = y_h + y_p$. This method works only for constant coefficients. meaning $a(x)$, $b(x)$, $c(x)$, $f(x)$ in equation (1) are constant.

(a) $y'' + 4y' + 4y = \pi$

(b) $2y'' + 3y' = -5$

(c) $3y'' = 20$

Solution: All equations have constant coefficients, therefore the method applies. The method selects a trial solution for y_p which makes all terms zero except the lowest derivative term. Then solve for the trial solution by quadrature to obtain y_p. The answer should be verified due to the possibility of integration and algebra errors.

(a) Truncate all but the lowest term to obtain $4y = \pi$, then $y_p(x) = \pi/4$. The homogeneous solution y_h is the solution of $y'' + 4y' + 4y = 0$ with characteristic equation $r^2 + 4r + 4 = 0$, factoring into $(r + 2)(r + 2) = 0$. Then the atoms are e^{-2x}, xe^{-2x} and $y_h(x) = c_1e^{-2x} + c_2xe^{-2x}$. The general solution is $y(x) = y_h(x) + y_p(x) = c_1e^{-2x} + c_2xe^{-2x} + \dfrac{\pi}{4}$.

(b) Truncate to $3y' = -5$ and integrate to obtain $y_p(x) = \frac{-5}{3}x$. The characteristic equation of $2y'' + 3y' = 0$ is $(2r + 3)r = 0$ with roots $r = 0, -3/2$. The atoms are e^{0x}, $e^{-3x/2}$ and then $y_h(x) = c_1e^{0x} + c_2e^{-3x/2}$. The general solution is $y(x) = y_h(x) + y_p(x) = c_1 + c_2e^{-3x/2} + \frac{-5}{2}x$, because e^{0x} is written as 1.

(c) The quadrature solution is $y_p(x) = \frac{20}{3}\frac{x^2}{2}$. The characteristic equation for $3y'' = 0$ is $3r^2 = 0$ with double root $r = 0, 0$. The atoms are e^{0x}, xe^{0x} and the homogeneous solution is $y_h(x) = c_1e^{0x} + c_2xe^{0x} = c_1 + c_2x$. Then the general solution is $y(x) = y_h(x) + y_p(x) = c_1 + c_2x + \frac{20}{3}\frac{x^2}{2}$.

Example 6.12 (Equilibrium Method Failure)

The equation $y'' + y' = 2x$ fails to have constant coefficients, meaning $a(x)$, $b(x)$, $c(x)$, $f(x)$ are not all constant. Blind application of the equilibrium method gives $y = x^2$, **not** a solution. Explain.

Solution: The error: $y'' + y' = 2x$ does not have constant coefficients, which is required to apply the equilibrium method. **What went wrong**? The equilibrium method blindly applied gives the equation $0 + y' = 2x$, which by quadrature implies $y(x) = x^2$. It appears to work! Let's test $y = x^2$. Insert $y = x^2$ into $y'' + y' = 2x$, then $(x^2)'' + (x^2)' = 2x$, which implies $2 + 2x = 2x$ and finally the false equation $2 = 0$. Therefore, $y = x^2$ is not a solution of $y'' + y' = 2x$.

Proofs and Details

Proof of Theorem 6.5: The three terms of the differential equation, $c(x)y$, $b(x)y'$ and $a(x)y''$, are computed using the expression $y = c_1y_1 + c_2y_2$. The formulas are added to obtain the left hand side LHS of the differential equation:

$\text{LHS} = c_1[ay_1'' + by_1' + cy_1]$	Add terms $c(x)y$, $b(x)y'$, $a(x)y''$
$\quad + c_2[ay_2'' + by_2' + cy_2]$	and then collect on c_1, c_2.
$= c_1[0] + c_2[0]$	Both y_1, y_2 satisfy $ay'' + by' + cy = 0$.
$= \text{RHS}$	The left and right sides match.

Proof of Theorem 6.6: The basic ideas for the proof appear already in the proof of the Picard-Lindelöf theorem, page 68. Additional proof is required, because the solution is supposed to be defined on all of J, whereas the basic Picard-Lindelöf theorem supplies only *local existence.*

Existence. Picard's ideas write the solution $y(x)$ on J as the sum of an infinite series of continuous functions. This is accomplished by using the **Position-Velocity substitution** $x = t$, $X = y(t)$, $Y = y'(t)$ and definitions $t_0 = x_0$, $X_0 = g_1$, $Y_0 = g_2$ to re-write the differential equation and initial conditions in the new form

$$\begin{aligned} &X' = Y, \quad Y' = (f(t) - b(t)Y - c(t)X)/a(t), \\ &X(t_0) = X_0, \quad Y(t_0) = Y_0. \end{aligned}$$

The *Picard iterates* are defined by

$$\begin{aligned} X_n(t) &= \int_{t_0}^{t} Y_{n-1}(x)dx, \\ Y_n(t) &= \int_{t_0}^{t} (f(x) - b(x)Y_{n-1}(x) - c(x)X_{n-1}(x))\frac{dx}{a(x)}. \end{aligned}$$

The new bit of information provided by these formulas is significant: because X_0 and Y_0 are defined everywhere on J, so also are X_n and Y_n. This explains why the series equality

$$y(x) = X_0 + \sum_{n=1}^{\infty}(X_n(x) - X_{n-1}(x))$$

provides a formula for $y(x)$ on *all of interval* J, instead of on just a local section of the interval.

The demand that the series converge on J creates new technical problems, to be solved by modifying Picard's proof. Suffice it to say that Picard's ideas are sufficient to give series convergence and hence existence of $y(x)$ on J.

Uniqueness. An independent proof of the uniqueness will be given, based upon calculus ideas only.

Let two solutions y_1 and y_2 of the differential equation be given, having the same initial conditions. Then their difference $y = y_1 - y_2$ satisfies the homogeneous differential equation $a(x)y'' + b(x)y' + c(x)y = 0$ and the initial conditions $y(x_0) = y'(x_0) = 0$. Some details:

$\text{LHS} = ay'' + by' + cy$	Left side of $ay'' + by' + cy = f$.
$= ay_1'' + by_1' + cy_1$	Substitute $y = y_1 - y_2$.
$\quad - (ay_2'' + by_2' + cy_2)$	
$= f(x) - f(x)$	Both y_1, y_2 satisfy $ay'' + by' + cy = f$.
$= 0$	The homogeneous equation is satisfied.

To prove $y_1 = y_2$, it suffices to show $y(x) \equiv 0$. This will be accomplished by showing that the non-negative function

$$z(t) = (y(t))^2 + (y'(t))^2$$

satisfies $z(t) \le 0$, which implies $z(t) \equiv 0$ and then $y(x) \equiv 0$. The argument depends upon the following inequality.

Lemma. The function $z(t)$ satisfies $|z'| \le Mz$ for some constant $M \ge 0$.

To finish the uniqueness proof, observe first that initial conditions $y(x_0) = y'(x_0) = 0$ imply $z(t_0) = 0$. By the lemma, $|z'| \le Mz$ for some constant M, or equivalently $-Mz \le z' \le Mz$. Multiply $z' \le Mz$ by the *integrating factor* e^{-Mt} to give $(e^{-Mt}z(t))' \le 0$. Integration over $[t_0, t]$ shows $e^{-Mt}z(t) \le 0$. Then $z(t) = 0$ for $t \ge t_0$. Similarly, $-z' \le Mz$ implies $z(t) = 0$ for $t \le t_0$. This concludes the uniqueness proof, except for the proof of the lemma.

Proof of the lemma. Compute the derivative z' as follows, using notation $X = y(t)$ and $Y = y'(t)$ to re-write $z(t) = (y(t))^2 + (y'(t))^2 = X^2 + Y^2$.

$z' = 2XX' + 2YY'$	Power and product rules.
$= 2XY + 2Y(-cX - bY)/a$	Use $X' = Y$ and the homogeneous equation $aY' + bY + cX = 0$.
$= (2 - 2c/a)XY + (-2b/a)Y^2$	Collect terms.

Let $M = 2\max_{A \le x \le B}\{|1 - c(x)/a(x)| + |-2b(x)/a(x)|\}$, where $[A, B]$ is an arbitrary subinterval of J containing x_0. The estimate $|z'| \le Mz$ will be established.

$\lvert z'\rvert = \lvert(2 - 2c/a)XY + (-2b/a)Y^2\rvert$	Estimate modulus of z'.
$\le \lvert 1 - c/a\rvert\lvert 2XY\rvert + \lvert -2b/a\rvert\lvert Y\rvert^2$	Apply $\lvert c + d\rvert \le \lvert c\rvert + \lvert d\rvert$ and $\lvert uv\rvert = \lvert u\rvert\lvert v\rvert$.
$\le (M/2)\lvert 2XY\rvert + (M/2)\lvert Y\rvert^2$	Definition of maximum M applied.
$\le Mz$	Use $\lvert 2XY\rvert \le X^2 + Y^2$, proved from $(\lvert X\rvert - \lvert Y\rvert)^2 \ge 0$.

Proof of Theorem 6.7: To define y_1 and y_2 requires application of Picard's existence-uniqueness Theorem 6.6, page 442. Select them by their initial conditions, $y_1(x_0) = 1$, $y_1'(x_0) = 0$ and $y_2(x_0) = 0$, $y_2'(x_0) = 1$.

To complete the proof, a given solution $y(x)$ must be expressed as a linear combination $y(x) = c_1y_1(x) + c_2y_2(x)$ for some values of c_1, c_2.

Define $c_1 = y(x_0)$, $c_2 = y'(x_0)$. Let $u(x) = y(x) - c_1y_1(x) - c_2y_2(x)$. The equation $y(x) = c_1y_1(x) + c_2y_2(x)$ will be verified by showing $u(x) \equiv 0$.

First, u is a solution of $a(x)y'' + b(x)y' + c(x)y = 0$, by the superposition principle, Theorem 6.5. It has initial conditions $u(x_0) = y(x_0) - c_1(1) - c_2(0) = 0$ and $u'(x_0) = y'(x_0) - c_1(0) - c_2(1) = 0$. By uniqueness of initial value problems, $u(x) \equiv 0$, which completes the proof.

Proof of Theorem 6.8: Let $y_p(x)$ be a given particular solution of $a(x)y'' + b(x)y' + c(x)y = f(x)$. Let $y(x)$ be any other solution of this equation and define $u(x) = y(x) - y_p(x)$. Subtract the two differential equations to verify that u is a solution of the homogeneous equation $a(x)u'' + b(x)u' + c(x)u = 0$. By Theorem 6.7, $u = y_h(x)$ for some choice of constants c_1, c_2. Then $y(x) = u(x) + y_p(x) = y_h(x) + y_p(x)$, as was to be shown, completing the proof.

Proof of Theorem 6.9: . Let $W(x) = e^{-\int_{x_0}^{x}(b(r)/a(r))dr}$. By the chain rule and the fundamental theorem of calculus, $W' = -bW/a$. Let $u(x) = 1/y_1^2(x)$ to simplify displays. The successive derivatives of y_2 are

$$\begin{aligned}
y_2(x) &= y_1(x)\int_{x_0}^{x} Wudt && \text{Definition of } y_2,\ u \text{ and } W. \\
y_2'(x) &= \left(y_1(x)\int_{x_0}^{x} Wudt\right)' && \text{Apply the product rule.} \\
&= y_1'(x)\int_{x_0}^{x} Wudt + y_1(x)W(x)u(x) && \text{Use } (\textstyle\int_{x_0}^{x} G(t)dt)' = G(x). \\
&= y_1'(x)\int_{x_0}^{x} Wudt + \frac{W(x)}{y_1(x)} && \\
y_2''(x) &= \left(y_1'(x)\int_{x_0}^{x} Wudt + \frac{W(x)}{y_1(x)}\right)' && \\
&= y_1''(x)\int_{x_0}^{x} Wudt + y_1'(x)W(x)u(x) && \text{Apply the sum and quotient rules.} \\
&\quad + \frac{W'(x)y_1(x) - W(x)y_1'(x)}{y_1^2(x)} && \\
&= y_1''(x)\int_{x_0}^{x} Wudt + \frac{W'(x)}{y_1(x)} && \text{Simplify non-integral terms.} \\
&= y_1''(x)\int_{x_0}^{x} Wudt - \frac{b(x)W(x)}{a(x)y_1(x)} && \text{Use } W' = -(b/a)W.
\end{aligned}$$

The derivative formulas are multiplied respectively by c, b and a to obtain an expression $\mathcal{E} = ay_2'' + by_2' + cy_2$, which must be shown to be zero. The details:

$$\begin{aligned}
\mathcal{E} &= cy_2 + by_2' + ay_2'' && \\
&= c\left(y_1\textstyle\int Wu\right) + b\left(y_1'\textstyle\int Wu + W/y_1\right) && \\
&\quad + a\left(y_1''\textstyle\int Wu - bW/(ay_1)\right) && \\
&= (cy_1 + by_1' + ay_1'')\textstyle\int Wu && \text{Collect all integral terms.} \\
&\quad + bW/y_1 - bW/y_1 && \\
&= 0 && \text{Because } ay_1'' + by_1' + cy_1 = 0.
\end{aligned}$$

General Solution. To show that $c_1y_1 + c_2y_2$ is the general solution, for this choice of y_1, y_2, let $y(x)$ be a solution of the homogeneous equation and define

$$c_1 = \frac{y(x_0)}{y_1(x_0)}, \quad c_2 = y_1(x_0)(y'(x_0) - c_1y_1'(x_0)).$$

It will be shown that $y(x) = c_1y_1(x) + c_2y_2(x)$ by verifying that $u(x) = y(x) - c_1y_1(x) - c_2y_2(x)$ is zero. Superposition implies u is a solution of the homogeneous equation. It has initial conditions $u(x_0) = 0$, $u'(x_0) = 0$, because $y_2'(x_0) = 1/y_1(x_0)$. Uniqueness of initial value problems implies $u(x) \equiv 0$, completing the proof.

Proof of Theorem 6.10, Equilibrium Method: In the case $c \neq 0$, find an equilibrium solution $y =$ constant by substitution of $y = k$ into the differential equation (the *equilibrium method*). Then $ck = f$ and $y_p(x) = \frac{f}{c}$.

For case $c = 0$, $b \neq 0$, observe that the differential equation in terms of the velocity $v = y'$ is $av' + bv = f$. Apply the equilibrium method to this equation to obtain $v = f/b$ and finally $y = \int vdx = \frac{f}{b}x$.

For the last case $b = c = 0$ and $a \neq 0$, then the equation is in terms of the acceleration $p = y''$ the new equation $ap = f$. Then $p = f/a$ is the quadrature equation $y'' = f/a$ with solution $y_p(x) = \frac{f}{a}\frac{x^2}{2}$.

Exercises 6.2

Continuous Coefficients

Determine all intervals J of existence of $y(x)$, according to Picard's theorem.

1. $y'' + y = \ln|x|$

2. $y'' = \ln|x - 1|$

3. $y'' + (1/x)y = 0$

4. $y'' + \frac{1}{1+x}y' + \frac{1}{x}y = 0$

5. $x^2y'' + y = \sin x$

6. $x^2y'' + xy' = 0$

Superposition

Verify that $y = c_1y_1 + c_2y_2$ is a solution.

7. $y'' = 0$, $y_1(x) = 1$, $y_2(x) = x$

8. $y'' = 0$, $y_1(x) = 1 + x$, $y_2(x) = 1 - x$

9. $y''' = 0$, $y_1(x) = x$, $y_2(x) = x^2$

10. $y''' = 0$, $y_1(x) = 1 + x$, $y_2(x) = x + x^2$

Structure

Verify that $y = y_h + y_p$ is a solution.

11. $y'' + y = 2$, $y_h(x) = c_1 \cos x + c_2 \sin x$, $y_p(x) = 2$

12. $y'' + 4y = 4$, $y_h(x) = c_1 \cos 2x + c_2 \sin 2x$, $y_p(x) = 1$

13. $y'' + y' = 5$, $y_h(x) = c_1 + c_2e^{-x}$, $y_p(x) = 5x$

14. $y'' + 3y' = 5$, $y_h(x) = c_1 + c_2e^{-3x}$, $y_p(x) = 5x/3$

15. $y'' + y' = 2x$, $y_h(x) = c_1 + c_2e^{-x}$, $y_p(x) = x^2 - 2x$

16. $y'' + 2y' = 4x$, $y_h(x) = c_1 + c_2e^{-2x}$, $y_p(x) = x^2 - x$

Initial Value Problems

Solve for constants c_1, c_2 in the general solution $y_h = c_1y_1 + c_2y_2$.

17. $y'' = 0$, $y_1 = 1$, $y_2 = x$, $y(0) = 1$, $y'(0) = 2$

18. $y'' = 0$, $y_1 = 1+x$, $y_2 = 1-x$, $y(0) = 1$, $y'(0) = 2$

19. $y'' + y = 0$, $y_1 = \cos x$, $y_2 = \sin x$, $y(0) = 1$, $y'(0) = -1$

20. $y'' + y = 0$, $y_1 = \sin x$, $y_2 = \cos x$, $y(0) = 1$, $y'(0) = -1$

21. $y'' + 4y = 0$, $y_1 = \cos 2x$, $y_2 = \sin 2x$, $y(0) = 1$, $y'(0) = -1$

22. $y'' + 4y = 0$, $y_1 = \sin 2x$, $y_2 = \cos 2x$, $y(0) = 1$, $y'(0) = -1$

23. $y''+y' = 0$, $y_1 = 1$, $y_2 = e^{-x}$, $y(0) = 1$, $y'(0) = -1$

24. $y''+y' = 0$, $y_1 = 1$, $y_2 = e^{-x}$, $y(0) = 2$, $y'(0) = -3$

25. $y'' + 3y' = 0$, $y_1 = 1$, $y_2 = e^{-3x}$, $y(0) = 1$, $y'(0) = -1$

26. $y'' + 5y' = 0$, $y_1 = 1$, $y_2 = e^{-5x}$, $y(0) = 1$, $y'(0) = -1$

Recognizing y_h

Extract from the given solution y a particular solution y_p with fewest terms.

27. $y'' + y = x$,
$y = c_1 \cos x + c_2 \sin x + x$

28. $y'' + y = x$,
$y = \cos x + x$

29. $y'' + y' = x$,
$y = c_1 + c_2e^{-x} + x^2/2 - x$

30. $y'' + y' = x$,
$y = e^{-x} - x + 1 + x^2/2$

31. $y'' + 2y' + y = 1 + x$,
$y = (c_1 + c_2x)e^{-x} + x - 1$

32. $y'' + 2y' + y = 1 + x$,
$y = e^{-x} + x + xe^{-x} - 1$

Reduction of Order

Given solution y_1, find an independent solution y_2 by reduction of order.

33. $y'' + 2y' = 0$, $y_1(x) = 1$

34. $y'' + 2y' = 0$, $y_1(x) = e^{-2x}$

35. $2y'' + 3y' + y = 0$, $y_1(x) = e^{-x}$

36. $2y'' - y' - y = 0$, $y_1(x) = e^x$

Equilibrium Method

Apply the equilibrium method to find y_p, then find the general solution $y = y_h + y_p$.

37. $2y'' = 3$

38. $y'' + 4y' = 5$

39. $y'' + 3y' + 2y = 3$

40. $y'' - y' - 2y = 2$

41. $y'' + y = 1$

42. $3y'' + y' + y = 7$

43. $6y'' + 7y' + 2y = 5$

44. $y'' - 2y' + 4y = 8$

45. $4y'' - 4y' + y = 8$

46. $4y'' - 12y' + 9y = 18$

6.3 Higher Order Linear Constant-Coefficient Equations

Discussed here are structure results for the n-th order linear differential equation

$$a_n y^{(n)} + \cdots + a_0 y = f(x).$$

It is assumed that each coefficient is **constant** and the leading coefficient a_n is **not zero**. The **forcing term** or **input** $f(x)$ is assumed to either be zero, in which case the equation is called **homogeneous**, or else $f(x)$ is nonzero and continuous, and then the equation is called **non-homogeneous**. The **characteristic equation** is

$$a_n r^n + \cdots + a_0 = 0.$$

It is obtained from Euler's substitution $y = e^{rx}$ or by the shortcut substitutions $y^{(k)} \to r^k$. The left side of the characteristic equation is called the **characteristic polynomial**.

Picard-Lindelöf Theorem

The foundation of the theory of linear constant coefficient differential equations is the existence-uniqueness result of Picard-Lindelöf, which says that, given constants g_1, ..., g_n, the initial value problem

$$\begin{aligned} &a_n y^{(n)} + \cdots + a_0 y = f(x), \\ &y(0) = g_1, \ldots, y^{n-1}(0) = g_n, \end{aligned}$$

has a unique solution $y(x)$ defined on each open interval for which $f(x)$ is defined and continuous.

General Solution

A linear homogeneous constant coefficient differential equation has a general solution $y_h(x)$ written in terms of n arbitrary constants $c_1, \ldots, c_n$ and n solutions $y_1(x)$, ..., $y_n(x)$ as the linear combination

$$y_h(x) = c_1 y_1(x) + \cdots + c_n y_n(x).$$

Discussed here is one way to define the solutions y_1, ..., y_n.

Consider the case of $n = 2$, already discussed. The Picard-Lindelöf theorem applies with initial values $y(0) = 1$, $y'(0) = 0$ to define solution $y_1(x)$. The initial values are changed to $y(0) = 0$, $y'(0) = 1$, then Picard-Lindelöf applies again to define solution $y_2(x)$. Solution $y(x) = c_1 y_1(x) + c_2 y_2(x)$ satisfies initial conditions $y(0) = g_1$, $y'(0) = g_2$ when $c_1 = g_1, c_2 = g_2$.

In the $n = 2$ case, solutions y_1, y_2 are defined using initial conditions which form the columns of the 2×2 identity matrix. In a similar way, for general n, solutions $y_1(x)$, …, $y_n(x)$ are defined by applying the Picard-Lindeöf theorem, with initial conditions g_1, …, g_n successively taken as the columns of the $n \times n$ identity matrix.

The expression $y_h(x)$ is called a general solution, because any solution of the differential equation is equal to $y_h(x)$ for a unique specialization of the constants c_1, …, c_n.

Solution Structure

An **Euler base atom** is one of the functions

$$e^{ax}, e^{ax}\cos bx, e^{ax}\sin bx$$

where a and b are real numbers, $b > 0$.

An **Euler solution atom** is a power x^n times a base atom, where $n \geq 0$ is an integer.

Complex Numbers and Atoms. An Euler solution atom can alternatively be defined as the nonzero real or imaginary part of $x^n e^{rx}$ where $r = a + ib$ with symbols a and $b \geq 0$ are real and $n \geq 0$ is an integer, provided minus signs are stripped off, leaving coefficient 1. Euler's formula

$$e^{i\theta} = \cos\theta + i\sin\theta$$

facilitates taking real and imaginary parts of the complex exponential term $x^n e^{rx}$. For instance,

$$x^7 e^{(2+3i)x} = x^7 e^{2x}\cos 3x + ix^7 e^{2x}\sin 3x$$

has real and imaginary parts $x^7 e^{2x}\cos 3x$, $x^7 e^{2x}\sin 3x$, which are themselves atoms.

A complete list of all possible atoms appears in the rightmost section of the table below, in which a, b are real, $b > 0$ and $n \geq 0$ is an integer.

$r = 0$	$1,$	$x,$	$x^2,$	$\ldots,$	$x^n,$	$\ldots$
$r = a$	$e^{ax},$	$xe^{ax},$	$x^2e^{ax},$	$\ldots,$	$x^ne^{ax},$	$\ldots$
$r = ib$	$\cos bx,$	$x\cos bx,$	$x^2\cos bx,$	$\ldots,$	$x^n\cos bx,$	$\ldots$
$r = ib$	$\sin bx,$	$x\sin bx,$	$x^2\sin bx,$	$\ldots,$	$x^n\sin bx,$	$\ldots$
$r = a + ib$	$e^{ax}\cos bx,$	$xe^{ax}\cos bx,$	$x^2e^{ax}\cos bx,$	$\ldots,$	$x^ne^{ax}\cos bx,$	$\ldots$
$r = a + ib$	$e^{ax}\sin bx,$	$xe^{ax}\sin bx,$	$x^2e^{ax}\sin bx,$	$\ldots,$	$x^ne^{ax}\sin bx,$	$\ldots$

The table only uses $b > 0$, because Euler atoms must have coefficient 1. For instance, $xe^{(1-2i)x} = xe^x\cos 2x - ixe^x\sin 2x$ does not have atoms for real and imaginary parts (coefficient -1 is the problem). Yes, stripping the minus sign gives $xe^x\sin 2x$, which is an atom (coefficient 1).

Detecting Euler Solution Atoms

A term that makes up an atom has coefficient 1, therefore 2 and $2e^x$ are not atoms, but the 2 can be stripped off to expose atoms 1 and e^x. Combinations like $2x + 3x^2$ are not atoms, but individual stripped terms x and x^2 are atoms. Terms like e^{x^2}, $\ln|x|$ and $x/(1+x^2)$ are not atoms, nor are they sums of constants times atoms. The expressions $\cosh x$, $\sinh x$ and $\sin^4 x$ are not atoms, but they are combinations of atoms. Fractional powers may not appear in atoms, for instance, neither x^π nor $x^{5/2}\sin x$ is an atom.

Linear Algebra Background

Borrowed from the subject of linear algebra is the terminology **linear combination**, which in the case of two functions f_1, f_2 is the expression $f = c_1 f_1 + c_2 f_2$. More generally, given functions $f_1, \ldots, f_k$, and constants $c_1, \ldots, c_k$, the expression $f = c_1 f_1 + \cdots + c_k f_k$ is called a linear combination of the functions $f_1, \ldots,$ f_k.

A function list $f_1, \ldots, f_k$ is called **linearly independent** provided every linear combination is uniquely represented by the constants $c_1, \ldots, c_k$.

Independence is tested by solving for constants $c_1, \ldots, c_k$ in the equation $c_1 f_1(x) + \cdots + c_k f_k(x) = 0$, assumed satisfied for all x in a common domain of $f_1, \ldots, f_k$. Independence holds if and only if the constants are all zero.

Theorem 6.11 (Independence and Euler Solution Atoms)
A list of finitely many distinct Euler solution atoms is linearly independent.

Outline of the proof on page 398.

Because subsets of independent sets are independent, then list x^2, x^5, x^8 is independent by virtue of independence of the powers $1, x, \ldots, x^n$.

Solution methods for linear constant differential equations implicitly use Theorem 6.11.

Fundamental Results

Theorem 6.12 (Homogeneous Solution y_h and Atoms)
Linear homogeneous differential equations with constant coefficients have general solution $y_h(x)$ equal to a linear combination of Euler atoms.

Theorem 6.13 (Particular Solution y_p and Atoms)
Linear non-homogeneous differential equations with constant coefficients having forcing term $f(x)$ equal to a linear combination of atoms have a particular solution $y_p(x)$ which is a linear combination of Euler atoms.

Theorem 6.14 (General Solution y and Atoms)
Linear non-homogeneous differential equations with constant coefficients having forcing term $f(x) =$ a linear combination of Euler atoms have general solution

$$y(x) = y_h(x) + y_p(x) = \text{ a linear combination of Euler atoms.}$$

The first result, for the special case of second order differential equations, can be justified from Theorem 6.1, page 431. The solutions $e^{r_1 x}$, $e^{r_2 x}$, $xe^{r_1 x}$, $e^{ax}\cos bx$ and $e^{ax}\sin bx$ in the theorem are Euler atoms.

The third theorem easily follows from the first two. The first and second theorems follow directly from Euler's Theorem 6.15 and the method of undetermined coefficients, *infra*.

How to Solve Equations of Order n

Picard's existence–uniqueness theorem says that $y''' + 2y'' + y = 0$ has general solution y constructed from linear combinations of 3 independent solutions of this differential equation. The general solution of an n-th order linear differential equation is constructed from linear combinations of n independent solutions of the equation.

Linear algebra defines the dimension of the solution set to be this same fixed number n. Once n independent solutions are found for the differential equation, the search for the general solution has ended: the general solution y must be a linear combination of these n independent solutions.

Because of the preceding structure theorems, we have reduced the search for the general solution to the following:

> Find n distinct Euler solution atoms of the nth order differential equation.

Euler's basic result tells us how to find the list of distinct atoms.

Theorem 6.15 (Euler's Theorem)
Assume r_0 is a real or complex root of the characteristic equation. If complex, write $r_0 = a + ib$ with a, b real.

(a) The functions $e^{r_0 x}$, $xe^{r_0 x}$, ..., $x^k e^{r_0 x}$ are solutions of a linear homogeneous constant–coefficient differential equation if and only if $(r - r_0)^{k+1}$ is a factor of the characteristic polynomial.

(b) Assume $b > 0$. Functions $e^{ax}\cos bx$, $xe^{ax}\cos bx$, ..., $x^k e^{ax}\cos bx$, $e^{ax}\sin bx$, $xe^{ax}\sin bx$, ..., $x^k e^{ax}\sin bx$ (a, b real, $b > 0$) are solutions of a linear homogeneous constant–coefficient differential equation if and only if $((r-a)^2 + b^2)^{k+1}$ is a factor of the characteristic polynomial.

Proof on page 459.

Theorem 6.16 (Real and Complex Solutions)
Let $y(x) = u(x) + iv(x)$ be a solution of a linear constant-coefficient differential equation (a_0,…,a_n assumed real), with $u(x)$ and $v(x)$ both real. Then $u(x)$ and $v(x)$ are both real solutions of the differential equation. Briefly stated, the real and imaginary parts of a solution are also solutions.

Proof on page 460

Root Multiplicity

A polynomial equation $p(r) = 0$ is defined in college algebra to have a root $r = r_0$ of multiplicity m provided $(r - r_0)^m$ divides $p(r)$ but $(r - r_0)^{m+1}$ does not. For instance, $(r-1)^3(r+2)(r^2+4)^2 = 0$ has roots 1, -2, $2i$, $-2i$ of multiplicity 3, 1, 2, 2, respectively.

Atom Lists

Let $r = r_0$ be a real root of the characteristic equation $p(r) = 0$, of multiplicity $k+1$. Then Euler's theorem finds a base atom solution $e^{r_0 x}$. A total of $k+1$ solutions are obtained from this base atom by **multiplying the base atom** by the powers $1, x, \ldots, x^k$:

$$e^{r_0 x}, \quad xe^{r_0 x}, \quad \ldots, \quad x^k e^{r_0 x}.$$

A special case occurs when $r_0 = 0$. Then $e^{0x} = 1$ is the base atom and the $k+1$ solution atoms are the powers

$$1, \quad x, \quad \ldots, \quad x^k.$$

The number of Euler solution atoms expected for a given root $r = r_0$ equals the multiplicity of the root r_0.

Let $r = a + ib$ be a complex root of the characteristic equation $p(r) = 0$, of multiplicity $k+1$. Euler's Theorem implies that e^{ax+ibx} is a solution, and the theorem on complex solutions implies that the differential equation has two base solution atoms

$$e^{ax}\cos bx, \quad e^{ax}\sin bx.$$

Euler's Theorem implies that we should multiply these base atoms by powers 1, $x, \ldots, x^k$ to obtain $k+1$ solution atoms for each of the base atoms, giving the atom list for a complex root

$$\begin{array}{llll} e^{ax}\cos(bx), & xe^{ax}\cos(bx), & \ldots, & x^k e^{ax}\cos(bx), \\ e^{ax}\sin(bx), & xe^{ax}\sin(bx), & \ldots, & x^k e^{ax}\sin(bx). \end{array}$$

A special case occurs when $a = 0$. Then the base atoms are pure harmonics $\cos bx$, $\sin bx$ and the list has no visible exponentials:

$$\begin{array}{llll}\cos(bx), & x\cos(bx), & \ldots, & x^k\cos(bx), \\ \sin(bx), & x\sin(bx), & \ldots, & x^k\sin(bx).\end{array}$$

Shortcut Explained. A remaining mystery is the skipped complex root $a - ib$. We explain why we focused on $a + ib$ with $b > 0$ and ignored its conjugate $a - ib$. Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$ using $\theta = rx = ax + ibx$ implies

$$x^j e^{rx} = \left(x^j e^{ax}\cos(bx)\right) + i\left(x^j e^{ax}\sin(bx)\right).$$

The real and imaginary parts of this complex linear combination are Euler atoms. If r is replaced by its complex conjugate $\overline{r} = a - ib$, then the same two atoms are distilled from the linear combination. Picard's Theorem dictates that we find $2k+2$ atoms from the pair of roots $a \pm ib$. Because the process above finds $2k+2$ atoms, the second conjugate root is ignored, as a *shortcut*.

Examples and Methods

Example 6.13 (First Order)
Solve $2y' + 5y = 0$, showing $y_h = c_1e^{-5x/2}$.

Solution: Euler's Theorem 6.15 will be applied. The characteristic equation is $2r+5 = 0$ with real root $r = -5/2$. The corresponding atom e^{rx} is given explicitly by $e^{-5x/2}$. Because the order of the differential equation is 1, then all atoms have been found. Write the general solution y_h by multiplying the atom list by constant c_1, then $y_h = c_1e^{-5x/2}$.

Example 6.14 (Second Order Distinct Real Roots)
Solve $y'' + 3y' + 2y = 0$, showing $y_h = c_1e^{-x} + c_2e^{-2x}$.

Solution: The factored characteristic equation is $(r + 1)(r + 2) = 0$. The distinct real roots are $r_1 = -1$, $r_2 = -2$. Euler's Theorem 6.15 applies to find the atom list e^{-x}, e^{-2x}. All atoms have been found, because the order of the differential equation is 2. The general solution y_h is written by multiplying the atom list by constants c_1, c_2, then $y_h = c_1e^{-x} + c_2e^{-2x}$.

Example 6.15 (Second Order Double Real Root)
Solve $y'' + 2y' + y = 0$, showing $y_h = c_1e^{-x} + c_2xe^{-x}$.

Solution: The factored characteristic equation is $(r+1)(r+1) = 0$, with double real root $r = -1, -1$. The root multiplicity is 2, so we must find two atoms for the root $r = -1$. Euler's Theorem 6.15 applies to find a base atom e^{-x}. Multiply the base atom by 1, x to find two solution atoms e^{-x}, xe^{-x}. Because the order of the differential equation is 2, then all atoms have been found. Write the general solution y_h by multiplying the atom list by constants c_1, c_2, then $y_h = c_1e^{-x} + c_2xe^{-x}$.

Example 6.16 (Second Order Complex Conjugate Roots)

Solve the differential equation $y'' + 2y' + 5y = 0$, verifying the equation $y_h = c_1 e^{-x} \cos 2x + c_2 e^{-x} \sin 2x$.

Solution: The characteristic equation $r^2 + 2r + 5 = 0$ factors into $(r+1)^2 + 4 = 0$, therefore it has complex conjugate roots $r_1 = -1 + 2i$, $r_2 = -1 - 2i$. There are two methods for finding the atoms associated with these roots. We discuss both possibilities.

Method 1. The first statement in Euler's Theorem 6.15 applies to report two complex solutions e^{-x+2xi}, e^{-x-2xi}. These solutions are not atoms, but linear combinations of atoms, from which a list of two atoms is determined. The atoms are $e^{-x} \cos 2x$, $e^{-x} \sin 2x$. This process uses the two identities

$$e^{i\theta} = \cos\theta + i \sin\theta, \quad e^{-i\theta} = \cos\theta - i \sin\theta.$$

Write

$$\begin{aligned} e^{-x+2xi} &= \left(e^{-x} \cos 2x\right) + i \left(e^{-x} \sin 2x\right), \\ e^{-x-2xi} &= \left(e^{-x} \cos 2x\right) - i \left(e^{-x} \sin 2x\right), \end{aligned}$$

then extract the two distinct atoms that appear in these two linear combinations:

$$e^{-x} \cos 2x, \quad e^{-x} \sin 2x.$$

Method 2. The second statement in Euler's Theorem 6.15 is more efficient. Characteristic equation root $r = -1 + 2i$ was found from the factorization $(r+1)^2 + 4 = 0$, which by Euler's theorem implies there are two distinct solution atoms

$$e^{-x} \cos 2x, \quad e^{-x} \sin 2x.$$

General Solution. Because the order of the differential equation is 2, then all atoms have been found. Write the general solution y_h by multiplying the atom list by constants c_1, c_2, then $y_h = c_1 e^{-x} \cos 2x + c_2 e^{-x} \sin 2x$.

The example uses a **shortcut**. Euler's theorem applied to the second conjugate root $-1 - 2i$ will produce no new atoms. The step of finding the distinct atoms can be shortened by observing that the outcome is exactly the real and imaginary parts of the first complex exponential e^{ax+ibx} with $b > 0$. The preferred method for finding the atoms is to use the second statement in Euler's theorem.

Example 6.17 (Third Order Distinct Roots)

Solve $y''' - y' = 0$, showing $y_h = c_1 + c_2 e^x + c_3 e^{-x}$.

Solution: The factored characteristic equation is $r(r-1)(r+1) = 0$ with real roots $r_1 = 0$, $r_2 = 1$, $r_3 = -1$. Euler's Theorem 6.15 applies to report the atom list e^{0x}, e^x, e^{-x}. The general solution y_h is written by multiplying the atom list by constants c_1, c_2, c_3, giving $y_h = c_1 e^{0x} + c_2 e^x + c_3 e^{-x}$. Convention replaces e^{0x} by 1 in the final equation.

Example 6.18 (Third Order with One Double Root)

Solve $y''' - y'' = 0$, verifying that $y_h = c_1 + c_2 x + c_3 e^x$.

Solution: The characteristic equation is $r^3 - r^2 = 0$. It factors into $r^2(r-1) = 0$ with real roots $r_1 = 0$, $r_2 = 0$, $r_3 = 1$. Euler's Theorem 6.15 applies to find the base atom list e^{0x}, e^x. Because root $r = 0$ has multiplicity 2, we must multiply base atom e^{0x} by 1 and x to find the required 2 atoms e^{0x}, xe^{0x}. Then the completed list of 3 atoms is 1, x, e^x. The general solution y_h is written by multiplying the atom list by constants c_1, c_2, c_3 to give $y_h = c_1 + c_2 x + c_3 e^x$.

Example 6.19 (Fourth Order)

Solve $y^{iv} - y'' = 0$, showing $y_h = c_1 + c_2x + c_3e^x + c_4e^{-x}$.

Solution: Notation: Define $y^{iv} = \dfrac{d^4y}{dx^4}$, the fourth derivative of y. The factored characteristic equation is $r^2(r-1)(r+1) = 0$ with real roots $r_1 = 0$, $r_2 = 0$, $r_3 = 1$, $r_4 = -1$. Euler's Theorem 6.15 applies to obtain the base atom list e^{0x}, e^x, e^{-x}. The first base atom e^{0x} comes from root $r = 0$, which has multiplicity 2. Euler's Theorem requires that this base atom be multiplied by 1, x. The atom list of 4 atoms is then 1, x, e^x, e^{-x}. All atoms have been found, because the order of the differential equation is 4. The general solution y_h is written by multiplying the atom list by constants c_1, c_2, c_3, c_4 to obtain the general solution $y_h = c_1 + c_2x + c_3e^x + c_4e^{-x}$.

Example 6.20 (Tenth Order)

A linear homogeneous constant coefficient differential equation has characteristic equation

$$r^2(r-1)^2(r^2-1)(r^2+1)^2 = 0.$$

Solve the differential equation, showing that

$$\begin{aligned} y_h &= c_1 + c_2x + c_3e^x + c_4xe^x + c_5x^2e^x + c_6e^{-x} \\ &\quad +c_7\cos x + c_8x\cos x + c_9\sin x + c_{10}x\sin x. \end{aligned}$$

Solution: The factored form of the characteristic equation is

$$r^3(r-1)^2(r-1)(r+1)(r-i)^2(r+i)^2 = 0.$$

The roots, listed according to multiplicity, make the *list of roots*

$$L = \{0, 0, \quad 1, 1, 1, \quad -1, \quad i, i, \quad -i, -i\}.$$

There are two methods for finding the atoms from list L.

Method 1. The first statement in Euler's theorem gives the exponential-type solutions

$$e^{0x}, xe^{0x}, e^x, xe^x, x^2e^x, e^{-x}, e^{ix}, xe^{ix}, e^{-ix}, xe^{-ix}.$$

The first six in the list are atoms, but the last four are not. Because $e^{ix} = \cos x + i\sin x$, we can distill from the complex exponentials the additional four atoms $\cos x$, $x\cos x$, $\sin x$, $x\sin x$. Then the *list of* 10 *distinct atoms* is

$$1, x, e^x, xe^x, x^2e^x, e^{-x}, \cos x, x\cos x, \sin x, x\sin x.$$

Method 2. The above list can be obtained directly from the second statement in Euler's theorem. The real exponential atoms are obtained from the first statement in Euler's theorem:

$$1, x, e^x, xe^x, x^2e^x, e^{-x}.$$

The second statement of Euler's theorem applies to the complex factor $(r^2+1)^2$ to obtain the trigonometric atoms

$$\cos x, x\cos x, \sin x, x\sin x.$$

General Solution. Then y_h is a linear combination of the 10 atoms:

$$\begin{aligned} y_h &= c_1 + c_2x + c_3e^x + c_4xe^x + c_5x^2e^x + c_6e^{-x} \\ &\quad +c_7\cos x + c_8x\cos x + c_9\sin x + c_{10}x\sin x. \end{aligned}$$

Example 6.21 (Differential Equation from General Solution)
A linear homogeneous constant coefficient differential equation has general solution

$$y_h = c_1 + c_2 x + c_3 e^x + c_4 x e^x + c_5 x^2 e^x + c_6 \cos x + c_7 \sin x.$$

Find the differential equation.

Solution: Take the partial derivative of y_h with respect to the symbols c_1, …, c_7 to give the atom list

$$1, x, \quad e^x, xe^x, x^2e^x, \quad \cos x, \sin x.$$

This atom list is constructed from exponential solutions obtained from Euler's theorem, applied to the root list

$$0, 0, \quad 1, 1, 1, \quad i, -i.$$

There are 7 roots, hence by the root-factor theorem of college algebra the characteristic polynomial has individual factors r, r, $r-1$, $r-1$, $r-1$, $r-i$, $r+i$. Then the differential equation is of order 7 with characteristic polynomial

$$\begin{aligned} p(r) &= (r-0)^2(r-1)^3(r-i)(r+i) \\ &= r^6 - 2r^5 + 2r^4 - 2r^3 + r^2. \end{aligned}$$

The differential equation is obtained by the translation $r^j \to y^{(j)}$:

$$y^{(6)} - 2y^{(5)} + 2y^{(4)} - 2y''' + y'' = 0.$$

Proofs and Details

Proof of Euler's Theorem 6.15: The first statement will be proved for $n = 2$. The details for the general case are left as an exercise.

Let $y = e^{rx}$. Then

$$y = e^{rx}, \quad y' = re^{rx}, \quad y'' = r^2e^{rx}.$$

Substitute into the differential equation to obtain the following.

$$a_2y'' + a_1y' + a_0y = 0$$
$$a_2r^2e^{rx} + a_1re^{rx} + a_0e^{rx} = 0$$
$$\left(a_2r^2 + a_1r + a_0\right)e^{rx} = 0$$

Then $y = e^{rx}$ is a solution if and only if $a_2r^2 + a_1r + a_0 = 0$, that is, the characteristic equation is satisfied.

To prove the second statement, assume a differential equation of order n

$$a_ny^{(n)} + \cdots + a_0y = 0.$$

Perform a change of variables $y = e^{cx}z$, which changes dependent variable y into z. If y is a solution, then

$$y = e^{cx}z, \quad y' = ce^{cx}z + e^{cx}z', \quad y'' = c^2e^{cx}z + 2ce^{cx}z' + e^{cx}z'', \cdots$$

Because each derivative of y is a multiple of e^{cx}, then, after substitution of the relations into the differential equation, the common factor e^{cx} cancels, giving a new constant coefficient differential equation for z.

To illustrate, in the case $n = 2$, the new differential equation for z is

$$a_2 z'' + (2a_2 c + a_1)z' + (a_2 c^2 + a_1 c + a_0)z = 0.$$

The coefficients of the z-equation are the Taylor series coefficients $\dfrac{p^k(0)}{k!}$ of the characteristic polynomial $p(r) = a_2 r^2 + a_1 r + a_0$:

$$\begin{aligned} a_2 &= \frac{p''(c)}{2!}, \\ (2a_2 c + a_1) &= \frac{p'(c)}{1!}, \\ (a_2 c^2 + a_1 c + a_0) &= \frac{p(c)}{0!}. \end{aligned}$$

By induction, the change of variables $y = e^{cx}z$ produces from $a_n y^{(n)} + \cdots + a_0 y = 0$ a new constant-coefficient differential equation $b_n z^{(n)} + \cdots + b_0 z = 0$ whose coefficients are given by

$$b_k = \frac{p^k(c)}{k!}.$$

Assume now characteristic polynomial $p(r) = a_n r^n + \cdots + a_0$ and let $r = c$ be a root of $p(r) = 0$ of algebraic multiplicity $k + 1$. Then $p(c) = p'(c) = \cdots = p^{(k)}(c) = 0$. This means that $b_0 = \cdots = b_k = 0$. Therefore, the z-equation is a differential equation in the variable $v = z^{(k+1)}$. Because the selections $z = 1, x, \ldots, x^k$ all imply $v = 0$, then the polynomials 1, x, $\ldots$, x^k are solutions of the z-equation. Hence, $y = e^{cx}z$ implies e^{cx}, xe^{cx}, $\ldots$, $x^k e^{cx}$ are solutions of the y-equation.

Conversely, assume that e^{cx}, xe^{cx}, $\ldots$, $x^k e^{cx}$ are solutions of the y-equation. We will verify that $r = c$ is a root of $p(r) = 0$ of algebraic multiplicity $k + 1$. First, 1, $\ldots$, x^k are solutions of the z-equation. Setting $z = 1$ implies $b_0 = 0$ Then setting $z = x$ implies $b_1 = 0$ (because $b_0 = 0$ already). Proceeding in this way, $b_0 = \cdots = b_k = 0$. Therefore, the characteristic polynomial of the z-equation is

$$q(r) = b_n r^n + \cdots + b_{k+1} r^{k+1}.$$

The reader can prove the following useful result; see the exercises.

Lemma 6.1 (Kümmer's Lemma) Under the change of variables $y = e^{cx}z$, the characteristic polynomials $p(r)$, $q(r)$ of the y-equation and the z-equation, respectively, satisfy the relation $q(r) = p(r + c)$.

Assuming Kümmer's Lemma, we can complete the proof. Already, we know that r^{k+1} divides $q(r)$. Then r^{k+1} divides $p(r + c)$, or equivalently, $(r - c)^{k+1}$ divides $p(r)$. This implies $r = c$ is a root of $p(r) = 0$ of algebraic multiplicity $k + 1$. ■

Proof of Theorem 6.16: Substitute $y = u + iv$ into the differential equation and separate terms as follows:

$$(a_n u^{(n)} + \cdots + a_0 u) + i(a_n v^{(n)} + \cdots + a_0 v) = 0.$$

For each x, the left side of the preceding relation is a complex number $a + ib$ with a, b real. The right side is $0 + 0i$. By equality of complex numbers, $a = 0$ and $b = 0$, which implies

$$a_n u^{(n)} + \cdots + a_0 u = 0,$$
$$a_n v^{(n)} + \cdots + a_0 v = 0.$$

Therefore, u and v are real solutions of the differential equation. ■

Exercises 6.3

Constant Coefficients

Solve for $y(x)$. Proceed as in Examples 6.13–6.20.

1. $3y' - 2y = 0$

2. $2y' + 7y = 0$

3. $y'' - y' = 0$

4. $y'' + 2y' = 0$

5. $y'' - y = 0$

6. $y'' - 4y = 0$

7. $y'' + 2y' + y = 0$

8. $y'' + 4y' + 4y = 0$

9. $y'' + 3y' + 2y = 0$

10. $y'' - 3y' + 2y = 0$

11. $y'' + y = 0$

12. $y'' + 4y = 0$

13. $y'' + y' + y = 0$

14. $y'' + 2y' + 2y = 0$

15. $y'' = 0$

16. $y''' = 0$

17. $\frac{d^4y}{dx^4} = 0$

18. $\frac{d^5y}{dx^5} = 0$

19. $y''' + 2y'' = 0$

20. $y''' + 4y' = 0$

21. $\frac{d^4y}{dx^4} + y'' = 0$

22. $\frac{d^5y}{dx^5} + y''' = 0$

Detecting Atoms

Decompose each atom into a base atom times a power of x. If the expression fails to be an atom, then explain the failure.

23. $-x$

24. x

25. $x^2 \cos \pi x$

26. $x^{3/2} \cos x$

27. $x^{1000} e^{-2x}$

28. $x + x^2$

29. $\dfrac{x}{1 + x^2}$

30. $\ln |xe^{2x}|$

31. $\sin x$

32. $\sin x - \cos x$

Higher Order

A homogeneous linear constant-coefficient differential equation can be defined by (1) coefficients, (2) the characteristic equation, (3) roots of the characteristic equation. In each case, solve the differential equation.

33. $y''' + 2y'' + y' = 0$

34. $y''' - 3y'' + 2y' = 0$

35. $y^{(4)} + 4y'' = 0$

36. $y^{(4)} + 4y''' + 4y'' = 0$

37. Order 5, $r^2(r - 1)^3 = 0$

38. Order 5, $(r^3 - r^2)(r^2 + 1) = 0$.

39. Order 6, $r^2(r^2 + 2r + 2)^2 = 0$.

40. Order 6, $(r^2 - r)(r^2 + 4r + 5)^2 = 0$.

41. Order 10, $(r^4+r^3)(r^2-1)^2(r^2+1) = 0$.

42. Order 10, $(r^3+r^2)(r-1)^3(r^2+1)^2 = 0$.

43. Order 5, roots $r = 0, 0, 1, 1, 1$.

44. Order 5, roots $r = 0, 0, 1, i, -i$.

45. Order 6, roots $r = 0, 0, i, -i, i, -i$.

46. Order 6, roots $r = 0, -1, 1 + i, 1 - i, 2i, -2i$.

47. Order 10, roots $r = 0, 0, 0, 1, 1, -1, -1, -1, i, -i$.

48. Order 10, roots $r = 0, 0, 1, 1, 1, -1, i, -i, i, -i$.

Initial Value Problems

Given in each case is a set of independent solutions of the differential equation. Solve for the coefficients c_1, c_2, … in the general solution, using the given initial conditions.

49. e^x, e^{-x}, $y(0) = 0$, $y'(0) = 1$

50. xe^x, e^x, $y(0) = 1$, $y'(0) = -1$

51. $\cos x, \sin x$, $y(0) = -1$, $y'(0) = 1$

52. $\cos 2x, \sin 2x$, $y(0) = 1$, $y'(0) = 0$

53. $e^x, \cos x, \sin x$, $y(0) = -1$, $y'(0) = 1$, $y''(0) = 0$

54. $1, \cos x, \sin x$, $y(0) = -1$, $y'(0) = 1$, $y''(0) = 0$

55. $e^x, xe^x, \cos x, \sin x$, $y(0) = -1$, $y'(0) = 1$, $y''(0) = 0$, $y'''(0) = 0$

56. $1, x, \cos x, \sin x$, $y(0) = 1$, $y'(0) = -1$, $y''(0) = 0$, $y'''(0) = 0$

57. $1, x, x^2, x^3, x^4$, $y(0) = 1$, $y'(0) = 2$, $y''(0) = 1$, $y'''(0) = 3$, $y^{(4)}(0) = 0$

58. $e^x, xe^x, x^2e^x, 1, x$, $y(0) = 1$, $y'(0) = 0$, $y''(0) = 1$, $y'''(0) = 0$, $y^{(4)}(0) = 0$

Inverse Problem

Find a linear constant-coefficient homogeneous differential equation from the given information. Follow Example 6.21.

59. The characteristic equation is $(r + 1)^3(r^2 + 4) = 0$.

60. The general solution is a linear combination of the Euler solution atoms $e^x, e^{2x}, e^{3x}, \cos x, \sin x$.

61. The roots of the characteristic polynomial are $0, 0, 2 + 3i, 2 - 3i$.

62. The equation has order 4. Known solutions are $e^x + 4 \sin 2x$, xe^x.

63. The equation has order 10. Known solutions are $\sin 2x$, x^7e^x.

64. The equation is $my''+cy'+ky = 0$ with $m = 1$ and c, k positive. A solution is $y(x) = e^{-x/5} \cos(2x - \theta)$ for some angle θ.

Independence of Euler Atoms

65. Apply the independence test page 378 to atoms 1 and x: form equation $0 = c_1 + c_2x$, then solve for $c_1 = 0$, $c_2 = 0$. This proves Euler atoms $1, x$ are independent.

66. Show that Euler atoms $1, x, x^2$ are independent using the independence test page 378,

67. A Taylor series is zero if and only if its coefficients are zero. Use this result to give a complete proof that the list $1, \ldots,$ x^k is independent. Hint: a polynomial is a Taylor series.

68. Show that Euler atoms e^x, xe^x, x^2e^x are independent using the independence test page 378.

Wronskian Test

Establish independence of the given lists of functions by using the Wronskian test page 385:

Functions $f_1, f_2, \ldots, f_n$ are independent if $W(x_0) \ne 0$ for some x_0, where $W(x)$ is the $n \times n$ determinant

$$\begin{vmatrix} f_1(x) & \cdots & f_n(x) \\ & \vdots & \\ f_1^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}$$

69. $1, x, e^x$

70. $1, x, x^2, e^x$

71. $\cos x, \sin x, e^x$

72. $\cos x, \sin x, \sin 2x$

Kümmer's Lemma

73. Compute the characteristic polynomials $p(r)$ and $q(r)$ for

$$\begin{aligned}&y'' + 3y' + 2y = 0 \text{ and}\\ &z'' + z' = 0.\end{aligned}$$

Verify the equations are related by $y = e^{-x}z$ and $p(r-1) = q(r)$.

74. Compute the characteristic polynomials $p(r)$ and $q(r)$ for

$$\begin{aligned}&ay'' + by' + cy = 0 \text{ and}\\ &az'' + (2ar_0 + b)z' +\\ &\quad (ar_0^2 + br_0 + c)z = 0.\end{aligned}$$

Verify the equations are related by $y = e^{r_0 x}z$ and $p(r + r_0) = q(r)$.

6.4 Variation of Parameters

The **Method of Variation of Parameters** applies to solve

$$a(x)y'' + b(x)y' + c(x)y = f(x). \tag{1}$$

Continuity of a, b, c and f is assumed, plus $a(x) \neq 0$. The method is important because it solves the largest class of equations. Specifically **included** are functions $f(x)$ like $\ln|x|$, $|x|$, e^{x^2}, $x/(1+x^2)$, which are excluded in the method of undetermined coefficients.

Homogeneous Equation

The method of variation of parameters uses facts about the homogeneous differential equation

$$a(x)y'' + b(x)y' + c(x)y = 0. \tag{2}$$

Success in the method depends upon a general solution expression for (2). Assumed are two *known solutions* y_1, y_2, Symbols c_1, c_2 represent arbitrary constants. The general solution:

$$y = c_1y_1(x) + c_2y_2(x) \tag{3}$$

If a, b, c are constants, then Theorem 6.1, page 431, applied to (2) implies y_1 and y_2 can be selected as *independent Euler solution atoms.*

Independence

Two solutions y_1, y_2 of (2) are called **independent** if neither is a constant multiple of the other. The term **dependent** means *not independent*, in which case either $y_1(x) = cy_2(x)$ or $y_2(x) = cy_1(x)$ holds for all x, for some constant c. Independence can be tested through the **Wronskian determinant** of y_1, y_2, defined by

$$W(x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1(x)y_2'(x) - y_1'(x)y_2(x).$$

Theorem 6.17 (Wronskian and Independence)
The Wronskian of two solutions satisfies $a(x)W' + b(x)W = 0$, which implies **Abel's identity**

$$W(x) = W(x_0)e^{-\int_{x_0}^{x}(b(t)/a(t))dt}.$$

Two solutions of (2) are independent if and only if $W(x) \neq 0$.

Proof on page 466.

Niels Henrik Abel (1802–1829) was born in Nedstrand, Norway. He made major contributions to mathematics, especially *elliptic functions*, dying from tuberculosis at age 26.

Theorem 6.18 (Variation of Parameters Formula)
Let a, b, c, f be continuous near $x = x_0$ and $a(x) \neq 0$. Let y_1, y_2 be two independent solutions of homogeneous equation $a(x)y'' + b(x)y' + c(x)y = 0$ and let $W(x) = y_1(x)y_2'(x) - y_1'(x)y_2(x)$. Then the non-homogeneous differential equation

$$a(x)y'' + b(x)y' + c(x)y = f$$

has a particular solution

$$(4) \qquad y_p(x) = \left(\int \frac{y_2(x)(-f(x))}{a(x)W(x)}dx\right) y_1(x) + \left(\int \frac{y_1(x)f(x)}{a(x)W(x)}dx\right) y_2(x).$$

If both integrals have limits x_0 and x, then $y_p(x_0) = 0$.

Proof on page 467.

History of Variation of Parameters

The solution y_p was discovered by varying the constants c_1, c_2 in the homogeneous solution $y_h = c_1y_1 + c_2y_2$, assuming c_1, c_2 depend on x. This results in formulas $c_1(x) = \int C_1F$, $c_2(x) = \int C_2F$ where $F(x) = f(x)/a(x)$, $C_1(t) = \dfrac{-y_2(t)}{W(t)}$,$C_2(t) = \dfrac{y_1(t)}{W(t)}$; see the historical details on page 467. Then

$$\begin{aligned}
y &= c_1y_1(x) + c_2y_2(x) && \text{Formula for } y_h. \\
y &= \left(\int C_1F\right) y_1(x) + \left(\int C_2F\right) y_2(x) && \text{Substitute for } c_1,\ c_2. \\
&= \left(\int -y_2\frac{F}{W}\right) y_1(x) + \left(\int y_1\frac{F}{W}\right) y_2(x) && \text{Use (2) for } C_1,\ C_2. \\
&= \int (y_2(x)y_1(t) - y_1(x)y_2(t))\frac{F(t)}{W(t)}dt && \text{Collect on } F/W. \\
&= \int \frac{y_1(t)y_2(x) - y_1(x)y_2(t)}{y_1(t)y_2'(t) - y_1'(t)y_2(t)}F(t)dt && \text{Expand } W = y_1y_2' - y_1'y_2.
\end{aligned}$$

Any one of the last three equivalent formulas is called a **Classical variation of parameters formula**. The fraction in the last integrand is called Cauchy's kernel. We prefer the first, equivalent to equation (4), for ease of use.

Examples and Methods

Example 6.22 (Independence)
Consider $y'' - y = 0$. Show the two solutions $\sinh(x)$ and $\cosh(x)$ are independent using Wronskians.

Solution: Let $W(x)$ be the Wronskian of $\sinh(x)$ and $\cosh(x)$. The calculation below shows $W(x) = -1$. By Theorem 6.17, the solutions are independent.

Background. The calculus *definitions* for hyperbolic functions are $\sinh x = (e^x - e^{-x})/2$, $\cosh x = (e^x + e^{-x})/2$. Their derivatives are $(\sinh x)' = \cosh x$ and $(\cosh x)' = \sinh x$. For instance, $(\cosh x)'$ stands for $\frac{1}{2}(e^x + e^{-x})'$, which evaluates to $\frac{1}{2}(e^x - e^{-x})$, or $\sinh x$.

Wronskian detail. Let $y_1 = \sinh x$, $y_2 = \cosh x$. Then

$$\begin{aligned} W &= y_1(x)y_2'(x) - y_1'(x)y_2(x) && \text{Definition of Wronskian } W. \\ &= \sinh(x)\sinh(x) - \cosh(x)\cosh(x) && \text{Substitute for } y_1,\ y_1',\ y_2,\ y_2'. \\ &= \tfrac{1}{4}(e^x - e^{-x})^2 - \tfrac{1}{4}(e^x + e^{-x})^2 && \text{Apply exponential definitions.} \\ &= -1 && \text{Expand and cancel terms.} \end{aligned}$$

Example 6.23 (Wronskian)

Given $2y'' - xy' + 3y = 0$, verify that a solution pair y_1, y_2 has Wronskian $W(x) = W(0)e^{x^2/4}$.

Solution: Let $a(x) = 2$, $b(x) = -x$, $c(x) = 3$. The Wronskian is a solution of $W' = -(b/a)W$, hence $W' = xW/2$. The solution is $W = W(0)e^{x^2/4}$, by the linear integrating factor method or the homogeneous equation shortcut.

Example 6.24 (Variation of Parameters)

Solve $y'' + y = \sec x$ by variation of parameters, verifying $y = c_1 \cos x + c_2 \sin x + x \sin x + \cos(x) \ln|\cos(x)|$.

Solution:

Homogeneous solution y_h. Theorem 6.1 is applied to the constant equation $y''+y = 0$. The characteristic equation $r^2 + 1 = 0$ has roots $r = \pm i$ and then $y_h = c_1 \cos x + c_2 \sin x$.

Wronskian. Suitable independent solutions are $y_1 = \cos x$ and $y_2 = \sin x$, taken from the formula for y_h. Then $W(x) = \cos^2 x + \sin^2 x = 1$.

Calculate y_p. The variation of parameters formula (4) is applied. The integration proceeds near $x = 0$, because $\sec(x)$ is continuous near $x = 0$.

$$\begin{aligned} y_p(x) &= -y_1(x) \textstyle\int y_2(x)\sec(x)dx + y_2(x) \int y_1(x) \sec x dx && \boxed{1} \\ &= -\cos x \textstyle\int \tan(x)dx + \sin x \int 1dx && \boxed{2} \\ &= x \sin x + \cos(x) \ln|\cos(x)| && \boxed{3} \end{aligned}$$

Details: $\boxed{1}$ Use equation (4). $\boxed{2}$ Substitute $y_1 = \cos x$, $y_2 = \sin x$. $\boxed{3}$ Integral tables applied. Integration constants set to zero.

Proofs and Details

Proof of Theorem 6.17: The function $W(t)$ given by Abel's identity is the unique solution of the growth-decay equation $W' = -(b(x)/a(x))W$; see page 3. It suffices then to show that W satisfies this differential equation. The details:

$W' = (y_1y_2' - y_1'y_2)'$	Definition of Wronskian.
$\quad = y_1y_2'' + y_1'y_2' - y_1''y_2 - y_1'y_2'$	Product rule; $y_1'y_2'$ cancels.
$\quad = y_1(-by_2' - cy_2)/a - (-by_1' - cy_1)y_2/a$	Both y_1, y_2 satisfy (2).
$\quad = -b(y_1y_2' - y_1'y_2)/a$	Cancel common cy_1y_2/a.
$\quad = -bW/a$	Verification completed.

The independence statement will be proved from the contrapositive: $W(x) = 0$ for all x if and only if y_1, y_2 are not independent. Technically, independence is defined relative to the common domain of the graphs of y_1, y_2 and W. Henceforth, *for all* x means for all x in the common domain.

Let y_1, y_2 be two solutions of (2), not independent. By re-labelling as necessary, $y_1(x) = cy_2(x)$ holds for all x, for some constant c. Differentiation implies $y_1'(x) = cy_2'(x)$. Then the terms in $W(x)$ cancel, giving $W(x) = 0$ for all x.

Conversely, let $W(x) = 0$ for all x. If $y_1 \equiv 0$, then $y_1(x) = cy_2(x)$ holds for $c = 0$ and y_1, y_2 are not independent. Otherwise, $y_1(x_0) \neq 0$ for some x_0. Define $c = y_2(x_0)/y_1(x_0)$. Then $W(x_0) = 0$ implies $y_2'(x_0) = cy_1'(x_0)$. Define $y = y_2 - cy_1$. By linearity, y is a solution of (2). Further, $y(x_0) = y'(x_0) = 0$. By uniqueness of initial value problems, $y \equiv 0$, that is, $y_2(x) = cy_1(x)$ for all x, showing y_1, y_2 are not independent.

Proof of Theorem 6.18: Let $F(t) = f(t)/a(t)$, $C_1(x) = -y_2(x)/W(x)$, $C_2(x) = y_1(x)/W(x)$. Then y_p as given in (4) can be differentiated twice using the product rule and the fundamental theorem of calculus rule $(\int g)' = g$. Because $y_1C_1 + y_2C_2 = 0$ and $y_1'C_1 + y_2'C_2 = 1$, then y_p and its derivatives are given by

$$\begin{array}{rcl} y_p(x) & = & y_1 \int C_1F dx + y_2 \int C_2F dx, \\ y_p'(x) & = & y_1' \int C_1F dx + y_2' \int C_2F dx, \\ y_p''(x) & = & y_1'' \int C_1F dx + y_2'' \int C_2F dx + F(x). \end{array}$$

Let $F_1 = ay_1'' + by_1' + cy_1$, $F_2 = ay_2'' + by_2' + cy_2$. Then

$$ay_p'' + by_p' + cy_p = F_1 \int C_1F dx + F_2 \int C_2F dx + aF.$$

Because y_1, y_2 are solutions of the homogeneous differential equation, then $F_1 = F_2 = 0$. By definition, $aF = f$. Therefore,

$$ay_p'' + by_p' + cy_p = f.$$

■

Historical Details. The original variation ideas, attributed to Joseph Louis Lagrange (1736-1813), involve substitution of $y = c_1(x)y_1(x) + c_2(x)y_2(x)$ into (1) plus imposing an extra unmotivated condition on the unknowns c_1, c_2:

$$c_1'y_1 + c_2'y_2 = 0.$$

The product rule gives $y' = c_1'y_1 + c_1y_1' + c_2'y_2 + c_2y_2'$, which then reduces to the two-termed expression $y' = c_1y_1' + c_2y_2'$. Substitution into (1) gives

$$a(c_1'y_1' + c_1y_1'' + c_2'y_2' + c_2y_2'') + b(c_1y_1' + c_2y_2') + c(c_1y_1 + c_2y_2) = f$$

which upon collection of terms becomes

$$c_1(ay_1'' + by_1' + cy_1) + c_2(ay_2'' + by_2' + cy_2) + ay_1'c_1' + ay_2'c_2' = f.$$

The first two groups of terms vanish because y_1, y_2 are solutions of the homogeneous equation, leaving just $ay_1'c_1'+ay_2'c_2' = f$. There are now two equations and two unknowns $X = c_1'$, $Y = c_2'$:

$$\begin{array}{rcrcl} ay_1'X & + & ay_2'Y & = & f, \\ y_1X & + & y_2Y & = & 0. \end{array}$$

Solving by elimination,

$$X = \frac{-y_2 f}{aW}, \quad Y = \frac{y_1 f}{aW}.$$

Then c_1 is the integral of X and c_2 is the integral of Y, which completes the historical account of the relations

$$c_1(x) = \int \frac{-y2(x)f(x)}{a(x)W(x)}dx, \quad c_2(x) = \int \frac{y_1(x)f(x)}{a(x)W(x)}dx.$$

Exercises 6.4

Independence: Constant Equation

Find solutions y_1, y_2 of the given homogeneous differential equation using Theorem 6.1 page 431. Then apply the Wronskian test page 464 to prove independence, following Example 6.22.

1. $y'' - y = 0$

2. $y'' - 4y = 0$

3. $y'' + y = 0$

4. $y'' + 4y = 0$

5. $4y'' = 0$

6. $y'' = 0$

7. $4y'' + y' = 0$

8. $y'' + y' = 0$

9. $y'' + y' + y = 0$

10. $y'' - y' + y = 0$

11. $y'' + 8y' + 2y = 0$

12. $y'' + 16y' + 4y = 0$

Independence for Euler's Equation

Change variables, $x = e^t$, $u(t) = y(x)$ in $Ax^2y''(x) + Bxy'(x) + Cy(x) = 0$ to obtain a constant-coefficient equation $A\left(\frac{d^2u}{dt^2} - \frac{du}{dt}\right) + B\frac{du}{dt} + Au = 0$. Solve for $u(t)$ and then substitute $t = \ln|x|$ to obtain $y(x)$. Find two solutions y_1, y_2 which are independent by the Wronskian test page 464.

13. $x^2y'' + y = 0$

14. $x^2y'' + 4y = 0$

15. $x^2y'' + 2xy' + y = 0$

16. $x^2y'' + 8xy' + 4y = 0$

Wronskian

Compute the Wronskian, up a constant multiple, without solving the differential equation: Example 6.23 page 466.

17. $y'' + y' - xy = 0$

18. $y'' - y' + xy = 0$

19. $2y'' + y' + \sin(x)y = 0$

20. $4y'' - y' + \cos(x)y = 0$

21. $x^2y'' + xy' - y = 0$

22. $x^2y'' - 2xy' + y = 0$

Variation of Parameters

Find the general solution $y_h + y_p$ by applying a variation of parameters formula: Example 6.24 page 466.

23. $y'' = x^2$

24. $y'' = x^3$

25. $y'' + y = \sin x$

26. $y'' + y = \cos x$

27. $y'' + y' = e^x$

28. $y'' + y' = -e^x$

29. $y'' + 2y' + y = e^{-x}$

30. $y'' - 2y' + y = e^x$

6.5 Undetermined Coefficients

The **method of undetermined coefficients** applies to solve constant-coefficient differential equations

(1) $$ay'' + by' + cy = f(x).$$

It finds a particular solution y_p *without* the integration steps present in variation of parameters. The method's importance is argued from its direct applicability to second order differential equations in mechanics and circuit theory. Requirements for $f(x)$ appear below.

Everything said here for second order differential equations applies unchanged to higher order differential equations

$$y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_0 y = f(x).$$

Definition 6.2 (Euler Solution Atom)
The term **atom** is an abbreviation for the phrase *Euler solution atom of a constant-coefficient linear homogeneous differential equation*. Assume symbols a and b are real constants with $b > 0$. Define an **Euler base atom** as one of the functions

$$e^{ax}, \quad e^{ax}\cos bx, \quad e^{ax}\sin bx.$$

Define an **Euler solution atom** as a power x^m times a base atom, for integers $m = 0, 1, 2, \ldots$:

$$\textbf{Euler solution atom} = x^m(\textbf{base atom}).$$

Requirements

The method of undetermined coefficients has special requirements:

- Equation $ay'' + by' + cy = f(x)$ has constant coefficients a, b, c.
- The function $f(x)$ is a sum of constants times Euler solution atoms.

Method of Undetermined Coefficients

Step 1. Define the list of k Euler atoms in a trial solution using Rule I and Rule II [details below]. Multiply these atoms by **undetermined coefficients** d_1, …, d_k, then add to define **trial solution** y.

Step 2. Substitute y into the differential equation.

Step 3. Match coefficients of Euler atoms left and right to write out linear algebraic equations for unknowns d_1, d_2, …, d_k. Solve the equations.

Step 4. The trial solution y with evaluated coefficients d_1, d_2, …, d_k becomes the particular solution y_p.

The Trial Solution Method

Central to the method of undetermined coefficients is the concept of a **trial solution** y, which is formally a linear combination of functions with coefficients yet to be determined. The method uses a *guess* of the form of a particular solution, then finds it explicitly *without actually solving the differential equation.* Knowing one particular solution y_p is enough to give the general solution of the differential equation (1), due to the superposition principle

$$y = y_h + y_p.$$

Example 6.25 (Trial Solution Illustration)

Consider the equation $y'' = 6x + e^x$ and a trial solution

$$y = d_1x^3 + d_2e^x.$$

Derive the equation

$$y_p = x^3 + e^x,$$

by calculating the **undetermined coefficients** d_1, d_2.

Solution: We first discuss how to solve the differential equation, because this background is needed to understand how the trial solution method works.

Answer check. The method of quadrature also applies to find $y = c_1 + c_2x + x^3 + e^x$ instead of $y = x^3 + e^x$. Superposition $y = y_h + y_p$ implies that the shortest answer for a particular solution is $y_p = x^3 + e^x$, obtained by dropping the homogeneous solution $c_1 + c_2x$.

Details.

We will show how to find d_1, d_2 in the trial solution $y = d_1x^3 + d_2e^x$ *without solving the differential equation*. The idea is to **substitute the trial solution into the differential equation**. This gives from equation $y'' = 6x + e^x$ the successive relations

$$\begin{array}{rcl} (d_1x^3 + d_2e^x)'' & = & 6x + e^x \\ 6d_1x + d_2e^x & = & 6x + e^x \end{array}$$

The last relation implies, by independence of the atoms x, e^x, the coefficient-matching equations [4]

$$\begin{array}{rcl} 6d_1 & = & 6, \\ d_2 & = & 1. \end{array}$$

The solution to this 2×2 linear system of equations is $d_1 = d_2 = 1$. Then the trial solution is

$$y = d_1x^3 + d_2e^x = x^3 + e^x.$$

We write $y_p = x^3 + e^x$.

That y_p is actually a solution of $y'' = 6x + e^x$ can be justified by computing the second derivative of $x^3 + e^x$.

[4] Euler atoms are independent in the sense of linear algebra. See Theorem 6.11, page 453. Independence means unique representation of linear combinations, which provides coefficient matching.

Why the Trial Solution has only Atoms x^3 and e^x

The differential equation $y'' = 6x + e^x$ can also be solved by answering this question:

> What expression y is differentiated twice to obtain $6x + e^x$?

Calculus suggests differentiating some cubic polynomial and some expression containing e^x. This is the central idea behind choosing a trial solution. Any trial solution, when substituted into the left side y'' of the differential equation, has to produce the terms in $6x + e^x$. Therefore, Euler atoms in the trial solution must have base atoms which appear in terms of the right side $6x + e^x$.

Explained is why terms in the trial solution $y = d_1x^3 + d_2e^x$ are limited to base atoms 1 and e^x.

Unexplained is why atoms 1, x, x^2 were not included in the trial solution. Insight can be gained by substitution of a combination $d_3 + d_4x + d_5x^2$ into the differential equation. Consider these steps:

$$\begin{array}{lcl}
(\text{trial solution})'' & = & 6x + e^x \\
(d_3 + d_4x + d_5x^2)'' & = & 6x + e^x \\
d_3(1)'' + d_4(x)'' + d_5(x^2)'' & = & 6x + e^x \\
d_3(0) + d_4(0) + d_5(2) & = & 6x + e^x
\end{array}$$

The coefficients d_3 and d_4 are multiplied by zero, because 1, x are solutions of the homogeneous equation $y'' = 0$. In general, homogeneous solution terms should not be added to a trial solution, because upon substitution these terms vanish from the left side of the differential equation. More succinctly, the missing variables d_3, d_4 are **free variables** in the language of linear algebra. We would choose $d_3 = d_4 = 0$ for simplicity. Term $2d_5$ is a multiple of base atom $1 = e^{0x}$. Because that atom does not appear on the right side $6x + e^x$, then $d_5 = 0$. The conclusion for this experiment: the trial solution $y = d_3 + d_4x + d_5x^2$ has three useless terms which do not contribute to terms on the right side of $y'' = 6x + e^x$.

Euler Solution Atoms in the General Solution

Superposition $y = y_h + y_p$ is used to describe the structure of solutions in differential equations solved by the method of undetermined coefficients. The homogeneous solution y_h of $ay'' + by' + cy = 0$ is constructed from atoms found by Euler's theorem. Therefore, y_h is a sum of constants times atoms. For the nonhomogeneous equation $ay'' + by' + cy = f(x)$, the method of undetermined coefficients finds y_p as a sum of constants times atoms. The plan here is to describe completely the atoms in solutions y_h and y_p.

Theorem 6.19 (Solution Structure)

A differential equation $ay'' + by' + cy = f(x)$ with constant coefficients a, b, c and right side $f(x)$ a sum of constants times Euler atoms has general solution $y = y_h + y_p$ which is a sum of constants times Euler atoms. In the language of linear algebra:

Solutions $y(x)$ of $ay'' + by' + cy = f(x)$ are a linear combination of Euler atoms.

Euler Atoms in the Homogeneous Solution

The atoms in y_h are found from Euler's theorem applied to the characteristic equation $ar^2 + br + c = 0$. To illustrate, the characteristic equation $r^2 + 2r + 1 = 0$ has double root -1, -1 and the corresponding atoms are e^{-x}, xe^{-x}.

Euler atoms can be extracted from a general solution $y_h = c_1 e^x + c_2 xe^x$ by taking partial derivatives on the symbols c_1, c_2. Conversely, two distinct Euler atoms are sufficient to form the general solution y_h. Euler atoms for the homogeneous equation can therefore be prescribed by any one of the following means:

1. The characteristic equation $ar^2 + br + c = 0$.

2. The roots of the characteristic equation.

3. The general solution expression y_h, with symbols c_1, c_2.

Euler Atoms in a Particular Solution y_p

The Euler atoms that appear in y_p may be assumed to not duplicate any atoms in y_h. The logic is that y_p can be shortened in length by moving any homogeneous solution into the terms of y_h, due to superposition $y = y_h + y_p$.

Explained below is how to construct the k atoms in y_p directly from the right side $f(x)$ of the differential equation. This is done by two rules, called **Rule I** and **Rule II**. We always proceed under the assumption that Rule I will work, and if it fails, then we go on to apply Rule II.

Undetermined Coefficients Rule I

Assume $f(x)$ in the equation $ay'' + by' + cy = f(x)$ is a sum of constants times Euler atoms. For each atom A appearing in $f(x)$, extract all distinct atoms that appear in A, A', A'', ..., then collect all computed atoms into a list of k distinct Euler atoms.

Test for a Valid Trial Solution

If the list **contains no solution of the homogeneous differential equation**, then multiply the k Euler atoms by undetermined coefficients d_1, ..., d_k to form the trial solution

$$y_p = d_1(\text{atom 1}) + d_2(\text{atom 2}) + \cdots + d_k(\text{atom k}).$$

Undetermined Coefficients Rule II

Assume Rule I constructed a list of k atoms, but Rule I **FAILED** the **TEST**. The particular solution y_p is still a sum of constants times k Euler atoms. Rule II changes some or all of the k atoms, by repeated multiplication by x.

The k-atom list is subdivided into groups with the same base atom, called **group 1**, **group 2**, and so on. Each group is tested for a solution of the homogeneous differential equation. If found, then multiply each Euler atom in the group by factor x. Repeat until **no group contains a solution of the homogeneous differential equation**. The final set of k Euler atoms is used to construct

$$y_p = d_1(\text{atom 1}) + d_2(\text{atom 2}) + \cdots + d_k(\text{atom k}).$$

Grouping Atoms

The Rule I process of finding derivatives A, A', A'',... can be replaced by the simpler task of forming **the group of each atom** A. The idea can be seen from the example $A = x^2e^x$. Each differentiation $A, A'A'', \ldots$ causes one lower power of x to appear, then we can predict that the distinct atoms that appear in the derivatives of A are

$$e^x, \quad xe^x, \quad x^2e^x.$$

This set is called the **group of Euler atom** A. In this example, $B = e^x$ is the **base atom** for atom $A = x^2e^x$ and the group is base atom B multiplied by the powers 1, x, x^2.

Assume Euler atom A is base atom B times a power x^m, for some integer $m \ge 0$. The **Group of Euler atom** A is the base atom B multiplied successively by the $m + 1$ powers 1, x, $\ldots$, x^m. The group starts with the base atom B and ends with the atom $A = x^mB$.

$$\begin{aligned} B &= \text{any base atom} \\ \text{group of } x^mB &\equiv B, \quad xB, \quad x^2B, \ldots, x^mB. \end{aligned}$$

Differentiation of an atom A with a sine or cosine factor produces two groups, not one. For example, $A = x^2 \sin x$ upon differentiation produces two groups

$$\begin{array}{ll} \textbf{cosine group}: & \cos x, \quad x\cos x, \quad x^2\cos x \\ \textbf{sine group}: & \sin x, \quad x\sin x, \quad x^2\sin x. \end{array}$$

Key Examples of Atom Grouping

1. The atom x^2e^{0x} has base atom $e^{0x} = 1$ and group 1, x, x^2. The group size is 3.

2. The atom $e^{-\pi x}$ has base atom $e^{-\pi x}$ and group $e^{-\pi x}$. A base atom has group size 1.

3. Atom $x^3e^x\cos x$ has base atom $e^x \cos x$ and two 4-element groups:
$e^x\cos x$, $xe^x\cos x$, $x^2e^x\cos x$, $x^3e^x\cos x$ and
$e^x\sin x$, $xe^x\sin x$, $x^2e^x\sin x$, $x^3e^x\sin x$.

4. Atom x^2e^x has base atom e^x. The group is the set of 3 atoms e^x, xe^x, x^2e^x.

5. If $A = xe^x \cos 2x$, then the Rule I process of extracting atoms from A, A', A'', …causes two groups to be formed, **group 1**: $e^x \cos 2x$, $xe^x \cos 2x$ and **group 2**: $e^x \sin 2x$, $xe^x \sin 2x$. A shortcut for writing the second group is to change cosine to sine in the first group.

Undetermined Coefficient Method Details

The undetermined coefficients trial solution y uses **Rule I** and **Rule II**. Then a particular solution, according to the method, is

$$y_p = \text{a linear combination of atoms.}$$

The discussion here is restricted to second order equations $n = 2$.

Superposition. The relation $y = y_h + y_p$ suggests solving $ay'' + by' + cy = f(x)$ in two stages:

(a) Apply **Euler's Theorem** to find y_h as a sum of constants times atoms.

(b) Apply the **method of undetermined coefficients** to find y_p as a sum of constants times atoms.

Symbols. The symbols c_1, c_2 are reserved for use as arbitrary constants in the general solution y_h of the homogeneous equation. Symbols d_1, d_2, …are reserved for use in the trial solution y of the non-homogeneous equation. Abbreviations: c = constant, d = determined.

Expect to find two arbitrary constants c_1, c_2 in the solution y_h, but in contrast, no arbitrary constants appear in y_p. The literature's terminology *undetermined coefficients* is misleading, because in fact symbols d_1, d_2, …are *determined.*

Algebra Background. The trial solution method requires background in the solution of simultaneous linear algebraic equations, as is often taught in college algebra. A linear algebra background will make the details seem even easier.

Example 6.26 (Undetermined Coefficients Illustration)

Solve the differential equation $y'' - y = x + xe^x$ by the method of undetermined coefficients, verifying

$$y_h = c_1e^x + c_2e^{-x}, \quad y_p = -x - \frac{1}{4}xe^x + \frac{1}{4}x^2e^x.$$

Solution:
Homogeneous Solution. The homogeneous equation

$$y'' - y = 0$$

has characteristic equation $r^2 - 1 = 0$. The roots $r = \pm 1$ produce by Euler's theorem the list of atoms e^x, e^{-x}. Then the homogeneous solution is a linear combination of the Euler atoms: $y_h = c_1 e^x + c_2 e^{-x}$.

Trial Solution. The shortest trial solution is

$$y = (d_1 + d_2 x) + (d_3 x e^x + d_4 x^2 e^x),$$

to be justified below.

Rule I. Let $f(x) = x + xe^x$. The derivatives $f, f', f'', \ldots$ are linear combinations of the four Euler atoms $1, x, e^x, xe^x$. Because e^x is a solution of the homogeneous equation $y'' - y = 0$, then Rule I FAILS the TEST.

Rule II. Divide the list $1, x, e^x, xe^x$ into two groups with identical base atom:

Group	Euler Atoms	Base Atom
group 1 :	$1, x$	1
group 2 :	e^x, xe^x	e^x

Group 1 contains no solution of the homogeneous equation $y'' - y = 0$, therefore Rule II changes nothing. **Group 2** contains solution e^x of the homogeneous equation. Rule II says to multiply group 2 by x, until the modified group contains no solution of the homogeneous differential equation $y'' - y = 0$ .Then

Group	Euler Atoms	Action
New group 1 :	$1, x$	no change
New group 2 :	xe^x, x^2e^x	multiplied once by x

In **New Group 2**, xe^x is not a solution of the homogeneous problem, because if it is, then 1 is a double root of the characteristic equation $r^2 - 1 = 0$ [it isn't].

The final groups have been found in Rule II. The shortest trial solution is

$$\begin{aligned} y &= \text{linear combination of Euler atoms in the new groups} \\ &= d_1 + d_2 x + d_3 x e^x + d_4 x^2 e^x. \end{aligned}$$

Equations for the undetermined coefficients. Substitute $y = d_1 + d_2 x + d_3 xe^x + d_4 x^2 e^x$ into $y'' - y = x + xe^x$. The details:

$$\begin{aligned}
\text{LHS} &= y'' - y && \text{Left side of the equation.} \\
&= [y_1'' - y_1] + [y_2'' - y_2] && \text{Let } y = y_1 + y_2,\ y_1 = d_1 + d_2 x,\ y_2 = d_3 xe^x + d_4 x^2 e^x. \\
&= [0 - y_1] + \\
&\quad [2d_3 e^x + 2d_4 e^x + 4d_4 xe^x] && \text{Use } y_1'' = 0 \text{ and } y_2'' = y_2 + 2d_3 e^x + 2d_4 e^x + 4d_4 xe^x. \\
&= (-d_1)1 + (-d_2)x + \\
&\quad (2d_3 + 2d_4)e^x + (4d_4)xe^x && \text{Collect on distinct Euler atoms.}
\end{aligned}$$

Then $y'' - y = f(x)$ simplifies to

$$(-d_1)1 + (-d_2)x + (2d_3 + 2d_4)e^x + (4d_4)xe^x = f(x).$$

Write out a 4×4 **system**. Because $f(x) = x + xe^x$, the last display gives the expansion below, which has been written with each side a linear combination of the atoms 1, x, e^x, xe^x.

$$\begin{aligned} &(-d_1)1 + (-d_2)x + \\ &(2d_3 + 2d_4)e^x + (4d_4)xe^x = (0)1 + (1)x + (0)e^x + (1)xe^x. \end{aligned} \tag{2}$$

Equate coefficients of matching atoms 1, x, e^x, xe^x left and right to give the system of equations

$$(3)\qquad \begin{array}{rrrrcll} -d_1 & & & & = & 0, & \text{match on } 1 \\ & -d_2 & & & = & 1, & \text{match on } x \\ & & 2d_3 & +2d_4 & = & 0, & \text{match on } e^x \\ & & & 4d_4 & = & 1. & \text{match on } xe^x \end{array}$$

Atom matching effectively removes x and changes the equation into a 4×4 linear algebraic nonhomogeneous system of equations for d_1, d_2, d_3, d_4.

The technique is **independence**. To explain, linear independence of atoms means that a linear combination of atoms is uniquely represented. Then two such equal representations must have matching coefficients. Relation (2) says that two linear combinations of the same list of atoms are equal. Then coefficients of 1, x, e^x, xe^x left and right in (2) must match, giving system (3).

Solve the equations. The 4×4 system by design always has a unique solution. In the language of linear algebra, there are zero free variables. In the present case, the system is triangular, solved by back-substitution to give the unique solution $d_1 = 0$, $d_2 = -1$, $d_4 = 1/4$, $d_3 = -1/4$.

Report y_p. The trial solution $y = d_1 + d_2x + d_3xe^x + d_4x^2e^x$ with determined coefficients $d_1 = 0$, $d_2 = -1$, $d_3 = -1/4$, $d_4 = 1/4$ becomes the particular solution

$$y_p = -x - \frac{1}{4}xe^x + \frac{1}{4}x^2e^x.$$

General solution. Superposition implies the general solution is $y = y_h + y_p$. From above, $y_h = c_1e^x + c_2e^{-x}$ and $y_p = -x - \frac{1}{4}xe^x + \frac{1}{4}x^2e^x$. Then $y = y_h + y_p$ is given by

$$y = c_1e^x + c_2e^{-x} - x - \frac{1}{4}xe^x + \frac{1}{4}x^2e^x.$$

Answer Check. Computer algebra system `maple` is used.

```
yh:=c1*exp(x)+c2*exp(-x);
yp:=-x-(1/4)*x*exp(x)+(1/4)*x^2*exp(x);
de:=diff(y(x),x,x)-y(x)=x+x*exp(x):
odetest(y(x)=yh+yp,de); # Success is a report of zero.
```

Further examples: pages 481, 482, 484, 484.

Constructing Euler Atoms from Roots

An Euler atom is constructed from a real number a or a complex number $a + ib$. The number used for the construction is called a **root** for the atom. Euler's theorem page 454 provides the rules:

> Real root $r = a$ constructs the exponential base atom e^{ax}. If $a = 0$, then the base atom is $e^{0x} = 1$.
>
> For a complex root $r = a + ib$, $b > 0$, construct two base atoms $e^{ax}\cos bx$ and $e^{ax}\sin bx$.

Atoms constructed from roots a or $a+ib$ using Euler's multiplicity theorem gives the complete list of all possible atoms:

Root	Euler Atoms ($b>0$)				
$r=a$	e^{ax},	xe^{ax},	x^2e^{ax},	$\ldots$,	x^ne^{ax}
$r=a+ib$	$e^{ax}\cos bx$,	$xe^{ax}\cos bx$,	$x^2e^{ax}\cos bx$,	$\ldots$,	$x^ne^{ax}\cos bx$
$r=a+ib$	$e^{ax}\sin bx$,	$xe^{ax}\sin bx$,	$x^2e^{ax}\sin bx$,	$\ldots$,	$x^ne^{ax}\sin bx$

Constructing Roots from Euler Atoms

An Euler atom can be viewed as having been constructed from a unique real root a or a unique pair of complex roots $a\pm ib$. The reverse process considers an atom and finds the possible root (or roots) used for its construction plus the root's multiplicity. Details in the following table:

Euler Atom	Base Atom	Root	Multiplicity
x^ne^{ax}	e^{ax}	a	$n+1$
$x^ne^{ax}\cos(bx)$	$e^{ax}\cos(bx)$	$a\pm ib$	$n+1$
$x^ne^{ax}\sin(bx)$	$e^{ax}\sin(bx)$	$a\pm ib$	$n+1$

Examples of Atoms and Roots

The atoms for root $r=0$ of multiplicity 4 are $1,x,x^2,x^3$. The atoms for root $r=2+3i$ of multiplicity 3 are

$$\begin{array}{lll} e^{2x}\cos(3x), & xe^{2x}\cos(3x), & x^2e^{2x}\cos(3x) \\ e^{2x}\sin(3x), & xe^{2x}\sin(3x), & x^2e^{2x}\sin(3x). \end{array}$$

The roots for atom x^3 are $r=0,0,0,0$ (quad root). The roots for atom xe^x are $r=1,1$. The roots for atom $x\cos x$ are $r=i,-i,i,-i$ (double complex root).

Polynomials and Root Multiplicity

In college algebra, roots of polynomials are studied through the theory of equations, which includes the root and factor theorem, the rational root theorem, the division algorithm and Descarte's rule of signs.

The multiplicity of a polynomial root $r=r_0$ is defined in college algebra to be the unique integer m such that $(r-r_0)^m$ divides the polynomial, but $(r-r_0)^{m+1}$ does not.

The algebra topic is enriched by calculus:

Theorem 6.20 (Multiplicity of a Root)
Let $p(r)$ be the characteristic polynomial for a given linear homogeneous differential equation with constant coefficients. The **Multiplicity** of a root $r=r_0$ of $p(r)=0$

can be determined by calculus as follows.

$$\begin{array}{ll}
\text{Multiplicity } 1 & p(r_0)=0,\ p'(r_0)\neq 0\\
\text{Multiplicity } 2 & p(r_0)=p'(r_0)=0,\ p''(r_0)\neq 0\\
\text{Multiplicity } 3 & p(r_0)=p'(r_0)=p''(r_0)=0,\ p'''(r_0)\neq 0\\
\vdots & \vdots\\
\text{Multiplicity } m & p(r_0)=\cdots=p^{(m-1)}(r_0)=0,\ p^{(m)}(r_0)\neq 0
\end{array}$$

Factorization of the characteristic polynomial may be possible. If so, then the roots and their multiplicities are all known at once. Factorization is not needed at all to test if $r = r_0$ is a root, and only basic calculus is required to determine the multiplicity of a root.

Computing the Shortest Trial Solution

Described here is are two alternatives to Rule I and Rule II, to construct the shortest trial solution in the method of undetermined coefficients. The first method uses Laplace theory. The second method uses differential operator techniques, presented here assuming minimal background.

Laplace's Method

Readers who are unfamiliar with Laplace theory should skip this subsection and go on to the next.

The idea will be communicated by example, which hopefully is enough for a reader already familiar with Laplace theory. Suppose we are going to solve the equation

$$\frac{d^2y}{dt^2}+y=t+e^t$$

using the theory of undetermined coefficients. Then Rule I applies and we don't need Rule II, giving $y = y_h + y_p$ where

$$y_h = c_1\cos t + c_2\sin t,\quad y_p = d_1 + d + 2t + d_3e^t.$$

Laplace theory can quickly find y_p by assuming zero initial data $y(0) = y'(0) = 0$, in which case another candidate y for y_p is found by the transfer function method:

$$\mathcal{L}(y) = (\text{Transfer function})(\text{Laplace of } t+e^t) = \frac{s^2+s-1}{s^2(s-1)(s^2+1)}.$$

Partial fraction theory applies:

$$\mathcal{L}(y) = \frac{a+bs}{s^2+1}+\frac{c}{s}+\frac{d}{s^2}+\frac{f}{s-1} = \mathcal{L}(a\sin t + b\cos t + c + dt + fe^t).$$

Lerch's theorem applies:

$$y = a\sin t + b\cos t + c + dt + fe^t.$$

The term $a\sin t + b\cos t$ represents a solution y_h of the homogeneous problem $y'' + y = 0$. Remove the homogeneous solution, then report a particular solution as having the form

$$y_p = c + dt + fe^t.$$

This is the shortest trial solution, obtained by Laplace theory.

The Method of Annihilators

Suppose that $f(x)$ is a sum of constants times Euler atoms. The **Annihilator** of $f(x)$ is the unique minimal-order homogeneous constant-coefficient higher order differential equation of leading coefficient one which has $f(x)$ as a particular solution.

For example, if $f(x) = x + e^x$, then the annihilator of $f(x)$ is the third order constant-coefficient homogeneous differential equation [details in examples below]

$$y''' - y'' = 0$$

Required is that $f(x)$ is a *particular solution* of the differential equation, related to the general solution $y(x)$ by specialization of constants.

Examples of annihilators: The differential equation $y''+y = 0$ is the annihilator for $\sin x$, but also the annihilator for $2\cos x - \sin x$. The differential equation $y''' + 4y' = 0$ is the annihilator for any of $\sin 2x$, $1 + \cos 2x$, $7 - 5\sin 2x$.

An annihilator can be given by its characteristic equation, e.g., $r^3 + 4r = 0$ generates annihilator $y''' + 4y' = 0$.

Characteristic Polynomial of the Annihilator

Let $f(x)$ be a given linear combination of atoms. The algorithm:

1. Determine the list of atoms for $f(x)$.
2. Find the root(s) for each base atom B. Then find the corresponding highest power real factors in the characteristic equation, using Euler's theorem.
3. The characteristic polynomial is the product of the highest power distinct factors so found.

For instance, $f(x) = 2e^x + \cos 3x - x - x^3$ has base atoms e^x, $\cos 3x$, 1 with corresponding roots 1, $\pm 3i$, $0, 0, 0, 0$, listed according to multiplicity. By Euler's theorem, the corresponding factors with highest powers are $r - 1$, $r^2 + 9$, $(r - 0)^4$, which implies the characteristic polynomial is $(r - 1)(r^2 + 9)r^4$.

Annihilator Method Algorithm

Assume that the non-homogeneous differential equation of order n has constant coefficients and the right side $f(x)$ is a linear combination of atoms. The method arises by applying the annihilator of f, as a differential operator, to the non-homogeneous differential equation

$$y^{(n)} + \sum_{k=0}^{n-1} a_k y^{(k)} = f(x).$$

Because the annihilator applied to $f(x)$ is zero, then any solution $y = y_p(x)$ satisfies a higher-order homogeneous equation, whose characteristic equation is known [see item 3 below].

1. Find the homogeneous equation characteristic polynomial $p(r)$.
2. Find the characteristic polynomial $q(r)$ for the annihilator of $f(x)$.
3. The shortest trial solution is a linear combination of the atoms obtained from $p(r)q(r) = 0$, after removing those atoms which correspond to the roots of $p(r) = 0$.

Further examples pages 486, 486.

Further study

The trial solution method is enriched by developing a **Library of Special Methods** for finding y_p, which includes Kümmer's method; see page 575. The library provides an independent justification of the trial solution method. The only background required is college algebra and polynomial calculus. The trademark of the library method is the *absence of linear algebra, tables or special cases*, that can be found in other literature on the subject.

Examples

Example 6.27 (Polynomial Trial Solution)

Solve for y_p in $y'' = 2 - x + x^3$ using the method of undetermined coefficients, verifying $y_p = x^2 - \frac{1}{6}x^3 + \frac{1}{20}x^5$.

Solution:

Homogeneous solution. The homogeneous equation $y'' = 0$ has characteristic equation $r^2 = 0$ with roots $r = 0, 0$. Euler's theorem generates the two atoms 1, x. Then the homogeneous solution is $y_h = c_1 + c_2 x$.

Trial solution. Let's justify the selection of the trial solution

$$y = d_1 x^2 + d_2 x^3 + d_3 x^4 + d_4 x^5.$$

Rule I applied to the right side $f(x) = 2 - x + x^3$ gives a single group of four atoms

$$\textbf{group 1}: \quad 1, x, x^2, x^3.$$

Because 1 is a solution of the homogeneous equation $y'' = 0$, then Rule I **FAILS** the **TEST**. Rule II is applied to **group 1**, which modifies the group by multiplication by x. The correction by x-multiplication must be applied twice, because both 1 and x are solutions of the homogeneous differential equation $y'' = 0$. Then the new group is

$$\textbf{New group 1}: \quad x^2, x^3, x^4, x^5.$$

The trial solution is then a linear combination of four Euler atoms from the new group, $y = d_1x^2 + d_2x^3 + d_3x^4 + d_4x^5$.

Equations for the undetermined coefficients. The details:

$$\begin{aligned} 2 - x + x^3 &= y'' && \textsf{Reverse sides.} \\ &= 2d_1 + 6d_2x + 12d_3x^2 + 20d_4x^3 && \textsf{Substitute } y. \end{aligned}$$

Equate coefficients of Euler atoms on each side of the equal sign to obtain the system of equations

$$\begin{array}{rcl} 2d_1 &=& 2, \\ 6d_2 &=& -1, \\ 12d_3 &=& 0, \\ 20d_4 &=& 1. \end{array}$$

Solve the equations. This is a triangular system of linear equations for unknowns d_1, d_2, d_3, d_4. Solving gives $d_1 = 1$, $d_2 = -1/6$, $d_3 = 0$, $d_4 = 1/20$.

Report y_p. The trial solution expression $y = d_1x^2+d_2x^3+d_3x^4+d_4x^5$ after substitution of the values found for d_1 to d_4 gives the particular solution

$$y_p = x^2 - \frac{1}{6}x^3 + \frac{1}{20}x^5.$$

Example 6.28 (Undetermined Coefficient Method)

Solve $y'' + y = 2 + e^x + \sin(x)$ by the trial solution method, verifying $y = c_1\cos(x) + c_2\sin(x) + 2 + \frac{1}{2}e^x - \frac{1}{2}x\sin x$.

Solution:

Homogeneous solution. The characteristic equation for the homogeneous equation $y'' + y = 0$ is $r^2 + 1 = 0$. It has roots $r = \pm i$ and atom list $\cos x, \sin x$. Then y_h is a linear combination of the two atoms:

$$y_h = c_1\cos x + c_2\sin x.$$

Symbols c_1 and c_2 are arbitrary constants.

Rule I. The right side $f(x) = 2 + e^x + \sin x$ of the differential equation is differentiated a few times to discover the atom list $1, e^x, \cos x, \sin x$. Because $\cos x$ is a solution of the homogeneous equation $y'' + y = 0$, then Rule I **FAILS** the **TEST**.

Rule II. The Euler atoms are grouped by equal base atom as follows.

Group	Euler Atoms	Rule II action	New Group
group 1:	1	no change	1
group 2:	e^x	no change	e^x
group 3:	$\cos x$	multiply once by x	$x\cos x$
group 4:	$\sin x$	multiply once by x	$x\sin x$

Group 1 and **Group 2** are unchanged by Rule II, because they do not contain a solution of the homogeneous equation $y''+y=0$. **Group 3** and **Group 4** do contain a homogeneous solution, therefore each group is multiplied by x. The resulting new groups 3 and 4 do not contain a homogeneous solution. It is expected, in general, to iterate on x-multiplication on a group until the first time that the new group contains no solution of the homogeneous equation.

The trial solution is a linear combination of the four Euler atoms in the new groups:

$$y = d_1 + d_2e^x + d_3x\cos x + d_4x\sin x.$$

Equations for the undetermined coefficients.

$$\begin{aligned}\text{LHS} &= y''+y && \text{Left side of the equation } y''+y = 2+e^x+\sin(x).\\ &= d_1 + 2d_2e^x - 2d_3\sin(x) + 2d_4\cos(x) && \text{Substitute trial solution } y.\end{aligned}$$

The equation $y''+y=2+e^x+\sin(x)$ becomes

$$d_1 + 2d_2e^x - 2d_3\sin(x) + 2d_4\cos(x) = 2+e^x+\sin(x).$$

Equating coefficients of atoms left and right implies the equations

$$\begin{array}{rcl} d_1 &=& 2,\\ 2d_2 &=& 1,\\ -2d_3 &=& 1,\\ 2d_4 &=& 0.\end{array}$$

Solve the equations. There are no details, because the system is diagonal. The displayed answers are $d_1=2$, $d_2=1/2$, $d_3=-1/2$, $d_4=0$.

Particular solution y_p. The particular solution is obtained from the trial solution $y=d_1+d_2e^x+d_3x\cos x+d_4x\sin x$ by replacing the undetermined coefficients d_1 to d_4 by their values determined above:

$$y_p = 2 + \frac{1}{2}e^x - \frac{1}{2}x\cos(x).$$

General Solution. Add y_h and y_p to obtain the general solution

$$y = c_1\cos(x) + c_2\sin(x) + 2 + \frac{1}{2}e^x - \frac{1}{2}x\cos(x).$$

Answer check. Computer algebra system `maple` checks the answer as follows.

```
dsolve(diff(y(x),x,x)+y(x)=2+exp(x)+sin(x),y(x));
# y(x) = 2+1/2*exp(x)-1/2*cos(x)*x+_C1*cos(x)+_C2*sin(x)
```

Example 6.29 (Two Methods)

Solve $y'' - y = e^x$ by undetermined coefficients and by variation of parameters. Explain any differences in the answers.

Solution: The general solution is reported to be $y = y_h + y_p = c_1 e^x + c_2 e^{-x} + xe^x/2$. Details follow.

Homogeneous solution. The characteristic equation $r^2 - 1 = 0$ for $y'' - y = 0$ has roots ± 1 with atom list e^x, e^{-x}. The homogeneous solution is $y_h = c_1 e^x + c_2 e^{-x}$.

Undetermined Coefficients Summary. The right side of the differential equation, $f(x) = e^x$, contains only the single atom e^x, therefore the Rule I atom list is e^x. Rule I FAILS the TEST, because e^x is a solution of the homogeneous equation. Rule II applies, then x multiplies the group e^x to obtain the new group xe^x. This atom is not a solution of the homogeneous equation, therefore the trial solution is $y = d_1 xe^x$. Substitute it into $y'' - y = e^x$ to obtain $2d_1 e^x + d_1 xe^x - d_1 xe^x = e^x$. Match coefficients of e^x to compute $d_1 = 1/2$. Then $y_p = xe^x/2$.

Variation of Parameters Summary. The homogeneous solution $y_h = c_1 e^x + c_2 e^{-x}$ found above implies $y_1 = e^x$, $y_2 = e^{-x}$ is a suitable independent pair of solutions. Their Wronskian is

$$W = \begin{vmatrix} e^x & e^{-x} \\ e^x & -e^{-x} \end{vmatrix} = -2.$$

The variation of parameters formula (6.18) applies:

$$y_p(x) = \left(\int \frac{-e^{-x}}{-2} e^x dx \right) e^x + \left(\int \frac{e^x}{-2} e^x dx \right) e^{-x}.$$

Integration with zero constants of integration gives $y_p(x) = xe^x/2 - e^x/4$.

Differences. The two methods give respectively $y_p = xe^x/2$ and $y_p = xe^x/2 - e^x/4$. The solutions $y_1 = xe^x/2$ and $y_2 = xe^x/2 - e^x/4$ differ by the homogeneous solution $y_h = y_2 - y_1 = -xe^x/4$. In both cases, the general solution is $y = c_1 e^x + c_2 e^{-x} + \frac{1}{2}xe^x$, because homogeneous solution terms can be absorbed into the arbitrary constants c_1, c_2.

Example 6.30 (Sine–Cosine Trial solution)

Verify for $y'' + 4y = \sin x - \cos x$ that $y_p(x) = 5\cos x + 3\sin x$, using trial solution $y = A\cos x + B\sin x$.

Solution: Let's justify the trial solution. Rule I differentiates $f(x) = \sin x - \cos x$ to determine the atom list $\cos x$, $\sin x$. Because $\cos x$ and $\sin x$ are not solutions of the homogeneous equation $y'' + 4y = 0$, then Rule I succeeds and the trial solution is $y = d_1 \cos x + d_2 \sin x$. Replace d_1, d_2 by symbols A, B to agree with the given trial solution.

Equations for the undetermined coefficients. Substitute $y = A\cos x + B\sin x$ into the differential equation and use $u'' = -u$ for $u = \sin x$ or $u = \cos x$ to obtain the relation

$$\begin{aligned} \sin x - \cos x &= y'' + 4y \\ &= (-A + 4)\cos x + (-B + 4)\sin x. \end{aligned}$$

Matching coefficients of sine and cosine terms on the left and right gives the system of equations

$$\begin{aligned} -A + 4 &= -1, \\ -B + 4 &= 1. \end{aligned}$$

Solve the equations. The system is diagonal, therefore $A = 5$ and $B = 3$.

Report y_p. The trial solution $y = A\cos x + B\sin x$ after substitution of found values for A, B becomes the particular solution $y_p(x) = 5\cos x + 3\sin x$.

Generally, the method of undetermined coefficients applied to similar second order problems produces linear algebraic equations that must be solved by linear algebra techniques. Sometimes, the most convenient is Cramer's 2 × 2 rule.

Example 6.31 (Exponential Trial Solution)

Solve for y_p in

$$y'' - 2y' + y = (1 + x - x^2)e^x$$

by the method of undetermined coefficients, verifying that

$$y_p = \frac{1}{2}x^2e^x + \frac{1}{6}x^3e^x - \frac{1}{12}x^4e^x.$$

Solution:

Homogeneous solution. The homogeneous equation is $y'' - 2y' + y = 0$. The characteristic equation $r^2 - 2r + 1 = 0$ has a double root $r = 1$ and by Euler's theorem the corresponding atom list is e^x, xe^x. Then the homogeneous general solution is $y_h = c_1e^x + c_2xe^x$, where c_1 and c_2 are arbitrary constants.

Trial solution. Let's apply Rule I. The derivatives of $f(x) = (1 + x - x^2)e^x$ are combinations of the list of distinct Euler atoms e^x, xe^x, x^2e^x. Because the first two atoms are solutions of the homogeneous equation, then Rule I FAILS the TEST. Rule II applies: the list of atoms for $f(x)$ has just one group:

$$\textbf{group 1}: \quad e^x, \quad xe^x, \quad x^2e^x.$$

Rule II modifies the list of three atoms by x-multiplication. It is applied twice, because both e^x and xe^x are solutions of the homogeneous differential equation. The new group of three atoms is

$$\textbf{New group 1}: \quad x^2e^x, \quad x^3e^x, \quad x^4e^x.$$

A trial solution according to Rule II is a linear combination of the new group atoms:

$$y = d_1x^2e^x + d_2x^3e^x + d_3x^4e^x.$$

Equations for the undetermined coefficients. Substitute the trial solution $y = d_1x^2e^x + d_2x^3e^x + d_3x^4e^x$ solution into $y'' - 2y' + y = (1 + x - x^2)e^x$, in order to find the undetermined coefficients d_1, d_2, d_3. To present the details, let $q(x) = d_1x^2 + d_2x^3 + d_3x^4$, then $y = q(x)e^x$ implies

$$\begin{aligned}
\text{LHS} &= y'' - 2y' + y \\
&= [q(x)e^x]'' - 2[q(x)e^x]' + q(x)e^x \\
&= q(x)e^x + 2q'(x)e^x + q''(x)e^x - 2q'(x)e^x - 2q(x)e^x + q(x)e^x \\
&= q''(x)e^x \\
&= [2d_1 + 6d_2x + 12d_2x^2]e^x.
\end{aligned}$$

Matching coefficients of Euler atoms left and right gives the 3 × 3 system of equations

$$\begin{aligned}
2d_1 &= 1, \\
6d_2 &= 1, \\
12d_3 &= -1.
\end{aligned}$$

Solve the equations. The 3 × 3 system is diagonal and needs no further analysis: $d_1 = 1/6$, $d_2 = 1/6$, $d_3 = -1/12$.

Report y_p. The trial solution after substitution of found coefficients d_1, d_2, d_3 becomes the particular solution

$$y_p = \frac{1}{2}x^2e^x + \frac{1}{6}x^3e^x - \frac{1}{12}x^4e^x.$$

General solution. The superposition relation $y = y_h + y_p$ is the general solution

$$y = c_1e^x + c_2xe^x + \frac{1}{2}x^2e^x + \frac{1}{6}x^3e^x - \frac{1}{12}x^4e^x.$$

Answer check. The `maple` code:

```
de:=diff(y(x),x,x)-2*diff(y(x),x)+y(x)=(1+x-x^2)*exp(x);
dsolve(de,y(x));
# y(x) = 1/2*exp(x)*x^2 + 1/6*exp(x)*x^3
#       -1/12*exp(x)*x^4+_C1*exp(x)+_C2*exp(x)*x
```

Example 6.32 (Annihilator)

Find the annihilator for $f(x) = x - 4\sin 3x$.

Solution: First, identify $f(x) = x - 4\sin 3x$ as a linear combination of the atoms x, $\sin 3x$. Euler's theorem implies that the characteristic polynomial must have roots $0, 3i, -3i$. Then the characteristic polynomial must contain these factors:

Roots $r = 0, 0$	Atoms $1, x$	Factor r^2
Roots $\pm 3i$	Atoms $\cos 3x, \sin 3x$	Factors $r - 3i$, $r + 3i$

Multiply the factors r^2 and $(r - 3i)(r + 3i) = r^2 + 9$ to generate the characteristic polynomial

$$(\text{factor } r^2) \text{ times } (\text{factor } r^2 + 9) = r^4 + 9r^2.$$

The annihilator is $y^{(4)} + 9y'' = 0$, obtained by translation of characteristic equation $r^4 + 9r^2 = 0$ into a differential equation.

Example 6.33 (Annihilator)

Find the annihilator for $f(x) = e^x(x^2 - 2\cos 3x)$.

Solution: Function $f(x) = e^x(x^2 - 2\cos 3x)$ is a linear combination of the atoms x^2e^x, $e^x \cos 3x$. Euler's theorem implies that the roots are $r = 1, 1, 1, 1 \pm 3i$. Then the characteristic polynomial must contain factors as follows.

Roots	Atoms	Factor
$r = 1, 1, 1$	e^x, xe^x, x^2e^x	$(r-1)^3$
$1 \pm 3i$	$e^x \cos 3x, e^x \sin 3x$	$(r - 1 - 3i)(r - 1 + 3i)$

Multiply the factors $(r - 1)^3$ and $(r - 1)^2 + 9$ to generate the characteristic equation

$$(r-1)^3((r-1)^2 + 9) = 0.$$

Expanding, the characteristic polynomial is $r^5 - 5\,r^4 + 19\,r^3 - 37\,r^2 + 32\,r - 10$. In applications, we would stop here, with the characteristic polynomial. If we continue, then the annihilator is the differential equation $y^{(5)} - 5y^{(4)} + 19y''' - 37y'' + 32y' - 10y = 0$.

Example 6.34 (Annihilator Method)

Find the shortest trial solution for the differential equation $y'' - y = x + xe^x$ using the Method of Annihilators.

Solution: The example was solved previously using Rule I and Rule II with answer

$$y_p = d_1 + d_2x + d_3xe^x + d_4x^2e^x.$$

Homogeneous equation. The characteristic polynomial for homogeneous equation $y'' - y = 0$ is $p(r) = r^2 - 1$. It has roots $r = 1$, $r = -1$ and corresponding atoms e^x, e^{-x}.

Annihilator for $f(x)$. The right side of the differential equation is $f(x) = x + xe^x$. We compute the characteristic polynomial $q(r)$ of an annihilator of $f(x)$. The atoms for $f, f', f'', \ldots$ are $1, x, e^x, xe^x$ with corresponding roots $0, 0, 1, 1$. The factors of the characteristic polynomial $q(r)$ are then r^2, $(r-1)^2$, by Euler's theorem. Specifically, we used these specialized conclusions from Euler's theorem:

1. Root $r = 0$ of $q(r) = 0$ has multiplicity 2 if and only if r^2 is a factor of $q(r)$;
2. Root $r = 1$ of $q(r) = 0$ has multiplicity 2 if and only if $(r-1)^2$ is a factor of $q(r)$.

The conclusion of this analysis is that $q(r) =$ product of the factors $= r^2(r-1)^2$.

Trial solution. Let

$$w(r) = p(r)q(r) = (r^2 - 1)r^2(r-1)^2 = r^2(r+1)(r-1)^3.$$

Then y_p must be a solution of the differential equation with characteristic equation $w(r) = 0$, which implies that y_p is a linear combination of the atoms

$$1, x, e^{-x}, e^x, xe^x, x^2e^x.$$

Atoms e^{-x} and e^x are solutions of the homogeneous equation, therefore they are removed. The shortest trial solution is a linear combination of the Euler atoms

$$1, x, xe^x, x^2e^x.$$

Then

$$y_p = d_1 + d_2x + d_5xe^x + d_6x^2e^x,$$

which agrees with the shortest trial solution obtained by Rule I and Rule II.

Exercises 6.5

Polynomial Solutions

Determine a polynomial solution y_p for the given differential equation.

1. $y'' = x$

2. $y'' = x - 1$

3. $y'' = x^2 - x$

4. $y'' = x^2 + x - 1$

5. $y'' - y' = 1$

6. $y'' - 5y' = 10$

7. $y'' - y' = x$

8. $y'' - y' = x - 1$

9. $y'' - y' + y = 1$

10. $y'' - y' + y = -2$

11. $y'' + y = 1 - x$

12. $y'' + y = 2 + x$

13. $y'' - y = x^2$

14. $y'' - y = x^3$

Polynomial-Exponential Solutions

Determine a solution y_p for the given differential equation.

15. $y'' + y = e^x$

16. $y'' + y = e^{-x}$

17. $y'' = e^{2x}$

18. $y'' = e^{-2x}$

19. $y'' - y = (x+1)e^{2x}$

20. $y'' - y = (x-1)e^{-2x}$

21. $y'' - y' = (x+3)e^{2x}$

22. $y'' - y' = (x-2)e^{-2x}$

23. $y'' - 3y' + 2y = (x^2+3)e^{3x}$

24. $y'' - 3y' + 2y = (x^2-2)e^{-3x}$

Sine and Cosine Solutions

Determine a solution y_p for the given differential equation.

25. $y'' = \sin(x)$

26. $y'' = \cos(x)$

27. $y'' + y = \sin(x)$

28. $y'' + y = \cos(x)$

29. $y'' = (x+1)\sin(x)$

30. $y'' = (x+1)\cos(x)$

31. $y'' - y = (x+1)e^x \sin(2x)$

32. $y'' - y = (x+1)e^x \cos(2x)$

33. $y'' - y' - y = e^x \sin(2x)$

34. $y'' - y' - y = (x^2+x)e^x \cos(2x)$

Undetermined Coefficients Algorithm

Determine a solution y_p for the given differential equation.

35. $y'' = x + \sin(x)$

36. $y'' = 1 + x + \cos(x)$

37. $y'' + y = x + \sin(x)$

38. $y'' + y = 1 + x + \cos(x)$

39. $y'' + y = \sin(x) + \cos(x)$

40. $y'' + y = \sin(x) - \cos(x)$

41. $y'' = x + xe^x + \sin(x)$

42. $y'' = x - xe^x + \cos(x)$

43. $y'' - y = \sinh(x) + \cos^2(x)$

44. $y'' - y = \cosh(x) + \sin^2(x)$

45. $y'' + y' - y = x^2 e^x$

46. $y'' + y' - y = xe^x \sin(2x)$

Roots and Related Atoms

Euler atoms A and B are said to be **related** if and only if the derivative lists A, A', … and B, B', … share a common Euler atom.

47. Find the roots, listed according to multiplicity, for the atoms 1, x, x^2, e^{-x}, $\cos 2x$, $\sin 3x$, $x\cos \pi x$, $e^{-x}\sin 3x$.

48. Find the roots, listed according to multiplicity, for the atoms 1, x^3, e^{2x}, $\cos x/2$, $\sin 4x$, $x^2 \cos x$, $e^{3x}\sin 2x$.

49. Let $A = xe^{-2x}$ and $B = x^2e^{-2x}$. Verify that A and B are related.

50. Let $A = xe^{-2x}$ and $B = x^2e^{2x}$. Verify that A and B are not related.

51. Prove that atoms A and B are related if and only if their base atoms have the same roots.

52. Prove that atoms A and B are related if and only if they are in the same **group**. See page 474 for the definition of a group of atoms.

Modify a Trial Solution

Apply Rule II to modify the given Rule I trial solution into the shortest trial solution.

53. The characteristic equation has factors r^3, (r^3+2r^2+2), $(r-1)^2$, $(r+1)$, $(r^2+4)^3$ and the Rule I trial solution is constructed from atoms 1, x, e^x, xe^x, e^{-x}, $\cos 2x$, $\sin 2x$, $\cos x$, $\sin x$.

54. The characteristic equation has factors r^2, (r^3+3r^2+2), $(r+1)$, $(r^2+4)^3$ and the Rule I trial solution is constructed from atoms 1, x, e^x, xe^x, e^{-x}, $\cos 2x$, $\sin 2x$.

Annihilators and Laplace Theory

Laplace theory can construct the annihilator of $f(t)$. The example $y''+4y=t+2t^3$ is used to discuss the techniques. Formulas to be justified: $p(s)=\mathcal{L}(f)/\mathcal{L}(y)$ and $q(s)=\textbf{denom}(\mathcal{L}(f(t)))$.

55. (Transfer Function) Find the characteristic polynomial $q(r)$ for the homogeneous equation $y''+4y=0$. The transfer function for $y''+4y=f(t)$ is $\mathcal{L}(y)/\mathcal{L}(f)$, which equals $1/q(s)$.

56. (Laplace of $y_p(t)$)
The Laplace of $y(t)$ for problem $y''+4y=f(t)$, $y(0)=y'(0)=0$ must equal the Laplace of $f(t)$ times the transfer function. Justify and explain what it has to do with finding y_p.

57. (Annihilator of $f(t)$)
Let $g(t)=t+2t^3$. Verify that $\mathcal{L}(g(t))=\dfrac{s^2+12}{s^4}$, which is a proper fraction with denominator s^4. Then explain why one annihilator of $g(t)$ has characteristic polynomial r^4. The result means that $y=g(t)=t+2t^3$ is a solution of $y''''=0$.

58. (Laplace Theory finds y_p)
Show that the problem $y''+4y=t+2t^3$, $y(0)=y'(0)=0$ has Laplace transform

$$\mathcal{L}(y)=\frac{s^2+12}{(s^2+4)s^4}.$$

Explain why $y(t)$ must be a solution of the constant-coefficient homogeneous differential equation having characteristic polynomial $w(r)=(r^2+4)r^4$.

Annihilator Method Justified

The method of annihilators can be justified by successive differentiation of a nonhomogeneous differential equation, then forming a linear combination of the resulting formulas. It is carried out here, for exposition efficiency, for the nonhomogeneous equation $y''+4y=x+2x^3$. The right side is $f(x)=x+2x^3$ and the homogeneous equation is $y''+4y=0$.

59. (Homogeneous equation)
Verify that $y''+4y=0$ has characteristic polynomial $q(r)=r^2+4$.

60. (Annihilator)
Verify that $y^{(4)}=0$ is an annihilator for $f(x)=x+2x^3$, with characteristic polynomial $q(r)=r^4$.

61. (Composite Equation)
Differentiate four times across the equation $y''+4y=f(x)$ to obtain $y^{(6)}+4y^{(4)}=f^{(4)}(x)$. Argue that $f^{(4)}(x)=0$ because $y^{(4)}=0$ is an annihilator of $f(x)$. This proves that y_p is a solution of higher order equation $y^{(6)}+4y^{(4)}=0$. Then argue that $w(r)=r^4(r^2+4)$ is the characteristic polynomial of the equation $y^{(6)}+4y^{(4)}=0$.

62. (General Solution)
Solve the homogeneous composite equation $y^{(6)}+4y^{(4)}=0$ using its characteristic polynomial $w(r)=r^4(r^2+4)$.

63. (Extraneous Atoms)
Argue that the general solution from the previous exercise contains two terms constructed from atoms derived from roots of the polynomial $q(r)=r^2+4$. Remove these terms to obtain the shortest expression for y_p and explain why it works.

64. (Particular Solution)
Report the form of the shortest particular solution of $y''+4y=f(x)$, according to the previous exercise.

6.6 Undamped Mechanical Vibrations

The study of vibrating mechanical systems begins here with examples for undamped systems with one degree of freedom. The main example is a mass on a spring. The undamped, unforced cases are considered in a number of physical examples, which include the following: simple pendulum, compound pendulum, swinging rod, torsional pendulum, shockless auto, sliding wheel, rolling wheel.

Simple Harmonic Motion

Consider the spring-mass system of Figure 2, where x measures the signed distance from the equilibrium position of the mass. The spring is assumed to exert a force under both compression and elongation. Such springs are commonly used in automotive suspension systems, notably coil springs and leaf springs. In the case of coil springs, there is normally space between the coils, allowing the spring to exert bidirectional forces.

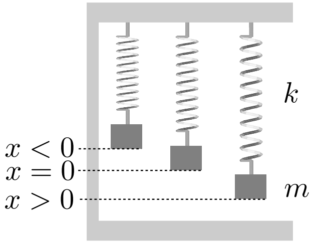


Figure 2. An Undamped Spring-Mass System.
Compression, equilibrium and elongation of the spring are shown with corresponding positions of the mass m.

Hooke's Law. The basic physical law to be applied is:

> The linear restoring force F exerted by a spring is proportional to the signed elongation X, briefly, $F = -kX$.

The number k is called **Hooke's constant** for the spring. In the model of Figure 2, $X = x(t)$ and $k > 0$. The minus sign accounts for the action of the force: the spring tries to **restore** the mass to the equilibrium state, so the vector force is directed toward the equilibrium position $x = 0$.

Newton's Second Law. Specialized to the model in Figure 2, Newton's second law says:

> The force F exerted by a mass m attached to a spring is $F = ma$ where $a = d^2x/dt^2$ is the acceleration of the mass.

The **Weight** $W = mg$ is defined in terms of the **Gravitational Constant** $g = 32$ ft/s^2, 9.8 m/s^2 or 980 cm/s^2 where the mass m is given respectively in **slugs**, **kilograms** or **grams**. The weight is the force due to gravity and it has the appropriate units for a force: **pounds** in the case of the fps system of units.

Method of Force Competition

Hooke's law $F = -kx(t)$ and Newton's second law $F = mx''(t)$ give two independent equations for the force acting on the system. Equating competing forces implies that the signed displacement $x(t)$ satisfies the **Free Vibration** equation

$$mx''(t) + kx(t) = 0.$$

It is also called the **Harmonic Oscillator** in its equivalent form

$$x''(t) + \omega^2 x(t) = 0, \quad \omega^2 = \frac{k}{m}.$$

In this context, ω is the **Natural Frequency** of the free vibration. The harmonic oscillator is said to describe a **Simple Harmonic Motion** $x(t)$. By Theorem 6.1 page 431:

$$x(t) = c_1 \cos \omega t + c_2 \sin \omega t$$

Background: Fundamental Trigonometric Identities

The identities used repeatedly in differential equations applications are:

$$\begin{aligned}&\cos^2\theta + \sin^2\theta = 1\\ &1 + \tan^2\theta = \sec^2\theta\\ &\cot^2\theta + 1 = \csc^2\theta\end{aligned}$$

Pythagorean identities. $\boxed{1}$

$$\begin{aligned}&\sin(-\theta) = -\sin(\theta)\\ &\cos(-\theta) = \cos(\theta)\end{aligned}$$

Odd-even identities. $\boxed{2}$

$\boxed{1}$: Divide the first by $\cos^2\theta$ or $\sin^2\theta$ to derive the others.

$\boxed{2}$: Identities like $\tan(-\theta) = -\tan(\theta)$ can be derived as needed from these two identities, e.g., $\tan\theta = \sin\theta/\cos\theta$.

$$\begin{aligned}&\sin(a+b) = \sin(a)\cos(b) + \sin(b)\cos(a)\\ &\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)\end{aligned}$$

Sum identities. $\boxed{3}$

$$\begin{aligned}&\sin(a-b) = \sin(a)\cos(b) - \sin(b)\cos(a)\\ &\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b)\end{aligned}$$

Difference identities. $\boxed{4}$

$\boxed{3}$: Obtain the second from the first by differentiation on symbol a, holding b constant.

$\boxed{4}$: Both follow from the sum identities by replacing symbol b by $-b$, then apply the even-odd relations.

Background: Harmonic Motion

It is known from trigonometry that

$$x(t) = A\cos(\omega t - \alpha)$$

has **Amplitude** A, **Period** $2\pi/\omega$ and **Phase shift** α/ω. A full period is called a **Cycle** and a half-period a **Semicycle**. The **Frequency** $\omega/(2\pi)$ is the number of complete cycles per second, or the reciprocal of the period.

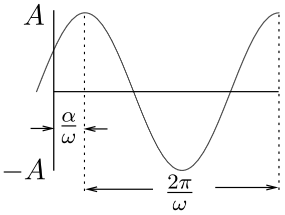


Figure 3. Simple Harmonic Motion. Shown is $x(t) = A\ \cos(\omega t - \alpha)$, period $2\pi/\omega$, phase shift α/ω and amplitude A.

Visualization of Harmonic Motion

A simple harmonic motion can be obtained graphically by means of the experiment shown in Figure 4, in which an undamped spring-mass system has an attached pen that writes on a moving paper chart. The chart produces the simple harmonic motion $x(t) = c_1\cos\omega t + c_2\sin\omega t$ or equivalently $x(t) = A\cos(\omega t - \alpha)$.

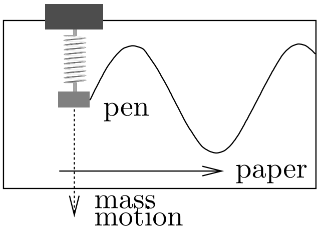


Figure 4. A Chart from Harmonic Motion. A moving paper chart records the vertical motion of a mass on a spring using an attached pen.

Phase-Amplitude Conversion

Given a simple harmonic motion $x(t) = c_1\cos\omega t + c_2\sin\omega t$, as in Figure 3, define **Amplitude** A and **Phase angle** α by the formulas

$$A = \sqrt{c_1^2 + c_2^2}, \quad c_1 = A\cos\alpha \quad \text{and} \quad c_2 = A\sin\alpha.$$

Then the simple harmonic motion has the **Phase-Amplitude form**

$$x(t) = A\cos(\omega t - \alpha). \tag{1}$$

Details. Equation (1) is derived from the cosine difference identity page 491 and basic triangle definitions of sine and cosine.

$x(t) = c_1\cos\omega t + c_2\sin\omega t$	Harmonic oscillator $x'' + \omega^2 x = 0$, general solution.

$x(t) = A\cos\alpha\cos\omega t + A\sin\alpha\sin\omega t$	Insert identities $c_1 = A\cos\alpha$ and $c_2 = A\sin\alpha$.
$x(t) = A\cos(\omega t - \alpha)$	Use $a = \omega t$ and $b = \alpha$ in the cosine difference identity.

Phase Shift Calculations. The phase shift is the amount of horizontal translation required to shift the cosine curve $\cos(\omega t - \alpha)$ so that its graph is atop $\cos(\omega t)$. To find the phase shift from equation (1), set the argument of the cosine term to zero, then solve for t.

To solve for $\alpha \ge 0$ and less than 2π, the expected range, form equations $c_1 = A\cos\alpha$, $c_2 = A\sin\alpha$, then compute numerically by calculator the radian angle $\phi = \arctan(c_2/c_1)$, $|\phi| < \pi/2$. Quadrantial angle rules are applied when $c_1 = 0$ or $c_2 = 0$. Calculators return a division by zero error for $c_1 = 0$ and maybe $\phi = 0$ for $c_2 = 0$, the latter incorrect if $c_1 < 0$. Computers should have **atan2**, a C library function that accepts c_1, c_2 and returns angle $|\phi| < \pi/2$. A calculator or computer answer that is negative requires correction by adding 2π to the radian answer. The corrected answer would give $\cos(\omega t - \alpha - 2\pi)$ instead of $\cos(\omega t - \alpha)$, however the cosine has period 2π: the phase-amplitude answers are equal.

Applications

Considered below are a variety of models with pendulum-like motion. The illustrations start with the simple pendulum and end with applications to auto suspension systems and rolling wheels.

Simple Pendulum

A pendulum is constructed from a thin massless wire or rod of length L and a body of mass m, as in Figure 5.

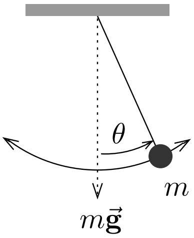


Figure 5. A Simple Pendulum

Derived below is the **Pendulum Equation**

$$\theta''(t) + \frac{g}{L}\sin\theta(t) = 0. \tag{2}$$

Details: Along the circular arc traveled by the mass, the velocity is ds/dt where $s = L\theta(t)$ is arclength. The acceleration is $L\theta''(t)$. Newton's second law for the force along this arc is $F = mL\theta''(t)$. Another relation for the force can be found by resolving the vector gravitational force $m\vec{\mathbf{g}}$ into its normal and tangential components. By trigonometry, the tangential component gives a second force equation $F = -mg\sin\theta(t)$. Equate competing forces and cancel m to obtain (2).

Because the mass m cancels from the equation, the pendulum oscillation depends only upon the length of the string and not upon the mass!

The **Linearized pendulum** equation is

$$\Theta''(t) + \frac{g}{L}\Theta(t) = 0. \tag{3}$$

Details: Approximation $\sin u \approx u$ is valid for small angles u. Apply the approximation to (2). The result is the **linearized pendulum** (3).

Equation (2) is indistinguishable from the classical harmonic oscillator, except for variable names. The solution of (3):

$$\Theta(t) = A\cos(\omega t - \alpha), \quad \omega^2 = g/L$$

Gymnast Swinging about a Horizontal Bar

The mass m of the gymnast is assumed concentrated at the center of the gymnast's physical height H. The problem is then simplified to a pendulum motion with $L = H/2$. The resulting equation of motion for the angle θ between the gravity vector and the gymnast is by equation (2) the **Gymnast's Equation**

$$\theta''(t) + \frac{2g}{H}\sin\theta(t) = 0. \tag{4}$$

The linearized version of this equation is not interesting, because the angle θ is never small. Commonly, $\theta(t)$ goes through many multiples of 2π radians, during an exercise.

Physical Pendulum

The **Compound Pendulum** or **Physical Pendulum** is a rigid body of total mass m having center of mass C which is suspended from a fixed origin O – see Figure 6.

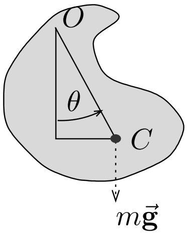


Figure 6. A Physical Pendulum

Derived by force competition is the **Compound Pendulum** equation

$$\theta''(t) + \frac{mgd}{I}\sin\theta(t) = 0. \tag{5}$$

Details: The vector $\vec{r}$ from O to C has magnitude $d = \|\vec{r}\| > 0$. The gravity force vector $\vec{\mathbf{G}} = m\vec{\mathbf{g}}$ (mass × acceleration due to gravity) makes angle θ with vector $\vec{r}$. The restoring torque $\vec{r} \times \vec{\mathbf{G}}$ has magnitude $F = -\|\vec{r} \times \vec{\mathbf{G}}\| = -\|\vec{r}\|\|\vec{\mathbf{G}}\| \sin\theta = -mgd \sin\theta$. Newton's second law gives a second force equation $F = I\theta''(t)$ where I is the torque of the rigid body about O. Force competition results in equation (5).

Approximation $\sin u \approx u$ applied to equation (5) gives a harmonic oscillator known as the **linearized compound pendulum**:

$$\Theta''(t) + \omega^2\Theta(t) = 0, \quad \omega = \sqrt{\frac{mgd}{I}}. \tag{6}$$

Swinging Rod

As depicted in Figure 7, a swinging rod is a special case of the compound pendulum. Assumed for the modeling is a rod of length L and mass m, with uniform mass density.

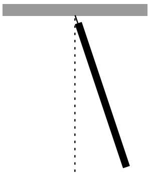


Figure 7. A Swinging Rod

The **Swinging Rod** equation

$$\theta''(t) + \frac{3g}{2L}\sin\theta(t) = 0 \tag{7}$$

will be derived from the compound pendulum equation (5).

Details: The center of mass distance $d = L/2$ appears in the calculus torque relation $I = mL^2/3$. Then:

$$\frac{mgd}{I} = \frac{3mgL}{2mL^2} = \frac{3g}{2L}$$

Insert this relation into the compound pendulum equation (5). The result is the **swinging rod** equation (7).

If equation (6) is used instead (5), then the result is the **linearized swinging rod** equation

$$\Theta''(t) + \omega^2\Theta(t) = 0, \quad \omega = \sqrt{\frac{3g}{2L}}. \tag{8}$$

Torsional Pendulum

A model for a balance wheel in a watch, a gavanometer or a Cavendish torsional balance is the torsional pendulum, which is a rigid body suspended by a solid wire – see Figure 8.

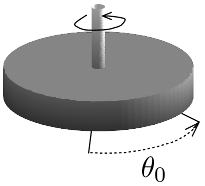


Figure 8. A Torsional Pendulum.
An example is a balance wheel in a watch. The wheel rotates angle θ_0 about the vertical axis, which acts as a spring, exerting torque I against the rotation.

The **Torsional Pendulum** equation:

$$\theta_0''(t) + \omega^2\theta_0(t) = 0, \quad \omega = \sqrt{\frac{\kappa}{I}}. \tag{9}$$

Details: The wire undergoes twisting, which exerts a restoring force $F = -\kappa\theta_0$ when the body is rotated through angle θ_0. There is no small angle restriction on this restoring force, because it acts in the spirit of Hooke's law like a linear spring restoring force. The model uses Newton's second law force relation $F = I\theta_0''(t)$, as in the physical pendulum. Force competition against the restoring force $F = -\kappa\theta_0$ gives the **torsional pendulum** equation (9).

Shockless Auto

An automobile loaded with passengers is supported by four coil springs, as in Figure 9, but all of the shock absorbers are worn out. The simplistic linear model $mx''(t) + kx(t) = 0$ will be applied. The plan is to estimate the number of seconds it takes for one complete oscillation. This is the time between two consecutive *bottom–outs* of the automobile.[5]

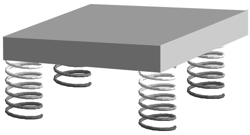


Figure 9. Car on Four Springs: Linear Model

Assume the car plus occupants has mass 1350 Kg. Let each coil spring have Hooke's constant $k = 20000$ Newtons per meter. The load is divided among the four springs equally, so each spring supports $m = 1350/4$ Kg. Let ω be the natural frequency of vibration. Then the number of seconds for one complete oscillation is the period $T = 2\pi/\omega$ seconds. The free vibration model for one spring is

$$\frac{1350}{4}x''(t) + 20000x(t) = 0.$$

The harmonic oscillator form is $x'' + \omega^2 x = 0$, where $\omega^2 = \frac{20000(4)}{1350} = 59.26$. Therefore, $\omega = 7.70$. Then the period is $T = 2\pi/\omega = 0.82$ seconds. The interpretation: the auto bottoms-out every 0.82 seconds.

[5] Teenagers popularized late-night cruising of Los Angeles boulevards in shockless 4-door sedans. They disabled the shock absorbers and modified the suspension to give a completely undamped ride.

Rolling Wheel on a Spring

A wheel of total mass m and radius R is attached at its center to a spring of Hooke's constant k, as in Figure 10. The wheel rolls without slipping. The spring is assumed to have negligible mass and zero kinetic energy. Let k be the Hooke's constant for the spring. Let $x(t)$ be the elongation of the spring from equilibrium, $x > 0$ corresponding to the wheel rolling to the right and $x < 0$ corresponding to the wheel rolling to the left.

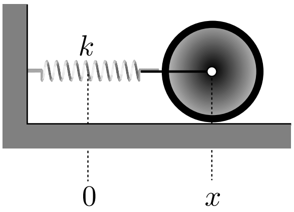


Figure 10. A Rolling Wheel on a Spring.

Derived below is the **Rolling Wheel Equation**

$$mx''(t) + \frac{2}{3}kx(t) = 0. \tag{10}$$

Details: The spring does not react only to tension, but it reacts like a coil spring with spacing that restores bi-directionally to equilibrium.

Figure 11. Restoring Force $F = kx$.
By Hooke's law, the spring restores to equilibrium for both compression and elongation.

If the wheel slides frictionless, then the model is the harmonic oscillator equation $mx''(t) + kx(t) = 0$. A wheel that rolls without slipping has inertia, and consideration of this physical difference will be shown to give equation (10).

A curious consequence is that $x(t)$ is identical to the frictionless sliding wheel with spring constant reduced from k to $2k/3$. This makes sense physically, because rolling wheel inertia is observed to reduce the apparent stiffness of the spring.

The derivation begins with the energy conservation law

$$\text{Kinetic} + \text{Potential} = \text{constant}.$$

The kinetic energy T is the sum of two energies, $T_1 = \frac{1}{2}mv^2$ for translation and $T_2 = \frac{1}{2}I\omega^2$ for the rolling wheel, whose inertia is $I = \frac{1}{2}mR^2$. The velocity is $v = R\omega = x'(t)$. Algebra gives $T = T_1 + T_2 = \frac{3}{4}mv^2$. The potential energy is $K = \frac{1}{2}kx^2$ for a spring of Hooke's constant k. Application of the energy conservation law $T + K = c$ gives the equation $\frac{3}{4}m(x'(t))^2 + \frac{1}{2}k(x(t))^2 = c$. Differentiate this equation on t to obtain $\frac{3}{2}mx'(t)x''(t) + kx(t)x'(t) = 0$, then cancel $x'(t)$ to give equation (10).

Examples and Methods

Example 6.35 (Harmonic Vibration)
A mass of $m = 250$ grams attached to a spring of Hooke's constant k undergoes free undamped vibration. At equilibrium, the spring is stretched 25 cm by a force of 8 Newtons. At time $t = 0$, the spring is stretched 0.5 m and the mass is set in motion with initial velocity 5 m/s directed downward from equilibrium. Find:

(a) The numerical value of Hooke's constant k.

(b) The initial value problem for vibration $x(t)$.

Solution:
(a): Hooke's law Force=k(elongation) is applied with force 8 Newtons and elongation $25/100 = 1/4$ meter. Equation $8 = k(1/4)$ implies $k = 32$ N/m.

(b): Given $m = 250/1000$ kg and $k = 32$ N/m from part (a), then the free vibration model $mx'' + kx = 0$ becomes $\frac{1}{4}x'' + 32x = 0$. Initial conditions are $x(0) = 0.5$ m and $x'(0) = 5$ m/s. The initial value problem is

$$\begin{cases} \dfrac{d^2x}{dt^2} + 128x &= 0, \\ x(0) &= 0.5, \\ x'(0) &= 5. \end{cases}$$

Example 6.36 (Phase-Amplitude Conversion)
Write the vibration equation

$$x(t) = 2\cos(3t) + 5\sin(3t)$$

in phase-amplitude form $x = A\cos(\omega t - \alpha)$. Create a graphic of $x(t)$ with labels for period, amplitude and phase shift.

Solution:
The answer and the graphic appear below.

$$x(t) = \sqrt{29}\cos(3t - 1.190289950) = \sqrt{29}\cos(3(t - 0.3967633167)).$$

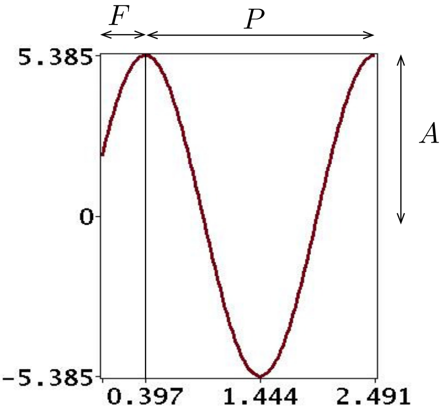


Figure 12. Harmonic Oscillation.

The graph of $2\cos(3t) + 5\sin(3t)$. It has amplitude $A = \sqrt{29} = 5.385$, period $P = 2\pi/3$ and phase shift $F = 0.3967633167$. The graph is on $0 \le t \le P + F$.

Algebra Details. The plan is to re-write $x(t)$ in the form $x(t) = A\cos(\omega t - \alpha)$, called the phase-amplitude form of the harmonic oscillation. The main tools from trigonometry appear on page 491.

Start with $x(t) = 2\cos(3t) + 5\sin(3t)$. Compare the expression for $x(t)$ with Trig identity $x(t) = A\cos(\omega t - \alpha) = A\cos(\alpha)\cos(\omega t) + A\sin(\alpha)\sin(\omega t)$. Then define accordingly

$$\omega = 3, \quad A\cos(\alpha) = 2, \quad A\sin(\alpha) = 5.$$

The Pythagorean identity $\cos^2\alpha + \sin^2\alpha = 1$ implies $A^2 = 2^2 + 5^2 = 29$ and then the amplitude is $A = \sqrt{29}$. Because $\cos\alpha = 2/A$, $\sin\alpha = 5/A$, then both the sine and cosine are positive, placing angle α in quadrant I. Divide equations $\cos\alpha = 2/A$, $\sin\alpha = 5/A$ to obtain $\tan(\alpha) = 5/2$, which by calculator implies $\alpha = \arctan(5/2) = 1.190289950$ radians or 68.19859051 degrees. Then $x(t) = A\cos(\omega t - \alpha) = \sqrt{29}\cos(3t - 1.190289950)$.

Computer Details. Either equation for $x(t)$ can be used to produce a computer graphic. A hand-drawn graphic would use only the phase-amplitude form. The period is $P = 2\pi/\omega = 2\pi/3$. The amplitude is $A = \sqrt{29} = 5.385164807$ and the phase shift is $F = \alpha/\omega = 0.3967633167$. The graph is on $0 \le t \le P + F$.

```
# Maple
F:=evalf(arctan(5/2)/3); P:=2*Pi/3;A:=sqrt(29);
X:=t->2*cos(3*t)+5*sin(3*t);
opts:=xtickmarks=[0,F,P/2+F,P+F],ytickmarks=[-A,0,A],
      axes=boxed,thickness=3,labels=["",""];
plot(X(t),t=0..P+F,opts);
```

Example 6.37 (Undamped Spring-Mass System)

A mass of 6 Kg is attached to a spring that elongates 20 centimeters due to a force of 12 Newtons. The motion starts at equilibrium with velocity -5 m/s. Find an equation for $x(t)$ using the free undamped vibration model $mx'' + kx = 0$.

Solution: The answer is $x(t) = -\sqrt{\frac{5}{2}}\sin(\sqrt{10}t)$.

The mass is $m = 6$ kg. Hooke's law $F = kx$ is applied with $F = 12$ N and $x = 20/100$ m. Then Hooke's constant is $k = 60$ N/m. Initial conditions are $x(0) = 0$ m (equilibrium) and $x'(0) = -5$ m/s. The model is

$$\begin{cases} 6\dfrac{d^2x}{dt^2} + 60x &= 0, \\ x(0) &= 0, \\ x'(0) &= -5. \end{cases}$$

Solve the Initial Value Problem. The characteristic equation $6r^2 + 60 = 0$ is solved for $r = \pm i\sqrt{10}$, then the Euler solution atoms are $\cos(\sqrt{10}t)$, $\sin(\sqrt{10}t)$. The general solution is a linear combination of Euler atoms:

$$x(t) = c_1\cos(\sqrt{10}t) + c_2\sin(\sqrt{10}t).$$

The task remaining is determination of constants c_1, c_2 subject to initial conditions $x(0) = 0$, $x'(0) = -5$. The linear algebra problem uses the derivative formula

$$x'(t) = -\sqrt{10}c_1\sin(\sqrt{10}t) + \sqrt{10}c_2\cos(\sqrt{10}t).$$

The 2×2 system of linear algebraic equations for c_1, c_2 is obtained from the two equations $x(0) = 0$, $x'(0) = -5$ as follows.

$$\left\{\begin{array}{rcrcrl} \cos(0)c_1 & + & \sin(0)c_2 & = & 0, & \text{Equation } x(0)=0 \\ -\sqrt{10}\sin(0)c_1 & + & \sqrt{10}\cos(0)c_2 & = & -5, & \text{Equation } x'(0)=-5 \end{array}\right.$$

Because $\cos(0) = 1$, $\sin(0) = 0$, then $c_1 = 0$ and $c_2 = -5/\sqrt{10} = -\sqrt{5/2}$. Insert answers c_1, c_2 into the general solution to find the answer to the initial value problem:

$$x(t) = -\sqrt{\frac{5}{2}}\sin(\sqrt{10}t).$$

Example 6.38 (Pendulum)

A simple linearized pendulum of length 2.5 m oscillates with angle variable $\theta(t)$ satisfying $\theta(0) = 0$ (equilibrium position) and $\theta'(0) = 3$ (radial velocity). Find $\theta(t)$ in phase-amplitude form and report the period, amplitude and phase shift.

Solution: The answer is $\theta(t) = 3\sqrt{\frac{25}{98}}\sin\left(\sqrt{\frac{98}{25}}\,t\right)$, which has amplitude $3\sqrt{\dfrac{25}{98}}$, period $2\pi\sqrt{\dfrac{25}{98}}$, phase shift zero.

The mass is not given, because we use model equation (3), $\theta''(t) + \frac{g}{L}\theta(t) = 0$, in which $g = 9.8$ and $L = 2.5$. Then the initial value problem is

$$\left\{\begin{array}{rcl} \dfrac{d^2\theta}{dt^2} + \dfrac{98}{25}\theta & = & 0, \\ \theta(0) & = & 0, \\ \theta'(0) & = & 3. \end{array}\right.$$

Solve the Initial Value Problem. The characteristic equation $r^2 + \frac{98}{25} = 0$ is solved for $r = \pm i\omega$ where $\omega = \sqrt{\frac{98}{25}}$. The Euler solution atoms are $\cos(\omega t)$, $\sin(\omega t)$. The general solution:

$$\theta(t) = c_1\cos(\omega t) + c_2\sin(\omega t).$$

The task remaining is determination of constants c_1, c_2 subject to initial conditions $\theta(0) = 0$, $\theta'(0) = 3$. The linear algebra problem uses the derivative formula

$$\theta'(t) = -\omega c_1\sin(\omega t) + \omega c_2\cos(\omega t).$$

The 2×2 system of equations for c_1, c_2 is obtained from equations $\theta(0) = 0$, $\theta'(0) = 3$ as follows.

$$\left\{\begin{array}{rcrcrl} \cos(0)c_1 & + & \sin(0)c_2 & = & 0, & \text{Equation } \theta(0)=0 \\ -\omega\sin(0)c_1 & + & \omega\cos(0)c_2 & = & 3, & \text{Equation } \theta'(0)=3 \end{array}\right.$$

Because $\cos(0) = 1$, $\sin(0) = 0$, then $c_1 = 0$ and $c_2 = 3/\omega = 3\sqrt{\frac{25}{98}}$. The solution to the initial value problem is

$$\theta(t) = 3\sqrt{\frac{25}{98}}\sin\left(\sqrt{\frac{98}{25}}\,t\right).$$

Example 6.39 (Gymnast)

Consider the change of variables $x(t) = \theta(t)$, $y(t) = \theta'(t)$, called the **position-velocity substitution**. Re-write the gymnast equation (4), $\theta'' + \frac{2g}{H}\sin\theta = 0$, in the form

$$\text{(11)} \qquad \begin{aligned} \frac{dx}{dt} &= y(t), \\ \frac{dy}{dt} &= -\frac{2g}{H}\sin(x(t)). \end{aligned}$$

Apply the method of quadrature to develop the equation for the **total mechanical energy**

$$\text{(12)} \qquad \frac{1}{2}y^2 + \frac{2g}{H}(1 - \cos x) = E.$$

Solution: The terms in the energy equation (12) are $\frac{1}{2}y^2$, called the **Kinetic Energy**, and $\omega^2(1 - \cos x)$, called the **Potential Energy**. We will show that $E = \frac{1}{2}y(0)^2$.

Details for (11): Define $x(t) = \theta(t)$ and $y(t) = \theta'(t)$. Then

$$\begin{aligned} x' &= \theta' & \\ &= y & \text{Used } x(t) = \theta(t) \text{ and } y(t) = \theta'(t). \\ y' &= \theta'' & \\ &= -\frac{2g}{H}\sin(\theta) & \text{Used } x(t) = \theta(t) \text{ and } \theta'' + \tfrac{2g}{H}\sin\theta = 0. \\ &= -\frac{2g}{H}\sin(x) & \end{aligned}$$

Details for (12): Because $y = x'$, we multiply the second equation in (11) by y and then re-write the resulting equation as

$$yy' = -\frac{2g}{H}x'\sin(x).$$

This is a quadrature equation. Integrate on variable t across the equation to obtain for some constant C the identity

$$\frac{1}{2}y^2 = \frac{2g}{H}\cos(x) + C.$$

Let $t = 0$ in this equation to evaluate $C = \frac{1}{2}(y(0))^2 - \frac{2g}{H}$. Then rearrange terms to obtain the equation

$$\frac{1}{2}y^2 + \frac{2g}{H}(1 - \cos(x)) = \frac{1}{2}(y(0))^2.$$

This is equation (12) with $E = \frac{1}{2}(y(0))^2$.

Example 6.40 (Swinging Rod)

A uniform rod of length 16 cm swings from a support at origin $\mathcal{O}$. The motion started at angle $\theta(0) = \pi/12$ radians with radial velocity zero. Find approximate equations for the motion at the extreme end of the rod in rectangular coordinates.

Solution: The answer is

$$\begin{aligned} x(t) &= \frac{16}{100}\cos(\theta(t)), \\ y(t) &= \frac{16}{100}\sin(\theta(t)), \\ \theta(t) &= \frac{\pi}{12}\cos\left(\frac{t}{2}\sqrt{735}\right). \end{aligned}$$

The mass is not given, because we use model equation (8), $\theta''(t) + \frac{3g}{2L}\sin(\theta(t)) = 0$, in which $g = 9.8$ m/s^2 and $L = 16/100$ m. Then the initial value problem is

$$\begin{cases} \dfrac{d^2\theta}{dt^2} + \dfrac{735}{4}\sin(\theta) &= 0, \\ \theta(0) &= \pi/2, \\ \theta'(0) &= 0. \end{cases}$$

The linearized equation will be used to find an approximate formula for the motion. The initial value problem is

$$(13) \quad \begin{cases} \dfrac{d^2\theta}{dt^2} + \dfrac{735}{4}\theta &= 0, \\ \theta(0) &= \pi/12, \\ \theta'(0) &= 0. \end{cases}$$

The rectangular coordinates for the end of the rod are

$$x(t) = L\cos(\theta(t)), \quad y(t) = L\sin(\theta(t)).$$

Solve the Initial Value Problem. As in two previous examples, system (8) is readily solved with general solution

$$\theta(t) = c_1\cos(\omega t) + c_2\sin(\omega t), \quad \omega = \frac{\sqrt{735}}{2}.$$

Initial conditions imply $c_1 = \frac{\pi}{12}, c_2 = 0$. Details not supplied. Then

$$\theta(t) = \frac{\pi}{12}\cos\left(\frac{\sqrt{735}}{2}t\right).$$

Final Answer. The formula for $\theta(t)$ is inserted into polar coordinate equations $x = r\cos\theta, y = r\sin\theta$ with $r = L$ to obtain the reported answers.

Example 6.41 (Torsional Pendulum)

The balance wheel of a classical watch oscillates with angular amplitude π radians and period 0.5 seconds. Find the following values.

(a) The maximum angular speed of the balance wheel.

(b) The angular speed when the angle equals $\pi/2$ radians.

(c) The angular acceleration when the angle equals $\pi/4$ radians.

Solution: The answers are (a) $4\pi^2$, (b) $-2\pi^2\sqrt{3}$, (c) $-4\pi^3$.

The model is equation (9), $\theta_0''(t) + \omega^2\theta_0(t) = 0$, where $\omega = \sqrt{\frac{\kappa}{I}}$. The general solution in phase-amplitude form is $\theta_0(t) = A\cos(\omega t - \alpha)$, with constants A, α replacing the constants c_1, c_2 in a general solution. We are given that $A = \pi$. The period $\frac{2\pi}{\omega}$ equals 0.5, which implies $\omega = 4\pi$. Then

$$\theta_0(t) = \pi\cos(4\pi t - \alpha).$$

The constant α is undetermined by the information supplied.

(a): The angular speed is $\theta_0'(t) = -4\pi^2\sin(4\pi t - \alpha)$. It is a maximum when the sine factor equals -1. Then $\theta_0'(t) = 4\pi^2$ is the maximum angular speed of the balance wheel.

(b): The angle $\theta_0(t) = \pi/2$ is valid only when the cosine factor in $\theta_0(t) = \pi\cos(4\pi t - \alpha)$ is equal to $1/2$. Then $\sin(4\pi t - \alpha) = \sqrt{3}/2$, from trigonometry. The angular speed at this moment is $\theta_0'(t) = -4\pi^2\sin(\omega t - \alpha) = -2\pi^2\sqrt{3}$.

(c): Apply the equation $\theta_0''(t) + \omega^2\theta_0(t) = 0$ to obtain the acceleration relation $\theta_0''(t) = -16\pi^2\theta_0(t)$. When $\theta_0(t) = \pi/4$, then the acceleration equals $-4\pi^3$.

Example 6.42 (Shockless Auto)

A shockless auto of total mass 1400 kg bounces on a level street, making 8 **bottom-outs** in 10 seconds. Estimate the Hooke's constant k for each of the four coil springs.

Solution: The answer is $k = 8750\pi^2 \approx 86596$.

The model equation $mx'' + kx = 0$ is used. Then $x(t) = A\cos(\omega t - \alpha)$ is a general solution, with A and α constant and $\omega^2 = \frac{k}{m}$. The mass is not 1400 kg, but 1/4 of that, because each of the four springs carries an equal load. Let $m = 1400/4$. The period of oscillation is $2\pi/\omega$, which has to equal $\frac{1}{2}\frac{8}{10}$, because two bottom-outs mark one complete cycle. Then $\frac{2\pi}{\omega} = \frac{4}{10}$ implies $\omega = 5\pi$. Finally, $k = m\omega^2 = \frac{1400}{4}(5\pi)^2 = 8750\pi^2$.

Example 6.43 (Rolling Wheel)

A wheel of mass 10 kg and radius 0.35 m rolls frictionless with attached coil spring as in Figure 23. The observed frequency of oscillation is 8 full cycles every 3 seconds. Estimate the Hooke's constant k of the spring.

Solution: The answer is $k = \frac{135\pi^2}{16}$.

The rolling wheel model (10) will be used, equation $mx''(t) + \frac{2}{3}kx(t) = 0$. Known is the mass $m = 10$ kg and the general solution $x(t) = A\cos(\omega t - \alpha)$ with A and α constant and natural frequency $\omega = \sqrt{\frac{2}{3}\frac{k}{m}}$. Then the period of oscillation is $\frac{2\pi}{\omega} = \frac{8}{3}$, because the cosine factor passes through 8 periods in 3 seconds. The equation determines $\omega = \frac{3}{8}(2\pi)$. Then $k = m\frac{3}{2}\omega^2 = 10\frac{3}{2}\left(\frac{6\pi}{8}\right)^2 = \frac{135\pi^2}{16}$.

Exercises 6.6

Simple Harmonic Motion

Determine the model equation $mx''(t) + kx(t) = 0$, the natural frequency $\omega = \sqrt{k/m}$, the period $2\pi/\omega$ and the solution $x(t)$ for the following spring–mass systems.

1. A mass of 4 Kg attached to a spring of Hooke's constant 20 Newtons per meter starts from equilibrium plus 0.05 meters with velocity 0.

2. A mass of 2 Kg attached to a spring of Hooke's constant 20 Newtons per meter starts from equilibrium plus 0.07 meters with velocity 0.

3. A mass of 2 Kg is attached to a spring that elongates 20 centimeters due to a force of 10 Newtons. The motion starts at equilibrium with velocity -5 meters per second.

4. A mass of 4 Kg is attached to a spring that elongates 20 centimeters due to a force of 12 Newtons. The motion starts at equilibrium with velocity -8 meters per second.

5. A mass of 3 Kg is attached to a coil spring that compresses 2 centimeters when 1 Kg rests on the top coil. The motion starts at equilibrium plus 3 centimeters with velocity 0.

6. A mass of 4 Kg is attached to a coil spring that compresses 2 centimeters when 2 Kg rests on the top coil. The motion starts at equilibrium plus 4 centimeters with velocity 0.

7. A mass of 5 Kg is attached to a coil spring that compresses 1.5 centimeters when 1 Kg rests on the top coil. The motion starts at equilibrium plus 3 centimeters with velocity -5 meters per second.

8. A mass of 4 Kg is attached to a coil spring that compresses 2.2 centimeters when 2 Kg rests on the top coil. The motion starts at equilibrium plus 4 centimeters with velocity -8 meters per second.

9. A mass of 5 Kg is attached to a spring that elongates 25 centimeters due to a force of 10 Newtons. The motion starts at equilibrium with velocity 6 meters per second.

10. A mass of 5 Kg is attached to a spring that elongates 30 centimeters due to a force of 15 Newtons. The motion starts at equilibrium with velocity 4 meters per second.

Phase–amplitude Form

Solve the given differential equation and report the general solution. Solve for the constants c_1, c_2. Report the solution in phase–amplitude form

$$x(t) = A\cos(\omega t - \alpha)$$

with $A > 0$ and $0 \le \alpha < 2\pi$.

11. $x'' + 4x = 0$,
$x(0) = 1$, $x'(0) = -1$

12. $x'' + 4x = 0$,
$x(0) = 1$, $x'(0) = 1$

13. $x'' + 16x = 0$,
$x(0) = 2$, $x'(0) = -1$

14. $x'' + 16x = 0$,
$x(0) = -2$, $x'(0) = -1$

15. $5x'' + 11x = 0$,
$x(0) = -4$, $x'(0) = 1$

16. $5x'' + 11x = 0$,
$x(0) = -4$, $x'(0) = -1$

17. $x'' + x = 0$,
$x(0) = 1$, $x'(0) = -2$

18. $x'' + x = 0$,
$x(0) = -1$, $x'(0) = 2$

19. $x'' + 36x = 0$,
$x(0) = 1$, $x'(0) = -4$

20. $x'' + 64x = 0$,
$x(0) = -1$, $x'(0) = 4$

Pendulum

The formula

$$\frac{P_1}{P_2} = \frac{R_1}{R_2}\sqrt{\frac{L_1}{L_2}}$$

is valid for the periods P_1, P_2 of two pendulums of lengths L_1, L_2 located at distances R_1, R_2 from the center of the earth. The formula implies that a pendulum can be used to find the radius of the earth at a location. It is also useful for designing a pendulum clock adjustment screw.

21. Derive the formula, using $\omega = \sqrt{g/L}$, period $P = 2\pi/\omega$ and the gravitational relation $g = GM/R^2$.

22. A pendulum clock taken on a voyage loses 2 minutes a day compared to its exact timing at home. Determine the altitude change at the destination.

23. A pendulum clock with adjustable length L loses 3 minutes per day when $L = 30$ inches. What length L adjusts the clock to perfect time?

24. A pendulum clock with adjustable length L loses 4 minutes per day when $L = 30$ inches. What fineness length F is required for a 1/4–turn of the adjustment screw, in order to have 1/4–turns of the screw set the clock to perfect time plus or minus one second per day?

Torsional Pendulum

Solve for $\theta_0(t)$.

25. $\theta_0''(t) + \theta_0(t) = 0$

26. $\theta_0''(t) + 4\theta_0(t) = 0$

27. $\theta_0''(t) + 16\theta_0(t) = 0$

28. $\theta_0''(t) + 36\theta_0(t) = 0$

Shockless Auto

Find the period and frequency of oscillation of the car on four springs. Use model $mx''(t) + kx(t) = 0$.

29. Assume the car plus occupants has mass 1650 Kg. Let each coil spring have Hooke's constant $k = 20000$ Newtons per meter.

30. Assume the car plus occupants has mass 1850 Kg. Let each coil spring have Hooke's constant $k = 20000$ Newtons per meter.

31. Assume the car plus occupants has mass 1350 Kg. Let each coil spring have Hooke's constant $k = 18000$ Newtons per meter.

32. Assume the car plus occupants has mass 1350 Kg. Let each coil spring have Hooke's constant $k = 16000$ Newtons per meter.

Rolling Wheel on a Spring

Solve the rolling wheel model $mx''(t) + \frac{2}{3}\,kx(t) = 0$ and also the frictionless model $mx''(t) + kx(t) = 0$, each with the given initial conditions. Graph the two solutions $x_1(t), x_2(t)$ on one set of axes.

33. $m = 1$, $k = 4$,
$x(0) = 1$, $x'(0) = 0$

34. $m = 5$, $k = 18$,
$x(0) = 1$, $x'(0) = 0$

35. $m = 11$, $k = 18$,
$x(0) = 0$, $x'(0) = 1$

36. $m = 7$, $k = 18$,
$x(0) = 0$, $x'(0) = 1$

6.7 Forced and Damped Vibrations

The study of vibrating mechanical systems continues. The main example is a system consisting of an externally forced mass on a spring with damping.[6] Both undamped and damped systems are studied. A few physical examples are included: clothes dryer, cafe door, pet door, bicycle trailer.

Forced Undamped Motion

The equation for study is a forced spring–mass system

$$mx''(t) + kx(t) = f(t).$$

The model originates by equating the Newton's second law force $mx''(t)$ to the sum of the Hooke's force $-kx(t)$ and the external force $f(t)$. The physical model is a laboratory box containing an undamped spring–mass system, transported on a truck as in Figure 13, with external force $f(t) = F_0 \cos \omega t$ induced by the speed bumps.

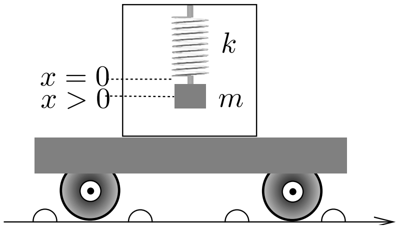


Figure 13. An undamped, forced spring-mass system.
A box containing a spring-mass system is transported on a truck. Speed bumps on the shoulder of the road transfer periodic vertical oscillations to the box.

The **forced spring-mass system** takes the form $x''(t) + \omega_0^2\, x(t) = \frac{F_0}{m} \cos \omega t$. Symbol $\omega_0 = \sqrt{k/m}$ is called the **Natural Frequency**. It is the number of full periods of free oscillation per second for the **unforced spring–mass system** $x''(t) + \omega_0^2\, x(t) = 0$. The **External Frequency** ω is the number of full periods of oscillation per second of the external force $f(t) = F_0 \cos \omega t$. In the case of Figure 13, $f(t)$ is the vertical force applied to the box containing the spring–mass system, due to the speed bumps. The general solution $x(t)$ always presents itself in two pieces, as the sum of the homogeneous solution x_h and a particular solution x_p. For $\omega \neq \omega_0$, the solution formulas are (full details on page 522)

$$\text{(1)}\qquad \begin{array}{lcl} \multicolumn{3}{l}{x''(t) + \omega_0^2\, x(t) = \dfrac{F_0}{m} \cos \omega t, \quad \omega_0 = \sqrt{\dfrac{k}{m}}\qquad ,} \\ x(t) & = & x_h(t) + x_p(t), \\ x_h(t) & = & c_1 \cos \omega_0 t + c_2 \sin \omega_0 t, \quad c_1,\ c_2 \text{ constants}, \\ x_p(t) & = & \dfrac{F_0/m}{\omega_0^2 - \omega^2} \cos \omega t. \end{array}$$

A general statement can be made about the solution decomposition:

[6] Damping is energy dissipation and dampening is making something wet.

> The solution is a sum of two harmonic oscillations, one of natural frequency ω_0 due to the spring and the other of natural frequency ω due to the external force $F_0 \cos \omega t$.

Beats

The physical phenomenon of **beats** refers to the periodic interference of two sound waves of slightly different frequencies. Human heartbeat uses the same terminology. Our pulse rate is $40 - 100$ **beats** per minute at rest. The phenomenon of **beats** will be explained mathematically *infra*. An illustration of the graphical meaning is in Figure 14.

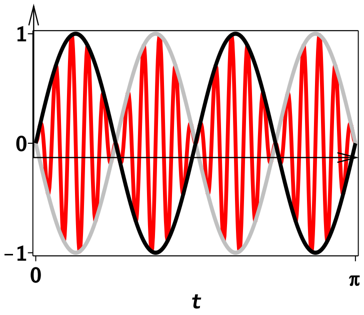


Figure 14. Beats.
Shown is a periodic oscillation

$$x(t) = 2 \sin 4t \sin 40t$$

with rapidly–varying factor $\sin 40t$ and the two slowly–varying envelope curves

$$x_1(t) = 2 \sin 4t, \quad x_2(t) = -2 \sin 4t.$$

A key example is piano tuning. A tuning fork is struck, then the piano string is tuned until the beats are not heard. The number of beats per second (unit Hz) is approximately the frequency difference between the two sources, e.g., two tuning forks of frequencies 440 Hz and 437 Hz would produce 3 beats per second.

The average human ear can detect beats only if the two interfering sound waves have a frequency difference of about 7 Hz or less. Ear-tuned pianos are subject to the same human ear limitations. Two piano keys are more than 7 Hz apart, even for a badly tuned piano, which is why simultaneously struck piano keys are heard as just one sound (no beats).

A **destructive interference** occurs during a very brief interval, so our impression is that the sound periodically stops, only briefly, and then starts again with a *beat*, a section of sound that is instantaneously loud again. The beat we hear corresponds to maxima in Figure 14.

In Figure 14, we see not the two individual sound waves, but their **superposition**, because $2 \sin(4t) \sin(40t) = \cos(36t) - \cos(44t) =$ sum of two harmonic oscillations of different frequencies. See equation 2 below for details. When the tuning fork and the piano string have the same exact frequency ω, then Figure 14 would show a simple harmonic wave, because the two sounds would **superimpose** to a graph that looks like $\cos(\omega t - \alpha)$.

The origin of the phenomenon of **beats** can be seen from the formula

$$x(t) = 2\sin at \sin bt.$$

There is no sound when $x(t) \approx 0$: this is when destructive interference occurs. When a is small compared to b, e.g., $a = 4$ and $b = 40$, then there are long intervals between the zeros of $A(t) = 2\sin at$, at which destructive interference occurs. Otherwise, the amplitude of the sound wave is the average value of $A(t)$, which is 1. The sound stops at a zero of $A(t)$ and then it is rapidly loud again, causing the beat.

Black Box in the Trunk

Return to the forced harmonic oscillator

$$x''(t) + \omega_0^2\, x(t) = \frac{F_0}{m}\cos\omega t, \quad \omega_0 = \sqrt{\frac{k}{m}},$$

whose solution $x(t)$ appears in equation (1). The expression for $x(t)$ will show the phenomenon of **beats** for certain choices of frequencies ω_0, ω and initial position and velocity $x(0)$, $x'(0)$.

For instance, consider one possible expression $x(t) = \cos(\omega_0 t) - \cos(\omega t)$. Use the trigonometric identity $2\sin c \sin d = \cos(c-d) - \cos(c+d)$, derived from identities on page 491, to write

$$(2) \qquad x(t) = \cos(\omega_0 t) - \cos(\omega t) = 2\sin\frac{1}{2}(\omega - \omega_0)t \sin\frac{1}{2}(\omega_0 + \omega)t.$$

If $\omega \approx \omega_0$, then the first factor $2\sin\frac{1}{2}(\omega-\omega_0)t$ has natural frequency $a = \frac{1}{2}(\omega-\omega_0)$ near zero. The natural frequency $b = \frac{1}{2}(\omega_0 + \omega)$ of the other factor can be relatively large and therefore $x(t)$ is a product of a **Slowly Varying** oscillation $2\sin at$ and a **Rapidly Varying** oscillation $\sin bt$. A graphic of $x(t)$ looks like Figure 14.

Rotating Drum on a Cart

Figure 15 shows a model for a rotating machine, like a front–loading clothes dryer.

For modeling purposes, the rotating drum with load is replaced by an idealized model: a mass $\mathcal{M}$ on a string of radius R rotating with angular speed ω. The center of rotation is located along the center–line of the cart. The total mass m of the cart includes the rotating mass $\mathcal{M}$, which we imagine to be an off–center lump of wet laundry inside the dryer drum. Vibrations cause the cart to skid left or right.

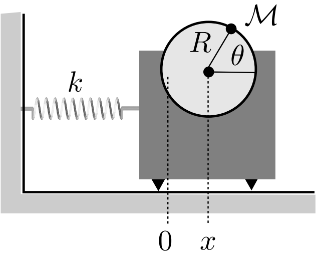


Figure 15. Rotating Vertical Drum.

Like a front-loading clothes dryer, or a washing machine, the drum is installed on a cart with skids. An internal spring restores the cart to equilibrium $x = 0$.

A spring of Hooke's constant k restores the cart to its equilibrium position $x = 0$. The cart has position $x > 0$ corresponding to skidding distance x to the right of the equilibrium position, due to the off-center load. Similarly, $x < 0$ means the cart skidded distance $|x|$ to the left.

Modeling. Friction ignored, Newton's second law gives force $F = m\overline{x}''(t)$, where $\overline{x}$ locates the cart's center of mass. Hooke's law gives force $F = -kx(t)$. The centroid $\overline{x}$ can be expanded in terms of $x(t)$ by using calculus moment of inertia formulas. Let $m_1 = m - \mathcal{M}$ be the cart mass, $m_2 = \mathcal{M}$ the drum mass, $x_1 = x(t)$ the moment arm for m_1 and $x_2 = x(t) + R\cos\theta$ the moment arm for m_2. Then $\theta = \omega t$ in Figure 15 gives

$$\begin{aligned} \overline{x}(t) &= \frac{m_1x_1 + m_2x_2}{m_1 + m_2} \\ &= \frac{(m - \mathcal{M})x(t) + \mathcal{M}(x(t) + R\cos\theta)}{m} \\ &= x(t) + \frac{R\mathcal{M}}{m}\cos\omega t. \end{aligned} \tag{3}$$

Force competition $m\overline{x}'' = -kx$ and derivative expansion results in the forced harmonic oscillator

$$mx''(t) + kx(t) = R\mathcal{M}\omega^2 \cos\omega t. \tag{4}$$

Forced Damped Motion

Real systems do not exhibit idealized harmonic motion, because **damping** occurs. A watch balance wheel submerged in oil is a key example: frictional forces due to the viscosity of the oil will cause the wheel to stop after a short time. The same wheel submerged in air will appear to display harmonic motion, but indeed there is friction present, however small, which slows the motion.

Consider a spring–mass system consisting of a mass m and a spring with Hooke's constant k, with an added **dashpot** or **damper**, depicted in Figure 16 as a piston inside a cylinder attached to the mass. A useful physical model, for purposes of intuition, is a screen door with USA hardware: the door is equipped with a spring to restore the door to the jamb position and an adjustable piston–cylinder style dashpot.

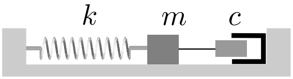


Figure 16. A spring-mass system with dashpot

The dashpot is assumed to operate in the **viscous domain**, which means that the force due to the damper device is proportional to the speed that the mass is moving: $F = cx'(t)$. The number $c \geq 0$ is called the **dashpot constant**. Three forces act: (1) Newton's second law $F_1 = mx''(t)$, (2) viscous damping $F_2 = cx'(t)$ and (3) the spring restoring force $F_3 = kx(t)$. The sum of the forces $F_1 + F_2 + F_3$ acting on the system must equal the **External Force** $f(t)$, which gives the equation for a **Forced Damped Spring–Mass System**

$$mx''(t) + cx'(t) + kx(t) = f(t). \tag{5}$$

If there is no external force, $f(t) = 0$, then the vibration is called **free** or **unforced** and otherwise it is called **forced**. Equation (5) is called **damped** if $c > 0$ and **undamped** if $c = 0$.

A useful visualization for a forced system is a vertical laboratory spring–mass system with dashpot placed inside a box, which is transported down a washboard road inside an auto trunk. The function $f(t)$ is the vertical oscillation of the auto trunk. The function $x(t)$ is the signed excursion of the mass in response to the washboard road. See Figure 17.

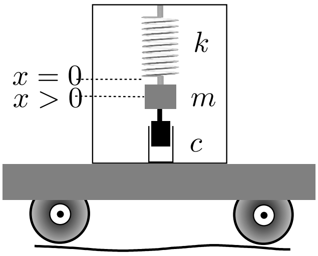


Figure 17. A Damped Spring-Mass System with External Forcing.
The apparatus is placed in a box, then transported in an auto trunk along a washboard road. Vertical excursion $x(t)$ of the mass is measured from equilibrium.

Seismoscope

The 1875 **horizontal motion seismoscope** of F. Cecchi (1822-1887) reacted to an earthquake. It started a clock, and then it started motion of a recording surface, which ran at a speed of 1cm per second for 20 seconds. The clock provided the observer with the earthquake hit time.

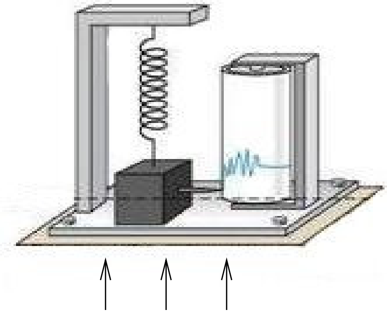


Figure 18. A Simplistic Vertical Motion Seismoscope.

The apparently stationary heavy mass on a spring writes with the attached stylus onto a rotating drum, as the ground moves up.

The motion of the heavy mass m in Figure 18 can be modeled by a forced spring-mass system with damping. The first model has the form

$$mx'' + cx' + kx = f(t)$$

where $f(t)$ is the vertical ground force due to the earthquake. In terms of the vertical ground motion $u(t)$, Newton's second law gives the force equation $f(t) = -mu''(t)$. The second model for the motion of the mass is then

$$\begin{aligned}
&x''(t) + 2\beta\Omega_0 x'(t) + \Omega_0^2 x(t) = -u''(t),\\
&\tfrac{c}{m} = 2\beta\Omega_0, \quad \tfrac{k}{m} = \Omega_0^2,\\
&x(t) = \text{mass position measured from equilibrium},\\
&u(t) = \text{vertical ground motion due to the earthquake.}
\end{aligned} \tag{6}$$

Some observations about equation (6):

Slow ground movement means $x' \approx 0$ and $x'' \approx 0$, then (6) implies $\Omega_0^2 x(t) = -u''(t)$. The seismometer records ground acceleration.

Fast ground movement means $x \approx 0$ and $x' \approx 0$, then (6) implies $x''(t) = -u''(t)$. The seismometer records ground displacement.

A **release test** will find β, Ω_0 experimentally. See the exercises for details.

The point of (6) is to determine $u(t)$, by knowing $x(t)$ from the seismograph.

Free damped motion

Consider the special case of no external force, $f(t) = 0$. The vibration $x(t)$ satisfies the homogeneous differential equation

$$mx''(t) + cx'(t) + kx(t) = 0. \tag{7}$$

Cafe Door

Restaurant waiters and waitresses are accustomed to the cafe door, which partially blocks the view of onlookers, but allows rapid, collision-free trips to the kitchen – see Figure 19. The door is equipped with a spring which tries to restore the door to the equilibrium position $x = 0$, which is the plane of the door frame. There is a dashpot attached, to keep the number of oscillations low.

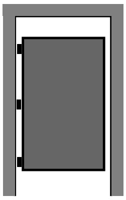


Figure 19. A Cafe Door.
There are three hinges with dashpot in the lower hinge.
The equilibrium position is the plane of the door frame.

The top view of the door, Figure 20, shows how the angle $x(t)$ from equilibrium $x = 0$ is measured from different door positions.

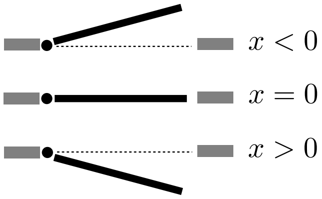


Figure 20. Top View of a Cafe Door.
The three possible door positions.

$x < 0$ kitchen
$x = 0$ door frame
$x > 0$ restaurant

The figure shows that, for modeling purposes, the cafe door can be reduced to a torsional pendulum with viscous damping. This results in the **cafe door** equation

$$Ix''(t) + cx'(t) + \kappa x(t) = 0. \tag{8}$$

The removal of the spring ($\kappa = 0$) causes the vibration $x(t)$ to be monotonic, which is a reasonable fit to a springless cafe door.

Pet Door

Designed for dogs and cats, the small door in Figure 21 permits free entry and exit.

Figure 21. A Pet Door.
The equilibrium position is the plane of the door frame.
The door swings from hinges on the top edge.
One hinge is spring-loaded with dashpot.

Like the cafe door, the spring restores the door to the equilibrium position while the dashpot acts to eventually stop the oscillations. However, there is one fundamental difference: if the spring–dashpot system is removed, then the door continues to oscillate! The cafe door model will not describe the pet door.

For modeling purposes, the door can be compressed to a linearized swinging rod of length L (the door height). The torque $I = mL^2/3$ of the door assembly becomes important, as well as the linear restoring force kx of the spring and the viscous damping force cx' of the dashpot. All considered, a suitable model is the **pet door** equation

$$I\,x''(t) + cx'(t) + \left(k + \frac{mgL}{2}\right)x(t) = 0. \tag{9}$$

Derivation of (9) is by equating to zero the algebraic sum of the forces.

Removing the dashpot and spring ($c = k = 0$) gives a harmonic oscillator $x''(t) + \omega^2 x(t) = 0$ with $\omega^2 = \dfrac{mgL}{2I}$, which matches physical intuition. Equation (9) is *formally* the cafe door equation with an added linearization term $\dfrac{mgL}{2}x(t)$ obtained from $\dfrac{mgL}{2}\sin x(t)$.

Modeling Unforced Damped Vibration

The cafe door (8) and the pet door (9) have equations in the same form as a damped spring–mass system (7), and all equations can be reduced, for suitable definitions of constants p and q, to the simplified second order differential equation

$$x''(t) + p\,x'(t) + q\,x(t) = 0. \tag{10}$$

The solution $x(t)$ of this equation is a linear combination of two Euler atoms determined by the roots of the characteristic equation

$$r^2 + pr + q = 0.$$

There are three types of solutions possible, organized by the sign of the discriminant

$$p^2 - 4q.$$

Positive Discriminant	Distinct real roots $r_1 \neq r_2$ $x = c_1 e^{r_1 t} + c_2 e^{r_2 t}$
Zero Discriminant	Double real root $r_1 = r_2$ $x = c_1 e^{r_1 t} + c_2\, t\, e^{r_1 t}$
Negative Discriminant	Complex conjugate roots $a \pm i\,b$ $x = e^{at}(c_1 \cos bt + c_2 \sin bt)$

Tuning a dashpot

The pet door and the cafe door have dashpots with an adjustment screw. The screw changes the dashpot coefficient c which in turn changes the size of coefficient p in (10). More damping c means p is larger.

There is a *critical damping effect* for a certain screw setting: if the setting is decreased more, then the door *oscillates*, whereas if the setting is increased, then the door has a *monotone non-oscillatory behavior.* The monotonic behavior can result in the door opening in one direction followed by slowly settling to exactly the door jamb position. If p is too large, then it could take 10 minutes for the door to close!

The critical case corresponds to the least $p > 0$ (the smallest damping constant $c > 0$) required to close the door with this kind of monotonic behavior. The same can be said about decreasing the damping: the more p is decreased, the more the door oscillations approach those of no dashpot at all, which is a pure harmonic oscillation.

As viewed from the characteristic equation $r^2 + pr + q = 0$, the change is due to a change in character of the roots from real to complex, which is measured by a sign change from positive to negative for the **Discriminant** $p^2 - 4q$. The physical response and the three cases of the constant–coefficient theorem, page 431, lead to the following terminology.

Classification	**Defining properties**
Overdamped	Distinct real roots $r_1 \neq r_2$ Positive discriminant $x = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ $= $ exponential × monotonic function

Critically damped	Double real root $r_1 = r_2$ Zero discriminant $x = c_1 e^{r_1 t} + c_2\, t\, e^{r_1 t}$ $=$ exponential × monotonic function
Underdamped	Complex conjugate roots $a \pm i\, b$ Negative discriminant $x = e^{at}(c_1 \cos bt + c_2 \sin bt)$ $=$ exponential × harmonic oscillation

Envelope Curves and Pseudo-Period

In the under-damped case the solution $x(t)$ of $mx''+cx'+kx = 0$ can be expressed in phase-amplitude form

$$\begin{aligned} x(t) &= e^{at}(c_1 \cos bt + c_2 \sin bt) \\ &= e^{at} C \cos(bt - \alpha). \end{aligned}$$

In this formula, $c_1 = C\cos\alpha$, $c_2 = C\sin\alpha$ and $C = \sqrt{c_1^2 + c_2^2}$. The **Pseudo-Period** is $T = \dfrac{2\pi}{b}$, so named because the harmonic factor $\cos(bt - \alpha)$ has period $2\pi/b$. The factor Ce^{at} generates the two **envelope curves**

$$y = Ce^{at}, \quad y = -Ce^{at}.$$

The solution $x(t)$ oscillates entirely inside the region defined by the envelope curves. Crossings of the t-axis happen at $bt = n\pi + \alpha$, $n = 0, \pm 1, \pm 2, \ldots$. Contact with the envelope curves happens at $bt = n\pi + \pi/2 + \alpha$, $n = 0, \pm 1, \pm 2, \ldots$.

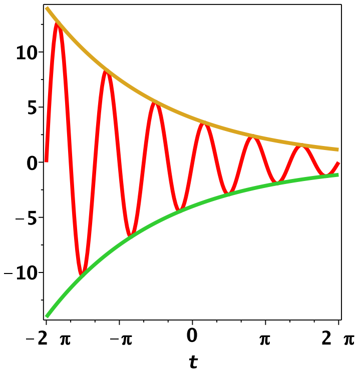


Figure 22. Envelope Curves.
A particular solution of the differential equation $25x'' + 10x' + 226x = 0$ is

$x(t) = 4e^{-t/5} \sin 3t$ red,

which has pseudo-period $T = \dfrac{2\pi}{3}$.
The envelope curves are

$x_1(t) = 4e^{-t/5}$ yellow,
$x_2(t) = -4e^{-t/5}$ green.

Bicycle trailer

An auto tows a one–wheel trailer over a washboard road. Shown in Figure 23 is the trailer strut, which has a single coil spring and two dampers. The mass m includes the trailer and the bicycles.

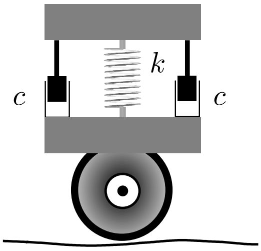


Figure 23. A trailer strut with dampers on a washboard road

Suppose a washboard dirt road has about 2 full oscillations (2 bumps and 2 valleys) every 3 meters and a full oscillation has amplitude 6 centimeters. Let s denote the horizontal distance along the road and let ω be the number of full oscillations of the roadway per unit length. The oscillation period is $2\pi/\omega$, therefore $2\pi/\omega = 3/2$ or $\omega = 4\pi/3$. A model for the road surface is

$$y = \frac{5}{100}\cos \omega s.$$

Let $x(t)$ denote the vertical elongation of the spring, measured from equilibrium. Newton's second law gives a force $F_1 = mx''(t)$ and the viscous damping force is $F_2 = 2cx'(t)$. The trailer elongates the spring by $x - y$, therefore the Hooke's force is $F_3 = k(x - y)$. The sum of the forces $F_1 + F_2 + F_3$ must be zero, which implies

$$mx''(t) + 2cx'(t) + k(x(t) - y(t)) = 0.$$

Write $s = vt$ where v is the speedometer reading of the car in meters per second. The expanded differential equation is the forced damped spring-mass system equation

$$mx''(t) + 2cx'(t) + kx(t) = \frac{k}{20}\cos(4\pi vt/3).$$

The solution $x(t)$ of this model, with $x(0)$ and $x'(0)$ given, describes the vertical excursion of the trailer bed from the roadway. The **observed oscillations** of the trailer are modeled by the steady-state solution

$$x_{\mathrm{SS}}(t) = A\cos(4\pi vt/3) + B\sin(4\pi vt/3),$$

where A, B are constants determined by the method of undetermined coefficients. From physical data, the amplitude $C = \sqrt{A^2 + B^2}$ of this oscillation might be 6cm or larger. The maximum amplitude C over all speedometer readings v can be found by calculus. The computation uses the formula

$$(11) \qquad C(v) = \frac{k/20}{\sqrt{(k - m\omega^2)^2 + (2c\omega)^2}}, \quad \omega = \frac{4\pi v}{3}.$$

Set $\frac{dC}{dv} = 0$ and then solve for the speed v^* which maximizes $C(v)$. The maximum excursion of the trailer is then

$$C(v^*) = \frac{km}{40c\sqrt{km - c^2}}.$$

The values of k, m, c can be found from an experiment: record $C(v)$ at three different speeds $v = v_1, v_2, v_3$. Then solve the system of three equations in three unknowns m, k, c, arising from (11).

Examples and Methods

Example 6.44 (Forced Undamped Vibration)

Solve the vibration equation

$$x'' + 225x = 209\cos(4t).$$

Solution: The answer is $x(t) = c_1\cos(15t) + c_2\sin(15t) + \cos(4t)$. The vibration is an example of **beats** for certain values of c_1, c_2. The solution is a superposition of two harmonic oscillations of frequencies 15 and 4. There are two ways to solve the problem, detailed below.

First Solution Details. A shortcut is to use equations (1), page 506. The given equation $x'' + 225x = 209\cos(4t)$ provides symbols $m = 1$, $k = 225$, $F_0 = 209$, $\omega = 4$. Then $\omega_0 = \sqrt{225} = 15$ is the unforced natural frequency of vibration. Substitution of the symbols into equations (1) gives $x_h = c_1\cos(15t) + c_2\sin(15t)$ and $x_p = F_1\cos(4t)$ with $F_1 = (209/1)/(225 - 4^2) = 1$. By superposition $x = x_h + x_p$. The reported solution is verified.

Second Solution Details. The characteristic equation $r^2 + 225 = 0$ of the homogeneous problem $x'' + 225x = 0$ has complex conjugate roots $\pm 15i$ and Euler solution atoms $\cos(15t), \sin(15t)$. Then $x_h(t) = c_1\cos(15t) + c_2\sin(15t)$.

A particular solution by Rule I of the method of undetermined coefficients is $x(t) = A\cos(4t) + B\sin(4t)$. Substitution into the non-homogeneous equation $x'' + 225x = 209\cos(4t)$ gives the relation

$$-16(A\cos(4t) + B\sin(4t)) + 225(A\cos(4t) + B\sin(4t)) = 209\cos(4t).$$

It reduces to the equation

$$209A\cos(4t) + 226B\sin(4t) = 209\cos(4t).$$

Independence of Euler atoms $\cos(4t)$, $\sin(4t)$ implies matching coefficients. Then $B = 0$ and $A = 1$. The trial solution $x(t) = A\cos(4t) + B\sin(4t)$ upon substitution of $A = 1, B = 0$ becomes particular solution $x_p(t) = \cos(4t)$.

Superposition gives general solution $x(t) = x_h(t) + x_p(t)$, therefore the answer reported has been verified.

Example 6.45 (Beats)

Write the linear combination $x(t) = \cos 10t - \cos 20t$ in the form $x(t) = C\sin at\sin bt$. Then graph the slowly-varying envelope curves and the curve $x(t)$.

Solution: The answer is $x(t) = 2\sin(5t)\sin(15t)$, which implies $C = 2, a = 5, b = 15$ with envelope curves $\pm 2\sin 5t$ (sine factor with longer period appears first). The graphic in Figure 24 is made from these formulas using a computer graphics program.

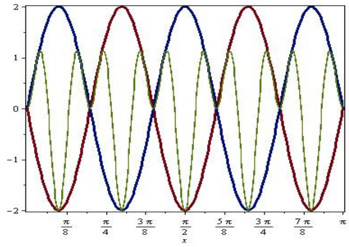


Figure 24. Beats Oscillation.
Plot of slowly-varying envelopes $\pm 2\sin(5t)$ and the oscillation $x(t) = 2\sin(5t)\sin(15t)$.

Details. The basic tool is the cosine sum formula from page 491. Let's assemble the formulas

$$\begin{aligned}\cos(A-B) &= \cos A\cos B + \sin A\sin B,\\ \cos(A+B) &= \cos A\cos B - \sin A\sin B.\end{aligned}$$

Because $x(t) = \cos 10t - \cos 20t = \cos(A-B) - \cos(A+B) = 2\sin A\sin B$, then choose $A - B = 10t$ and $A + B = 20t$. Then the unique solution is $A = 15t, B = 5t$, which implies the formula

$$x(t) = 2\sin A\sin B = 2\sin(15t)\sin(5t)\,.$$

The slowly-varying envelope curves are $\pm 2\sin(5t)$, because the sine factor periods are $2\pi/15$ and $2\pi/5$, the second being the longer period.

Example 6.46 (Rotating Drum)

An unloaded European-style washing machine weighs 156 lbs. When loaded with an off-center wet mass of 4 kg, it has horizontal excursions x from equilibrium satisfying approximately the rotating drum equation (4):

$$mx''(t) + kx(t) = R\mathcal{M}\omega^2\cos\omega t.$$

Assume Hooke's spring constant $k = 10$ slugs per foot. The drum has diameter 30 in and during a water extraction cycle it rotates at 600 rpm. Discuss assumptions and computations for the values $\mathcal{M} = 0.275$, $m = 5.15$, $R = 1.25$ and $\omega = 20\pi$. Then compute the approximate expression

$$(12) \qquad x\,(t) = c_1\cos\left(\frac{20t}{\sqrt{206}}\right) + c_2\sin\left(\frac{20t}{\sqrt{206}}\right) - 55\,\frac{\pi^2\cos(20\,\pi\,t)}{824\,\pi^2 - 4}.$$

Solution:

Details. Central to the mathematical formulation is Newton's formula $W = mg$, which in words is *weight W (a force) equals mass m times gravitational acceleration g.* Use $g = 32$ ft/sec per second, for simplicity of discussion. Using $g = 32.2$ changes constants in a minor way.

Basic plan. Use model (4). After, we will be tormented and humiliated by closer analysis of the physical problem. Let's assume the centroid of the wet load is approximately on the edge of the rotating drum, in order to simplify the formulas and use model (4). The rotating machine in the absence of the wet load is assumed to operate at equilibrium $x = 0$. Issues like additional internal damping and frictional forces on the mounting surface will be patently ignored with no apologies.

Wet load mass $\mathcal{M}$: A unit conversion is required for the wet load mass: 4 kg represents 4(2.2) lbs. Then $W = 8.8$ lbs is the wet load weight and its mass is $\mathcal{M} = W/g = 8.8/32 = 0.275$ slugs.

Total machine mass m: Total machine weight is $W = 156 + 8.8 = 164.8$ lbs, then formula $W = mg$ implies the total mass is $m = 164.8/32 = 5.15$ slugs.

Drum radius R: A conversion to feet is required, giving $R = \frac{1}{2}(30)$ in $= \frac{15}{12}$ in $= 1.25$ ft.

Natural frequency of rotation ω: Supplied is the rotational period $2\pi/\omega$, which is equal to 1/10 second (600 revolutions in 60 seconds). Solve $2\pi/\omega = 1/10$ for $\omega = 20\pi$.

Solution $x(t)$: We'll use equation (4) with the constants inserted:

$$5.15x''(t) + 10x(t) = 1.25(0.275)(20\pi)^2 \cos(20\pi t).$$

Without machine assist, the homogeneous equation $5.15x''(t) + 10x(t) = 0$ is solved as $x_h = c_1 \cos(bt) + c_2 \sin(bt)$ where $b = \sqrt{k/m} = 20/\sqrt{206}$. Then undetermined coefficients is applied with (shortcut) trial solution $x = A\cos(20\pi t)$ to the non-homogeneous problem, giving

$$A = \frac{-55\,\pi^2}{824\,\pi^2 - 4}, \quad x_p = \frac{-55\,\pi^2}{824\,\pi^2 - 4}\cos(20\pi t).$$

The reported answer in equation (12) is $x = x_h + x_p$.

Answer check: Computer algebra system `maple` solves the equation using this code:

```
f:=t->1.25*(0.275)*(20*Pi)^2*cos (20*Pi* t);
de:=5.15*diff(x(t),t,t)+10*x(t)=f(t);
dsolve(de,x(t));
```

Vibrations of x_p have amplitude about 0.13 cm and period 0.1. The harmonic vibrations of x_h have a longer period of about 4.5. For example, if the spin cycle starts from rest, then $x(t)$ will have amplitude of about 0.13 and its graphic on $0 < t < 4.5$ will look like a beats figure, with slow oscillation envelope of approximate period 4.5.

Example 6.47 (Damped Spring-Mass System)

Let $x(t)$ be the defected distance from equilibrium in a damped spring-mass system with free oscillation equation

$$4x''(t) + 3x'(t) + 17x(t) = 0.$$

Find an expression for $x(t)$.

Solution: The answer is

$$x(t) = c_1 e^{-3t/8}\cos(\sqrt{263}t/8) + c_2 e^{-3t/8}\sin(\sqrt{263}t/8).$$

Details. The homogeneous solution $x(t)$ is a linear combination of two Euler solution atoms found from the characteristic equation $4r^2 + 3r + 17 = 0$. The roots according to the quadratic formula are $-\frac{3}{8} \pm \frac{i}{8}\sqrt{263}$. Then the two Euler solution atoms are

$$e^{-3t/8}\cos(\sqrt{263}t/8), \quad e^{-3t/8}\sin(\sqrt{263}t/8),$$

from which the solution formula follows.

Remarks. The oscillation is classified as **under-damped**, because of the presence of sine and cosine oscillatory factors in the Euler solution atoms. Any solution is the product of an exponential factor and a harmonic oscillation, therefore the solution is **pseudo-periodic** with **pseudo-period** $16\pi/\sqrt{263}$.

Example 6.48 (Seismoscope)

Consider the seismoscope equation

$$x''(t) + 12x'(t) + 100x(t) = -u''(t).$$

Find an expression for the seismoscope stylus record $x(t)$ in terms of the ground motion $u(t)$.

Solution: In terms of particular solution $x_p(t)$, defined below in integral equation (14) or (15), the answer is

$$x(t) = c_2 e^{-6t}\cos(8t) + c_2 e^{-6t}\sin(8t) + x_p(t). \tag{13}$$

Details. The solution method is superposition $x(t) = x_h(t) + x_p(t)$ where x_h is the solution of the homogeneous equation $x''(t) + 12x'(t) + 100x(t) = 0$ and x_p is a variation of parameters solution of the non-homogeneous equation $x''(t)+12x'(t)+100x(t) = f(t)$, where $f(t) = -u''(t)$.

Homogeneous solution x_h. The characteristic equation $r^2 + 12r + 100 = 0$ has factorization $(r+6)^2 + 64 = 0$, hence complex conjugate roots $r = -6 \pm 8i$. The Euler solution atoms are $e^{-6t}\cos(8t)$, $e^{-6t}\sin(8t)$, from which we construct the general solution

$$x_h(t) = c_2 e^{-6t}\cos(8t) + c_2 e^{-6t}\sin(8t).$$

Non-homogeneous solution x_p. Let's start by writing the variation of parameters formula in the different form

$$\begin{aligned} x_p(t) &= y_1(t)\left(\int_0^t -\frac{y_2(x)f(x)}{W(x)}dx\right) + y_2(t)\left(\int_0^t \frac{y_1(x)f(x)}{W(x)}dx\right) \\ &= \int_0^t \frac{W_1(t,x)}{W(x)} f(x)dx \end{aligned}$$

where

$$\begin{aligned} f(x) &= -u''(x), \\ y_1(t) &= e^{-6t}\cos(8t), \\ y_2(t) &= e^{-6t}\sin(8t), \\ W(x) &= 8e^{-12x}, && \text{Details below in } \boxed{\mathbf{1}}. \\ W_1(t,x) &= -y_1(t)y_2(x) + y_2(t)y_1(x) \\ &= e^{-6t-6x}(\sin 8t\cos 8x - \cos 8t \sin 8x) \\ &= e^{-6t-6x}\sin(8t-8x). && \text{Trig identity.} \end{aligned}$$

Condensing the definitions gives the final formula

$$x_p(t) = -\int_0^t e^{-6t+6x}\sin(8t-8x)u''(x)dx. \tag{14}$$

It is possible to integrate this equation by parts and express the answer entirely in terms of $u(t)$. Some integration by parts free terms are collected into $x_h(t)$ to produce the replacement formula

$$\begin{aligned} x_p^*(t) &= -u(t) + \int_0^t K(t-x)u(x)dx, \\ K(w) &= 12\,e^{-6w}\cos(8w) + \frac{7}{2}\,e^{-6w}\sin(8w). \end{aligned} \tag{15}$$

Laplace theory can derive formula (15) using the convolution theorem. Generally, (14) and (15) are different answers.

$\boxed{\mathbf{1}}$ **Wronskian determinant details**.

A shortcut is to use Theorem 6.17, page 464. The answer is $W(x) = W(0)e^{-12x}$ where $W(0) = 8$ is computed from the first line of the determinant expansion below. Details below compute $W(x)$ directly from the definition.

$$W(x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}$$

Variation of parameters definition of the Wronskian of y_1, y_2.

$$= \begin{vmatrix} y_1 & y_2 \\ -6y_1 - 8y_2 & -6y_2 + 8y_1 \end{vmatrix}$$

Because $y_1' = -6y_1 - 8y_2$ and $y_2' = -6y_2 + 8y_1$.

$$= \begin{vmatrix} y_1 & y_2 \\ -8y_2 & 8y_1 \end{vmatrix}$$

Combination rule `combo(1,2,6)`.

$$= 8(y_1^2 + y_2^2)$$

Sarrus' Rule.

$$= 8e^{-12x}(\cos^2(8x) + \sin^2(8x))$$

Expand $y_1(x) = e^{-6x}\cos(8x)$ and $y_2(x) = e^{-6x}\sin(8x)$.

$$= 8e^{-12x}.$$

Pythagorean identity.

Example 6.49 (Cafe Door)
Consider the cafe door equation (8):

$$Ix''(t) + cx'(t) + \kappa x(t) = 0.$$

Find an expression for $x(t)$. Then show details for why the motion $x(t)$ is eventually monotonic when the spring is removed.

Solution:

First, divide by torque $I > 0$ to obtain equation $x'' + 2ax' + bx = 0$ with new symbols $2a = c/I$, $b = \kappa/I$. The characteristic equation is $(r + a)^2 + b - a^2 = 0$. There are three cases determined by the sign of $b - a^2$ for the form of the solution. Because $b - a^2 = \frac{4I\kappa - c^2}{4I^2}$, then $b - a^2$ has sign determined by $4I\kappa - c^2$.

Case $4I\kappa - c^2 > 0$.
Then the characteristic equation roots are complex conjugates $-a \pm i\sqrt{b - a^2}$. The solution is **under-damped**, oscillatory and given by

$$\begin{aligned} x(t) &= c_1 e^{-at}\cos(\sqrt{b - a^2}\, t) + c_2 e^{-at}\sin(\sqrt{b - a^2}\, t) \\ &= c_1 e^{\frac{ct}{2I}}\cos\left(\sqrt{4I\kappa - c^2}\,\frac{t}{2I}\right) + c_2 e^{\frac{ct}{2I}}\sin\left(\sqrt{4I\kappa - c^2}\,\frac{t}{2I}\right). \end{aligned}$$

Case $4I\kappa - c^2 = 0$.
Then the characteristic equation roots are equal, $-a, -a$. The solution is **critically damped**, non-oscillatory and given by

$$x(t) = c_1 e^{-at} + c_2\, t\, e^{-at} = c_1 e^{\frac{ct}{2I}} + c_2\, t\, e^{\frac{ct}{2I}}.$$

Case $4I\kappa - c^2 < 0$.
Then the characteristic equation roots are real and unequal, $-a \pm \sqrt{a^2 - b}$. The solution is **over-damped**, non-oscillatory and given by

$$\begin{aligned} x(t) &= c_1 e^{-at - \sqrt{a^2 - b}\, t} + c_2\, t\, e^{-at + \sqrt{a^2 - b}\, t} \\ &= c_1\, e^{\left(c - \sqrt{c^2 - 4I\kappa}\right)\frac{t}{2I}} + c_2\, e^{\left(c + \sqrt{c^2 - 4I\kappa}\right)\frac{t}{2I}}. \end{aligned}$$

Cafe door with no spring. This event is defined by $\kappa = 0$, which eliminates the under-damped case $4I\kappa - c^2 > 0$. Suppose hereafter that $x(t)$ is a nonzero solution. The critically damped case is $a = 0$. Then the solution can be written as $x(t) = c_1 + c_2\,t$, which crosses the axis $x = 0$ at most once. The over-damped case $4I\kappa - c^2 < 0$ can be written $x(t) = \left(c_1 + c_2\,e^{B}\right)\,e^{At}$ where $B > 0$. Similarly, it crosses the axis $x = 0$ at most once, due to the factor $c_1 + c_2\,e^{Bt}$.

Example 6.50 (Pet Door)

A pet door of height $L = 1.5$ feet and weight 8 pounds oscillates freely because the dashpot has been removed. Assume Hooke's spring constant $k = 10$. Find an expression for the angular motion $x(t)$ using equation (9) with torque $I = mL^2/3$:

$$I\,x''(t) + cx'(t) + \left(k + \frac{mgL}{2}\right)x(t) = 0.$$

Solution:
Removal of the dashpot corresponds to $c = 0$. The mass m satisfies $W = mg$, which from $W = 8$ and $g = 32$ gives $m = 0.25$ slugs. Then the torque is $I = mL^2/3 = L^2/12 = 3/16$ and $mgL/2 = 3g/16 = 6$. Equation (9) becomes

$$\frac{3}{16}\,x''(t) + 16x(t) = 0.$$

This is the classical harmonic oscillator $x'' + \omega^2 x = 0$ with $\omega^2 = 16^2/3$. Then $\omega = 16/\sqrt{3}$ and

$$x(t) = c_1\cos\left(\frac{16\,t}{\sqrt{3}}\right) + c_2\sin\left(\frac{16\,t}{\sqrt{3}}\right).$$

Example 6.51 (Tuning a Dashpot)

Classify the following equations as over-damped, critically damped or under-damped free vibrations.

(a) $x'' + 2x' + 3x = 0$

(b) $x'' + 4x' + 3x = 0$

(c) $x'' + 2x' + x = 0$

Solution: The answers: (a) Under-damped, (b) Over-damped, (c) Critically damped. Definitions on page 513.

Details (a). The characteristic equation $r^2 + 2r + 3 = 0$ factors into $(r + 1)^2 + 2 = 0$ with complex conjugate roots $-1 \pm i\sqrt{2}$. The Euler solution atoms contain sines and cosines, therefore (a) is oscillatory, classified as under-damped.

Details (b). The characteristic equation $r^2+4r+3 = 0$ factors into $(r+3)(r+1) = 0$ with distinct real roots $-3, -1$. Therefore, (b) is non-oscillatory, classified as over-damped because of distinct roots.

Details (c). The characteristic equation $r^2 + 2r + 1 = 0$ factors into $(r + 1)(r + 1) = 0$ with equal real roots $-1, -1$. Therefore, (b) is non-oscillatory, classified as critically damped because of equal roots.

Summary of Methods. Classification requires only the roots of the characteristic equation.

Over-damped means *too much damping*. In the screen door example, the tuning screw has made the dashpot constant c large, which means an overly-aggressive dashpot that halts motion. This means the screen door hangs open. Then the screen door has no oscillations, equivalently, $x(t)$ has no sines or cosines.

Critically damped is an unstable state. In the screen door example, it is the impossible to achieve the ideal dashpot tuning screw setting on a screen door: the door opens and then slowly closes to the jamb position, the door hardware making a single click as it locks the door on the jamb. A turn of the tuning screw in either direction jumps between oscillation and non-oscillation of the screen door.

Under-damped means *not enough damping effect*. Physically, the dashpot is not effective. In the screen door example this means the screen door oscillates and bangs repeatedly on the door jamb. Detection in $x(t)$ is the presence of oscillating sines and cosines. Solution $x(t)$ is called *oscillatory*.

Example 6.52 (Pseudo-Period)

Find the pseudo-period and time-varying amplitude for the free damped vibration

$$4x'' + 2x' + 3x = 0, \quad x(0) = 1, \quad x'(0) = -1.$$

Solution: The answers: Pseudo period $8\pi/\sqrt{11}$ and amplitude $4e^{-t/4}$ are obtained from the solution $x(t) = 4e^{-t/4}\cos\left(\sqrt{11}\,\frac{t}{4}\right)$.

Details. The characteristic equation $4r^2 + 2r + 3 = 0$ has complex conjugate roots $-\frac{1}{4} \pm i\,\frac{\sqrt{11}}{4}$, obtained from the quadratic formula. Then the general solution is

$$x(t) = c_1\, e^{-t/4}\cos\left(\sqrt{11}\,\frac{t}{4}\right) + c_2\, e^{-t/4}\sin\left(\sqrt{11}\,\frac{t}{4}\right).$$

Initial conditions $x(0) = 4, x'(0) = -1$ give the two equations

$$\begin{array}{rcrcr} (1)c_1 & + & (0)c_2 & = & 4, \\ \left(\frac{-1}{4}\right)c_1 & + & \left(\frac{\sqrt{11}}{4}\right)c_2 & = & -1, \end{array}$$

with unique solution $c_1 = 4, c_2 = 0$. The pseudo-period is the period $2\pi/\omega$ of the trig factor $\cos(\omega t)$, where $\omega = \frac{1}{4}\sqrt{11}$. The time-varying amplitude is the factor in front of the cosine factor, namely $4e^{-t/4}$.

Remark on Method. If both c_1, c_2 are nonzero, then a trig identity is applied first to write $x(t) = Ae^{-t/4}\cos(\omega t - \alpha)$. The amplitude is then $Ae^{-t/4}$. The period is unchanged.

Proofs and Details

Details for equation (1), page 506:

Homogeneous solution x_h. The characteristic equation for $x'' + \omega_0^2 x = 0$ is $r^2 + \omega_0^2 = 0$ with complex conjugate roots $r = \pi i\omega_0$. Then the Euler solution atoms are $\cos(\omega_0 t), \sin(\omega_0 t)$. The general solution is a linear combination of the Euler solution atoms, as displayed in equation (1).

Particular solution x_p. The method of undetermined coefficients applies, because the equation has constant coefficients and the forcing term $f(t) = (F_0/m)\cos(\omega t)$ is a linear combination of Euler solution atoms. Derivatives of $f(t)$ are linear combinations of the two atoms $\cos(\omega t), \sin(\omega t)$ and therefore the initial trial solution in the method of undetermined coefficients is $x(t) = d_1\cos(\omega t) + d_2\sin(\omega t)$. Neither of the two atoms appearing in the trial solution are solutions of the unforced equation $x'' + \omega_0^2 x = 0$, because that would require the false equation $\omega_0 = \omega$). Therefore, the initial trial solution is the final trial solution, no changes made, no Rule II applied.

The trial solution $x(t) = d_1\cos(\omega t) + d_2\sin(\omega t)$ is substituted into $x'' + \omega_0^2 x = \frac{F_0}{m}\cos\omega t$ in order to determine d_1, d_2. The calculation uses the equation $x'' + \omega^2 x = 0$, satisfied by $\cos\omega t, \sin\omega t$ and the trial solution $x(t)$. Then

$$\begin{aligned}
x'' + \omega_0^2 x &= \tfrac{F_0}{m}\cos(\omega t), \\
-\omega^2 x + \omega_0^2 x &= \tfrac{F_0}{m}\cos(\omega t), \\
\left(\omega_0^2 - \omega^2\right) x &= \tfrac{F_0}{m}\cos(\omega t), \\
Cd_1\cos(\omega t) + Cd_2\sin(\omega t) &= \tfrac{F_0}{m}\cos(\omega t),
\end{aligned}$$

where $C = \left(\omega_0^2 - \omega^2\right)$. Matching coefficients of the Euler atoms $\cos(\omega t), \sin(\omega t)$ then implies

$$\begin{aligned}
Cd_1 &= \tfrac{F_0}{m}, \\
Cd_2 &= 0.
\end{aligned}$$

Division by C gives $d_1 = \frac{F_0}{mC}$ and $d_2 = 0$, which implies $x(t) = \frac{F_0}{mC}\cos(\omega t)$. This is the answer for x_p reported in equation (1).

Exercises 6.7

Forced Undamped Vibration

Solve the given equation.

1. $x'' + 100x = 20\cos(5t)$

2. $x'' + 16x = 100\cos(10t)$

3. $x'' + \omega_0^2 x = 100\cos(\omega t)$, when the internal frequency ω_0 is twice the external frequency ω.

4. $x'' + \omega_0^2 x = 5\cos(\omega t)$, when the internal frequency ω_0 is half the external frequency ω.

Black Box in the Trunk

5. Construct an example $x'' + \omega_0^2 x = F_1\cos(\omega t)$ with a solution $x(t)$ having beats every two seconds.

6. A solution $x(t)$ of $x'' + 25x = 100\cos(\omega t)$ has beats every two seconds. Find ω.

Rotating Drum

Solve the given equation.

7. $x'' + 100x = 500\,\omega^2\cos(\omega t)$, $\omega \ne 10$.

8. $x'' + \omega_0^2 x = 5\,\omega^2\cos(\omega t)$, $\omega \ne \omega_0$.

Harmonic Oscillations

Express the general solution as a sum of two harmonic oscillations of different frequencies, each oscillation written in phase-amplitude form.

9. $x'' + 9x = \sin 4t$

10. $x'' + 100x = \sin 5t$

11. $x'' + 4x = \cos 4t$

12. $x'' + 4x = \sin t$

Beats: Convert and Graph

Write each linear combination as $x(t) = C\sin at\sin bt$. Then graph the slowly-varying envelope curves and the curve $x(t)$.

13. $x(t) = \cos 4t - \cos t$

14. $x(t) = \cos 10t - \cos t$

15. $x(t) = \cos 16t - \cos 12t$

16. $x(t) = \cos 25t - \cos 23t$

Beats: Solve, find Envelopes

Solve each differential equation with $x(0) = x'(0) = 0$ and determine the slowly-varying envelope curves.

17. $x'' + x = 99 \cos 10t$.

18. $x'' + 4x = 252 \cos 10t$.

19. $x'' + x = 143 \cos 12t$.

20. $x'' + 256x = 252 \cos 2t$.

Waves and Superposition

Graph the individual waves x_1, x_2 and then the superposition $x = x_1 + x_2$. Report the apparent period of the superimposed waves.

21. $x_1(t) = \sin 22t$, $x_2(t) = 2 \sin 20t$

22. $x_1(t) = \cos 16t$, $x_2(t) = 4 \cos 20t$

23. $x_1(t) = \cos 16t$, $x_2(t) = 4 \sin 16t$

24. $x_1(t) = \cos 25t$, $x_2(t) = 4 \cos 27t$

Periodicity

25. Let $x_1(t) = \cos 25t$, $x_2(t) = 4 \cos 27t$. Their sum has period $T = m\frac{2\pi}{25} = n\frac{2\pi}{27}$ for some integers m, n. Find all m, n and the least period T.

26. Let $x_1(t) = \cos \omega_1 t$, $x_2(t) = \cos \omega_2 t$. Find a condition on ω_1, ω_2 which implies that the sum $x_1 + x_2$ is periodic.

27. Let $x(t) = \cos(t) - \cos(\sqrt{2}t)$. Explain without proof, from a graphic, why $x(t)$ is not periodic.

28. Let $x(t) = \cos(5t) + \cos(5\sqrt{2}t)$. Is $x(t)$ is periodic? Explain without proof.

Rotating Drum

Let $x(t)$ and $x_p(t)$ be defined as in Example 4, page 509. Replace Hooke's constant $k = 10$ by $k = 1$, all other constants unchanged.

29. Re-compute the amplitude $A(t)$ of solution $x_p(t)$. Find the decimal value for the maximum of $|A(t)|$.

30. Find $x(t)$ when $x(0) = x'(0) = 0$. It is known that $x(t)$ fails to be periodic. Let $t_1 = 0, \ldots, t_{29}$ be the consecutive extrema on $0 \le t \le 1.4$. Verify graphically or by computation that $|x(t_{i+1}) - x(t_i)| \approx 0.133$ for $i = 1, \ldots, 28$.

Musical Instruments

Melodious tones are superpositions of harmonics $\sin(n\omega t)$, with $n =$ an integer, $\omega =$ fundamental frequency.

In 1885 Alexander J. Ellis introduced a measurement unit **Cent** by the equation **one cent** $= 2^{\frac{1}{12}} \approx 1.0005777895$. On most pianos, the frequency ratio between two adjacent keys equals 100 **cents**, called an **equally tempered semitone**. Two piano keys of frequencies 480 Hz and 960 Hz span 1200 **cents** and have tones $\sin(\omega t)$ and $\sin(2\omega t)$ with $\omega = 480$. A span of 1200 **cents** between two piano key frequencies is called an **Octave**.

31. **(Equal Temperament)** Find the 12 frequencies of equal temperament for octave 480 Hz to 960 Hz. The first two frequencies are 480, 508.5422851.

32. **(Flute or Noise)** Equation $x(t) = \sin 220\pi t + 2 \sin 330\pi t$ could represent a tone from a flute or just a dissonant, unpleasing sound. Discuss the impossibility of answering the question with a simple yes or no.

33. **(Guitar)** Air inside a guitar vibrates a little like air in a bottle when you blow across the top. Consider a flask of volume $V = 1$ liter, neck length $L = 5$ cm and neck cross-section $S = 3$ cm^2. The vibration has model $x'' + f^2 x = 0$ with $f = c\sqrt{\frac{S}{VL}}$, where $c = 343$ m/s is the speed of sound in air. Compute $\frac{f}{2\pi}$ and $\lambda = \frac{2\pi c}{f}$, the frequency and wavelength. The answers are about 130 Hz and $\lambda = 2.6$ meters, a low sound.

34. **(Helmholtz Resonance)** Repeat the previous exercise calculations, using a flask with neck diameter 2.0 cm and neck length 3 cm. The tone should be lower, about 100 Hz, and the wavelength λ should be longer.

Seismoscope

35. Verify that x_p given in (14) and x_p^* given by (15), page 519, have the same initial conditions when $u(0) = u'(0) = 0$, that is, the ground does not move at $t = 0$. Conclude that $x_p = x_p^*$ in this situation.

36. A **release test** begins by starting a vibration with $u = 0$. Two successive maxima $(t_1, x_1), (t_2, x_2)$ are recorded. Explain how to find β, Ω_0 in the equation $x'' + 2\beta\Omega_0 x' + \Omega_0^2 x = 0$, using Exercises 69 and 70, *infra*.

Free Damped Motion

Classify the homogeneous equation $mx'' + cx' + kx = 0$ as **over-damped**, **critically damped** or **under-damped**. Then solve the equation for the general solution $x(t)$.

37. $m = 1, c = 2, k = 1$

38. $m = 1, c = 4, k = 4$

39. $m = 1, c = 2, k = 3$

40. $m = 1, c = 5, k = 6$

41. $m = 1, c = 2, k = 5$

42. $m = 1, c = 12, k = 37$

43. $m = 6, c = 17, k = 7$

44. $m = 10, c = 31, k = 15$

45. $m = 25, c = 30, k = 9$

46. $m = 9, c = 30, k = 25$

47. $m = 9, c = 24, k = 41$

48. $m = 4, c = 12, k = 34$

Cafe and Pet Door

Classify as a cafe door model and/or a pet door model. Solve the equation for the general solution and identify as oscillatory or non-oscillatory.

49. $x'' + x' = 0$

50. $x'' + 2x' + x = 0$

51. $x'' + 2x' + 5x = 0$

52. $x'' + x' + 3x = 0$

53. $9x'' + 24x' + 41x = 0$

54. $6x'' + 17x' = 0$

55. $9x'' + 24x' = 0$

56. $6x'' + 17x' + 7x = 0$

Classification

Classify $mx'' + cx' + kx = 0$ as **over-damped**, **critically damped** or **under-damped** without solving the differential equation.

57. $m = 5, c = 12, k = 34$

58. $m = 7, c = 12, k = 19$

59. $m = 5, c = 10, k = 3$

60. $m = 7, c = 12, k = 3$

61. $m = 9, c = 30, k = 25$

62. $m = 25, c = 80, k = 64$

Critically Damped

The equation $mx'' + cx' + kx = 0$ is critically damped when $c^2 - 4mk = 0$. Establish the following results for $c > 0$.

63. The mass undergoes no oscillations, because

$$x(t) = (c_1 + c_2 t)e^{-\frac{ct}{2m}}.$$

64. The mass passes through $x = 0$ at most once.

Over-Damped

Equation $mx'' + cx' + kx = 0$ is defined to be over-damped when $c^2 - 4mk > 0$. Establish the following results for $c > 0$.

65. The mass undergoes no oscillations, because if r_1, r_2 are the roots of $mr^2 + cr + c = 0$, then

$$x(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}.$$

66. The mass passes through equilibrium position $x = 0$ at most once.

Under-Damped

Equation $mx'' + cx' + kx = 0$ is defined to be under-damped when $c^2 - 4mk < 0$. Establish the following results.

67. The mass undergoes infinitely many oscillations. If $c = 0$, then the oscillations are harmonic.

68. The solution $x(t)$ can be factored as an exponential function $e^{-\frac{ct}{2m}}$ times a harmonic oscillation. In symbols:

$$x(t) = e^{-\frac{ct}{2m}} \left(A\cos(\omega t - \alpha)\right).$$

Experimental Methods

Assume model $mx''+cx'+kx = 0$ is oscillatory. The results apply to find nonnegative constants m, c, k from one experimentally known solution $x(t)$. Provide details.

69. Let $x(t)$ have consecutive maxima at $t = t_1$ and $t = t_2 > t_1$. Then $t_2 - t_1 = T = \frac{2\pi}{\omega}$ = pseudo period of $x(t)$.

70. Let (t_1, x_1) and (t_2, x_2) be two consecutive maximum points of the graph of a solution $x(t) = Ce^{-ct/(2m)}\cos(\omega t - \alpha)$ of $mx'' + cx' + kx = 0$. Let $a \pm \omega i$ be the two complex roots of $mr^2 + cr + k = 0$ where $a = -c/(2m)$ and $\omega = \frac{1}{2m}\sqrt{4mk - c^2}$. Then

$$\ln\frac{x_1}{x_2} = \frac{c\pi}{m\omega},$$

71. **(Bike Trailer)** Assume *fps* units. A trailer equipped with one spring and one shock has mass $m = 100$ in the model $mx'' + cx' + kx = 0$. Find c and k from this experimental data: two consecutive maxima of $x(t)$ are $(0.35, 10/12)$ and $(1.15, 8/12)$.
Hint: Use exercises 69 and 70.

72. **(Auto)** Assume *fps* units. An auto weighing 2.4 tons is equipped with four identical springs and shocks. Each spring-shock module has damped oscillations satisfying $mx'' + cx' + kx = 0$. Find m. Then find c and k from this experimental data: two consecutive maxima of $x(t)$ are $(0.3, 3/12)$ and $(0.7, 2/12)$.
Hint: Use exercises 69 and 70.

Structure of Solutions

Establish these results for the damped spring-mass system $mx'' + cx' + kx = 0$. Assume $m > 0$, $c > 0$, $k > 0$.

73. **(Monotonic Factor)** Let the equation be critically damped or over-damped. Prove that

$$x(t) = e^{-pt} f(t)$$

where $p \ge 0$ and $f(t)$ is monotonic (f' one-signed).

74. **(Harmonic Factor)** Let the equation be under-damped. Prove that

$$x(t) = e^{-at} f(t)$$

where $a > 0$ and $f(t) = c_1\cos\omega t + c_2\sin\omega t = A\cos(\omega t - \alpha)$ is a harmonic oscillation.

75. **(Limit Zero and Transients)** A term appearing in a solution is called **transient** if it has limit zero at $t = \infty$. Prove that positive damping $c > 0$ implies that the homogeneous solution satisfies $\lim_{t\to\infty} x(t) = 0$.

76. **(Steady-State)** An **observable** or **steady-state** is expression obtained from a solution by excluding all terms with limit zero at $t = \infty$. The **Transient** is the expression excluded to obtain the steady state. Assume $mx'' + cx' + kx = 25\cos 2t$ has a solution

$$x(t) = 2te^{-t} - \cos 2t + \sin 2t.$$

Find the transient and steady-state terms.

Damping Effects

Construct a figure on $0 \le t \le 2$ with two curves, to illustrate the effect of removing the dashpot. Curve 1 is the solution of $mx'' + cx' + kx = 0$, $x(0) = x_0$, $x'(0) = v_0$. Curve 2 is the solution of $my'' + ky = 0$, $y(0) = x_0$, $y'(0) = v_0$.

77. $m = 2, c = 12, k = 50,$
$x_0 = 0, v_0 = -20$

78. $m = 1, c = 6, k = 25,$
$x_0 = 0, v_0 = 20$

79. $m = 1, c = 8, k = 25,$
$x_0 = 0, v_0 = 60$

80. $m = 1, c = 4, k = 20,$
$x_0 = 0, v_0 = 4$

Envelope and Pseudo-period

Plot on one graphic the envelope curves and the solution $x(t)$, over two pseudo-periods. Use initial conditions $x(0) = 0$, $x'(0) = 4$.

81. $x'' + 2x' + 5x = 0$

82. $x'' + 2x' + 26x = 0$

83. $2x'' + 12x' + 50x = 0$

84. $4x'' + 8x' + 20x = 0$

6.8 Resonance

A highlight in the study of vibrating mechanical systems is the theory of pure and practical resonance.

Pure Resonance and Beats

The notion of **pure resonance** in the differential equation

$$x''(t) + \omega_0^2\, x(t) = F_0 \cos(\omega t) \tag{1}$$

is the existence of a solution that is unbounded as $t \to \infty$. Unbounded means *not bounded.* Bounded means a constant M exists such that $|x(t)| \le M$ for all values of t. Already known, The theory of **Beats** page 507 solves (1) for $\omega \ne \omega_0$. The solution is the sum of two harmonic oscillations, hence it is bounded. Equation (1) for $\omega = \omega_0$ has by the method of undetermined coefficients the unbounded oscillatory solution $x(t) = \dfrac{F_0}{2\omega_0}\, t\, \sin(\omega_0\, t)$. Technical details are similar to Example 6.53, *infra.*

Pure resonance occurs exactly when the natural internal frequency ω_0 matches the natural external frequency ω, in which case all solutions of the differential equation are unbounded.

Figure 25 illustrates pure resonance for $x''(t)+16x(t) = 8\cos 4t$, which in equation (1) corresponds to $\omega = \omega_0 = 4$ and $F_0 = 8$.

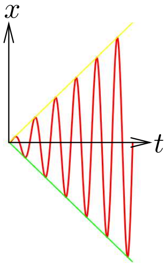


Figure 25. Pure resonance.
Equation $x''(t) + 16x(t) = 8\cos\omega t$, $\quad \omega = 4$.
Graphs:
envelope curve $x = t$ yellow
envelope curve $x = -t$ green
solution $x(t) = t\sin 4t$ red

Resonance and Undetermined Coefficients

An explanation of resonance can be based upon the theory of undetermined coefficients. An initial trial solution for

$$x''(t) + 16x(t) = 8\cos\omega t$$

is $x = d_1 \cos\omega t + d_2 \sin\omega t$. The homogeneous solution is $x_h = c_1\cos 4t + c_2 \sin 4t$. Euler atoms in $x_h(t)$ match Euler atoms in the trial solution $x = d_1\cos\omega t +$

$d_2 \sin \omega t$ exactly when $\omega = 4$. RULE II in undetermined coefficients applies exactly for $\omega = 4$. The two cases $\omega \neq 4$ and $\omega = 4$ give final trial solution

$$(2) \qquad x(t) = \begin{cases} d_1 \cos \omega t + d_2 \sin \omega t & \omega \neq 4, \\ t(d_1 \cos \omega t + d_2 \sin \omega t) & \omega = 4. \end{cases}$$

Even before the undetermined coefficients d_1, d_2 are evaluated, it is decided that unbounded solutions occur exactly when frequency matching $\omega = 4$ occurs, because of the amplitude factor t. If $\omega \neq 4$, then $x_p(t)$ is a pure harmonic oscillation, hence bounded. If $\omega = 4$, then amplitude factor t times a pure harmonic oscillation makes x_p unbounded.

Practical Resonance

The notion of pure resonance is easy to understand both mathematically and physically, because frequency matching characterizes the event. This ideal situation never happens in the physical world, because *damping is always present.* In the presence of damping $c > 0$, it will be established below that *only bounded solutions exist* for the forced spring-mass system

$$(3) \qquad mx''(t) + cx'(t) + kx(t) = F_0 \cos \omega t.$$

Our intuition about resonance seems to vaporize in the presence of damping effects. But not completely. Most would agree that the undamped intuition is correct when the damping effects are nearly zero.

Practical resonance is said to occur when the external frequency ω has been tuned to produce the largest possible solution (a more precise definition appears below). It will be shown that the steady-state solution $x_{\text{SS}}(t)$ has maximum amplitude, over all possible input frequencies ω, at the precise tuned frequency $\omega = \Omega$ given by the equation

$$(4) \qquad \Omega = \sqrt{\frac{k}{m} - \frac{c^2}{2m^2}}.$$

The equation only makes sense when $\frac{k}{m} - \frac{c^2}{2m^2} > 0$. Pure resonance $\omega = \sqrt{k/m}$ is the limiting case obtained by setting the damping constant c to zero in condition (4). This strange but predictable interaction exists between the damping constant c and the magnitude of a solution, relative to the external frequency ω, even though all solutions remain bounded.

The decomposition of $x(t)$ into homogeneous solution $x_h(t)$ and particular solution $x_p(t)$ gives some intuition into the complex relationship between the input frequency ω and the size of the solution $x(t)$.

Homogeneous Solution $x_h(t)$

Solution $x_h(t)$ for homogeneous equation $mx''(t)+cx'(t)+kx(t)=0$ for positive constants m, c, k will be shown to have limit zero at $t=\infty$, which means the graph of $x_h(t)$ follows the t-axis to $t=\infty$. An inequality of the form $|x_h(t)| \le e^{-qt}$ holds as $t \to \infty$, for some $q>0$: see the proof of Theorem 6.21. Figure 26 shows that the graph of $x_h(t)$ can cross the t-axis infinitely often, even though it is trapped between envelope curves $x=\pm e^{-qt}$ near $t=\infty$.[7]

Theorem 6.21 (Transient Solution)
Assume positive values for m, c, k. The solution $x_h(t)$ of the homogeneous equation $mx''(t)+cx'(t)+kx(t)=0$ has limit zero at $t=\infty$:

$$\lim_{t\to\infty} x_h(t) = 0 \qquad \text{for positive } m,\ c,\ k$$

Proof on page 542.

Definition 6.3 (Transient Solution)
A solution $x(t)$ of a differential equation is called a **transient solution** provided it satisfies the relation $\lim_{t\to\infty} x(t)=0$.

A transient solution $x(t)$ for large t has its graph atop the axis $x=0$, as in Figure 26.

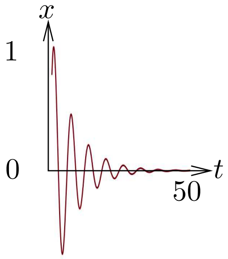


Figure 26. Transient Oscillatory Solution.
Shown is solution $x = e^{-t/8}(\cos t + \sin t)$ of differential equation $64x''+16x'+65x=0$.

Particular Solution $x_p(t)$

Let's find $x_p(t)$ for $mx''(t)+cx'(t)+kx(t)=F_0\cos\omega t$ by the method of undetermined coefficients. It will be found that $x_p(t)$ equals $x_{SS}(t)$ defined in Definition 6.4 and explicitly given in equation (5) *infra*.

Definition 6.4 (Steady-State Solution)
Assume for non-homogeneous equation $mx''(t)+cx'(t)+kx(t)=F_0\cos\omega t$ that m, c, k are all positive values. The **steady–state** solution $x_{SS}(t)$ is a particular solution $x_p(t)$ in superposition $x(t)=x_p(t)+x_h(t)$, found from any general solution $x(t)$ by removing all terms containing negative exponentials. The terms removed add to some homogeneous solution $x_h(t)$.

[7] A funnel in first order theory for $y'=f(y)$ may also have limit $y=0$ at infinity, but the funnel graph cannot cross $y=0$.

Steady-state solution $x_{SS}(t)$ is **observable**, because it is visible as the graph of $x(t)$ for t large enough for the negative exponential terms become zero to pixel resolution. Uniqueness of $x_{SS}(t)$ implies Definition 6.4 is sensible, details in the proof of Theorem 6.22.

Theorem 6.22 (Steady-State Solution)
Assume positive values for m, c, k. The unique **steady-state solution** $x_{SS}(t)$ of the non-homogeneous equation $mx''(t) + cx'(t) + kx(t) = F_0 \cos \omega t$ with period $2\pi/\omega$ is given by

$$(5) \qquad \begin{aligned} x_{SS}(t) &= \frac{F_0}{(k - m\omega^2)^2 + (c\omega)^2}\left((k - m\omega^2)\cos\omega t + (c\omega)\sin\omega t\right) \\ &= \frac{F_0}{\sqrt{(k - m\omega^2)^2 + (c\omega)^2}} \cos(\omega t - \alpha), \end{aligned}$$

where α is defined by the phase–amplitude relations (see page 492)

$$(6) \qquad \begin{aligned} &C\cos\alpha = k - m\omega^2, \quad C\sin\alpha = c\omega, \\ &C = F_0/\sqrt{(k - m\omega^2)^2 + (c\omega)^2}. \end{aligned}$$

Proof on page 543.

It is possible to be mislead by the method of undetermined coefficients, in which it turns out that $x_p(t)$ and $x_{SS}(t)$ are the same. Alternatively, a particular solution $x_p(t)$ can be calculated by variation of parameters, a method which produces in $x_p(t)$ extra terms containing negative exponentials. These extra terms come from the homogeneous solution – their appearance cannot always be avoided. This justifies the careful definition of steady–state solution, in which the transient terms are removed from a general solution $x(t)$ to produce $x_{SS}(t)$.

Definition 6.5 (Practical Resonance)
Assume positive values for m, c, k in non-homogeneous equation $mx''(t) + cx'(t) + kx(t) = F_0 \cos \omega t$. **Practical resonance** occurs if there is a value of external frequency $\omega > 0$ in which produces the largest possible steady-state amplitude $C(\omega)$ in the steady-state periodic solution x_{SS} defined by equation (5) in Theorem 6.22.

Theorem 6.23 (Practical Resonance Identity)
Assume positive values for m, c, k in non-homogeneous equation $mx''(t) + cx'(t) + kx(t) = F_0 \cos \omega t$. Practical resonance for $mx''(t) + cx'(t) + kx(t) = F_0 \cos \omega t$ occurs precisely when the external frequency ω is tuned to

$$\Omega = \sqrt{\frac{k}{m} - \frac{c^2}{2m^2}}$$

and the square root argument $\frac{k}{m} - \frac{c^2}{2m^2}$ is positive.

Proof on page 543.

Theorem 6.24 (Pure Resonance Identity)
Assume m and k are positive in non-homogeneous equation $mx''(t) + kx(t) = F_0 \cos \omega t$. **Pure resonance** results from tuned external frequency value

$$\omega = \sqrt{\frac{k}{m}} = \left(\sqrt{\frac{k}{m} - \frac{c^2}{2m^2}} \right) \Bigg|_{c=0}$$

This value is the limiting case $c = 0$ in Theorem 6.23. If $\omega = \frac{k}{m}$ is inserted into $mx''(t) + kx(t) = F_0 \cos \omega t$, then $x_p(t) = \dfrac{F_0}{2m\omega} t \sin(\omega t)$ is an unbounded solution, causing all solutions $x(t)$ to be unbounded. Proof on page 543.

An Illustration. Figure 27 illustrates practical resonance for $x'' + cx' + 26x = 10 \cos \omega t$. The amplitude $C(\omega)$ of the steady–state periodic solution is graphed against the external natural frequency ω for damping constants $c = 1, 2, 3$. The practical resonance condition is $\Omega = \sqrt{26 - c^2/2}$. As c increases from 1 to 3, the maximum point $(\Omega, C(\Omega))$ satisfies a monotonicity condition: both Ω and $C(\Omega)$ decrease as c increases. The maxima for the three curves in the figure occur at $\omega = \sqrt{25.5}$, $\sqrt{24}$, $\sqrt{21.5}$. Pure resonance occurs when $c = 0$ and $\omega = \sqrt{26}$.

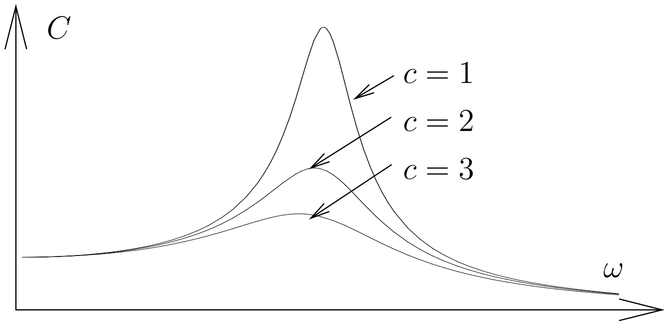


Figure 27. Practical resonance for $x'' + cx' + 26x = 10 \cos \omega t$.
The amplitude $C(\omega) = 10/\sqrt{(26 - \omega^2)^2 + (c\omega)^2}$ is plotted versus external frequency ω for $c = 1, 2, 3$.

Uniqueness of the Steady–State Periodic Solution

Any two solutions of the nonhomogeneous differential equation (3) which are periodic of period $2\pi/\omega$ must be identical by Theorem 6.22. A more general statement is true:

Theorem 6.25 (Uniqueness of a T-Periodic Solution)
Assume m, c, k positive. Consider the equation $mx''(t) + cx'(t) + kx(t) = f(t)$ with f continuous and T-periodic: $f(t + T) = f(t)$. Then a T–periodic solution is unique. Proof on page 544.

An Illustration. In Figure 28, the unique steady–state periodic solution is graphed for the differential equation $x'' + 2x' + 2x = \sin t + 2 \cos t$. The transient solution of the homogeneous equation and the steady–state solution appear in Figure 29. In Figure 30, several solutions are shown for the differential equation $x'' + 2x' + 2x = \sin t + 2\cos t$, all of which reproduce at $t = \infty$ the steady–state solution $x = \sin t$.

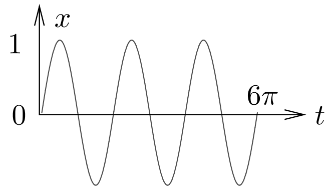


Figure 28. Steady-state solution.
Differential equation $x'' + 2x' + 2x = \sin t + 2\cos t$.
Periodic steady-state solution $x_{\text{SS}} = \sin t$.

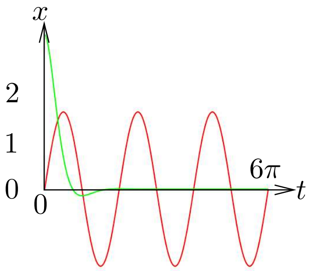


Figure 29. Transient and Steady-state.
General solution $x(t)$ is the graphical sum of x_h (green) and x_{SS} (red):
Transient Green $x_h = e^{-t}(2 * \cos t + 2\sin t)$
Steady-state Red $x_{\text{SS}} = \sin t$

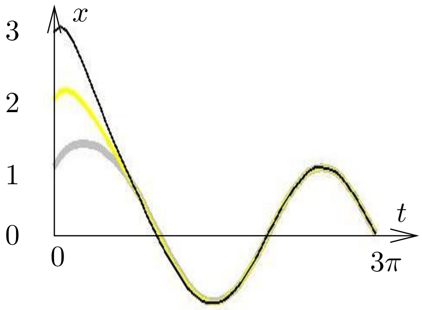


Figure 30. Steady-state.
Initial value problem solutions of $x'' + 2x' + 2x = \sin t + 2\cos t$ with $x'(0) = 1$ and $x(0) = 1, 2, 3$.
All graphically coincide with the steady-state solution $x = \sin t$ for $t \geq \pi$.

Pseudo–Periodic Solution

Resonance gives rise to solutions of the form $x(t) = A(t)\sin(\omega t - \alpha)$ where $A(t)$ is a time–varying amplitude. Figure 31 shows such a solution, which is called a **pseudo–periodic solution** because it has a natural period $2\pi/\omega$ arising from the trigonometric factor $\sin(\omega t - \alpha)$. The only requirement on $A(t)$ is that it be non–vanishing, so that it acts like an amplitude. The **pseudo–period** of a pseudo–periodic solution can be determined graphically, by computing the length of time it takes for $x(t)$ to vanish three times.

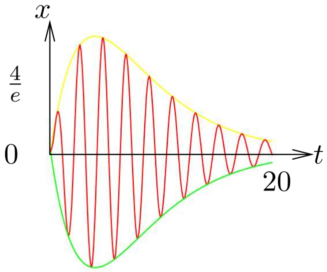


Figure 31. Pseudo-periodic solution.
Equation $16x'' + 8x' + 145x = 96e^{-t/4}\cos 3t$.
Legend for the graphic:
Envelope $x = te^{-t/4}$ Yellow
Envelope $x = -te^{-t/4}$ Green
Solution $x = te^{-t/4}\sin(3t)$ Red

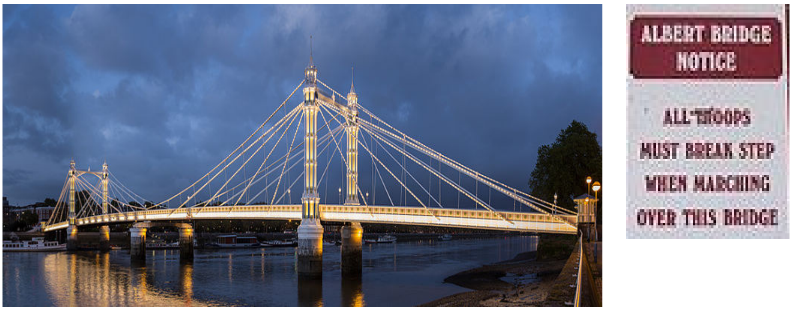
The pseudo-period $2\pi/3$ of $x = te^{-t/4}\sin(3t)$ is found by solving for t in $x(t) = 0$, equivalently $te^{-t/4}\sin(3t) = 0$. Then $3t = 0, \pi, 2\pi$ are the first three crossings of $x(t)$ with the t-axis. The pseudo-period is $2\pi/3$. The terminology does not mean that $x(t)$ is periodic, but pseudo-periodic, which is a periodic function multiplied by a nonzero amplitude function.

Resonance History

Soldiers Breaking Cadence, 1831

Figure 32. The Rebuilt Broughton Suspension Bridge.
On 12 April 1831, the original bridge collapsed, blamed on mechanical resonance from troops marching in cadence. The bridge spans the River Irwell between Broughton and Pendleton near Manchester, England. Photo from 1883.

The collapse of the Broughton suspension bridge in 1831 reportedly caused the now–standard military rule of breaking cadence when soldiers cross a bridge. Bridges like the Broughton bridge have many natural low frequencies of vibration, so it is possible for a column of soldiers to vibrate the bridge at one of the bridge's natural frequencies. The bridge locks onto the frequency while the soldiers continue to add to the excursions with every step, causing larger and larger bridge oscillations.



ALBERT BRIDGE
NOTICE

ALL TROOPS
MUST BREAK STEP
WHEN MARCHING
OVER THIS BRIDGE

Figure 33. The London Albert Bridge.
A sign added in 1973 warns marching ranks of soldiers to break cadence.

The Tacoma Narrows Bridge, 1940

The literature is rich with accounts of the November 7, 1940 Tacoma bridge disaster, the date when the bridge fell into the Tacoma Narrows.

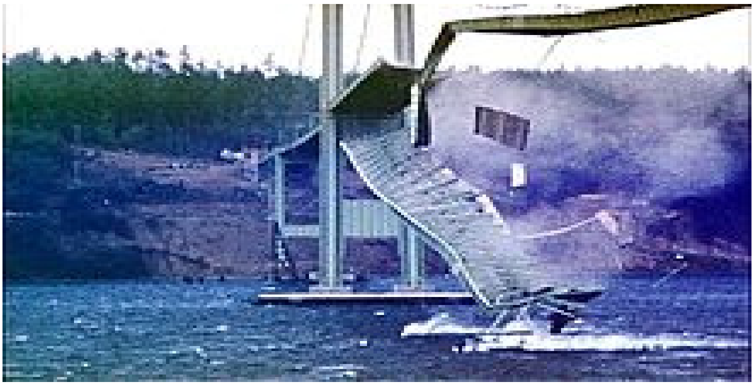


Figure 34. The Tacoma Narrows Bridge, 1940.
Historically, the disaster has been presented as an instance of **resonance**, a technical term which requires a periodic input of energy. No observer witnessed a **periodic** input of energy, and this is the source of the controversy over the cause of the bridge failure.

The bridge disaster has been blamed on **Aeroelastic Flutter**, a term used for aircraft:

> If energy input by aerodynamic excitation is larger than what is dissipated by system damping, then the amplitude of vibration will increase, resulting in self-exciting oscillation.

The Tacoma bridge was injected with energy from a 40 mph wind. The energy did not dissipate through the damping properties of the bridge structure. The energy was dissipated by the formation of longitudinal and transverse vibrations of the roadway, which eventually lead to failure.

There have been other explanations, none of which are more popular than aeroelastic flutter.

1940 Theodore von Karman proposed that **vortex shedding** had created a (periodic) force in its wake that excited the bridge into resonant oscillations. This resonance theory requires a periodic input caused by a 40 mph wind acting on the bridge structure. Wind tunnel experiments seemed to verify the explanation. The final *Federal Works Administration* report rejected the explanation.

2000 The resonance model was re-visited, because the hanging bridge suspension cables produce a force only in one direction. Using a modification of the classical linear resonance model, simulations reproduced oscillation magnitudes seen in the 1940 film of the bridge failure.

The Wine Glass Experiment, 1985

The equation $mx'' + cx' + kx = F_0 \cos(\omega t)$ with c replaced by zero is advertised as the basis for a physics experiment to break a wine glass with resonant sound waves.

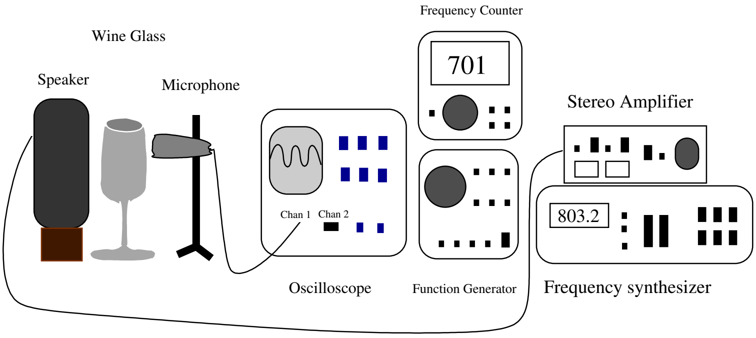


Figure 35. The Wine Glass Experiment Lab Table.
Equipment: A wine glass, a stereo amplifier, a speaker for sound waves, a frequency generator and a microphone connected to an oscilloscope.

The *wine glass experiment* is a portion of a film produced in 1985 by the Annenberg/CPB Project in Episode 17, **Resonance**, which is one of 52 episodes in **The Mechanical Universe** series. A **synopsis** appears below for a portion of episode 17, with parenthetical remarks inserted for the model equation $mx'' + kx = F_0 \cos \omega t$.

> A physicist in front of an audience of physics students equips a lab table with a frequency generator, an amplifier and an audio speaker. The *valuable* wine glass is replaced by a glass beaker. The frequency generator is tuned to the natural frequency of the glass beaker ($\omega \approx \omega_0$), then the volume knob on the amplifier is suddenly turned up (F_0 adjusted larger), whereupon the sound waves emitted from the speaker break the glass beaker.

The glass itself will vibrate at a certain frequency, as can be determined experimentally by *pinging* the glass rim. This vibration operates within elastic limits of the glass and the glass will not break under these circumstances. A physical explanation for the breakage is that an incoming sound wave from the speaker is timed to add to the glass rim excursion. After enough amplitude additions, the glass rim moves beyond the elastic limit and the glass breaks. The explanation implies that the external frequency from the speaker has to match the natural frequency of the glass. But there is more to it: the glass has some natural damping that nullifies feeble attempts to increase the glass rim amplitude. The physicist uses to great advantage this natural damping to *tune* the external frequency to the glass. The reason for turning up the volume on the amplifier is to nullify the damping effects of the glass. The amplitude additions then build rapidly and the glass breaks.

The London Millennium Foot-Bridge, 2000

Figure 36. The London Millennium Foot-Bridge.
Opened June 10, 2000 and closed two days later, London visitors nicknamed it the **Wobbly Bridge**. The reconstruction finished in 2002 added 5M pounds to the initial cost of 18M.

The opening of the bridge brought crowds of 90,000 people per day. The natural swaying motion of people walking across the span caused small sideways bridge oscillations, which in turn caused people on the bridge to sway in step, adding to the amplitude of the bridge oscillations.

Engineers fixed the vibration problem by retrofitting 37 energy dissipating viscous fluid dashpots to control horizontal movement and 52 tuned inertial mass dampers to control vertical movement.

Examples and Methods

Example 6.53 (Beats and Pure Resonance)

Solve by undetermined coefficients for a particular solution of the equation $x''(t) + 16x(t) = 8\cos\omega t$ for all values of $\omega > 0$, verifying that

$$x_p(t) = \begin{cases} \dfrac{8}{16-\omega^2}\cos(\omega t) & \omega \neq 4, \\ t\sin(4t) & \omega = 4. \end{cases}$$

Solution:

Trial solution details. Rule I of undetermined coefficients requires derivatives of $f(t) = 8\cos(\omega t)$, which are linear combinations of Euler atoms $\cos(\omega t)$, $\sin(\omega t)$. Then the Rule I trial solution is $x = d_1\cos(\omega t) + d_2\sin(\omega t)$.

The homogeneous solution solves $x'' + 16x = 0$, then $x_h = c_1\cos(4t) + c_2\sin(4t)$. Euler atoms $\cos(\omega t), \sin(\omega t)$ will be homogeneous solutions if and only if $\omega = 4$. Rule II applies only in the case $\omega = 4$, in which case the trial solution is $x = d_1 t\cos(4t) + d_2 t\sin(4t)$ (ωt equals $4t$).

Details for Beats, $\omega \neq 4$: Write $u = \cos(\omega t), v = \sin(\omega t)$ and $x(t) = d_1u + d_2v$. Then $x(t) = d_1u + d_2v$. Because $u'' + \omega^2u = 0$ and $v'' + \omega^2v = 0$, then $x'' + \omega^2x = 0$.

$$\begin{array}{ll}
x'' + 16x = 8u & \text{Original equation, } u = \cos(\omega t). \\
-\omega^2x + 16x = 8u & \text{Substitute from } x'' + \omega^2x = 0. \\
(16-\omega^2)(d_1u + d_2v) = 8u & \text{Collect on } x\text{. Substitute } x = d_1u + d_2v. \\
\left|\begin{array}{l} (16-\omega^2)d_1 = 8, \\ (16-\omega^2)d_2 = 0. \end{array}\right| & \text{Independence. Match coefficients of } u, v. \\
d_1 = \dfrac{8}{16-\omega^2},\ d_2 = 0 & \text{Solve for } d_1, d_2.
\end{array}$$

Details for Pure Resonance, $\omega = 4$: Define $u = \cos(4t), v = \sin(4t)$. The modified trial solution $x(t)$ then satisfies

$$\begin{array}{rcl}
x(t) & = & d_1tu + d_2tv, \\
x'(t) & = & d_1u + d_2v - 4d_1tv + 4d_2tu, \\
x''(t) & = & -8d_1v + 8d_2u - 16x(t).
\end{array} \tag{7}$$

Then

$$\begin{array}{ll}
x'' + 16x = 8u & \text{Original equation, } u = \cos(\omega t). \\
-8d_1v + 8d_2u = 8u & \text{Use equation (7), then cancel } 16x(t). \\
\left|\begin{array}{rcl} -8d_1 & = & 0, \\ 8d_2 & = & 8. \end{array}\right| & \text{Independence of } u, v \text{ implies matching coefficients.} \\
\begin{aligned} x(t) &= d_1tu + d_2tv \\ &= tv, \\ &= t\sin(4t). \end{aligned} & \text{Insert answers } d_1 = 0, d_2 = 1. \text{ Answer found.}
\end{array}$$

Example 6.54 (Damped Forced Spring-Mass System Trial Solution)

To equation $mx'' + cx' + kx = F_0 \cos(\omega t)$ with all coefficients positive apply undetermined coefficients to obtain trial solution

$$x(t) = A\cos\omega t + B\sin\omega t.$$

Solution: The derivatives of $f(t) = F_0\cos(\omega t)$ are linear combinations of Euler solution atoms $\cos(\omega t), \sin(\omega t)$. Rule I of the method of undetermined coefficients gives trial solution $x(t) = d_1\cos(\omega t) + d_2\sin(\omega t)$.

For characteristic equation $mr^2 + cr + k = 0$ with positive m, c, k, there are 3 cases to consider, based on the sign of the discriminant. In all 3 cases, equation $mr^2 + cr + k = 0$ has roots with nonzero real part. For instance, the real part is $-\frac{c}{2m}$ for a negative discriminant. Then the trial solution is not a solution of the homogeneous differential equation $mx'' + cx' + kx = 0$. Rule I in the method of undetermined coefficients does not fail and Rule II is not applied.

The reported trial solution is the final trial solution. To agree with notation, replace symbols d_1, d_2 by symbols A, B and report trial solution $x(t) = A\cos(\omega t) + B\sin(\omega t)$.

Example 6.55 (Undetermined Coefficients Calculation)

Substitute the trial solution $x(t) = A\cos(\omega t) + B\sin(\omega t)$ into the equation $mx'' + cx' + kx = F_0\cos(\omega t)$ to obtain the system of equations

$$\text{(8)} \qquad \begin{array}{rcrcl} (k - m\omega^2)A & + & (c\omega)B & = & F_0, \\ (-c\omega)A & + & (k - m\omega^2)B & = & 0. \end{array}$$

Solution: Define $u = \cos(\omega t), v = \sin(\omega t)$, to simplify the displays. Equations $u'' + \omega^2 u = 0$ and $v'' + \omega^2 v = 0$ are valid. By superposition, $x'' + \omega^2 x = 0$ holds for the trial solution $x(t) = A\cos(\omega t) + B\sin(\omega t)$.

$mx'' + cx' + kx = F_0 u$	Original differential equation.
$-m\omega^2 x + cx' + kx = F_0 u$	Use $x'' + \omega^2 x = 0$.
$(k - m\omega^2)x + cx' = F_0 u$	Collect on x and x'.
$\begin{array}{l}(k - m\omega^2)(Au + Bv) + \\ \quad c(-A\omega v + B\omega u) = F_0 u\end{array}$	Expand with $x = Au + Bv$ and $x' = -A\omega v + B\omega u$.
$\begin{array}{l}\left((k - m\omega^2)A + c\omega B\right) u + \\ \left(-c\omega A + (k - m\omega^2)B\right) v = F_0 u\end{array}$	Collect on u, v.
$\begin{array}{rcl}\left((k - m\omega^2)A + c\omega B\right) & = & F_0, \\ \left(-c\omega A + (k - m\omega^2)B\right) & = & 0.\end{array}$	Independence of u, v implies their coefficients match.
$\begin{array}{rrl}(k - m\omega^2)A & +c\omega B & = F_0, \\ -c\omega A & +(k - m\omega^2)B & = 0.\end{array}$	Linear equations in unknowns A, B. System (8) found.

Example 6.56 (Cramer's Rule Solution for A, B)

Verify using Cramer's determinant rule the formulas

$$A = \frac{(k - m\omega^2)F_0}{(k - m\omega^2)^2 + (c\omega)^2}, \quad B = \frac{c\omega F_0}{(k - m\omega^2)^2 + (c\omega)^2}$$

for the answers A, B to the system of equations (8).

Solution: Cramer's 2×2 rule for system $a_{11}x_1 + a_{12}x_2 = b_1$, $a_{21}x_1 + a_{22}x_2 = b_2$ is the set of equations

$$x_1 = \frac{\Delta_1}{\Delta},\ x_2 = \frac{\Delta_2}{\Delta},\ \Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ \Delta_1 = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix},\ \Delta_2 = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}.$$

Apply these formulas to system (8). Then

$$\Delta = \begin{vmatrix} k - m\omega^2 & c\omega \\ -c\omega & k - m\omega^2 \end{vmatrix},\ \Delta_1 = \begin{vmatrix} F_0 & c\omega \\ 0 & k - m\omega^2 \end{vmatrix},\ \Delta_2 = \begin{vmatrix} k - m\omega^2 & F_0 \\ -c\omega & 0 \end{vmatrix}.$$

Sarrus' 2×2 rule is applied to evaluate the determinants. Then

$$\Delta = (k - m\omega^2)^2 + (c\omega)^2,\ \Delta_1 = (k - m\omega^2)F_0,\ \Delta_2 = c\omega F_0.$$

Cramer's rule formulas $A = \frac{\Delta_1}{\Delta}$, $B = \frac{\Delta_2}{\Delta}$ give the reported answers.

Example 6.57 (Transient and Steady-State Solutions)

Compute the transient and steady-state solutions x_{tr} and x_{ss} for the equation $2x'' + 3x' + 2x = 174\cos(4t)$, verifying the formulas

$$\begin{aligned} x_{\text{tr}} &= e^{-3t/4}\left(c_1\cos(kt) + c_2\sin(kt)\right),\ k = \tfrac{\sqrt{7}}{4}, \\ x_{\text{ss}} &= -5\cos(4t) + 2\sin(4t). \end{aligned}$$

Solution:
Homogeneous Solution: The characteristic equation $2r^2 + 3r + 2 = 0$ has complex conjugate roots $-\frac{3}{4} \pm \frac{\sqrt{7}}{4}i$. Then the Euler solution atoms are $e^{-3t/4}\cos(kt)$, $e^{-3t/4}\sin(kt)$ where $k = \frac{\sqrt{7}}{4}$. The homogeneous solution is then

$$x_h = e^{-3t/4}\left(c_1\cos(kt) + c_2\sin(kt)\right).$$

Particular Solution: The method of undetermined coefficients applies with Rule I trial solution $x = A\cos(4t) + B\sin(4t)$. Let's justify this statement. The right side $f(t) = 174\cos(4t)$ has derivatives a linear combination of the Euler solution atoms $\cos(4t), \sin(4t)$. Rule I does not fail, because these Euler atoms are not solutions of the homogeneous equation. Rule II does not apply, and the final trial solution is $x = A\cos(4t) + B\sin(4t)$.

Let $u = \cos(4t), v = \sin(4t)$. Then $u'' + 16u = 0, v'' + 16v = 0$. Superposition implies $x'' + 16x = 0$. The following steps find the undetermined coefficients $A = -5, B = 2$.

$2x'' + 3x' + 2x = 174u$	Original differential equation, $u = \cos 4t$.
$-32x + 3x' + 2x = 174u$	Substitute $x'' + 16x = 0$, where $x = Au + Bv$.
$-30(Au + Bv) +$ $3(-4Av + 4Bu) = 174u$	Substitute $x = Au + Bv$, $x' = -4Av + 4Bu$.
$(-30A + 12B)u +$ $(-12A - 30B)v = 174u$	Collect on u, v.
$\left\lvert\begin{array}{rl} -30A + 12B = & 174, \\ -12A - 30B = & 0. \end{array}\right\rvert$	Independence of u, v implies matching coefficients (independent Euler atoms).

$$\left|\begin{array}{l} A=-5, \\ B=2. \end{array}\right| \qquad \text{Solve for } A, B \text{ by elimination or Cramer's rule.}$$

The particular solution is $x_p = -5\cos 4t + 2\sin 4t$.

General Solution: Superposition gives general solution

$$x = x_h + x_p = e^{-3t/4}\left(c_1\cos(kt) + c_2\sin(kt)\right) - 5\cos 4t + 2\sin 4t.$$

Transient Solution: This is the part of the general solution with negative exponential terms (terms that limit to zero at infinity). Then

$$x_{\text{tr}} = e^{-3t/4}\left(c_1\cos(kt) + c_2\sin(kt)\right).$$

Steady-State Solution: This is the part of the solution left over after the transients are removed. Then

$$x_{\text{SS}} = -5\cos 4t + 2\sin 4t.$$

Example 6.58 (Pseudo-periodic solution)

Derive the pseudo-periodic solution $x = te^{-t/4}\sin(3t)$ and its envelope curves $x = \pm te^{-t/4}$ for the equation $16x'' + 8x' + 145x = 96e^{-t/4}\cos 3t$.

Solution:

Envelope Curves. For damped oscillations, a solution of the form $x(t) = e^{at}(c_1\cos(bt) + c_2\sin(bt))$ has to be re-written in phase-amplitude form, using the formulas from page 492. Then $x(t) = Ce^{at}\cos(bt - \alpha)$ and by definition the envelope curves are $x = \pm Ce^{at}$, because the cosine factor has extreme values ± 1.

In the present example, the pseudo-periodic solution is $x(t) = te^{-t/4}\sin(3t)$. The same logic applies. The sine factor has extreme values ± 1, then the envelope curves are $x = \pm te^{-t/4}$.

Pseudo-periodic Solution. Undetermined coefficients will be applied to find a particular solution x_p of $16x'' + 8x' + 145x = 96e^{-t/4}\cos 3t$. It turns out that the desired pseudo-periodic solution is the undetermined coefficients answer $x = te^{-t/4}\sin 3t$. This is because the method subtracts all homogeneous terms from the particular solution. Superposition $x = x_h + x_p$ was invisibly used here. If x_p was found from another method, then homogeneous terms should be removed from the answer, before reporting the pseudo-periodic solution.

Homogeneous Solution. It is found from $16x'' + 8x' + 145x = 0$. The Euler solution atoms are $e^{-t/4}\cos(3t), e^{-t/4}\sin(3t)$, found from the characteristic equation $16r^2 + 8r + 145 = 0$, which has complex conjugate roots $r = -\frac{1}{4} \pm 3i$. Then

$$x_h(t) = c_1e^{-t/4}\cos(3t) + c_2e^{-t/4}\sin(3t).$$

Particular solution x_p. It is found by undetermined coefficients. The answer to be justified below is

$$x_p(t) = te^{-t/4}\sin(3t).$$

Differentiate the right side $f(t) = 96e^{-t/4}\cos 3t$ of the non-homogeneous equation to identify the Euler atoms $e^{-t/4}\cos 3t, e^{-t/4}\sin 3t$. Rule I of undetermined coefficients fails, because these atoms are solutions of the homogeneous equation. Then Rule II is applied to find the final trial solution

$$\begin{aligned} x &= t\left(d_1 e^{-t/4}\cos 3t + d_2 e^{-t/4}\sin 3t\right) \\ &= t(d_1 u + d_2 v) \end{aligned}$$

where $u = e^{-t/4}\cos 3t$ and $v = e^{-t/4}\sin 3t$. Then u, v are solutions of $16x'' + 8x' + 145x = 0$. Define $w = d_1 u + d_2 v$. Superposition implies w is also a solution of $16x'' + 8x' + 145x = 0$.

Compute the derivatives of the trial solution:

$$\begin{aligned} x &= t\,(d_1 u + d_2 v) = tw, \\ x' &= w + tw' \\ x'' &= 2w' + tw''. \end{aligned} \tag{9}$$

$16x'' + 8x' + 145x = 96u$	Original equation, $u = e^{-t/4}\cos 3t$.
$16(2w' + tw'') +$ $8(w + tw') + 145tw = 96u$	Use equations (9).
$32w' + 8w +$ $t(16w'' + 8w' + 145w) = 96u$	Collect terms on factor t.
$32w' + 8w = 96u$	Use homogeneous equation $16w'' + 8w' + 145w = 0$.
$-96d_1 v + 96 d_2 u = 96u$	Expand $w = d_1 u + d_2 v$, $w' = -\frac{1}{4}w - 3d_1 v + 3d_2 u$. Cancel $8w$.
$d_1 = 0, d_2 = 1$	Independence of u, v implies matching coefficients.

The trial solution $x = tw$ becomes $x = te^{-t/4}\sin(3t)$.

Other Methods to Find x_p. The possible methods are variation of parameters, Laplace theory and a computer algebra system. Below is sample maple code to check the answer given above.

```
de:=16*diff(x(t),t,t)+8*diff(x(t),t)+145*x(t)=
    96*exp(-t/4)*cos(3*t);
dsolve(de,x(t));
```

The answer involves homogeneous terms with arbitrary constants _C1, _C2. These terms must be removed to check the answer, $x_p = te^{-t/4}\sin(3t)$.

The example is complete.

Proofs and Technical Details

Proof of Theorem 6.21, Transient Solution:

For positive damping $c > 0$, equation (3) has homogeneous solution $x_h(t) = c_1 x_1(t) + c_2 x_2(t)$ where Euler atoms x_1 and x_2 are according to Theorem 6.1 page 431 given in terms of the roots of the characteristic equation $mr^2 + cr + k = 0$ as follows:

Let $D = c^2 - 4mk$.	The discriminant of $mr^2 + cr + k = 0$.
Case 1, $D > 0$	$x_1 = e^{r_1 t}$, $x_2 = e^{r_2 t}$ with r_1 and r_2 negative.

Case 2, $D = 0$	$x_1 = e^{r_1 t}$, $x_2 = te^{r_1 t}$ with r_1 negative.
Case 3, $D < 0$	$x_1 = e^{at}\cos bt$, $x_2 = e^{at}\sin bt$ with $b > 0$ and a negative.

Let's verify that $x_h(t) = e^{-qt}$(bounded function) for some $q > 0$, regardless of the positive values of m, c, k. For instance, Case 2 implies $x_h = e^{r_1 t/2}(c_1 e^{r_1 t/2} + c_2 t e^{r_1 t/2})$ and $(c_1 e^{r_1 t/2} + c_2 t e^{r_1 t/2})$ is bounded by some number M. Let $-q = r_1/2 < 0$. Then $|x_h(t)| \le Me^{-qt}$, which proves $x_h(t)$ has limit zero at $t = \infty$. A similar analysis applied to cases 1,2,3 reveals that $|x_h(t)| \le Me^{-qt}$ holds if q is smaller than $|\mathcal{R}\mathrm{e}(\lambda)|$ for all roots λ of the characteristic equation. ■

Proof of Theorem 6.22, Steady-State Solution

Uniqueness. Assume (5) has been proved. Suppose $x(t)$ is a periodic solution of period $2\pi/\omega$. Superposition implies $x(t) = x_{SS}(t) + x_h(t)$ for some homogeneous solution $x_h(t)$. Then $x(t) - x_{SS}(t)$ has period $2\pi/\omega$ and equals some $x_h(t)$, which has limit zero at $t = \infty$ by Theorem 6.21. Because a nonzero periodic function cannot have limit zero at $t = \infty$, then $x_h(t) = 0$, proving uniqueness $x(t) = x_{SS}(t)$.

Details for (6). The method of undetermined coefficients applies to $mx''(t) + cx'(t) + kx(t) = F_0 \cos\omega t$ with trial solution $x(t) = A\cos\omega t + B\sin\omega t$. The TEST succeeds, because by Theorem 6.21 the Euler atoms in $x_h(t)$ cannot match $\cos\omega t$ or $\sin\omega t$. More details in Example 6.54 page 539. Substitution of $x(t)$ into $mx''(t) + cx'(t) + kx(t) = F_0\cos\omega t$ produces a linear combination of Euler atoms on the left. Match the coefficients of the atoms left and right to verify the equations

$$\begin{array}{rcrcl} (k - m\omega^2)A & + & (c\omega)B & = & F_0, \\ (-c\omega)A & + & (k - m\omega^2)B & = & 0. \end{array} \tag{10}$$

Details in Example 6.55 page 539. Solve (10) for A, B with Cramer's rule or elimination. Then:

$$A = \frac{(k - m\omega^2)F_0}{(k - m\omega^2)^2 + (c\omega)^2}, \quad B = \frac{c\omega F_0}{(k - m\omega^2)^2 + (c\omega)^2}. \tag{11}$$

Details in Example 6.56 page 539. Substitute the answers in (11) into trial solution $x(t) = A\cos\omega t + B\sin\omega t$. Convert this solution to phase-amplitude form using formulas on page 492. Then (6) holds. ■

Proof of Theorem 6.23, Practical Resonance Identity:

Mathematically, a maximum happens exactly when the amplitude function $C = C(\omega)$ defined in (6) has a maximum. If a maximum exists on $0 < \omega < \infty$, then $C'(\omega) = 0$ at the maximum. The derivative is computed by the power rule:

$$\begin{aligned} C'(\omega) &= \frac{-F_0}{2}\,\frac{2(k - m\omega^2)(-2m\omega) + 2c^2\omega}{((k - m\omega^2)^2 + (c\omega)^2)^{3/2}} \\ &= \omega\left(2mk - c^2 - 2m^2\omega^2\right)\frac{C(\omega)^3}{F_0^2} \end{aligned} \tag{12}$$

If $2km - c^2 \le 0$, then $C'(\omega)$ does not vanish for $0 < \omega < \infty$. Then $C'(\omega)$ is one-signed and there is no maximum. If $2km - c^2 > 0$, then $2km - c^2 - 2m^2\omega^2 = 0$ has exactly one root $\omega = \sqrt{k/m - c^2/(2m^2)}$ in $0 < \omega < \infty$. Because $C(\infty) = 0$, then $C(\omega)$ is a maximum. ■

Proof of Theorem 6.24, Pure Resonance Identity:

The details follow Example 6.53 page 538. Let $\omega = \frac{k}{m}$. The homogeneous equation

$mx''(t)+kx(t)=0$ has general solution $x=c_1x_1+c_2x_2$ given by Euler atoms $x_1=\cos\omega t$, $x_2=\sin\omega t$. Undetermined coefficients applies, RULE II giving modified trial solution $X=t(d_1\cos\omega t+d_2\sin\omega t)$. Like Example 6.53, the trial solution is inserted into $mx''(t)+kx(t)=F_0\cos\omega t$, then Euler atom coefficients are matched left and right to obtain a diagonal system of linear algebraic equations for d_1,d_2. The answer: $d_1=0$, $d_2=\frac{F_0}{2m\omega}$. Insert the answers into the trial solution to find $x_p(t)=0+d_2t\sin\omega t=\frac{F_0}{2m\omega}t\sin\omega t$. ■

Proof of Theorem 6.25, Uniqueness T-periodic Solution:
The vehicle of proof is to show that the difference $x(t)$ of two T-periodic solutions is zero. Difference $x(t)$ is a solution of the homogeneous equation, it is T–periodic and it has limit zero at infinity. A periodic function with limit zero must be zero, therefore $x(t)=0$, which proves the two solutions are identical. ■

Exercises 6.8

Beats

Each equation satisfies the beats relation $\omega\neq\omega_0$. Find the general solution. See Example 6.53, page 538.

1. $x''+100x=10\sin 9t$

2. $x''+100x=5\sin 9t$

3. $x''+25x=5\sin 4t$

4. $x''+25x=5\cos 4t$

Pure Resonance

Each equation satisfies the pure resonance relation $\omega=\omega_0$. Find the general solution. See Example 6.53, page 538.

5. $x''+4x=10\sin 2t$

6. $x''+4x=5\sin 2t$

7. $x''+16x=5\sin 4t$

8. $x''+16x=10\sin 4t$

Practical Resonance

For each model, find the **tuned practical resonance** frequency Ω and the **resonant amplitude** C:

$$\Omega=\sqrt{k/m-c^2/(2m^2)},$$
$$C=F_0/\sqrt{(k-m\Omega^2)^2+(c\Omega)^2}$$

9. $x''+2x'+17x=100\cos(4t)$

10. $x''+2x'+10x=100\cos(4t)$

11. $x''+4x'+5x=10\cos(2t)$

12. $x''+2x'+6x=10\cos(2t)$

Transient Solution

Identify from superposition $x=x_h+x_p$ a shortest particular solution, given one particular solution.

13. $x''+2x'+10x=26\cos(3t)$,
$x=100e^{-t}\cos(3t)+3\cos(2t)+2\sin(2t)$

14. $x''+4x'+13x=920\cos(3t)$,
$x=5\,e^{-2t}\cos(3t)+23\cos(3t)+69\sin(3t)$

15. $x''+2x'+2x=2\cos(t)$,
$x=3\,e^{-t}\sin(t)+5\,e^{-t}\cos(t)+\cos(t)+2\sin(t)$

16. $x''+2x'+17x=65\cos(4t)$,
$x=-2\,e^{-t}\sin(4t)+7\,e^{-t}\cos(4t)+\cos(4t)+8\sin(4t)$

Steady-State Periodic Solution

Consider the model $mx''+cx'+kx=F_0\cos(\omega t)$ of external frequency ω. Compute the unique steady-state solution $A\cos(\omega t)+B\sin(\omega t)$ and its amplitude $C(\omega)=\sqrt{A^2+B^2}$. Graph the ratio $100C(\omega)/C(\Omega)$ on $0<\omega<\infty$, where Ω is the tuned practical resonance frequency.

17. $x''+2x'+17x=100\cos(4t)$

18. $x''+2x'+10x=100\cos(4t)$

19. $x''+4x'+5x=10\cos(2t)$

20. $x'' + 2x' + 6x = 10\cos(2t)$

21. $x'' + 4x' + 5x = 5\cos(2t)$

22. $x'' + 2x' + 5x = 5\cos(1.5t)$

Phase-Amplitude

Solve for a particular solution in the form $x(t) = C\cos(\omega t - \alpha)$.

23. $x'' + 6x' + 13x = 174\sin(5t)$

24. $x'' + 8x' + 25x = 100\cos(t) + 260\sin(t)$

6.9 Kepler's laws

Kepler's empirical laws of planetary motion are:

1. All planets move in elliptical orbits with the sun at one focus.
2. The radius vector from the sun to any planet sweeps out equal areas in equal times.
3. The square of the orbital period is proportional to the cube of the major semi-axis of its elliptical orbit.

Precise observations over 20 years on the planets and 777 stars visible to the naked eye were made by the Danish astronomer Tycho Brahe (1546-1601), who was a teacher of the German astronomer Johannes Kepler (1571-1630). It is Kepler who is credited with analyzing his teacher's observations, from which he deduced the three laws of planetary motion, about 1605. The results were published in 1609 and 1618.

About 100 years after Kepler, Isaac Newton formulated his renowned universal gravitation law. Newton showed in his *Principia Mathematica* (1687) that Kepler's laws implied his universal gravitation law. Newton also showed that Kepler's first two laws were a consequence of the universal gravitation law.

The purpose of this section is to establish Kepler's first two laws from Newton's universal gravitation law. Modern calculus courses provide the differential equations background outlined below.

Background

The derivation of Kepler's first two laws from Newton's law requires diverse background from calculus, analytic geometry, physics and differential equations. Outlined here is the material required to understand the derivation.

Analytic Geometry

An ellipse or circle equation in standard form is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

The numbers $a > 0$, $b > 0$ are called the major and minor semi-axis lengths, respectively. They are related by $b = a\sqrt{1 - e^2}$, where $0 \le e < 1$ is called the eccentricity. The equation is a circle if and only if $e = 0$.

Polar Coordinates

A point (r, θ) in polar coordinates is related to its rectangular coordinates (x, y) defined by the equations

$$x = r\cos\theta, \quad y = r\sin\theta, \quad x^2 + y^2 = 1, \quad \tan\theta = y/x.$$

Circles and ellipses have respectively the polar equations

$$r = 2a\cos(\theta - \theta_0), \quad r = \frac{ed}{1 + e\cos(\theta - \theta_0)}.$$

The number $a > 0$ is the radius of the circle. The number $d > 0$ is the distance to the directrix. The eccentricity satisfies $0 < e < 1$.

Calculus

The area of a sector in polar coordinates is given by

$$A = \frac{1}{2}r^2\theta.$$

A polar equation $r = f(\theta)$ encloses on the interval $\theta_1 \le \theta \le \theta_2$ the area

$$A = \int_{\theta_1}^{\theta_2} |f(\theta)|^2 \, d\theta.$$

Physics

Newton's universal gravitation law is given by the formula

$$F = G\,\frac{m_1 m_2}{r^2},$$

where $G = 6.672 \times 10^{-11}\frac{\text{N}\cdot\text{m}^2}{\text{kg}^2}$ is the universal gravitation constant and r is the distance between the two masses m_1, m_2. This equation gives only the magnitude of the force. Implied by the formula is the value of the fundamental constant $g \approx 9.80$ meters per second, the acceleration due to gravity:

$$g = G\frac{M}{R},$$

where $M \approx 5.98 \times 10^{24}$ kilograms and $R \approx 6.38 \times 10^6$ meters are respectively the mass of the earth and the radius of the earth. A similar formula applies for any planet. While g is computed for sea level, it varies significantly with altitude, e.g., 7.33 to 0.13 at altitudes from 1000 to $50,000$ kilometers.

Differential Equations

The second order differential equation

$$u'' + u = 0$$

is called the **harmonic oscillator**. It's solution is $u = c_1 \cos x + c_2 \sin x$, by the classical constant-coefficient Theorem 6.1. The forced equation $u'' + u = c$, where c is a constant, has a particular solution $u = c$, obtained by the equilibrium method. Therefore, the forced equation has the general solution

$$u = c_1 \cos x + c_2 \sin x + c.$$

Derivation of Kepler's First Two Laws

The second law will be derived first, then the details are used to derive the first law. The third law is not discussed here.

Kepler's Second Law

Assumed is the sun at the origin in the plane of motion of the planet. The position of the planet is written in vector form in polar coordinates by the formula

$$\vec{\mathbf{r}}(t) = \begin{pmatrix} r(t)\cos\theta(t) \\ r(t)\sin\theta(t) \end{pmatrix}.$$

Newton's universal gravitation law implies that the acceleration vector $\vec{\mathbf{r}}''(t)$ satisfies

$$\vec{\mathbf{r}}''(t) = -\frac{k}{|\vec{\mathbf{r}}(t)|^3}\,\vec{\mathbf{r}}(t).$$

The planet's motion can be expanded by the product rule and chain rule of calculus to give the relation

$$\vec{\mathbf{r}}'(t) = \begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix} r'(t) + \begin{pmatrix} -\sin\theta(t) \\ \cos\theta(t) \end{pmatrix} r(t)\theta'(t).$$

The column vectors in this formula are orthogonal hence independent. One more application of the product and chain rules gives

$$\begin{aligned} \vec{\mathbf{r}}''(t) &= \left(r''(t) - r(t)(\theta'(t))^2\right)\begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix} \\ &\quad + \left(\frac{1}{r(t)}\left(r^2(t)\theta'(t)\right)'\right)\begin{pmatrix} -\sin\theta(t) \\ \cos\theta(t) \end{pmatrix}. \end{aligned}$$

The independent vectors appearing in the formula happen to be the normal and tangential components of the acceleration, although we don't use this fact.

Newton's law expansion of $\vec{\mathbf{r}}''(t)$ requires that corresponding vector components must match, giving the relations

$$\begin{aligned}
&r''(t) - r(t)(\theta'(t))^2 = -\frac{k}{r^2(t)}, \\
&\frac{1}{r(t)}\left(r^2(t)\theta'(t)\right)' = 0.
\end{aligned} \tag{1}$$

The second formula in (1) implies that $dA(t) = 0$, where $dA(t)$ is the polar area increment swept out by the planet. Kepler's second law is proved.

Kepler's First Law

Write the second equation in (1) in integrated form $r^2(t)\theta'(t) = h$. Combine the first formula in (1) with the second (in integrated form) to obtain the nonlinear second order differential equation

$$r''(t) - \frac{h^2}{r^3(t)} = -\frac{k}{r^2(t)}. \tag{2}$$

Because $\theta'(t) = h/r^2(t) \neq 0$, then a variable change $t = t(\theta)$ is possible: $r(t) = r(t(\theta))$ is a function of θ. Let $u(\theta) = 1/r(t(\theta))$, then by the chain rule

$$\begin{aligned}
r'(t) &= -\frac{du/dt}{u^2(\theta)} \\
&= -r^2(t)u'(\theta)\theta'(t) \\
&= -hu'(\theta).
\end{aligned}$$

Differentiate again on θ and use (2) to obtain

$$u''(\theta) + u(\theta) = c,$$

where $c = k/h^2$. Solving gives

$$u(\theta) = c_1\cos\theta + c_2\sin\theta + c\,.$$

Use $u = 1/r$ to re-write this formula in the new form

$$r(t(\theta)) = \frac{1}{c + c_1\cos\theta + c_2\sin\theta}\,.$$

Define angle θ_0 and amplitude R by the formulas $R\cos\theta_0 = c_1$, $R\sin\theta_0 = c_2$. The sum formula for the cosine implies

$$\begin{aligned}
R\cos(\theta - \theta_0) &= R\cos\theta\cos\theta_0 + R\sin\theta\sin\theta_0 \\
&= c_1\cos\theta + c_2\sin\theta.
\end{aligned}$$

Substitution gives the ellipse equation in polar coordinates

$$\begin{aligned}
r(t(\theta)) &= \frac{1}{c + R\cos(\theta - \theta_0)} \\
&= \frac{\ell}{1 + e\,\cos(\theta - \theta_0)}\,.
\end{aligned}$$

Here, $\ell = h^2/k$ is half the latus rectum and $e = R\ell$ is the eccentricity of the ellipse. Initially, we don't know that $0 \leq e < 1$, but the requirement that a planetary orbit be bounded discards the possibility $e \geq 1$ (parabola or hyperbola). This completes the proof of Kepler's first law.

Chapter 7

Topics in Linear Differential Equations

Contents

Developed here is the theory for higher order linear constant-coefficient differential equations. Besides a basic formula for the solution of such equations, extensions are developed for the topics of variation of parameters and undetermined coefficients.

Enrichment topics include the Cauchy-Euler differential equation, the Cauchy kernel for second order linear differential equations, and a library of special methods for undetermined coefficients methods, the latter having prerequisites of only basic calculus and college algebra. Developed within the library methods is a verification of the method of undetermined coefficients, via Kümmer's change of variable.

7.1 Higher Order Homogeneous

Presented here is a solution method for higher order linear differential equations with real constant coefficients

$$y^n + a_{n-1}y^{(n-1)} + \cdots + a_0 y = 0. \tag{1}$$

This topic was covered earlier, therefore the central purpose of this section is the collection of additional exercises. The only new topics have to do with factorization of polynomials and differential operators. The first subject has to do with efficiency, a shortcut to speed up the process of solving a constant-coefficient linear homogeneous differential equation.

How to Solve Higher Order Equations

The **Characteristic Equation** of (1) is the polynomial equation

$$r^n + a_{n-1}r^{n-1} + \cdots + a_0 = 0. \tag{2}$$

The left side of (2) is called the **Characteristic Polynomial**. We assume the coefficients are real numbers.

For a real root $r = a$ of the characteristic equation, symbol k equals its **Algebraic Multiplicity**. Then k is the maximum power such that $(r-a)^k$ divides the characteristic polynomial.

The same symbol k is used for the **algebraic multiplicity** of a complex root $r = a + ib$. Complex roots always come in pairs, $a \pm ib$, because the coefficients of the characteristic polynomial are *real*. This means k is the maximum power such that $((r-a)^2 + b^2)^k$ divides the characteristic polynomial.

Constructing the General Solution

The general solution y of (1) is constructed as a linear combination of n Euler atoms. The list of n Euler atoms is found from the roots of the characteristic equation, by iterating on Step I and Step II below.

Step I: Real Roots

Each multiplicity k real root $r = a$ of the characteristic equation produces a group of k Euler atoms

$$e^{rx},\ xe^{rx},\ \ldots,\ x^{k-1}e^{rx}$$

which are solutions of (1). Append the group to the list of Euler atoms for equation (1).

Step II: Complex Root pairs

Each multiplicity k pair of complex roots $z = a + ib$ and $\overline{z} = a - ib$ of the characteristic equation produces two groups of k distinct Euler atoms

group 1: $e^{ax}\cos bx,\ xe^{ax}\cos bx,\ \ldots,\ x^{k-1}e^{ax}\cos bx,$
group 2: $e^{ax}\sin bx,\ xe^{ax}\sin bx,\ \ldots,\ x^{k-1}e^{ax}\sin bx,$

which are solutions to the differential equation. Append the two groups to the list of Euler atoms for equation (1).

Exponential Solutions and Euler's Theorem

Characteristic equation (2) is formally obtained from the differential equation by replacing $y^{(k)}$ by r^k. This device for remembering how to form the characteristic equation is attributed to **Euler**, because of the following fact.

Theorem 7.1 (Euler's Exponential Substitution)
Let w be a real or complex number. The function $y(x) = e^{wx}$ is a solution of (1) if and only if $r = w$ is a root of the characteristic equation (2).

Steps I and II above are justified from Euler's basic result:

Theorem 7.2 (Euler's Multiplicity Theorem)
Function $y(x) = x^p e^{wx}$ is a solution of (1) if and only if $(r - w)^{p+1}$ divides the characteristic polynomial.

An Illustration of the Higher Order Method

Consider the problem of solving a constant coefficient linear differential equation (1) of order 11 having factored characteristic equation

$$(r-2)^3(r+1)^2(r^2+4)^2(r^2+4r+5) = 0.$$

To be applied is the solution method for higher order equations. Then **Step I** loops on the two linear factors $r - 2$ and $r + 1$, while **Step 2** loops on the two real quadratic factors $r^2 + 4$ and $r^2 + 4r + 5$.

Hand solutions can be organized by a tabular method for generating the general solution y. The key element is that rows are distinct factors of the characteristic polynomial. This feature insures that each row contains distinct atoms not duplicated in another row.

Factor	Roots	Multiplicity	Atom Groups
$(r-2)^3$	$r = 2, 2, 2$	3	$e^{2x}, xe^{2x}, x^2e^{2x}$
$(r+1)^2$	$r = 1, 1$	2	e^x, xe^x
$(r^2+4)^2$	$r = \pm 2i, \pm 2i$	2	$\cos 2x, x\cos 2x$ $\sin 2x, x\sin 2x$
$(r+2)^2+1$	$r = -2 \pm i$	1	$e^{2x}\cos x$ $e^{2x}\sin x$

The equation has order $n = 11$. Symbols $c_1, \ldots, c_n$ will represent arbitrary constants in the general solution y. A real root of multiplicity k will consume k of these symbols, while a complex conjugate pair of roots of multiplicity k consumes $2k$ symbols. The number of terms added in Step I equals the multiplicity of the root, or twice that in Step II, the case of complex roots. The symbols are used in order, as the general solution is constructed, as follows.

Root(s)	**Count**	**Solution Terms Added**
$r = 2, 2, 2$	3	$(c_1 + c_2x + c_3x^2)e^{2x}$
$r = -1, -1$	2	$(c_4 + c_5x)e^{-x}$
$r = \pm 2i, \pm 2i$	4	$(c_6 + c_7x)\cos 2x + (c_8 + c_9x)\sin 2x$
$r = -2 \pm i$	2	$c_{10}e^{-2x}\cos x + c_{11}e^{-2x}\sin x$

Then the general solution is

$$\begin{aligned} y &= (c_1 + c_2x + c_3x^2)e^{2x} \\ &\quad +(c_4 + c_5x)e^{-x} \\ &\quad +(c_6 + c_7x)\cos 2x + (c_8 + c_9x)\sin 2x \\ &\quad +c_{10}e^{-2x}\cos x + c_{11}e^{-2x}\sin x. \end{aligned}$$

Computer Algebra System Solution

The system `maple` can symbolically solve a higher order equation. Below, `@` is the function composition operator, `@@` is the repeated composition operator and `D` is the differentiation operator. The coding writes the factors of

$$(r-2)^3(r+1)^2(r^2+4)^2(r^2+4r+5)$$

as differential operators $(D-2)^3$, $(D+1)^2$, $(D^2+4)^2$, D^2+4D+5. Then the differential equation is the composition of the component factors. See the next section for details about differential operators.

```
id:=x->x;
F1:=(D-2*id) @@ 3;
F2:=(D+id) @@ 2;
F3:=(D@D+4*id) @@ 2;
F4:=D@D+4*D+5*id;
de:=(F1@F2@F3@F4)(y)(x)=0:
dsolve({de},y(x));
```

Exercises 7.1

Higher Order Factored

Solve the higher order equation with the given characteristic equation. Display the roots according to multiplicity and list the corresponding solution atoms.

1. $(r-1)(r+2)(r-3)^2 = 0$

2. $(r-1)^2(r+2)(r+3) = 0$

3. $(r-1)^3(r+2)^2r^4 = 0$

4. $(r-1)^2(r+2)^3r^5 = 0$

5. $r^2(r-1)^2(r^2+4r+6) = 0$

6. $r^3(r-1)(r^2+4r+6)^2 = 0$

7. $(r-1)(r+2)(r^2+1)^2 = 0$

8. $(r-1)^2(r+2)(r^2+1) = 0$

9. $(r-1)^3(r+2)^2(r^2+4) = 0$

10. $(r-1)^4(r+2)(r^2+4)^2 = 0$

Higher Order Unfactored

Completely factor the given characteristic equation, then the roots according to multiplicity and the solution atoms.

11. $(r-1)(r^2-1)^2(r^2+1)^3 = 0$

12. $(r+1)^2(r^2-1)^2(r^2+1)^2 = 0$

13. $(r+2)^2(r^2-4)^2(r^2+16)^2 = 0$

14. $(r+2)^3(r^2-4)^4(r^2+5)^2 = 0$

15. $(r^3-1)^2(r-1)^2(r^2-1) = 0$

16. $(r^3-8)^2(r-2)^2(r^2-4) = 0$

17. $(r^2-4)^3(r^4-16)^2 = 0$

18. $(r^2+8)(r^4-64)^2 = 0$

19. $(r^2-r+1)(r^3+1)^2 = 0$

20. $(r^2+r+1)^2(r^3-1) = 0$

Higher Order Equations

The exercises study properties of Euler atoms and nth order linear differential equations.

21. (Euler's Theorem)
Explain why the derivatives of atom x^3e^x satisfy a higher order equation with characteristic equation $(r-1)^4 = 0$.

22. (Euler's Theorem)
Explain why the derivatives of atom $x^3\sin x$ satisfy a higher order equation with characteristic equation $(r^2+1)^4 = 0$.

23. (Kümmer's Change of Variable)
Consider a fourth order equation with characteristic equation $(r+a)^4 = 0$ and general solution y. Define $y = ue^{-ax}$. Find the differential equation for u and solve it. Then solve the original differential equation.

24. (Kümmer's Change of Variable)
A polynomial $u = c_0 + c_1x + c_2x^2$ satisfies $u''' = 0$. Define $y = ue^{ax}$. Prove that y satisfies a third order equation and determine its characteristic equation.

25. (Ziebur's Derivative Lemma)
Let y be a solution of a higher order constant-coefficient linear equation. Prove that the derivatives of y satisfy the same differential equation.

26. (Ziebur's Lemma: atoms)
Let $y = x^3e^x$ be a solution of a higher order constant-coefficient linear equation. Prove that Euler atoms e^x, xe^x, x^2e^x are solutions of the same differential equation.

27. (Ziebur's Atom Lemma)
Let y be an Euler atom solution of a higher order constant-coefficient linear equation. Prove that the Euler atoms extracted from the expressions $y, y', y'', \ldots$ are solutions of the same differential equation.

28. (Differential Operators)
Let y be a solution of a differential equation with characteristic equation $(r-1)^3(r+2)^6(r^2+4)^5 = 0$. Explain why y''' is a solution of a differential equation with characteristic equation $(r-1)^3(r+2)^6(r^2+4)^5r^3 = 0$.

29. (Higher Order Algorithm)
Let atom $x^2\cos x$ appear in the general solution of a linear higher order equation. Find the pair of complex conjugate roots that constructed this atom, and the multiplicity k. Report the $2k$ atoms which must also appear in the general solution.

30. (Higher Order Algorithm)
Let Euler atom $xe^x\cos 2x$ appear in the general solution of a linear higher order equation. Find the pair of complex conjugate roots that constructed this atom and estimate the multiplicity k. Report the $2k$ atoms which are expected to appear in the general solution.

31. (Higher Order Algorithm)
Let a higher order equation have characteristic equation $(r-9)^3(r-5)^2(r^2+4)^5 = 0$. Explain precisely using existence-uniqueness theorems why the general solution is a sum of constants times Euler atoms.

32. (Higher Order Algorithm)
Explain why any higher order linear homogeneous constant-coefficient differential equation has general solution a sum of constants times Euler atoms.

7.2 Differential Operators

A polynomial in the symbol $D = d/dx$ is called a **Differential Operator** and the formal manipulation of these expressions is called an **Operational Calculus**.

The meaning of an expression such as $D^2 + 3D + 2$ is through linearity, $[D^2 + 3D+2]y$ meaning $D^2y+3Dy+2y$, and each term has the corresponding meaning

$$Dy = y'(x), \quad D^2y = y''(x), \quad \cdots$$

Products of the expressions are defined through composition. For example, $(D+1)(D+2)y$ means $(D+1)(y'+2y)$, which in turn is defined to be $(y'+2y)'+(y'+2y)$. This example suggests that expansion of such factored products is identical to expansion of polynomial $(x + 1)(x + 2)$ into $x^2 + 3x + 2$.

Theorem 7.3 (Commuting Operators)
Let $P = p_0 + \cdots + p_nD^n$ and $Q = q_0 + \cdots + q_mD^m$ be two differential operators with constant coefficients. Define $R = r_0+\cdots+r_kD^k$ to be the polynomial product expansion of P and Q. Then for every infinitely differentiable function $y(x)$,

$$P(Qy) = Q(Py) = Ry.$$

In short, P and Q commute and their product in either order is the formal expanded polynomial product.

Proof: Define $p_i = 0$ for $i > n$ and $q_j = 0$ for $j > m$, so that P and Q can be written as infinite series. The Cauchy product theorem from series implies that $r_\ell = p_0q_\ell+\cdots+p_\ell q_0$. By definitions, and the Cauchy product theorem,

$$\begin{aligned}
P(Qy) &= \textstyle\sum_{i=0}^{\infty} p_iD^i(Qy) \\
&= \textstyle\sum_{i=0}^{\infty} p_iD^i\left(\sum_{j=0}^{\infty} q_jy^{(j)}\right) \\
&= \textstyle\sum_{i=0}^{\infty} p_i\left(\sum_{j=0}^{\infty} q_jy^{(i+j)}\right) \\
&= \textstyle\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_iq_jy^{(i+j)} \\
&= \textstyle\sum_{\ell=0}^{\infty}\sum_{j=0}^{\ell} p_{\ell-j}q_jy^{(\ell)} \\
&= \textstyle\sum_{\ell=0}^{\infty} r_\ell y^{(\ell)} \\
&= Ry
\end{aligned}$$

Because the series product in reverse order gives the identical answer, the proof is complete.

Factorization

The fundamental theorem of algebra implies that the characteristic equation of a real nth order linear constant-coefficient differential equation has exactly n roots, counted according to multiplicity. Some number of the roots are real and the remaining roots appear in complex conjugate pairs. This implies that every characteristic equation has a **factored form**

$$(r-a_1)^{k_1}\cdots(r-a_q)^{k_q}Q_1(r)^{m_1}\cdots Q_p(r)^{m_p}=0$$

where a_1, ..., a_q are the **distinct real roots** of the characteristic equation of algebraic multiplicities $k_1, \ldots, k_q$, respectively. Factors $Q_1(r), \ldots, Q_p(r)$ are the distinct real quadratic factors of the form $(r-z)(r-\overline{z})$. Symbol z exhausts the **distinct complex roots** $z=a+ib$ with $b>0$, having corresponding algebraic multiplicities m_1, ..., m_p. The quadratic $(r-z)(r-\overline{z})$ is normally written $(r-a)^2+b^2$.

General Solution

An nth order linear homogeneous differential equation with real constant coefficients can be written in D-operator notation via the distinct real linear and quadratic factors of the characteristic equation as

$$\left((D-a_1)^{k_1}\cdots(D-a_q)^{k_q}Q_1(D)^{m_1}\cdots Q_p(D)^{m_p}\right)y=0.$$

For $Q=(r-a)^2+b^2$, symbol $Q(D)=(D-a)^2+b^2$.

Picard's theorem on existence-uniqueness fixes the possible number of independent solutions at exactly n, the order of the differential equation. Each factor, real or quadratic, generates a certain number of distinct Euler solution atoms, the union of which counts to exactly n independent atoms, forming a solution basis for the differential equation.

Specifically, the general solution of

$$(D-a)^{k+1}y=0$$

is a polynomial $u=c_0+c_1x+\cdots+c_kx^k$ with $k+1$ terms times e^{ax}. This fact is proved by Kümmer's change of variable $y=e^{ax}u$, which finds an equivalent equation $D^{k+1}u=0$, solvable by quadrature. Details in the exercises.

The general solution of

$$((D-a)^2+b^2)^{k+1}y=0$$

is a real polynomial $u_1=a_0+\cdots+a_kx^k$ with $k+1$ terms times $e^{ax}\cos(bx)$ plus a real polynomial $u_2=b_0+\cdots+b_kx^k$ with $k+1$ terms times $e^{ax}\sin(bx)$.

Technical details: Kümmer's change of variable $y=e^{ax}u$ transforms to the equation $(D^2+b^2)^{k+1}u=0$. Because $D^2+b^2=(D-ib)(D+ib)$, the work done in the preceding paragraph applies, resulting in solutions that are polynomials with $k+1$ terms times e^{ibx} and e^{-ibx}. Taking real and imaginary parts of these solutions give the real solutions $u_1\cos(bx)$, $u_2\sin(bx)$. Transforming back multiplies these answers by e^{ax}.

Exercises 7.2

Operator Arithmetic

Compute the operator and solve the corresponding differential equation.

1. $D(D+1)+D$

2. $D(D+1)+D(D+2)$

3. $D(D+1)^2$

4. $D(D^2+1)^2$

5. $D^2(D^2+4)^2$

6. $(D-1)((D-1)^2+1)^2$

Operator Properties.

7. **(Operator Composition)** Multiply $P = D^2 + D$ and $Q = 2D + 3$ to get $R = 2D^3 + 5D^2 + 3D$. Then compute $P(Qy)$ and $Q(Py)$ for $y(x)$ 3-times differentiable, and show both equal Ry.

8. **(Kernels)**
The operators $(D-1)^2(D+2)$ and $(D-1)(D+2)^2$ share common factors. Find the Euler solution atoms shared by the corresponding differential equations.

9. **(Operator Multiply)**
Let differential equation $(D^2 + 2D + 1)y = 0$ be formally differentiated four times. Find its operator and solve the equation. What does this have to do with operator multiply?

10. **(Non-homogeneous Equation)** The differential equation $(D^5 + 4D^3)y = 0$ can be viewed as $(D^2 + 4)u = 0$ and $u = D^3y$. On the other hand, y is a linear combination of the atoms generated from the characteristic equation $r^3(r^2 + 4) = 0$. Use these facts to find a particular solution of the non-homogeneous equation $y''' = 3\cos 2x$.

Kümmer's Change of Variable

Kümmer's change of variable $y = ue^{ax}$ changes a y-differential equation into a u-differential equation. It can be used as a basis for solving homogeneous nth order linear constant coefficient differential equations.

11. Supply details: $y = ue^{ax}$ changes $y'' = 0$ into $u'' + 2au' + a^2u = 0$.

12. Supply details: $y = ue^{ax}$ changes $(D^2 + 4D)y = 0$ into $((D+a)^2 + 4(D+a))u = 0$.

13. Supply details: $y = ue^{ax}$ changes the differential equation $D^ny = 0$ into $(D+a)^nu = 0$.

14. Kümmer's substitution $y = ue^{ax}$ changes the differential equation $(D^n + a_{n-1}D^{n-1} + \cdots + a_0)y = 0$ into $(F^n + a_{n-1}F^{n-1} + \cdots + a_0)u = 0$, where $F = D + a$. Write the proof.

7.3 Higher Order Non-Homogeneous

Continued here is the study of higher order linear differential equations with real constant coefficients.

The **homogeneous equation** is

$$y^n + a_{n-1}y^{(n-1)} + \cdots + a_0 y = 0. \tag{1}$$

The variation of parameters formula and the method of undetermined coefficients are discussed for the associated **non-homogeneous equation**

$$y^n + a_{n-1}y^{(n-1)} + \cdots + a_0 y = f(x). \tag{2}$$

Variation of Parameters Formula

The Picard-Lindelöf theorem implies that on $(-\infty,\infty)$ there a unique solution of the initial value problem

$$\begin{array}{l} y^n + a_{n-1}y^{(n-1)} + \cdots + a_0 y = 0, \\ y(0) = \cdots = y^{(n-2)}(0) = 0, \quad y^{(n-1)}(0) = 1. \end{array} \tag{3}$$

The unique solution is called **Cauchy's kernel**, written $\mathcal{K}(x)$.

To illustrate, Cauchy's kernel $\mathcal{K}(x)$ for $y''' - y'' = 0$ is obtained from its general solution $y = c_1 + c_2 x + c_3 e^x$ by computing the values of the constants c_1, c_2, c_3 from initial conditions $y(0) = 0$, $y'(0) = 0$, $y''(0) = 1$, giving $\mathcal{K}(x) = e^x - x - 1$.

Theorem 7.4 (Higher Order Variation of Parameters)
Let $y^n + a_{n-1}y^{(n-1)} + \cdots + a_0 y = f(x)$ have constant coefficients a_0, ..., a_{n-1} and continuous forcing term $f(x)$. Denote by $\mathcal{K}(x)$ Cauchy's kernel for the homogeneous differential equation. Then a particular solution is given by the **Variation of Parameters Formula**

$$y_p(x) = \int_0^x \mathcal{K}(x-u) f(u) du. \tag{4}$$

This solution has zero initial conditions $y(0) = \cdots = y^{(n-1)}(0) = 0$.

Proof: Define $y(x) = \int_0^x \mathcal{K}(x-u)f(u)du$. Compute by the 2-variable chain rule applied to $F(x,y) = \int_0^x \mathcal{K}(y-u)f(u)du$ the formulas

$$\begin{array}{rcl} y(x) & = & F(x,x) \\ & = & \int_0^x \mathcal{K}(x-u)f(u)du, \\ y'(x) & = & F_x(x,x,) + F_y(x,x) \\ & = & \mathcal{K}(x-x)f(x) + \int_0^x \mathcal{K}'(x-u)f(u)du \\ & = & 0 + \int_0^x \mathcal{K}'(x-u)f(u)du. \end{array}$$

The process can be continued to obtain for $0 \le p < n-1$ the general relation

$$y^{(p)}(x) = \int_0^x \mathcal{K}^{(p)} f(u) du.$$

The relation justifies the initial conditions $y(0) = \cdots = y^{(n-1)}(0) = 0$, because each integral is zero at $x = 0$. Take $p = n - 1$ and differentiate once again to give

$$y^{(n)}(x) = \mathcal{K}^{(n-1)}(x-x)f(x) + \int_0^x \mathcal{K}^{(n)} f(u)du.$$

Because $\mathcal{K}^{(n-1)}(0) = 1$, this relation implies

$$y^{(n)} + \sum_{p=0}^{n-1} a_p y^{(p)} = f(x) + \int_0^x \left(\mathcal{K}^{(n)}(x-u) + \sum_{p=0}^{n-1} a_p \mathcal{K}^{(p)}(x-u)\right) f(u)du.$$

The sum under the integrand on the right is zero, because Cauchy's kernel satisfies the homogeneous differential equation. This proves $y(x)$ satisfies the nonhomogeneous differential equation. ■

Undetermined Coefficients Method

The method applies to higher order nonhomogeneous linear differential equations with real constant coefficients

$$(5) \qquad y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0 y = f(x).$$

It finds a particular solution y_p of (5) *without* the integration steps present in variation of parameters. The theory was already presented earlier, for the special case of second order differential equations. The contribution of this section is a higher order example and more exercises. The term **Euler atom** is an abbreviation for the phrase *Euler solution atom of a constant-coefficient linear homogeneous differential equation*. A **base atom** is one of e^{ax}, $e^{ax}\cos bx$, $e^{ax}\sin bx$ where symbols a and b are real constants with $b > 0$. Euler atoms are x^n times a base atom $n = 0, 1, 2, 3, \ldots$.

Requirements and limitations:

1. The coefficients on the left side of (5) are constant.
2. The function $f(x)$ is a sum of constants times atoms.

Method of Undetermined Coefficients

Step 1. Define the list of k atoms in a trial solution using Rule I and Rule II [details below]. Multiply these atoms by **undetermined coefficients** $d_1, \ldots, d_k$, then add to define **trial solution** y.

Step 2. Substitute y into the differential equation.

Step 3. Match coefficients of Euler atoms left and right to write out linear algebraic equations for unknowns $d_1, d_2, \ldots, d_k$. Solve the equations.

Step 4. The trial solution y with evaluated coefficients $d_1, d_2, \ldots, d_k$ becomes the particular solution y_p.

Undetermined Coefficients Rule I

Assume $f(x)$ in the equation $y^{(n)} + \cdots + a_0 y = f(x)$ is a sum of constants times Euler atoms. For each atom A appearing in $f(x)$, extract all distinct atoms that appear in A, A', A'', ..., then collect all these computed atoms into a list of k distinct Euler atoms.

If the list **contains a solution of the homogeneous differential equation**, then Rile I FAILS. Otherwise, multiply the k atoms by undetermined coefficients d_1, ..., d_k to form trial solution

$$y = d_1(\text{atom 1}) + d_2(\text{atom 2}) + \cdots + d_k(\text{atom k}).$$

Undetermined Coefficients Rule II

Assume Rule I constructed a list of k Euler atoms but **FAILED**. The particular solution y_p is still a sum of constants times k atoms. Rule II changes some or all of the k atoms, by repeated multiplication by x.

The k-atom list is subdivided into groups with the same base atom, called **group 1**, **group 2**, and so on. Test each group for a solution of the homogeneous differential equation. If found, then multiply each atom in the group by factor x. Repeat until **no group contains a solution of the homogeneous differential equation**. The final set of k Euler atoms is used to construct trial solution

$$y = d_1(\text{atom 1}) + d_2(\text{atom 2}) + \cdots + d_k(\text{atom k}).$$

A Common Difficulty

An able and earnest student working on undetermined coefficients writes:

> *I substituted trial solution y into the differential equation, but then I couldn't solve the equations. What's wrong?*

Trial solution substitution can result in a missing variable d_p on the left. It happens *exactly* when the trial solution contains a term d_pA, where A is an Euler solution atom of the homogeneous equation.

To illustrate, suppose $y = d_1x + d_2x^2$ is substituted into left side of the differential equation $y''' - y'' = x + x^2$ to get

$$\begin{array}{lcll} d_1[(x)''' - (x)''] & + & d_2[(x^2)''' - (x^2)''] & = & x + x^2, \\ d_1[0] & + & d_2[-2] & = & x + x^2. \end{array}$$

Then d_1 vanishes from the left side, because $(x)''' - (x)''$ evaluates to zero! Equation $(x)''' - (x)'' = 0$ means function $y(x) = x$ is a solution of the homogeneous differential equation for $y''' - y'' = f(x)$. Then d_1 is a **free variable** in the linear algebra problem. The other coefficient d_2 is determined to be zero. The nonsense equation $0 = x + x^2$ tells us **we chose the wrong trial solution**.

What caused the missing variable? Function $y = x$ was a solution of the homogeneous differential equation for $y''' - y'' = x + x^2$.

To prevent the error, test the trial solution before substitution:

> *Search the Euler atom list for trial solution y for a solution of the homogeneous equation – there shouldn't be any*!

The *test* should be used before embarking upon the time–consuming task of writing the linear algebraic equations and solving them.

Illustration: nth Order Undetermined Coefficients

Let's solve

$$y''' - y'' = xe^x + 2x + 1 + 3\sin x$$

Answer:

$$y_p(x) = -\frac{3}{2}x^2 - \frac{1}{3}x^3 - 2xe^x + \frac{1}{2}x^2e^x + \frac{3}{2}\cos x + \frac{3}{2}\sin x.$$

Solution:
Check Applicability. The right side $f(x) = xe^x + 2x + 1 + 3\sin x$ is a sum of terms constructed from Euler atoms xe^x, x, 1, $\sin x$. The left side has constant coefficients. Therefore, the method of undetermined coefficients applies to find a particular solution y_p.

Homogeneous solution. The equation $y''' - y'' = 0$ has general solution y_h equal to a linear combination of Euler atoms 1, x, e^x.

Rule I. The Euler atoms found in $f(x)$ are subjected to repeated differentiation. The six distinct atoms so found are 1, x, e^x, xe^x, $\cos x$, $\sin x$ (drop coefficients to identify new atoms). Three of these are solutions of the homogeneous equation: **Rule I FAILS**.

Rule II. Divide the list of six atoms 1, x, e^x, xe^x, $\cos x$, $\sin x$ into four groups with identical base atom:

Group	Euler Atoms	Base Atom
group 1 :	$1, x$	1
group 2 :	e^x, xe^x	e^x
group 3 :	$\cos x$	$\cos x$
group 4 :	$\sin x$	$\sin x$

Group 1 contains a solution of the homogeneous equation $y''' - y'' = 0$. Rule II says to multiply **group 1** by x. Rule II is repeated, because the new group x, x^2 still contains a solution of the homogeneous equation. The process stops with new group x^2, x^3. **Group 2** contains solution e^x of the homogeneous equation. Rule II says to multiply group 2 by x. The new group xe^x, x^2e^x contains no solution of the homogeneous differential equation $y'' - y = 0$ .The last two groups are unchanged, because neither contains a solution of the homogeneous equation. Then

Group	Atoms	Action
New group 1 :	x^2, x^3	multiplied by x twice
New group 2 :	xe^x, x^2e^x	multiplied once by x
group 3 :	$\cos x$	unchanged
group 4 :	$\sin x$	unchanged

The final groups have been found. The shortest trial solution is

$$\begin{aligned} y &= \text{linear combination of atoms in the new groups} \\ &= d_1x^2 + d_2x^3 + d_3xe^x + d_4x^2e^x + d_5\cos x + d_6\sin x. \end{aligned}$$

Equations for d_1 to d_6. Substitution of trial solution y into $y''' - y''$ requires formulas for y', y'', y''':

$$\begin{aligned} y' &= 2\,d_1x + 3\,d_2x^2 + d_3e^xx + d_3e^x + 2\,d_4xe^x + d_4x^2e^x \\ &\quad - d_5\sin(x) + d_6\cos(x), \\ y'' &= 2\,d_1 + 6\,d_2x + d_3e^xx + 2\,d_3e^x + 2\,d_4e^x + 4\,d_4xe^x + d_4x^2e^x \\ &\quad - d_5\cos(x) - d_6\sin(x), \\ y''' &= 6\,d_2 + d_3e^xx + 3\,d_3e^x + 6\,d_4e^x + 6\,d_4xe^x + d_4x^2e^x \\ &\quad + d_5\sin(x) - d_6\cos(x) \end{aligned}$$

Then

$$\begin{aligned} f(x) &= y''' - y'' && \text{Given equation.} \\ &= 6d_2 - 2d_1 - 6d_2x + (d_3 + 4d_4)e^x + 2d_4xe^x && \text{Substitute, then} \\ &\quad + (d_5 - d_6)\cos(x) + (d_5 + d_6)\sin(x) && \text{collect on atoms.} \end{aligned}$$

Because $f(x) \equiv 1 + 2x + xe^x + 3\sin x$, then two linear combinations of the same set of six Euler atoms are equal:

$$\begin{aligned} 1 + 2x + xe^x + 3\sin x &= (6d_2 - 2d_1)(1) + (-6d_2)x \\ &\quad + (d_3 + 4d_4)e^x + (2d_4)xe^x \\ &\quad + (d_5 - d_6)\cos(x) + (d_5 + d_6)\sin(x). \end{aligned}$$

Coefficients of Euler atoms on the left and right must match, by independence of atoms. Write out the equations for matching coefficients:

$$\begin{aligned} -2d_1 + 6d_2 &= 1, \\ -6d_2 &= 2, \\ d_3 + 4d_4 &= 0, \\ 2d_4 &= 1, \\ d_5 - d_6 &= 0, \\ d_5 + d_6 &= 3. \end{aligned}$$

Solve. The first four equations can be solved by back-substitution to give $d_2 = -1/3$, $d_1 = -3/2$, $d_4 = 1/2$, $d_3 = -2$. The last two equations are solved by elimination or Cramer's rule to give $d_5 = 3/2$, $d_6 = 3/2$.

Report y_p. The corrected trial solution y with evaluated coefficients d_1 to d_6 becomes the particular solution

$$y_p(x) = -\frac{3}{2}x^2 - \frac{1}{3}x^3 - 2xe^x + \frac{1}{2}x^2e^x + \frac{3}{2}\cos x + \frac{3}{2}\sin x.$$

Exercises 7.3

Variation of Parameters

Solve the higher order equation given by its characteristic equation and right side $f(x)$. Display the Cauchy kernel $\mathcal{K}(x)$ and a particular solution $y_p(x)$ with fewest terms. Use a computer algebra system to evaluate integrals, if possible.

1. $(r-1)(r+2)(r-3)^2 = 0$,
$f(x) = e^x$

2. $(r-1)^2(r+2)(r+3) = 0$,
$f(x) = e^x$

3. $(r-1)^3(r+2)^2r^4 = 0$,
$f(x) = x + e^{-2x}$

4. $(r-1)^2(r+2)^3r^5 = 0$,
$f(x) = x + e^{-2x}$

5. $r^2(r-1)^2(r^2+4r+6) = 0$,
$f(x) = x + e^x$

6. $r^3(r-1)(r^2+4r+6)^2 = 0$,
$f(x) = x^2 + e^x$

7. $(r-1)(r+2)(r^2+1)^2 = 0$,
$f(x) = \cos x + e^{-2x}$

8. $(r-1)^2(r+2)(r^2+1) = 0$,
$f(x) = \sin x + e^{-2x}$

9. $(r-1)^3(r+2)^2(r^2+4) = 0$,
$f(x) = \cos 2x + e^x$

10. $(r-1)^4(r+2)(r^2+4)^2 = 0$,
$f(x) = \sin 2x + e^x$

Undetermined Coefficient Method

A higher order equation is given by its characteristic equation and right side $f(x)$. Display (a) a trial solution, (b) a system of equations for the undetermined coefficients, and (c) a particular solution $y_p(x)$ with fewest terms. Use a computer algebra system to solve for undetermined coefficients, if possible.

11. $(r-1)(r+2)(r-3)^2 = 0$,
$f(x) = e^x$

12. $(r-1)^2(r+2)(r+3) = 0$,
$f(x) = e^x$

13. $(r-1)^3(r+2)^2r^4 = 0$,
$f(x) = x + e^{-2x}$

14. $(r-1)^2(r+2)^3r^5 = 0$,
$f(x) = x + e^{-2x}$

15. $r^2(r-1)^2(r^2+4r+6) = 0$,
$f(x) = x + e^x$

16. $r^3(r-1)(r^2+4r+6)^2 = 0$,
$f(x) = x^2 + e^x$

17. $(r-1)(r+2)(r^2+1)^2 = 0$,
$f(x) = \cos x + e^{-2x}$

18. $(r-1)^2(r+2)(r^2+1) = 0$,
$f(x) = \sin x + e^{-2x}$

19. $(r-1)^3(r+2)^2(r^2+4) = 0$,
$f(x) = \cos 2x + e^x$

20. $(r-1)^4(r+2)(r^2+4)^2 = 0$,
$f(x) = \sin 2x + e^x$

7.4 Cauchy-Euler Equation

The differential equation

$$a_n x^n y^{(n)} + a_{n-1} x^{n-1} y^{(n-1)} + \cdots + a_0 y = 0$$

is called the **Cauchy-Euler** differential equation of order n. The symbols a_i, $i = 0, \ldots, n$ are constants and $a_n \neq 0$.

The Cauchy-Euler equation is important in the theory of linear differential equations because it has direct application to **Fourier's method** in the study of partial differential equations. In particular, the second order Cauchy-Euler equation

$$ax^2 y'' + bxy' + cy = 0$$

accounts for the bulk of such applications in applied literature.

A second argument for studying the Cauchy-Euler equation is theoretical: it is a single example of a differential equation with non-constant coefficients that has a known closed-form solution. This fact is due to a change of variables $(x, y) \longrightarrow (t, z)$ given by equations

$$x = e^t, \quad z(t) = y(x),$$

which changes the Cauchy-Euler equation into a constant-coefficient differential equation. Since the constant-coefficient equations have closed-form solutions, so also do the Cauchy-Euler equations.

Theorem 7.5 (Cauchy-Euler Equation)
The change of variables $x = e^t$, $z(t) = y(e^t)$ transforms the Cauchy-Euler equation

$$ax^2 y'' + bxy' + cy = 0$$

into its equivalent constant-coefficient equation

$$a \frac{d}{dt}\left(\frac{d}{dt} - 1\right) z + b \frac{d}{dt} z + cz = 0.$$

The result is memorized by the general differentiation formula

$$x^k y^{(k)}(x) = \frac{d}{dt}\left(\frac{d}{dt} - 1\right) \cdots \left(\frac{d}{dt} - k + 1\right) z(t). \tag{1}$$

Proof: The equivalence is obtained from the formulas

$$y(x) = z(t), \quad xy'(x) = \frac{d}{dt} z(t), \quad x^2 y''(x) = \frac{d}{dt}\left(\frac{d}{dt} - 1\right) z(t)$$

by direct replacement of terms in $ax^2 y'' + bxy' + cy = 0$. It remains to establish the general identity (1), from which the replacements arise.

The method of proof is mathematical induction. The induction step uses the chain rule of calculus, which says that for $y = y(x)$ and $x = x(t)$,

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx}.$$

The identity (1) reduces to $y(x) = z(t)$ for $k = 0$. Assume it holds for a certain integer k; we prove it holds for $k + 1$, completing the induction.

Let us invoke the induction hypothesis LHS = RHS in (1) to write

$$\begin{aligned}
\frac{d}{dt}\mathsf{RHS} &= \frac{d}{dt}\mathsf{LHS} && \text{Reverse sides.}\\
&= \frac{dx}{dt}\frac{d}{dx}\mathsf{LHS} && \text{Apply the chain rule.}\\
&= e^t\frac{d}{dx}\mathsf{LHS} && \text{Use } x = e^t,\ dx/dt = e^t.\\
&= x\frac{d}{dx}\mathsf{LHS} && \text{Use } e^t = x.\\
&= x\left(x^k y^{(k)}(x)\right)' && \text{Expand with } ' = d/dx.\\
&= x\left(kx^{k-1}y^{(k)}(x) + x^k y^{(k+1)}(x)\right) && \text{Apply the product rule.}\\
&= k\,\mathsf{LHS} + x^{k+1}y^{(k+1)}(x) && \text{Use } x^k y^{(k)}(x) = \mathsf{LHS}.\\
&= k\,\mathsf{RHS} + x^{k+1}y^{(k+1)}(x) && \text{Use hypothesis LHS} = \text{RHS.}
\end{aligned}$$

Solve the resulting equation for $x^{k+1}y^{(k+1)}$. The result completes the induction. The details, which prove that (1) holds with k replaced by $k + 1$:

$$\begin{aligned}
x^{k+1}y^{(k+1)} &= \frac{d}{dt}\mathsf{RHS} - k\,\mathsf{RHS}\\
&= \left(\frac{d}{dt} - k\right)\mathsf{RHS}\\
&= \left(\frac{d}{dt} - k\right)\frac{d}{dt}\left(\frac{d}{dt} - 1\right)\cdots\left(\frac{d}{dt} - k + 1\right)z(t)\\
&= \frac{d}{dt}\left(\frac{d}{dt} - 1\right)\cdots\left(\frac{d}{dt} - k\right)z(t)
\end{aligned}$$

Example 7.1 (How to Solve a Cauchy-Euler Equation)
Show the solution details for the equation

$$2x^2y'' + 4xy' + 3y = 0,$$

verifying general solution

$$y(x) = c_1 x^{-1/2}\cos\left(\frac{\sqrt{5}}{2}\ln|x|\right) + c_2 e^{-t/2}\sin\left(\frac{\sqrt{5}}{2}\ln|x|\right).$$

Solution: The characteristic equation $2r(r-1) + 4r + 3 = 0$ can be obtained as follows:

$2x^2y'' + 4xy' + 3y = 0$	Given differential equation.
$2x^2r(r-1)x^{r-2} + 4xrx^{r-1} + 3x^r = 0$	Use **Euler's substitution** $y = x^r$.
$2r(r-1) + 4r + 3 = 0$	Cancel x^r. **Characteristic equation** found.
$2r^2 + 2r + 3 = 0$	Standard quadratic equation.
$r = -\frac{1}{2} \pm \frac{\sqrt{5}}{2}i$	Quadratic formula complex roots.

Cauchy-Euler Substitution. The second step is to use $y(x) = z(t)$ and $x = e^t$ to transform the differential equation. By Theorem 7.5,

$$2(d/dt)^2z + 2(d/dt)z + 3z = 0,$$

a constant-coefficient equation. Because the roots of the characteristic equation $2r^2 + 2r + 3 = 0$ are $r = -1/2 \pm \sqrt{5}i/2$, then the Euler solution atoms are

$$e^{-t/2}\cos\left(\frac{\sqrt{5}}{2}t\right), \quad e^{-t/2}\sin\left(\frac{\sqrt{5}}{2}t\right).$$

Back-substitute $x = e^t$ and $t = \ln|x|$ in this equation to obtain two independent solutions of $2x^2y'' + 4xy' + 3y = 0$:

$$x^{-1/2}\cos\left(\frac{\sqrt{5}}{2}\ln|x|\right), \quad e^{-t/2}\sin\left(\frac{\sqrt{5}}{2}\ln|x|\right).$$

Substitution Details. Because $x = e^t$, the factor $e^{-t/2}$ is re-written as $(e^t)^{-1/2} = x^{-1/2}$. Because $t = \ln|x|$, the trigonometric factors are back-substituted like this: $\cos\left(\frac{\sqrt{5}}{2}t\right) = \cos\left(\frac{\sqrt{5}}{2}\ln|x|\right)$.

General Solution. The final answer is the set of all linear combinations of the two preceding independent solutions.

Exercises 7.4

Cauchy-Euler Equation

Find solutions y_1, y_2 of the given homogeneous differential equation which are independent by the Wronskian test, page 464.

1. $x^2y'' + y = 0$

2. $x^2y'' + 4y = 0$

3. $x^2y'' + 2xy' + y = 0$

4. $x^2y'' + 8xy' + 4y = 0$

Variation of Parameters

Find a solution y_p using a variation of parameters formula.

5. $x^2y'' = x$

6. $x^3y'' = e^x$

7. $y'' + 9y = \sec 3x$

8. $y'' + 9y = \csc 3x$

7.5 Variation of Parameters Revisited

The independent functions y_1 and y_2 in the general solution $y_h = c_1y_1 + c_2y_2$ of a homogeneous linear differential equation

$$a(x)y'' + b(x)y' + c(x)y = 0$$

are used to define **Cauchy's kernel**[1]

$$\text{(1)} \qquad \mathcal{K}(x,t) = \frac{y_1(t)y_2(x) - y_1(x)y_2(t)}{y_1(t)y_2'(t) - y_1'(t)y_2(t)}.$$

The denominator is the *Wronskian* $W(t)$ of y_1, y_2. Define

$$\text{(2)} \qquad C_1(t) = \frac{-y_2(t)}{W(t)}, \quad C_2(t) = \frac{y_1(t)}{W(t)}.$$

Then Cauchy's kernel $\mathcal{K}$ has these properties (proved on page 571):

$$\begin{array}{ll} \mathcal{K}(x,t) = C_1(t)y_1(x) + C_2(t)y_2(x), & \mathcal{K}(x,x) = 0, \\ \mathcal{K}_x(x,t) = C_1(t)y_1'(x) + C_2(t)y_2'(x), & \mathcal{K}_x(x,x) = 1, \\ \mathcal{K}_{xx}(x,t) = C_1(t)y_1''(x) + C_2(t)y_2''(x), & a\mathcal{K}_{xx} + b\mathcal{K}_x + c\mathcal{K} = 0. \end{array}$$

Theorem 7.6 (Cauchy Kernel Shortcut)
Let a, b, c be constants and let U be the unique solution of $aU'' + bU' + cU = 0$, $U(0) = 0$, $U'(0) = 1$. Then Cauchy's kernel is $\mathcal{K}(x,t) = U(x-t)$.

Proof on page 572.

Theorem 7.7 (Variation of Parameters Formula: Cauchy's Kernel)
Let a, b, c, f be continuous near $x = x_0$ and $a(x) \neq 0$. Let $\mathcal{K}$ be Cauchy's kernel for $ay'' + by' + cy = 0$. Then the non-homogeneous initial value problem

$$ay'' + by' + cy = f, \quad y(x_0) = y'(x_0) = 0$$

has solution

$$y_p(x) = \int_{x_0}^{x} \frac{\mathcal{K}(x,t)f(t)}{a(t)}dt.$$

Proof on page 572.

Specific initial conditions $y(x_0) = y'(x_0) = 0$ imply that y_p can be determined in a laboratory with just one experimental setup. The integral form of y_p shows that it depends *linearly* on the input $f(x)$.

Example 7.2 (Cauchy Kernel)
Verify that the equation $2y'' - y' - y = 0$ has Cauchy kernel $\mathcal{K}(x,t) = \frac{2}{3}(e^{x-t} - e^{-(x-t)/2})$.

[1] Pronunciation *ko–she.*

Solution: The two independent solutions y_1, y_2 are calculated from Theorem 6.1, which uses the characteristic equation $2r^2 - r - 1 = 0$. The roots are $-1/2$ and 1. The general solution is $y = c_1 e^{-x/2} + c_2 e^x$. Therefore, $y_1 = e^{-x/2}$ and $y_2 = e^x$.

The Cauchy kernel is the quotient

$$\begin{aligned} \mathcal{K}(x,t) &= \frac{y_1(t)y_2(x) - y_1(x)y_2(t)}{y_1(t)y_2'(t) - y_1'(t)y_2(t)} && \text{Definition page 569.} \\ &= \frac{e^{-t/2}e^x - e^{-x/2}e^t}{e^{-t/2}e^t + 0.5e^{-t/2}e^t} && \text{Substitute } y_1 = e^{-x/2},\ y_2 = e^x. \\ &= \frac{2}{3}(e^{-t}e^x - e^{-x/2}e^{t/2}) && \text{Simplify.} \\ &= \frac{2}{3}(e^{x-t} - e^{(t-x)/2}) && \text{Final answer.} \end{aligned}$$

An alternative method to determine the Cauchy kernel is to apply the shortcut Theorem 7.6. We will apply it to check the answer. Solution U must be $U(x) = Ay_1(x) + By_2(x)$ for some constants A, B, determined by the conditions $U(0) = 0$, $U'(0) = 1$. The resulting equations for A, B are $A+B = 0$, $-A/2+B = 1$. Solving gives $-A = B = 2/3$ and then $U(x) = \frac{2}{3}(e^x - e^{-x/2})$. The kernel is $\mathcal{K}(x,t) = U(x-t) = \frac{2}{3}(e^{x-t} - e^{-(x-t)/2})$.

Example 7.3 (Variation of Parameters)

Solve $y'' = |x|$ by Cauchy kernel methods, verifying $y = c_1 + c_2 x + |x|^3/6$.

Solution: First, an independent method will be described, in order to provide a check on the solution. The method involves splitting the equation into two problems $y'' = x$ and $y'' = -x$. The two polynomial answers $y = x^3/6$ on $x > 0$ and $y = -x^3/6$ on $x < 0$, obtained by quadrature, are re-assembled to obtain a single formula $y_p(x) = |x|^3/6$ valid on $-\infty < x < \infty$.

The Cauchy kernel method will be applied to verify the general solution $y = c_1 + c_2 x + |x|^3/6$.

Homogeneous solution. Theorem 6.1 for constant equations, applied to $y'' = 0$, gives $y_h = c_1 + c_2 x$. Suitable independent solutions are $y_1(x) = 1$, $y_2(x) = x$.

Cauchy kernel for $y'' = 0$. It is computed by formula, $\mathcal{K}(x,t) = ((1)(x) - (t)(1))/(1)$ or $\mathcal{K}(x,t) = x - t$.

Variation of parameters. The solution is $y_p(x) = |x|^3/6$, by Theorem 7.7, details below.

$$\begin{aligned} y_p(x) &= \textstyle\int_0^x \mathcal{K}(x,t)|t|dt && \text{Theorem 7.7, page 569.} \\ &= \textstyle\int_0^x (x-t)tdt && \text{Substitute } \mathcal{K} = x - t \text{ and } |t| = t \text{ for } x > 0. \\ &= x\textstyle\int_0^x tdt - \int_0^x t^2 dt && \text{Split into two integrals.} \\ &= x^3/6 && \text{Evaluate for } x > 0. \end{aligned}$$

If $x < 0$, then the evaluation differs only by $|t| = -t$ in the integrand. This gives $y_p(x) = -x^3/6$ for $x < 0$. The two formulas can be combined into $y_p(x) = |x|^3/6$, valid for $-\infty < x < \infty$.

Example 7.4 (Two Methods)

Solve $y'' - y = e^x$ by undetermined coefficients and by variation of parameters. Explain any differences in the answers.

Solution:

Homogeneous solution. The characteristic equation $r^2 - 1 = 0$ for $y'' - y = 0$ has roots ± 1. The homogeneous solution is $y_h = c_1e^x + c_2e^{-x}$.

Undetermined Coefficients Summary. The general solution is reported to be $y = y_h + y_p = c_1e^x + c_2e^{-x} + xe^x/2$.

Kümmer's polynomial × exponential method applies to give $y = e^xY$ and $[(D+1)^2 - 1]Y = 1$. The latter simplifies to $Y'' + 2Y' = 1$, which has polynomial solution $Y = x/2$. Then $y_p = xe^x/2$.

Variation of Parameters Summary. The homogeneous solution $y_h = c_1e^x + c_2e^{-x}$ found above implies $y_1 = e^x$, $y_2 = e^{-x}$ is a suitable independent pair of solutions, because their Wronskian is $W = -2$

The Cauchy kernel is given by $\mathcal{K}(x,t) = \frac{1}{2}(e^{x-t} - e^{t-x})$, details below. The shortcut Theorem 7.6 also applies with $U(x) = \sinh(x) = (e^x - e^{-x})/2$. Variation of parameters Theorem 7.7 gives $y_p(x) = \int_0^x \mathcal{K}(x,t)e^t dt$. It evaluates to $y_p(x) = xe^x/2 - (e^x - e^{-x})/4$, details below.

Differences. The two methods give respectively $y_p = xe^x/2$, and $y_p = xe^x/2 - (e^x - e^{-x})/4$. The solutions $y_p = xe^x/2$ and $y_p = xe^x/2 - (e^x - e^{-x})/4$ differ by the homogeneous solution $(e^x - e^{-x})/4$. In both cases, the general solution is

$$y = c_1e^x + c_2e^{-x} + \frac{1}{2}xe^x,$$

because terms of the homogeneous solution can be absorbed into the arbitrary constants c_1, c_2.

Computational Details.

$$\begin{aligned}
\mathcal{K}(x,t) &= \frac{y_1(t)y_2(x) - y_1(x)y_2(t)}{y_1(t)y_2'(t) - y_1'(t)y_2(t)} && \text{Definition page 569.} \\
&= \frac{e^te^{-x} - e^xe^{-t}}{e^t(-e^{-t}) - e^te^{-t}} && \text{Substitute.} \\
&= \frac{1}{2}(e^{x-t} - e^{t-x}) && \text{Cauchy kernel found.}
\end{aligned}$$

$$\begin{aligned}
y_p(x) &= \int_0^x \mathcal{K}(x,t)e^t dt && \text{Theorem 7.7, page 569.} \\
&= \frac{1}{2}\int_0^x (e^{x-t} - e^{t-x})e^t dt && \text{Substitute } \mathcal{K} = \tfrac{1}{2}(e^{x-t} - e^{t-x}). \\
&= \frac{1}{2}e^x \int_0^x dt - \frac{1}{2}\int_0^x e^{2t-x} dt && \text{Split into two integrals.} \\
&= \frac{1}{2}xe^x - \frac{1}{4}(e^x - e^{-x}) && \text{Evaluation completed.}
\end{aligned}$$

Proofs and Technical Details

Proofs for page 569, Cauchy Kernel Properties:
The equation $\mathcal{K}(x,t) = C_1(t)y_1(x) + C_2(t)y_2(x)$ is an algebraic identity, using the definitions of C_1 and C_2. Then $\mathcal{K}(x,x)$ is a fraction with numerator $y_1(x)y_2(x) - y_1(x)y_2(x) = 0$, giving the second identity $\mathcal{K}(x,x) = 0$.

The partial derivative formula $\mathcal{K}_x(x,t) = C_1(t)y_1'(x)+C_2(t)y_2'(x)$ is obtained by ordinary differentiation on x in the previous identity. Then $\mathcal{K}_x(x,x)$ is a fraction with numerator $y_1(x)y_2'(x) - y_1'(x)y_2(x)$, which exactly cancels the denominator, giving the identity $\mathcal{K}_x(x,x) = 1$.

The second derivative formula $\mathcal{K}_{xx}(x,t) = C_1(t)y_1''(x) + C_2(t)y_2''(x)$ results by ordinary differentiation on x in the formula for $\mathcal{K}_x$. The differential equation $a\mathcal{K}_{xx}+b\mathcal{K}_x+c\mathcal{K} = 0$ is satisfied, because $\mathcal{K}$ in the variable x is a linear combination of y_1 and y_2, which are given to be solutions.

Proof of Theorem 7.6, Cauchy Kernel Shortcut:
Let $y(x) = \mathcal{K}(x,t) - U(x-t)$ for fixed t. It will be shown that y is a solution and $y(t) = y'(t) = 0$. Already known from page 569 is the relation $a\mathcal{K}_{xx}(x,t) + b\mathcal{K}_x(x,t) + c\mathcal{K}(x,t) = 0$. By assumption, $aU''(x-t) + bU'(x-t) + cU(x-t) = 0$. By the chain rule, both terms in y satisfy the differential equation, hence y is a solution. At $x = t$, $y(t) = \mathcal{K}(t,t) - U(0) = 0$ and $y'(t) = K_x(t,t) - U'(0) = 0$ (see page 569). Then y is a solution of the homogeneous equation with zero initial conditions. By uniqueness, $y(x) \equiv 0$, which proves $\mathcal{K}(x,t) = U(x-t)$.

Proof of Theorem 7.7, Variation of Parameters:
Let $F(t) = f(t)/a(t)$. It will be shown that y_p as given has two continuous derivatives given by the integral formulas

$$y_p'(x) = \int_{x_0}^{x} \mathcal{K}_x(x,t)F(t)dt, \quad y_p''(x) = \int_{x_0}^{x} \mathcal{K}_{xx}(x,t)F(t)dt + F(x).$$

Then

$$ay_p'' + by_p' + cy_p = \int_{x_0}^{x} (a\mathcal{K}_{xx} + b\mathcal{K}_x + c\mathcal{K})F(t)dt + aF.$$

The equation $a\mathcal{K}_{xx} + b\mathcal{K}_x + c\mathcal{K} = 0$, page 569, shows the integrand on the right is zero. Therefore $ay_p'' + by_p' + cy_p = f(x)$, which would complete the proof.

Needed for the calculation of the derivative formulas is the fundamental theorem of calculus relation $\left(\int_{x_0}^{x} G(t)dt\right)' = G(x)$, valid for continuous G. The product rule from calculus can be applied directly, because y_p is a sum of products:

$$\begin{aligned} y_p' &= \left(y_1(x)\int_{x_0}^{x} C_1F dt + y_2(x)\int_{x_0}^{x} C_2F dt\right)' \\ &= y_1'\int_{x_0}^{x} C_1F dt + y_2'\int_{x_0}^{x} C_2F dt + y_1(x)C_1(x)F(x) + y_2(x)C_2(x)F(x) \\ &= y_1'\int_{x_0}^{x} C_1F dt + y_2'\int_{x_0}^{x} C_2F dt + \mathcal{K}(x,x)F(x) \\ &= \int_{x_0}^{x} \mathcal{K}_x(x,t)F(t)dt \end{aligned}$$

The terms contributed by differentiation of the integrals add to zero because $\mathcal{K}(x,x) = 0$ (page 569).

$$\begin{aligned} y_p'' &= \left(y_1'(x)\int_{x_0}^{x} C_1F dt + y_2'(x)\int_{x_0}^{x} C_2F dt\right)' \\ &= y_1''\int_{x_0}^{x} C_1F dt + y_2''\int_{x_0}^{x} C_2F dt + y_1'(x)C_1(x)F(x) + y_2'(x)C_2(x)F(x) \\ &= y_1''\int_{x_0}^{x} C_1F dt + y_2''\int_{x_0}^{x} C_2F dt + \mathcal{K}_x(x,x)F(x) \\ &= \int_{x_0}^{x} \mathcal{K}_{xx}(x,t)F(t)dt + F(x) \end{aligned}$$

The terms contributed by differentiation of the integrals add to $F(x)$ because $\mathcal{K}_x(x,x) = 1$ (page 569).

Exercises 7.5

Cauchy Kernel

Find the Cauchy kernel $\mathcal{K}(x,t)$ for the given homogeneous differential equation.

1. $y'' - y = 0$

2. $y'' - 4y = 0$

3. $y'' + y = 0$

4. $y'' + 4y = 0$

5. $4y'' + y' = 0$

6. $y'' + y' = 0$

7. $y'' + y' + y = 0$

8. $y'' - y' + y = 0$

Variation of Parameters

Find the general solution $y_h + y_p$ by applying a variation of parameters formula.

9. $y'' = x^2$

10. $y'' = x^3$

11. $y'' + y = \sin x$

12. $y'' + y = \cos x$

13. $y'' + y' = \ln |x|$

14. $y'' + y' = -\ln |x|$

15. $y'' + 2y' + y = e^{-x}$

16. $y'' - 2y' + y = e^{x}$

7.6 Undetermined Coefficients Library

The study of undetermined coefficients continues for the problem

$$ay'' + by' + cy = f(x). \tag{1}$$

As in previous sections, $f(x)$ is assumed to be a sum of constants times **Euler solution atoms** and the symbols a, b, c are constants. Recorded here are special methods for efficiently solving (1). **Linear algebra is not required** in any of the special methods: only calculus and college algebra are assumed as background.

The special methods provide a justification for the trial solution method, presented earlier in this text.

The Easily-Solved Equations

The algebra problem for undetermined coefficients can involve many unknowns. It is recommended to reduce the size of the algebra problem by breaking the differential equation into several simpler differential equations. A particular solution y_p of (1) can be expressed as a sum

$$y_p = y_1 + \cdots + y_n$$

where each y_k solves a related easily-solved differential equation.

The idea can be quickly communicated for $n = 3$. The superposition principle applied to the three equations

$$\begin{aligned} ay_1'' + by_1' + cy_1 &= f_1(x), \\ ay_2'' + by_2' + cy_2 &= f_2(x), \\ ay_3'' + by_3' + cy_3 &= f_3(x) \end{aligned} \tag{2}$$

shows that $y = y_1 + y_2 + y_3$ is a solution of

$$ay'' + by' + cy = f_1 + f_2 + f_3. \tag{3}$$

If each equation in (2) is easily solved, then solving equation (3) is also easy: *add the three answers for the easily solved problems.*

To use the idea, it is necessary to start with $f(x)$ and determine a decomposition $f = f_1 + f_2 + f_3$ so that equations (2) are easily solved.

Each **Easily-Solved equation** is engineered to have right side in one of the four forms below:

$$\begin{array}{ll} p(x) & \textbf{polynomial}, \\ p(x)e^{kx} & \textbf{polynomial} \times \textbf{exponential}, \\ p(x)e^{kx}\cos mx & \textbf{polynomial} \times \textbf{exponential} \times \textbf{cosine}, \\ p(x)e^{kx}\sin mx & \textbf{polynomial} \times \textbf{exponential} \times \textbf{sine}. \end{array} \tag{4}$$

To illustrate, consider

$$ay'' + by' + cy = x + xe^x + x^2 \sin x - \pi e^{2x} \cos x + x^3. \tag{5}$$

The right side is decomposed as follows, in order to define the easily solved equations:

$ay_1'' + by_1' + cy_1 = x + x^3$	Polynomial.
$ay_2'' + by_2' + cy_2 = xe^x$	Polynomial × exponential.
$ay_3'' + by_3' + cy_3 = x^2 \sin x$	Polynomial × exponential × sine.
$ay_4'' + by_4' + cy_4 = -\pi e^{2x} \cos x$	Polynomial × exponential × cosine.

There are $n = 4$ equations. In the illustration, x^3 is included with x, but it could have caused creation of a fifth equation. To decrease effort, minimize the number n of easily solved equations. *One final checkpoint*: the right sides of the n equations must add to the right side of (5).

Library of Special Methods

It is assumed that the differential equation is already in easily-solved form: the library methods are designed to apply directly. If an equation requires a decomposition into easily-solved equations, then the desired solution is then the sum of the answers to the decomposed equations.

Equilibrium and Quadrature Methods

The special case of $ay'' + by' + cy = k$ where k is a constant occurs so often that an efficient method has been isolated to find y_p. It is called the **equilibrium method**, because in the simplest case y_p is a constant solution or an *equilibrium solution.* The method in words:

> Verify that the right side of the differential equation is constant. Cancel on the left side all derivative terms except for the lowest order and then solve for y by quadrature.

The method works to find a solution, because if a derivative $y^{(n)}$ is constant, then all higher derivatives $y^{(n+1)}$, y^{n+2}, etc., are zero. A precise description follows for second order equations.

Differential Equation	Cancelled DE	Particular Solution
$ay'' + by' + cy = k$, $c \neq 0$	$cy = k$	$y_p = \dfrac{k}{c}$
$ay'' + by' = k$, $b \neq 0$	$by' = k$	$y_p = \dfrac{k}{b}x$
$ay'' = k$, $a \neq 0$	$ay'' = k$	$y_p = \dfrac{k}{a}\dfrac{x^2}{2}$

The equilibrium method also applies to nth order linear differential equations $\sum_{i=0}^{n} a_i y^{(i)} = k$ with constant coefficients $a_0, \ldots, a_n$ and constant right side k.

A special case of the equilibrium method is the *simple quadrature method*, illustrated in Example 7.5 page 582. The method is used in elementary physics to solve falling body problems.

The Polynomial Method

The method applies to find a particular solution of $ay'' + by' + cy = p(x)$, where $p(x)$ represents a polynomial of degree $n \geq 1$. Such equations **always have a polynomial solution**; see Theorem 7.8 page 581.

Let a, b and c be given with $a \neq 0$. Differentiate the differential equation successively until the right side is constant:

$$(6) \qquad \begin{array}{ccccccc} ay'' & + & by' & + & cy & = & p(x), \\ ay''' & + & by'' & + & cy' & = & p'(x), \\ ay^{iv} & + & by''' & + & cy'' & = & p''(x), \\ & & & & & \vdots & \\ ay^{(n+2)} & + & by^{(n+1)} & + & cy^{(n)} & = & p^{(n)}(x). \end{array}$$

Apply the equilibrium method to the *last equation* in order to find a polynomial **trial solution**

$$y(x) = d_m \frac{x^m}{m!} + \cdots + d_0.$$

It will emerge that $y(x)$ always has $n + 1$ terms, but its degree can be either n, $n + 1$ or $n + 2$. The **undetermined coefficients** $d_0, \ldots, d_m$ are resolved by setting $x = 0$ in equations (6). The Taylor polynomial relations $d_0 = y(0), \ldots, d_m = y^{(m)}(0)$ give the equations

$$(7) \qquad \begin{array}{ccccccc} ad_2 & + & bd_1 & + & cd_0 & = & p(0), \\ ad_3 & + & bd_2 & + & cd_1 & = & p'(0), \\ ad_4 & + & bd_3 & + & cd_2 & = & p''(0), \\ & & & & & \vdots & \\ ad_{n+2} & + & bd_{n+1} & + & cd_n & = & p^{(n)}(0). \end{array}$$

These equations can always be solved by **back-substitution**; linear algebra is not required. Three cases arise, according to the number of zero roots of the characteristic equation $ar^2+br+c=0$. The values $m=n, n+1, n+2$ correspond to zero, one or two roots $r=0$.

No root $r=0$. Then $c\neq 0$. There were n integrations to find the trial solution, so $d_{n+2}=d_{n+1}=0$. The unknowns are d_0 to d_n. The system can be solved by back-substitution to uniquely determine $d_0, \ldots, d_n$. The resulting polynomial $y(x)$ is the desired solution $y_p(x)$.

One root $r=0$. Then $c=0$, $b\neq 0$. The unknowns are $d_0, \ldots, d_{n+1}$. There is no condition on d_0; simplify the trial solution by taking $d_0=0$. Solve (7) for unknowns d_1 to d_{n+1} as in the no root case.

Double root $r=0$. Then $c=b=0$ and $a\neq 0$. The equilibrium method gives a polynomial trial solution $y(x)$ involving $d_0, \ldots, d_{n+2}$. There are no conditions on d_0 and d_1. Simplify y by taking $d_0=d_1=0$. Solve (7) for unknowns d_2 to d_{n+2} as in the no root case.

College algebra back-substitution applied to (7) is illustrated in Example 7.7, page 583. A complete justification of the polynomial method appears in the proof of Theorem 7.8, page 588.

Recursive Polynomial Hybrid

A *recursive method* based upon quadrature appears in Example 7.8, page 584. This method, independent from the *polynomial method* above, is useful when the number of equations in (6) is two or three.

Some researchers (see [Gupta]) advertise the recursive method as easy to remember, easy to use and faster than other methods. In this textbook, the method is advertised as a **hybrid**: equations in (6) are written down, but equations (7) are not. Instead, the undetermined coefficients are found recursively, by repeated quadrature and back-substitution.

Classroom testing of the recursive polynomial method reveals it is best suited to algebraic helmsmen with flawless talents. The method should be applied when conditions suggest rapid and reliable computation details. Error propagation possibilities dictate that polynomial solutions of degree 4 or larger be suspect and subjected to an answer check.

Polynomial × Exponential Method

The method applies to special equations $ay''+by'+cy=p(x)e^{kx}$ where $p(x)$ is a polynomial. The idea, due to Kümmer, uses the transformation $y=e^{kx}Y$ to obtain the auxiliary equation

$$[a(D+k)^2 + b(D+k) + c]Y = p(x), \quad D = \frac{d}{dx}.$$

The polynomial method applies to find Y. Multiplication by e^{kx} gives y. Computational details are in Example 7.9, page 584. Justification appears in Theorem 7.9. In words, to find the differential equation for Y:

> In the differential equation, replace D by $D+k$ and cancel e^{kx} on the RHS.

Polynomial × Exponential × Cosine Method

The method applies to equations $ay'' + by' + cy = p(x)e^{kx}\cos(mx)$ where $p(x)$ is a polynomial. Kümmer's transformation $y = e^{kx}\mathcal{R}\text{e}(e^{imx}Y)$ gives the auxiliary problem

$$[a(D+z)^2 + b(D+z) + c]Y = p(x), \quad z = k + im, \quad D = \frac{d}{dx}.$$

The polynomial method applies to find Y. Symbol $\mathcal{R}\text{e}$ extracts the real part of a complex number. Details are in Example 7.10, page 585. The formula is justified in Theorem 7.10. In words, to find the equation for Y:

> In the differential equation, replace D by $D+k+im$ and cancel $e^{kx}\cos mx$ on the RHS.

Polynomial × Exponential × Sine Method

The method applies to equations $ay'' + by' + cy = p(x)e^{kx}\sin(mx)$ where $p(x)$ is a polynomial. Kümmer's transformation $y = e^{kx}\mathcal{I}\text{m}(e^{imx}Y)$ gives the auxiliary problem

$$[a(D+z)^2 + b(D+z) + c]Y = p(x), \quad z = k + im, \quad D = \frac{d}{dx}.$$

The polynomial method applies to find Y. Symbol $\mathcal{I}\text{m}$ extracts the imaginary part of a complex number. Details are in Example 7.11, page 586. The formula is justified in Theorem 7.10. In words, to find the equation for Y:

> In the differential equation, replace D by $D+k+im$ and cancel $e^{kx}\sin mx$ on the RHS.

Kümmer's Method

The methods known above as the polynomial × exponential method, the polynomial × exponential × cosine method, and the polynomial × exponential × sine method, are collectively called **Kümmer's method**, because of their origin.

Trial Solution Shortcut

The library of special methods leads to a justification for the **trial solution method**, a method which has been popularized by leading differential equation textbooks published over the past 50 years.

Trial Solutions and Kümmer's Method

Assume given $ay'' + by' + cy = f(x)$ where $f(x) =$(polynomial)$e^{kx}\cos mx$, then the method of Kümmer predicts

$$y = e^{kx}\,\mathcal{R}\mathrm{e}\left(Y(x)(\cos mx + i\sin mx)\right),$$

where $Y(x)$ is a polynomial solution of a different, **associated differential equation**. In the simplest case, $Y(x) = \sum_{j=0}^{n} A_j x^j + i\sum_{j=0}^{n} B_j x^j$, a polynomial of degree n with complex coefficients, matching the degree of the polynomial in $f(x)$. Expansion of the Kümmer formula for y plus definitions $a_j = A_j - B_j$, $b_j = B_j + A_j$ gives a **trial solution**

$$y = \left(\cos(mx)\sum_{j=0}^{n} a_j x^j + \sin(mx)\sum_{j=0}^{n} b_j x^j\right) e^{kx}. \tag{8}$$

The undetermined coefficients are $a_0, \ldots, a_n, b_0, \ldots, b_n$. Exactly the same trial solution results when $f(x) =$(polynomial)$e^{kx}\sin mx$. If $m = 0$, then the trigonometric functions do not appear and the trial solution is either a polynomial $(k = 0)$ or else a polynomial times an exponential.

The characteristic equation for the associated differential equation has root $r = 0$ exactly when $r = k + m\sqrt{-1}$ is a root of $ar^2 + br + c = 0$. Therefore, Y, and hence y, must be multiplied by x for each time $k + m\sqrt{-1}$ is a root of $ar^2 + br + c = 0$. In the trial solution method, this requirement is met by multiplication by x until the trial solution no longer contains a term of the homogeneous solution. Certainly both correction rules produce exactly the same final trial solution.

Shortcuts using (8) have been known for some time. The results can be summarized in words as follows.

> If the right side of $ay'' + by' + cy = f(x)$ is a polynomial of degree n times $e^{kx}\cos(mx)$ or $e^{kx}\sin(mx)$, then an initial trial solution y is given by relation (8), with undetermined coefficients $a_0, \ldots, a_n, b_0,$

> $\ldots$, b_n. Correct the trial solution y by multiplication by x, once for each time $r = k + m\sqrt{-1}$ is a root of the characteristic equation $ar^2 + br + c = 0$.

The Correction Rule

The **Final Trial Solution** is found by this rule:

> Given an initial trial solution y for $au'' + by' + cy = f(x)$, from Table 1 below, correct y by multiplication by x, once for each time that $r = k + m\sqrt{-1}$ is a root of the characteristic equation $ar^2 + br + c = 0$. This is equivalent to multiplication by x until the trial solution no longer contains a term of the homogeneous solution.

Once the **final trial solution** y is determined, then y is substituted into the differential equation. The undetermined coefficients are found by matching terms of the form $x^j e^{kx}\cos(mx)$ and $x^j e^{kx}\sin(mx)$, which appear on the left and right side of the equation after substitution.

A Table Lookup Method

Table 1 below summarizes the form of an initial trial solution in special cases, according to the form of $f(x)$.

Table 1. A Table Method for Trial Solutions.
The table predicts the **Initial Trial Solution** y in the method of undetermined coefficients. Then the **Correction Rule** is applied to find the **final trial solution**. Symbol n is the degree of the polynomial in column 1.

Form of $f(x)$	Values	Initial Trial Solution
constant	$k = m = 0$	$y = a_0$ =constant
polynomial	$k = m = 0$	$y = \sum_{j=0}^n a_j x^j$
combination of $\cos mx$ and $\sin mx$	$k = 0,\ m > 0$	$y = a_0 \cos mx + b_0 \sin mx$
(polynomial)e^{kx}	$m = 0$	$y = \left(\sum_{j=0}^n a_j x^j\right) e^{kx}$
(polynomial)$e^{kx}\cos mx$ or (polynomial)$e^{kx}\sin mx$	$m > 0$	$y = \left(\sum_{j=0}^n a_j x^j\right) e^{kx}\cos mx$ $+ \left(\sum_{j=0}^n b_j x^j\right) e^{kx}\sin mx$

Details for lines 2-3 of Table 1 appear in Examples 7.6 and 7.13.

Alternate Trial Solution Shortcut

The method avoids the root testing of the correction rule, at the expense of repeated substitutions. The simplicity of the method is appealing, but a few computations reveal that the correction rule is a more practical method.

> Let y be the initial trial solution of Table 1. Substitute it into the differential equation. It will either compute y_p, or else some coefficients cannot be determined. In the latter case, multiply y by x and repeat, until a solution y_p is found.

Key Theorems

Theorem 7.8 (Polynomial Solutions)
Assume a, b, c are constants, $a \neq 0$. Let $p(x)$ be a polynomial of degree d. Then $ay'' + by' + cy = p(x)$ has a polynomial solution y of degree d, $d+1$ or $d+2$. Precisely, these three cases hold:

Case 1. $ay'' + by' + cy = p(x)$ $c \neq 0$.　Then $y = y_0 + \cdots + y_d \dfrac{x^d}{d!}$.

Case 2. $ay'' + by' = p(x)$ $b \neq 0$.　Then $y = \left(y_0 + \cdots + y_d \dfrac{x^d}{d!}\right) x$.

Case 3. $ay'' = p(x)$ $a \neq 0$.　Then $y = \left(y_0 + \cdots + y_d \dfrac{x^d}{d!}\right) x^2$.

Proof on page 588.

Theorem 7.9 (Polynomial × Exponential)
Assume a, b, c, k are constants, $a \neq 0$, and $p(x)$ is a polynomial. If Y is a solution of $[a(D+k)^2 + b(D+k) + c]Y = p(x)$, then $y = e^{kx}Y$ is a solution of $ay'' + by' + cy = p(x)e^{kx}$.

Proof on page 588.

Theorem 7.10 (Polynomial × Exponential × Cosine or Sine)
Assume a, b, c, k, m are real, $a \neq 0$, $m > 0$. Let $p(x)$ be a real polynomial and $z = k + im$. If Y is a solution of $[a(D+z)^2 + b(D+z) + c]Y = p(x)$, then $y = e^{kx}\mathcal{R}\mathrm{e}(e^{imx}Y)$ is a solution of $ay'' + by' + cy = p(x)e^{kx}\cos(mx)$ and $y = e^{kx}\mathcal{I}\mathrm{m}(e^{imx}Y)$ is a solution of $ay'' + by' + cy = p(x)e^{kx}\sin(mx)$.

Proof on page 589.

The theorems form the theoretical basis for the method of undetermined coefficients.

Examples and Methods

Example 7.5 (Simple Quadrature)

Solve for y_p in $y'' = 2 - x + x^3$ using the fundamental theorem of calculus, verifying $y_p = x^2 - x^3/6 + x^5/20$.

Solution: Two integrations using the fundamental theorem of calculus give $y = y_0 + y_1 x + x^2 - x^3/6 + x^5/20$. The terms $y_0 + y_1 x$ represent the homogeneous solution y_h. Therefore, $y_p = x^2 - x^3/6 + x^5/20$ is reported. The method works in general for $ay'' + by' + cy = p(x)$, provided $b = c = 0$, that is, in **case 3** of Theorem 7.8. Some explicit details:

$$\begin{aligned}
&\int y''(x)dx = \int (2 - x + x^3)dx && \text{Integrate across on } x. \\
&y' = y_1 + 2x - x^2/2 + x^4/4 && \text{Fundamental theorem.} \\
&\int y'(x)dx = \int (y_1 + 2x - x^2/2 + x^4/4)dx && \text{Integrate across again on } x. \\
&y = y_0 + y_1 x + x^2 - x^3/6 + x^5/20 && \text{Fundamental theorem.}
\end{aligned}$$

Example 7.6 (Undetermined Coefficients: A Hybrid Method)

Solve for y_p in the equation $y'' - y' + y = 2 - x + x^3$ by the method of undetermined coefficients, verifying $y_p = -5 - x + 3x^2 + x^3$.

Solution: Let's begin by *calculating the trial solution* $y = d_0 + d_1 x + d_2 x^2/2 + x^3$. This is done by differentiation of $y'' - y' + y = 2 - x + x^3$ until the right side is constant:

$$y^v - y^{iv} + y''' = 6.$$

The equilibrium method solves the truncated equation $0 + 0 + y''' = 6$ by quadrature to give $y = d_0 + d_1 x + d_2 x^2/2 + x^3$.

The **undetermined coefficients** d_0, d_1, d_2 will be found by a classical technique in which the trial solution y is back-substituted into the differential equation. We begin by computing the derivatives of y:

$$\begin{aligned}
&y = d_0 + d_1 x + d_2 x^2/2 + x^3 && \text{Calculated above; see Theorem 7.8.} \\
&y' = d_1 + d_2 x + 3x^2 && \text{Differentiate.} \\
&y'' = d_2 + 6x && \text{Differentiate again.}
\end{aligned}$$

The relations above are back-substituted into the differential equation $y'' - y' + y = 2 - x + x^3$ as follows:

$$\begin{aligned}
2 - x + x^3 &= y'' - y' + y && \text{Write the DE backwards.} \\
&= \quad [d_2 + 6x] \\
&\quad - [cd_1 + d_2 x + 3x^2] && \text{Substitute for } y,\ y',\ y''. \\
&\quad + [d_0 + d_1 x + d_2 x^2/2 + x^3] \\
&= \quad [c_2 - c_1 + c_0] \\
&\quad + [6 - d_2 + c_1]x && \text{Collect on powers of } x. \\
&\quad + [-3 + d_2/2]x^2 \\
&\quad + [1]x^3
\end{aligned}$$

The coefficients d_0, d_1, d_2 are found by **matching powers** on the LHS and RHS of the expanded equation:

$$\begin{array}{rcll} 2 & = & [d_2 - d_1 + c_0] & \text{match constant term,} \\ -1 & = & [6 - d_2 + d_1] & \text{match } x\text{-term,} \\ 0 & = & [-3 + d_2/2] & \text{match } x^2\text{-term.} \end{array} \tag{9}$$

These equations are solved by back-substitution, starting with the last equation and proceeding to the first equation. The answers are successively $d_2 = 6$, $d_1 = -1$, $d_0 = -5$. For more detail on back-substitution, see the next example. Substitution into $y = d_0 + d_1x + d_2x^2/2 + x^3$ gives the particular solution $y_p = -5 - x + 3x^2 + x^3$.

Example 7.7 (Undetermined Coefficients: Taylor's Method)

Solve for y_p in the equation $y'' - y' + y = 2 - x + x^3$ by Taylor's method, verifying $y_p = -5 - x + 3x^2 + x^3$.

Solution: Theorem 7.8 implies that there is a polynomial solution $y = d_0 + d_1x + d_2x^2/2 + d_3x^3/6$. The **undetermined coefficients** d_0, d_1, d_2, d_3 will be found by a technique related to **Taylor's method** in calculus. The Taylor technique requires differential equations obtained by successive differentiation of $y'' - y' + y = 2 - x + x^3$, as follows.

$$\begin{array}{ll} y'' - y' + y = 2 - x + x^3 & \text{The original.} \\ y''' - y'' + y' = -1 + 3x^2 & \text{Differentiate the original once.} \\ y^{iv} - y''' + y'' = 6x & \text{Differentiate the original twice.} \\ y^{v} - y^{iv} + y''' = 6 & \text{Differentiate the original three times. The process stops} \\ & \text{when the right side is constant.} \end{array}$$

Set $x = 0$ in the above differential equations. Then substitute the Taylor polynomial derivative relations

$$y(0) = d_0,\ y'(0) = d_1,\ y''(0) = d_2,\ y'''(0) = d_3.$$

It is also true that $y^{iv}(0) = y^{v}(0) = 0$, since y is a cubic. This produces the following equations for *undetermined coefficients* d_0, d_1, d_2, d_3:

$$\begin{array}{rcr} d_2 - d_1 + d_0 & = & 2 \\ d_3 - d_2 + d_1 & = & -1 \\ -d_3 + d_2 & = & 0 \\ d_3 & = & 6 \end{array}$$

These equations are solved by **back-substitution**, working in reverse order. No experience with linear algebra is required, because this is strictly a low-level college algebra method. Successive back-substitutions, working from the last equation in reverse order, give the answers

$$\begin{array}{ll} d_3 = 6, & \text{Use the fourth equation first.} \\ d_2 = d_3 & \text{Solve for } d_2 \text{ in the third equation.} \\ \quad = 6, & \text{Back-substitute } d_3. \\ d_1 = -1 + d_2 - d_3 & \text{Solve for } d_1 \text{ in the second equation.} \end{array}$$

$= -1,$	Back-substitute d_2 and d_3.
$d_0 = 2 + d_1 - d_2$	Solve for d_0 in the first equation.
$= -5.$	Back-substitute d_1 and d_2.

The result is $d_0 = -5$, $d_1 = -1$, $d_2 = 6$, $d_3 = 6$. Substitution into $y = d_0 + d_1x + d_2x^2/2 + d_3x^3/6$ gives the particular solution $y_p = -5 - x + 3x^2 + x^3$.

Example 7.8 (Polynomial Method: Recursive Hybrid)

In the equation $y'' - y' = 2 - x + x^3$, verify $y_p = -7x - 5x^2/2 - x^3 - x^4/4$ by the polynomial method, using a recursive hybrid.

Solution: A **Recursive Method** will be applied, based upon the fundamental theorem of calculus, as in Example 7.5.

Step 1. Differentiate $y'' - y' = 2 - x + x^3$ until the right side is constant, to obtain

Equation 1: $y'' - y' = 2 - x + x^3$	The original.
Equation 2: $y''' - y'' = -1 + 3x^2$	Differentiate the original once.
Equation 3: $y^{iv} - y''' = 6x$	Differentiate the original twice.
Equation 4: $y^{v} - y^{iv} = 6$	Differentiate the original three times. The process stops when the right side is constant.

Step 2. There are 4 equations. Theorem 7.8 implies that there is a polynomial solution y of degree 4. Then $y^v = 0$.

The last equation $y^v - y^{iv} = 6$ then gives $y^{iv} = -6$, which can be solved for y''' by the fundamental theorem of calculus. Then $y''' = -6x + c$. Evaluate c by requiring that y satisfy equation 3: $y^{iv} - y''' = 6x$. Substitution of $y''' = -6x + c$, followed by setting $x = 0$ gives $-6 - c = 0$. Hence $c = -6$. The conclusion: $y''' = -6x - 6$.

Step 3. Solve $y''' = -6x - 6$, giving $y'' = -3x^2 - 6x + c$. Evaluate c as in *Step 2* using equation 2: $y''' - y'' = -1 + 3x^2$. Then $-6 - c = -1$ gives $c = -5$. The conclusion: $y'' = -3x^2 - 6x - 5$.

Step 4. Solve $y'' = -3x^2 - 6x - 5$, giving $y' = -x^3 - 3x^2 - 5x + c$. Evaluate c as in *Step 2* using equation 1: $y'' - y' = 2 - x + x^3$. Then $-5 - c = 2$ gives $c = -7$. The conclusion: $y' = -x^3 - 3x^2 - 5x - 7$.

Step 5. Solve $y' = -x^3 - 3x^2 - 5x - 7$, giving $y = -x^4/4 - x^3 - 5x^2/2 - 7x + c$. Just one solution is sought, so take $c = 0$. Then $y = -7x - 5x^2/2 - x^3 - x^4/4$. Theorem 7.8 also drops the constant term, because it is included in the homogeneous solution y_h. While this method duplicates all the steps in Example 7.7, it remains attractive due to its simplistic implementation. The method is best appreciated when it terminates at step 2 or 3.

Example 7.9 (Polynomial × Exponential)

Solve for y_p in $y'' - y' + y = (2 - x + x^3)e^{2x}$, verifying that $y_p = e^{2x}(x^3/3 - x^2 + x + 1/3)$.

Solution: Let $y = e^{2x}Y$ and $[(D+2)^2 - (D+2) + 1]Y = 2 - x + x^3$, as per the *polynomial* × *exponential method*, page 577. The equation $Y'' + 3Y' + 3Y = 2 - x + x^3$ will be solved by the polynomial method of Example 7.7.

Differentiate $Y'' + 3Y' + 3Y = 2 - x + x^3$ until the right side is constant.

$$\begin{aligned}
&Y'' + 3Y' + 3Y = 2 - x + x^3\\
&Y''' + 3Y'' + 3Y' = -1 + 3x^2\\
&Y^{iv} + 3Y''' + 3Y'' = 6x\\
&Y^{v} + 3Y^{iv} + 3Y''' = 6
\end{aligned}$$

The last equation, by the equilibrium method, implies Y is a polynomial of degree 4, $Y = d_0 + d_1 x + d_2 x^2/2 + d_3 x^3/6$. Set $x = 0$ and $d_i = Y^{(i)}(0)$ in the preceding equations to get the system

$$\begin{array}{rcr}
d_2 + 3d_1 + 3d_0 & = & 2\\
d_3 + 3d_2 + 3d_1 & = & -1\\
d_4 + 3d_3 + 3d_2 & = & 0\\
d_5 + 3d_4 + 3d_3 & = & 6
\end{array}$$

in which $d_4 = d_5 = 0$. Solving by back-substitution gives the answers $d_3 = 2$, $d_2 = -2$, $d_1 = 1$, $d_0 = 1/3$. Then $Y = x^3/3 - x^2 + x + 1/3$.

Finally, Kümmer's transformation $y = e^{2x}Y$ implies $y = e^{2x}(x^3/3 - x^2 + x + 1/3)$.

Example 7.10 (Polynomial × Exponential × Cosine)

Solve in $y'' - y' + y = (3 - x)e^{2x}\cos(3x)$ for y_p, verifying that $y_p = \frac{1}{507}((26x - 107)e^{2x}\cos(3x) + (115 - 39x)e^{2x}\sin(3x))$.

Solution: Let $z = 2 + 3i$. If Y satisfies $[(D+z)^2 - (D+z) + 1]Y = 3 - x$, then $y = e^{2x}\mathcal{R}\text{e}(e^{3ix}Y)$, by the method on page 578. The differential equation simplifies into $Y'' + (3+6i)Y' + (9i-6)Y = 3 - x$. It will be solved by the recursion method of Example 7.8.

Step 1. Differentiate $Y'' + (3+6i)Y' + (9i-6)Y = 3 - x$ until the right side is constant, to obtain $Y''' + (3+6i)Y'' + (9i-6)Y' = -1$. The conclusion: $Y' = 1/(6-9i)$.

Step 2. Solve $Y' = 1/(6-9i)$ for $Y = x/(6-9i) + c$. Evaluate c by requiring Y to satisfy the original equation $Y'' + (3+6i)Y' + (9i-6)Y = 3 - x$. Substitution of $Y' = x/(6-9i) + c$, followed by setting $x = 0$ gives $0 + (3+6i)/(6-9i) + (9i-6)c = 3$. Hence $c = (-15+33i)/(6-9i)^2$. The conclusion: $Y = x/(6-9i) + (-15+33i)/(6-9i)^2$.

Step 3. Use variable $y = e^{2x}\mathcal{R}\text{e}(e^{3ix}Y)$ to complete the solution. This is the point where complex arithmetic must be used. Let $y = e^{2x}\mathcal{Y}$ where $\mathcal{Y} = \mathcal{R}\text{e}(e^{3ix}Y)$. Some details:

$$\begin{aligned}
Y &= \frac{x}{6-9i} + \frac{-15+33i}{(6-9i)^2} && \text{The plan: write } Y = Y_1 + iY_2.\\
&= x\frac{6+9i}{6^2+9^2} + \frac{(-15+33i)(6+9i)^2}{(6^2+9^2)^2} && \text{Use } 1/Z = \overline{Z}/|Z|^2,\ Z = a+ib,\ \overline{Z} = a - ib,\ |Z| = a^2+b^2.\\
&= \frac{2x}{39} + \frac{xi}{13} + \frac{-2889-3105i}{117^2} && \text{Use } 6^2+9^2 = 117 = (9)(13).\\
&= \frac{26x-107}{507} + i\frac{39x-115}{507} && \text{Split off real and imaginary.}
\end{aligned}$$

$$Y_1 = \frac{26x-107}{507}, \quad Y_2 = \frac{39x-115}{507}$$ Decomposition found.

$$\begin{aligned}\mathcal{Y} &= \mathcal{R}\mathrm{e}((\cos 3x + i\sin 3x)(Y_1 + iY_2)) && \text{Use } e^{3ix} = \cos 3x + i\sin 3x.\\ &= Y_1\cos 3x - Y_2 \sin 3x && \text{Take the real part.}\\ &= \frac{26x-107}{507}\cos 3x + \frac{115-39x}{507}\sin 3x && \text{Substitute for } Y_1,\ Y_2.\end{aligned}$$

The solution $y = e^{2x}\mathcal{Y}$ multiplies the above display by e^{2x}. This verifies the formula $y_p = \frac{1}{507}((26x-107)e^{2x}\cos(3x) + (115-39x)e^{2x}\sin(3x))$.

Example 7.11 (Polynomial × Exponential × Sine)

Solve in $y'' - y' + y = (3-x)e^{2x}\sin(3x)$ for y_p, verifying that a particular solution is $y_p = \frac{1}{507}\left((39x-115)e^{2x}\cos(3x) + (26x-107)e^{2x}\sin(3x)\right)$.

Solution: Let $z = 2+3i$. Kümmer's transformation $y = e^{2x}\,\mathcal{I}\mathrm{m}(e^{3ix}Y)$ as on page 578 implies that Y satisfies $[(D+z)^2-(D+z)+1]Y = 3-x$. This equation has been solved in the previous example: $Y = Y_1+iY_2$ with $Y_1 = (26x-107)/507$ and $Y_2 = (39x-115)/507$. Let $\mathcal{Y} = \mathcal{I}\mathrm{m}(e^{3ix}Y)$. Then

$$\begin{aligned}\mathcal{Y} &= \mathcal{I}\mathrm{m}((\cos 3x + i\sin 3x)(Y_1 + iY_2)) && \text{Expand complex factors.}\\ &= Y_2\cos 3x + Y_1\sin 3x && \text{Extract the imaginary part.}\\ &= \frac{(39x-115)\cos 3x + (26x-107)\sin 3x}{507} && \text{Substitute for } Y_1 \text{ and } Y_2.\end{aligned}$$

The solution $y = e^{2x}\mathcal{Y}$ multiplies the display by e^{2x}. This verifies the formula $y = \frac{1}{507}\left((39x-115)e^{2x}\cos(3x) + (26x-107)e^{2x}\sin(3x)\right)$.

Example 7.12 (Undetermined Coefficient Library Methods)

Solve $y'' - y' + y = 1 + e^x + \cos(x)$, verifying

$$y = c_1 e^{x/2}\cos(\sqrt{3}x/2) + c_2 e^{x/2}\sin(\sqrt{3}x/2) + 1 + e^x - \sin(x).$$

Solution: There are $n = 3$ easily solved equations: $y_1'' - y_1' + y_1 = 1$, $y_2'' - y_2' + y_2 = e^x$ and $y_3'' - y_3' + y_3 = \cos(x)$. The plan is that each such equation is solvable by one of the **library methods**. Then $y_p = y_1 + y_2 + y_3$ is the sought particular solution.

Equation 1: $y_1'' - y_1' + y_1 = 1$. It is solved by *the equilibrium method*, which gives immediately solution $y_1 = 1$.

Equation 2: $y_2'' - y_2' + y_2 = e^x$. Then $y_2 = e^xY$ and $[(D+1)^2 - (D+1) + 1]Y = 1$, by the *polynomial × exponential method*. The equation simplifies to $Y'' + Y' + Y = 1$. Obtain $Y = 1$ by the *equilibrium method*, then $y_2 = e^x$.

Equation 3: $y_3'' - y_3' + y_3 = \cos(x)$. Then $[(D+i)^2 - (D+i) + 1]Y = 1$ and $y_3 = \mathcal{R}\mathrm{e}(e^{ix}Y)$, by the *polynomial × exponential × cosine method*. The equation simplifies to $Y'' + (2i-1)Y' - iY = 1$. Obtain $Y = i$ by the *equilibrium method*. Then $y_3 = \mathcal{R}\mathrm{e}(e^{ix}Y)$ implies $y_3 = -\sin(x)$.

Solution y_p. The particular solution is given by addition, $y_p = y_1 + y_2 + y_3$. Therefore, $y_p = 1 + e^x - \sin(x)$.

Solution y_h. The homogeneous solution y_h is the linear equation solution for $y''-y'+y=0$, obtained from Theorem 6.1, which uses the characteristic equation $r^2-r+1=0$. The latter has roots $r=(1\pm i\sqrt{3})/2$ and then $y_h=c_1e^{x/2}\cos(\sqrt{3}x/2)+c_2e^{x/2}\sin(\sqrt{3}x/2)$ where c_1 and c_2 are arbitrary constants.

General Solution. Add y_h and y_p to obtain the general solution

$$y=c_1e^{x/2}\cos(\sqrt{3}x/2)+c_2e^{x/2}\sin(\sqrt{3}x/2)+1+e^x-\sin(x).$$

Example 7.13 (Sine–Cosine Trial solution)

Verify for $y''+4y=\sin x-\cos x$ that $y_p(x)=5\cos x+3\sin x$.

Solution: The lookup table method suggests to substitute $y=d_1\cos x+d_2\sin x$ into the differential equation. The correction rule does not apply, because the homogeneous solution terms involve $\cos 2x$, $\sin 2x$. Use $u''=-u$ for $u=\sin x$ or $u=\cos x$ to obtain the relation

$$\begin{aligned}\sin x-\cos x &= y''+4y\\ &= (-d_1+4)\cos x+(-d_2+4)\sin x.\end{aligned}$$

Comparing sides, matching sine and cosine terms, gives

$$\begin{aligned}-d_1+4 &= -1,\\ -d_2+4 &= 1.\end{aligned}$$

Solving, $d_1=5$ and $d_2=3$. The trial solution $y=d_1\cos x+d_2\sin x$ becomes $y_p(x)=5\cos x+3\sin x$.

Historical Notes

The method of undetermined coefficients presented on page 104 uses the idea of a **trial solution**. Textbooks that present this method appear in the references, especially Edwards–Penney [EP2] and Kreyszig [Kreyszig].

If the right side $f(x)$ is a polynomial, then the trial solution is a polynomial $y=d_0+\cdots+d_kx^k$ with unknown coefficients. It is substituted into the non-homogeneous differential equation to determine the coefficients $d_0,\ldots,d_k$, as in Example 7.6. The Taylor method in Example 7.7 implements the same ideas. In the some textbook presentations, the three key theorems of this section are replaced by Table 1 and the **Correction Rule** on page 580. Attempts have been made to integrate the correction rule into the table itself; see Edwards–Penney [EP], [EP2].

The *method of annihilators* has been used as an alternative approach; see Kreider–Kuller–Ostberg–Perkins [KKOP]. The approach gives a deeper insight into higher order differential equations. It requires knowledge of linear algebra and a small nucleus of differential operator calculus.

The idea to employ a recursive polynomial method seems to appear first in a paper by Love [Love1989]. A generalization and expansion of details appears in [Gupta]. The method is certainly worth learning, but doing so does not excuse one from learning other methods. The recursive method is a worthwhile hybrid method for special circumstances.

Proofs and Technical Details

Proof of Theorem 7.8: The three cases correspond to zero, one or two roots $r = 0$ for the characteristic equation $ar^2 + br + c = 0$. The missing constant and x-terms in **case 2** and **case 3** are justified by including them in the homogeneous solution y_h, instead of in the particular solution y_p.

Assume $p(x)$ has degree d and succinctly write down the successive derivatives of the differential equation as

$$ay^{(2+k)} + by^{(1+k)} + cy^{(k)} = p^{(k)}(x), \quad k = 0, \ldots, d. \tag{10}$$

Assume, to consider simultaneously all three cases, that

$$y = y_0 + y_1 + \cdots + y_{m+d}\frac{x^{m+d}}{(m+d)!}$$

where $m = 0, 1, 2$ corresponding to cases 1,2,3, respectively. It has to be shown that there are coefficients y_0, …, y_{m+d} such that y is a solution of $ay'' + by' + cy = p(x)$.

Let $x = 0$ in equations (10) and use the definition of polynomial y to obtain the equations

$$ay_{2+k} + by_{1+k} + cy_k = p^{(k)}(0), \quad k = 0, \ldots, d. \tag{11}$$

In **case 1** ($c \neq 0$), $m = 0$ and the last equation in (11) gives $y_{m+d} = p^{(d)}(0)/c$. Back-substitution succeeds in finding the other coefficients, in reverse order, because $y^{(d+1)}(0) = y^{(d+2)}(0) = 0$, in this case. *Define* the constants y_0 to y_d to be the solutions of (11). Define $y_{d+1} = y_{d+2} = 0$.

In **case 2** ($c = 0$, $b \neq 0$), $m = 1$ and the last equation in (11) gives $y_{m+d} = p^{(d)}(0)/b$. Back-substitution succeeds in finding the other coefficients, in reverse order, because $y^{(d+2)}(0) = 0$, in this case. However, y_0 is undetermined. Take it to be zero, then *define* y_1 to y_{d+1} to be the solutions of (11). Define $y_{d+2} = 0$.

In **case 3** ($c = b = 0$), $m = 2$ and the last equation in (11) gives $y_{m+d} = p^{(d)}(0)/a$. Back-substitution succeeds in finding the other coefficients, in reverse order. However, y_0 and y_1 are undetermined. Take them to be zero, then *define* y_2 to y_{d+2} to be the solutions of (11).

It remains to prove that the polynomial y so defined is a solution of the differential equation $ay'' + by' + cy = p(x)$. Begin by applying quadrature to the last differentiated equation $ay^{(2+d)} + by^{(1+d)} + cy^{(d)} = p^{(d)}(x)$. The result is $ay^{(1+d)} + by^{(d)} + cy^{(d-1)} = p^{(d-1)}(x) + C$ with C undetermined. Set $x = 0$ in this equation. Then relations (11) say that $C = 0$. This process can be continued until $ay'' + by' + cy = q(x)$ is obtained, hence y is a solution.

Proof of Theorem 7.9: Kümmer's transformation $y = e^{kx}Y$ is differentiated twice to give the formulas

$$\begin{array}{rcl} y &=& e^{kx}Y, \\ y' &=& ke^{kx}Y + e^{kx}Y' \\ &=& e^{kx}(D+k)Y, \\ y'' &=& k^2e^{kx}Y + 2ke^{kx}Y' + e^{kx}Y'' \\ &=& e^{kx}(D+k)^2Y. \end{array}$$

Insert them into the differential equation $a(D+k)^2Y + b(D+k)Y + cY = p(x)$. Then multiply through by e^{kx} to remove the common factor e^{-kx} on the left, giving $ay'' + by' + cy = p(x)e^{kx}$. This completes the proof.

Proof of Theorem 7.10: Abbreviate $ay'' + by' + cy$ by Ly. Consider the complex equation $Lu = p(x)e^{zx}$, to be solved for $u = u_1 + iu_2$. According to Theorem 7.9, u can be computed as $u = e^{zx}Y$ where $[a(D+z)^2 + b(D+z) + c]Y = p(x)$. Take the real and imaginary parts of $u = e^{zx}Y$ and $Lu = p(x)e^{zx}$. Then $u_1 = \mathcal{R}\mathrm{e}(e^{zx}Y)$ and $u_2 = \mathcal{I}\mathrm{m}(e^{zx}Y)$ satisfy $Lu_1 = \mathcal{R}\mathrm{e}(p(x)e^{zx}) = p(x)\cos(mx)e^{kx}$ and $Lu_2 = \mathcal{I}\mathrm{m}(p(x)e^{zx}) = p(x)\sin(mx)e^{kx}$. ■

Exercises 7.6

Polynomial Solutions

Determine a polynomial solution y_p for the given differential equation. Apply Theorem 7.8, page 581, and model the solution after Examples 7.5, 7.6, 7.7 and 7.8.

1. $y'' = x$

2. $y'' = x - 1$

3. $y'' = x^2 - x$

4. $y'' = x^2 + x - 1$

5. $y'' - y' = 1$

6. $y'' - 5y' = 10$

7. $y'' - y' = x$

8. $y'' - y' = x - 1$

9. $y'' - y' + y = 1$

10. $y'' - y' + y = -2$

11. $y'' + y = 1 - x$

12. $y'' + y = 2 + x$

13. $y'' - y = x^2$

14. $y'' - y = x^3$

Polynomial-Exponential Solutions

Determine a solution y_p for the given differential equation. Apply Theorem 7.9, page 581, and model the solution after Example 7.9.

15. $y'' + y = e^x$

16. $y'' + y = e^{-x}$

17. $y'' = e^{2x}$

18. $y'' = e^{-2x}$

19. $y'' - y = (x+1)e^{2x}$

20. $y'' - y = (x-1)e^{-2x}$

21. $y'' - y' = (x+3)e^{2x}$

22. $y'' - y' = (x-2)e^{-2x}$

23. $y'' - 3y' + 2y = (x^2+3)e^{3x}$

24. $y'' - 3y' + 2y = (x^2-2)e^{-3x}$

Sine and Cosine Solutions

Determine a solution y_p for the given differential equation. Apply Theorem 7.10, page 581, and model the solution after Examples 7.10 and 7.11.

25. $y'' = \sin(x)$

26. $y'' = \cos(x)$

27. $y'' + y = \sin(x)$

28. $y'' + y = \cos(x)$

29. $y'' = (x+1)\sin(x)$

30. $y'' = (x+1)\cos(x)$

31. $y'' - y = (x+1)e^x\sin(2x)$

32. $y'' - y = (x+1)e^x\cos(2x)$

33. $y'' - y' - y = (x^2+x)e^x\sin(2x)$

34. $y'' - y' - y = (x^2+x)e^x\cos(2x)$

Undetermined Coefficients Algorithm

Determine a solution y_p for the given differential equation. Apply the polynomial algorithm, page 576, and model the solution after Example 7.12.

35. $y'' = x + \sin(x)$

36. $y'' = 1 + x + \cos(x)$

37. $y'' + y = x + \sin(x)$

38. $y'' + y = 1 + x + \cos(x)$

39. $y'' + y = \sin(x) + \cos(x)$

40. $y'' + y = \sin(x) - \cos(x)$

41. $y'' = x + xe^x + \sin(x)$

42. $y'' = x - xe^x + \cos(x)$

43. $y'' - y = \sinh(x) + \cos^2(x)$

44. $y'' - y = \cosh(x) + \sin^2(x)$

45. $y'' + y' - y = x^2e^x + xe^x\cos(2x)$

46. $y'' + y' - y = x^2e^{-x} + xe^x\sin(2x)$

Additional Proofs

The exercises below fill in details in the text. The hints are in the proofs in the textbook. No solutions will be given for the odd exercises.

47. (Theorem 7.8)
Supply the missing details in the proof of Theorem 7.8 for case 1. In particular, give the details for back-substitution.

48. (Theorem 7.8)
Supply the details in the proof of Theorem 7.8 for case 2. In particular, give the details for back-substitution and explain fully why it is possible to select $y_0 = 0$.

49. (Theorem 7.8)
Supply the details in the proof of Theorem 7.8 for case 3. In particular, explain why back-substitution leaves y_0 and y_1 undetermined, and why it is possible to select $y_0 = y_1 = 0$.

50. (Superposition)
Let Ly denote $ay''+by'+cy$. Show that solutions of $Lu = f(x)$ and $Lv = g(x)$ add to give $y = u + v$ as a solution of $Ly = f(x) + g(x)$.

51. (Easily Solved Equations)
Let Ly denote $ay'' + by' + cy$. Let $Ly_k = f_k(x)$ for $k = 1, \ldots, n$ and define $y = y_1 + \cdots + y_n$, $f = f_1 + \cdots + f_n$. Show that $Ly = f(x)$.

Chapter 8

Laplace Transform

Contents

The Laplace transform solves differential equations. Besides being a different and efficient alternative to variation of parameters and undetermined coefficients, Laplace's method is especially advantageous for a forcing term that is piecewise–defined, periodic or impulsive.

The Laplace method. It has humble beginnings as an extension of *the method of quadrature* to higher order differential equations and systems. The method is based upon ordinary calculus integrals:

> Multiply the differential equation by the Laplace integrator $dx = e^{-st}dt$ and integrate across the equation from $t = 0$ to $t = \infty$. Isolate left the Laplace integral $\int_{t=0}^{t=\infty} y(t)e^{-st}dt$. Look up the answer $y(t)$ in a Laplace integral table.

Definition 8.1 (Laplace Integral)

The **Laplace integral** or the **direct Laplace transform** of a function $f(t)$ defined for $0 \le t < \infty$ is the answer to the Newton calculus integration problem $\int_0^\infty f(t)e^{-st}dt$. Special notation replaces the integral notation in literature:

$$\mathcal{L}(f(t)) \quad \text{replaces} \quad \int_0^\infty f(t)\, e^{-st}\, dt.$$

Decoding and Encoding. The $\mathcal{L}$–notation recognizes that integration always proceeds over $t = 0$ to $t = \infty$ and that the integral always has a fixed *integrator* $e^{-st}dt$ instead of the expected dx. These minor differences distinguish *Laplace integrals* from the *ordinary integrals* found on the inside covers of calculus texts.

When reading mathematical text, replace symbol $\mathcal{L}$ by these words: **Laplace of**. Notation $\mathcal{L}(f(t))$ decodes into calculus by replacing $\mathcal{L}$ by $\int_0^\infty$, then append the Laplace integrator $e^{-st}dt$. For instance, notation $\mathcal{L}(t^2)$ decodes to $\int_0^\infty (t^2)\, e^{-s\,t}\, dt$. To encode $\int_0^\infty (\sin t)\, e^{-s\,t}\, dt$ to $\mathcal{L}(\sin t)$, replace $\int_0^\infty$ by $\mathcal{L}$, then erase Laplace integrator $e^{-st}dt$.

History. The first application of the Laplace method might have been in the 1910 work of H. Bateman [Bateman], who transformed Rutherford's radioactive decay equation $\frac{d}{dt}A(t) = -hA(t)$ by setting $a(x) = \int_0^\infty e^{-xt}A(t)dt$, thereby obtaining an equation in variable x (Laplace theory uses s instead of x). The first example presented here will parallel Bateman's 1910 exposition, in which he derived several properties of the Laplace integral as well as isolating what is today called *Laplace's method.* He used Lerch's 1903 theorem published in *Acta Mathematica.* The name *Laplace Transform* dates back to Euler 1763 and Spitzer 1878, which nowadays refers to the linear map $f \to \mathcal{L}(f(t)) \equiv \int_0^\infty e^{-st}f(t)dt$.

8.1 Laplace Method Introduction

The foundation of Laplace theory is **Lerch's 1903 cancellation law**

$$(1)\qquad \begin{array}{lcl} \int_0^\infty y(t)e^{-st}dt = \int_0^\infty f(t)e^{-st}dt & \text{implies} & y(t) = f(t), \\ & \text{or} & \\ \mathcal{L}(y(t) = \mathcal{L}(f(t)) & \text{implies} & y(t) = f(t). \end{array}$$

In differential equation applications, $y(t)$ is the unknown appearing in the equation while $f(t)$ is an explicit expression extracted or computed from Laplace integral tables. See page 596.

An Illustration. Laplace's method will be applied to solve the initial value problem[1]

$$\frac{dy}{dt} = -1, \quad y(0) = 0.$$

No background in Laplace theory is assumed here, only a calculus background is used. Calculus verifies the answer $y(t) = -t$.

The Plan. The method obtains an equation $\mathcal{L}(y(t)) = \mathcal{L}(-t)$, then Lerch's cancellation law implies that the $\mathcal{L}$-symbols cancel, which gives the differential equation solution $y(t) = -t$.

The Laplace method is advertised as a *generalization of the method of quadrature* to higher order differential equations and systems of differential equations. In addition to quadrature, the method uses *table lookup*: solution $y(t)$ is found from a special integral table.

[1] Laplace theory uses t instead of x. Prime notation y' means $\frac{dy}{dt}$.

Laplace Integral

The integral $\int_0^\infty g(t)e^{-st}dt$ is called the **Laplace integral** of the function $g(t)$. It is defined by $\lim_{N\to\infty}\int_0^N g(t)e^{-st}dt$ and depends on variable s. The ideas will be illustrated for $g(t)=1$, $g(t)=t$ and $g(t)=t^2$, producing the integral formulas in Table 1, *infra*.

$$\begin{array}{lll}
\boxed{1} & \int_0^\infty (1)e^{-st}dt = -(1/s)e^{-st}\Big|_{t=0}^{t=\infty} & \text{Laplace integral of } g(t)=1. \\
 & \qquad\qquad\qquad = 1/s & \text{Assumed } s>0. \\
\boxed{2} & \int_0^\infty (t)e^{-st}dt = \int_0^\infty -\frac{d}{ds}(e^{-st})dt & \text{Laplace integral of } g(t)=t. \\
 & \qquad\qquad\qquad = -\frac{d}{ds}\int_0^\infty (1)e^{-st}dt & \int \frac{d}{ds}F(t,s)dt = \frac{d}{ds}\int F(t,s)dt. \\
 & \qquad\qquad\qquad = -\frac{d}{ds}(1/s) & \text{By } \boxed{1}. \\
 & \qquad\qquad\qquad = 1/s^2 & \text{Differentiate.} \\
\boxed{3} & \int_0^\infty (t^2)e^{-st}dt = \int_0^\infty -\frac{d}{ds}(te^{-st})dt & \text{Laplace integral of } g(t)=t^2. \\
 & \qquad\qquad\qquad = -\frac{d}{ds}\int_0^\infty (t)e^{-st}dt & \\
 & \qquad\qquad\qquad = -\frac{d}{ds}(1/s^2) & \text{By } \boxed{2}. \\
 & \qquad\qquad\qquad = 2/s^3 &
\end{array}$$

Table 1. The Laplace Integral $\int_0^\infty g(t)e^{-st}dt$ for $g(t)=1$, t and t^2.

$$\int_0^\infty (1)e^{-st}\,dt = \frac{1}{s}, \qquad \int_0^\infty (t)e^{-st}\,dt = \frac{1}{s^2}, \qquad \int_0^\infty (t^2)e^{-st}\,dt = \frac{2}{s^3}.$$

$$\text{In summary,}\quad \mathcal{L}(t^n) = \frac{n!}{s^{1+n}}$$

Illustration Details for $y'=-1$, $y(0)=0$

The **Laplace method** will be applied to find the solution $y(t)=-t$ of the problem

$$y'=-1, \quad y(0)=0.$$

Laplace's method in Table 2 is entirely different from variation of parameters or undetermined coefficients. The method uses only basic calculus and college algebra. In the second Table 3, a succinct version of the first Table 2 is given, using $\mathcal{L}$-notation. The briefer exposition is a model for Laplace Method details as found in references.

Table 2. Laplace Method Details for Illustration $y' = -1$, $y(0) = 0$.

$y'(t)e^{-st}dt = -e^{-st}dt$	Multiply $y' = -1$ by $e^{-st}dt$.
$\int_0^\infty y'(t)e^{-st}dt = \int_0^\infty -e^{-st}dt$	Integrate $t = 0$ to $t = \infty$.
$\int_0^\infty y'(t)e^{-st}dt = -1/s$	Use Table 1 forwards.
$s\int_0^\infty y(t)e^{-st}dt - y(0) = -1/s$	Integrate by parts on the left.
$\int_0^\infty y(t)e^{-st}dt = -1/s^2$	Use $y(0) = 0$ and divide.
$\int_0^\infty y(t)e^{-st}dt = \int_0^\infty (-t)e^{-st}dt$	Use Table 1 backwards.
$y(t) = -t$	Apply Lerch's cancellation law. Solution found.

Table 3. Laplace Method $\mathcal{L}$-notation
Details for $y' = -1$, $y(0) = 0$ translated from Table 2.

$\mathcal{L}(y'(t)) = \mathcal{L}(-1)$	Apply $\mathcal{L}$ across $y' = -1$, or multiply $y' = -1$ by $e^{-st}dt$, integrate $t = 0$ to $t = \infty$.
$\mathcal{L}(y'(t)) = -1/s$	Use Table 1 forwards.
$s\mathcal{L}(y(t)) - y(0) = -1/s$	Integrate by parts on the left.
$\mathcal{L}(y(t)) = -1/s^2$	Use $y(0) = 0$ and divide.
$\mathcal{L}(y(t)) = \mathcal{L}(-t)$	Apply Table 1 backwards.
$y(t) = -t$	Invoke Lerch's cancellation law.

In Lerch's law, the formal rule of erasing the integral signs is valid *provided* the integrals are equal for large s and certain conditions hold on y and f — see Theorem 8.2. The illustration in Table 2 shows that Laplace theory requires an in-depth study of a **special integral table**, a table which is a true extension of the usual table found on the inside covers of calculus books; see Table 1 and section 8.2, Table 4 page 601.

The $\mathcal{L}$-notation for the direct Laplace transform produces briefer details, as witnessed by the translation of Table 2 into Table 3. It is advised to move from Laplace integral notation to the $\mathcal{L}$–notation as soon as possible, in order to highlight goalposts in the method.

Some Transform Rules

The formal properties of calculus integrals plus the integration by parts formula used in Tables 2 and 3 leads to these **rules** for the Laplace transform:

$\mathcal{L}(f(t) + g(t)) = \mathcal{L}(f(t)) + \mathcal{L}(g(t))$	The integral of a sum is the sum of the integrals.

$\mathcal{L}(cf(t)) = c\mathcal{L}(f(t))$	Constants c pass through the integral sign.
$\mathcal{L}(y'(t)) = s\mathcal{L}(y(t)) - y(0)$	The t-derivative rule, or integration by parts. See Theorem 8.3.
$\mathcal{L}(y(t)) = \mathcal{L}(f(t))$ implies $y(t) = f(t)$	Lerch's cancellation law. See Theorem 8.2.

The four rules above appear in Bateman's 1910 publication [Bateman]. The first two rules are referenced as **linearity of the Laplace transform**, which allow manipulation of the symbol $\mathcal{L}$ with rules known from calculus and matrix algebra. Laplace symbol $\mathcal{L}$ manipulates like matrix multiply.

Existence of the Transform

The Laplace integral $\int_0^\infty e^{-st} f(t)\,dt$ is known to exist in the sense of the improper integral definition[2]

$$\int_0^\infty g(t)dt = \lim_{N\to\infty} \int_0^N g(t)dt$$

provided $f(t)$ belongs to a class of functions known in the literature as functions of **exponential order**. For this class of functions the relation

$$\lim_{t\to\infty} \frac{f(t)}{e^{\alpha t}} = 0 \tag{2}$$

is required to hold for some real number α, or equivalently, for some constants M and α,

$$|f(t)| \le Me^{\alpha t}. \tag{3}$$

In addition, $f(t)$ is required to be **piecewise continuous** on each finite subinterval of $0 \le t < \infty$, a term defined as follows.

Definition 8.2 (Piecewise Continuous)
A function $f(t)$ is **piecewise continuous** on a finite interval $[a, b]$ provided there exists a partition $a = t_0 < \cdots < t_n = b$ of the interval $[a, b]$ and functions f_1, f_2, ..., f_n continuous on $(-\infty, \infty)$ such that for t not a partition point

$$f(t) = \begin{cases} f_1(t) & t_0 < t < t_1, \\ \vdots & \vdots \\ f_n(t) & t_{n-1} < t < t_n. \end{cases} \tag{4}$$

The values of f at partition points are undecided by equation (4). In particular, equation (4) implies that $f(t)$ has one-sided limits at each point of $a < t < b$ and appropriate one-sided limits at the endpoints. Therefore, f has at worst a **jump discontinuity** at each partition point.

[2] An advanced calculus background is assumed for the Laplace transform existence proof. Applications of Laplace theory require only a calculus background.

Theorem 8.1 (Existence of $\mathcal{L}(f)$)
Let $f(t)$ be piecewise continuous on every finite interval in $t \geq 0$ and satisfy $|f(t)| \leq Me^{\alpha t}$ for some constants M and α. Then:

1. Laplace integral $\mathcal{L}(f(t))$ exists for $s > \alpha$.
2. **Laplace is zero at** $s = \infty$: $\lim_{s\to\infty} \mathcal{L}(f(t)) = 0$.[3]

Proof on page 598.

Theorem 8.2 (Lerch 1903)
If $f_1(t)$ and $f_2(t)$ are continuous, of exponential order and for all $s > s_0$

$$\int_0^\infty f_1(t)e^{-st}dt = \int_0^\infty f_2(t)e^{-st}dt,$$

then $f_1(t) = f_2(t)$ for $t \geq 0$.[4]

Proofs in French: Lerch (1903) [Lerch] and English: Widder [Widd1941]. See also [Weis].

Theorem 8.3 (Parts Rule or t-Derivative Rule)
Let $f(t)$ be continuous and of exponential order. Let $f'(t)$ be piecewise continuous and of exponential order. Then $\mathcal{L}(f'(t))$ exists for all large s and $\mathcal{L}(f'(t)) = s\mathcal{L}(f(t)) - f(0)$.

Proof on page 639.

Theorem 8.4 (Euler Solution Atoms have Laplace Integrals)
Let $f(t)$ be $t^n e^{at}$ or the real or imaginary part of $t^n e^{at+ibt}$ where a, b are real, $b > 0$ and $n \geq 0$ is an integer. Briefly, f is an **Euler solution atom**. Then f is of exponential order and $\mathcal{L}(f(t))$ exists. Further, if $g(t)$ is a linear combination of Euler atoms, then $\mathcal{L}(g(t))$ exists.

Proof on page 598.

Remark. Because solutions to undetermined coefficient problems are a linear combination of Euler solution atoms, then Laplace's method applies to all such differential equations. This is the class of all constant-coefficient higher order linear differential equations, and all systems of differential equations with constant coefficients, having a forcing term which is a linear combination of Euler solution atoms.

[3] Literature might write $F(s)$ for $\mathcal{L}(f(f))$ and $\lim_{s\to\infty} F(s) = 0$

[4] The result extends to piecewise continuous functions provided the conclusion is weakened to: *at points where both f_1, f_2 are continuous, $f_1(t) = f_2(t)$*. Reference: *CRC Concise Encyclopedia of Mathematics* by Weisstein.

Examples and Methods

Example 8.1 (Laplace Method)

Solve the initial value problem $y' = 5 - 2t$, $y(0) = 1$ by the Laplace method to obtain $y(t) = 1 + 5t - t^2$.

Solution: Laplace's method is outlined in Tables 2 and 3. The $\mathcal{L}$-notation of Table 3 will be used to find the solution $y(t) = 1 + 5t - t^2$.

$$\begin{aligned}
\mathcal{L}(y'(t)) &= \mathcal{L}(5-2t) && \text{Apply } \mathcal{L} \text{ across } y' = 5-2t. \\
&= 5\mathcal{L}(1) - 2\mathcal{L}(t) && \text{Linearity of the transform.} \\
&= \frac{5}{s} - \frac{2}{s^2} && \text{Use Table 1 forwards.} \\
s\mathcal{L}(y(t)) - y(0) &= \frac{5}{s} - \frac{2}{s^2} && \text{Apply the parts rule, Theorem 8.3.} \\
\mathcal{L}(y(t)) &= \frac{1}{s} + \frac{5}{s^2} - \frac{2}{s^3} && \text{Use } y(0)=1 \text{ and divide.} \\
\mathcal{L}(y(t)) &= \mathcal{L}(1) + 5\mathcal{L}(t) - \mathcal{L}(t^2) && \text{Apply Table 1 backwards.} \\
&= \mathcal{L}(1 + 5t - t^2) && \text{Linearity of the transform.} \\
y(t) &= 1 + 5t - t^2 && \text{Use Lerch's cancellation law.}
\end{aligned}$$

Example 8.2 (Laplace Method)

Solve by Laplace's method the initial value problem $y'' = 10$, $y(0) = y'(0) = 0$ to obtain $y(t) = 5t^2$.

Solution: The $\mathcal{L}$-notation of Table 3 will be used to find the solution $y(t) = 5t^2$.

$$\begin{aligned}
\mathcal{L}(y''(t)) &= \mathcal{L}(10) && \text{Apply } \mathcal{L} \text{ across } y'' = 10. \\
s\mathcal{L}(y'(t)) - y'(0) &= \mathcal{L}(10) && \text{Apply the parts rule to } y'\text{, that is, replace } f \text{ by } y' \text{ in Theorem 8.3.} \\
s[s\mathcal{L}(y(t)) - y(0)] - y'(0) &= \mathcal{L}(10) && \text{Repeat the parts rule, on } y. \\
s^2\mathcal{L}(y(t)) &= \mathcal{L}(10) && \text{Use } y(0) = y'(0) = 0. \\
\mathcal{L}(y(t)) &= \frac{10}{s^3} && \text{Use Table 1 forwards. Then divide.} \\
\mathcal{L}(y(t)) &= \mathcal{L}(5t^2) && \text{Apply Table 1 backwards.} \\
y(t) &= 5t^2 && \text{Invoke Lerch's cancellation law.}
\end{aligned}$$

Example 8.3 (Exponential Order)

Show that $f(t) = e^t \cos t + t$ is of exponential order.

Solution: The proof must show that $f(t)$ is piecewise continuous on every interval $[a, b]$ and then find an $\alpha > 0$ such that $\lim_{t\to\infty} f(t)/e^{\alpha t} = 0$.

The given $f(t)$ is continuous on $(-\infty, \infty)$. Given interval $[a, b]$, define $t_0 = a, t_1 = b$ and $f_1(t) = f(t)$. Then (4) holds. Definition 8.2 implies f is piecewise continuous.

From L'Hospital's rule in calculus, $\lim_{t\to\infty} p(t)/e^{\alpha t} = 0$ for any polynomial p and any $\alpha > 0$. Choose $\alpha = 2$, then

$$\lim_{t\to\infty} \frac{f(t)}{e^{2t}} = \lim_{t\to\infty} \frac{\cos t}{e^t} + \lim_{t\to\infty} \frac{t}{e^{2t}} = 0.$$

Proofs and Technical Details

Proof of Theorem 8.1, Existence Laplace Integral:

Details 1. It has to be shown that the Laplace integral of f is finite for $s > \alpha$. Advanced calculus implies that it is sufficient to show that the integrand is absolutely bounded above by an integrable function $g(t)$. Take $g(t) = Me^{-(s-\alpha)t}$. Then $g(t) \ge 0$. Furthermore, g is integrable, because

$$\int_0^\infty g(t)dt = \frac{M}{s-\alpha}.$$

Inequality $|f(t)| \le Me^{\alpha t}$ implies the absolute value of the Laplace transform integrand $f(t)e^{-st}$ is estimated by

$$\left|f(t)e^{-st}\right| \le Me^{\alpha t}e^{-st} = g(t).$$

Details 2. The limit statement $\lim_{s\to\infty} \mathcal{L}(f(t)) = 0$ follows from $|\mathcal{L}(f(t))| \le \int_0^\infty g(t)dt = \dfrac{M}{s-\alpha}$, because the right side of this inequality has limit zero at $s = \infty$. ■

Proof of Theorem 8.4, Euler Atoms:

Function $f(t) = t^n e^{at}$ is everywhere continuous. By calculus, $\ln|x| \le 2x$ for $x \ge 1$. Define $c = 2|n| + |a|$. Then $|f(t)| = e^{n\ln|t|+at} \le e^{ct}$ for $t \ge 1$, which proves f is of exponential order. Similarly, $f(t) = \mathcal{R}\text{e}(t^n e^{at+ibt})$ is everywhere continuous and $|f(t)| \le |t^n e^{at+ibt}| = |t^n e^{at}| \le e^{ct}$. Details for $f(t) = \mathcal{I}\text{m}(t^n e^{at+ibt})$ are similar. Then f is of exponential order in all three cases. The Laplace integral exists by Theorem 8.1 page 596. ■

Exercises 8.1

Laplace method

Solve the given initial value problem using Laplace's method.

1. $y' = -2$, $y(0) = 0$.

2. $y' = 1$, $y(0) = 0$.

3. $y' = -t$, $y(0) = 0$.

4. $y' = t$, $y(0) = 0$.

5. $y' = 1 - t$, $y(0) = 0$.

6. $y' = 1 + t$, $y(0) = 0$.

7. $y' = 3 - 2t$, $y(0) = 0$.

8. $y' = 3 + 2t$, $y(0) = 0$.

9. $y'' = -2$, $y(0) = y'(0) = 0$.

10. $y'' = 1$, $y(0) = y'(0) = 0$.

11. $y'' = 1 - t$, $y(0) = y'(0) = 0$.

12. $y'' = 1 + t$, $y(0) = y'(0) = 0$.

13. $y'' = 3 - 2t$, $y(0) = y'(0) = 0$.

14. $y'' = 3 + 2t$, $y(0) = y'(0) = 0$.

Exponential order

Show that $f(t)$ is of exponential order, by finding a constant $\alpha \ge 0$ in each case such that $\displaystyle\lim_{t\to\infty} \frac{f(t)}{e^{\alpha t}} = 0$.

15. $f(t) = 1 + t$

16. $f(t) = e^t \sin(t)$

17. $f(t) = \sum_{n=0}^N c_n t^n$, for any choice of the constants $c_0, \ldots, c_N$.

18. $f(t) = \sum_{n=1}^N c_n \sin(nt)$, for any choice of the constants $c_1, \ldots, c_N$.

Existence of transforms

Let $f(t) = te^{t^2}\sin(e^{t^2})$. Establish these results.

19. The function $f(t)$ is not of exponential order.

20. The Laplace integral of $f(t)$, $\int_0^\infty f(t)e^{-st}dt$, converges for all $s > 0$.

Jump Magnitude

For f piecewise continuous, define the **jump** at t by

$$J(t) = \lim_{h\to 0+} f(t+h) - \lim_{h\to 0+} f(t-h).$$

Compute $J(t)$ for the following f.

21. $f(t) = 1$ for $t \ge 0$, else $f(t) = 0$

22. $f(t) = 1$ for $t \ge 1/2$, else $f(t) = 0$

23. $f(t) = t/|t|$ for $t \ne 0$, $f(0) = 0$

24. $f(t) = \sin t/|\sin t|$ for $t \ne n\pi$, $f(n\pi) = (-1)^n$

Taylor series

The series relation $\mathcal{L}(\sum_{n=0}^\infty c_n t^n) = \sum_{n=0}^\infty c_n \mathcal{L}(t^n)$ often holds, in which case the result $\mathcal{L}(t^n) = n!s^{-1-n}$ can be employed to find a series representation of the Laplace transform. Use this idea on the following to find a series formula for $\mathcal{L}(f(t))$.

25. $f(t) = e^{2t} = \sum_{n=0}^\infty (2t)^n/n!$

26. $f(t) = e^{-t} = \sum_{n=0}^\infty (-t)^n/n!$

Transfer of Radiance

The differential equation $\frac{d}{dr}N + \alpha N = N^*$ models laser beam radiance (absorption and scattering out of the beam) in a medium like water, where r is the distance from the source.

27. Solve $\frac{d}{dr}N + 2N = 1, N(0) = 20$ by Laplace's method.
Ans: $N(r) = \frac{1}{2} + \frac{39}{2}\,\mathrm{e}^{-2\,r}$.
Hint: Obtain $\mathcal{L}(N(t)) = \frac{1+20\,s}{s(s+2)} = \frac{1}{2s} + \frac{39}{2(s+2)}$ using $\mathcal{L}(e^{at}) = \frac{1}{s-a}$ from the Forward Table page 601.

28. Solve $\frac{d}{dr}N + 2N = 1 - e^{-r}, N(0) = 25$ by any method.
Ans: $N(r) = \frac{1}{2} - \mathrm{e}^{-r} + \frac{51}{2}\,\mathrm{e}^{-2\,r}$.

Hint: A particular solution is $N_p = \frac{1}{2} - e^{-r}$. Superposition applies. See also Example 8.11 page 609.

Piecewise-Defined Functions

29. Define a piecewise continuous function $f(t)$ on $[-1,1]$ that agrees with $\frac{\sin(t)}{|t|}$ except at $t = 0$. Suggestion: use Taylor expansion $\sin(t) = t - t^3/6 + \cdots$ to define continuous functions f_1, f_2 on $-\infty < t < \infty$.

30. Explain in detail why $1/t$ is not piecewise continuous on $[-1,1]$. ■

31. Find $\mathcal{L}(f(t))$, given

$$f(t) = \begin{cases} 1 & 1 \le t < 2, \\ 0 & \text{otherwise.} \end{cases}$$

32. Find $\mathcal{L}(\mathbf{pulse}(t,a,b))$, given

$$\mathbf{pulse}(t,a,b) = \begin{cases} 1 & a \le t < b, \\ 0 & \text{otherwise.} \end{cases}$$

33. Define

$$f(t) = \begin{cases} 1 & 1 \le t < 2, \\ 2 & 3 \le t < 4, \\ 0 & \text{otherwise.} \end{cases}$$

Find the weights c_1, c_2 such that
$f(t) = c_1\,\mathbf{pulse}(t,1,2) +$
$c_2\,\mathbf{pulse}(t,3,4)$.

34. Let
$f(t) = \cos(t)\,\mathbf{pulse}(t,0,\pi) +$
$(\sin(t) - 1)\,\mathbf{pulse}(t,\pi,2\pi)$
Write f as a piecewise-defined function and graph it.

Piecewise Continuous Definition

Let $g(t)$ be zero for $t < 0$ and have on $t \ge 0$ at most finitely many points of discontinuity, at which finite right and left hand limits exist.

This definition is an alternative way to define *piecewise continuous*, crafted for Laplace theory.

35. Let t_1, t_2 be consecutive points of discontinuity of g. Define a function $g_1(t)$ continuous on $-\infty < t < \infty$ such that $g(t) = g_1(t)$ on $t_1 \le t \le t_2$.

The whole real line is the required domain of g_1, which must be defined using g itself and right and left hand limit values of g.

36. Let t_1, t_2, t_3 be consecutive points of discontinuity of g. Invent functions $g_1(t)$, $g_2(t)$ continuous on $-\infty < t < \infty$ such that $g(t) = g_1(t)$ on $t_1 \le t \le t_2$ and $g(t) = g_2(t)$ on $t_2 \le t \le t_3$.

37. Define g_1, g_2 as in Exercise 36 above. Compute the **jump** at $t = t_2$, $J(t_2) = g(t_2 + 0) - g(t_2 - 0)$, in terms of g_1, g_2.

38. Using the preceding steps, prove that g is piecewise continuous according to the definition given in the text.

8.2 Laplace Integral Table

The objective in developing Laplace integral Table 4 and Table 6 is to keep the table size small. Table manipulation rules in Table 5 page 601 effectively increase the table size manyfold, making it possible to solve typical differential equations from electrical and mechanical models. The combination of Laplace tables plus the table manipulation rules is called the **Laplace transform calculus**.

Table 4 is considered to be a table of minimum size. Table 6 adds a number of special-use entries.

Derivations are postponed to page 650. The theory of the generalized factorial function, the **gamma function** $\Gamma(x)$, is on page 603. The **Dirac impulse** $\delta(t)$ is defined in Section 8.6 page 644.

Table 4. Minimal Forward Laplace Integral Table with $\mathcal{L}$-notation

$\int_0^\infty (t^n)e^{-st}\,dt = \frac{n!}{s^{1+n}}$	$\mathcal{L}(t^n) = \frac{n!}{s^{1+n}}$
$\int_0^\infty (e^{at})e^{-st}\,dt = \frac{1}{s-a}$	$\mathcal{L}(e^{at}) = \frac{1}{s-a}$
$\int_0^\infty (\cos bt)e^{-st}\,dt = \frac{s}{s^2+b^2}$	$\mathcal{L}(\cos bt) = \frac{s}{s^2+b^2}$
$\int_0^\infty (\sin bt)e^{-st}\,dt = \frac{b}{s^2+b^2}$	$\mathcal{L}(\sin bt) = \frac{b}{s^2+b^2}$

Table 5. Minimal Forward and Backward Laplace Integral Tables

Forward Table	Backward Table
$\mathcal{L}(t^n) = \frac{n!}{s^{1+n}}$	$\frac{1}{s^{1+n}} = \mathcal{L}\left(\frac{t^n}{n!}\right)$
$\mathcal{L}(e^{at}) = \frac{1}{s-a}$	$\frac{1}{s-a} = \mathcal{L}\left(e^{at}\right)$
$\mathcal{L}(\cos bt) = \frac{s}{s^2+b^2}$	$\frac{s}{s^2+b^2} = \mathcal{L}\left(\cos bt\right)$
$\mathcal{L}(\sin bt) = \frac{b}{s^2+b^2}$	$\frac{1}{s^2+b^2} = \mathcal{L}\left(\frac{\sin bt}{b}\right)$

On first reading of LaPlace theory, learn Table 5 and back-burner the other tables. To fully understand Table 6 *below* requires hours of Laplace use.

Table 6. Extended Laplace Integral Table

Forward Table

$\mathcal{L}(t^n)$	$=$	$\dfrac{n!}{s^{1+n}}$
$\mathcal{L}(e^{at})$	$=$	$\dfrac{1}{s-a}$
$\mathcal{L}(\cos bt)$	$=$	$\dfrac{s}{s^2+b^2}$
$\mathcal{L}(\sin bt)$	$=$	$\dfrac{b}{s^2+b^2}$

Conventions and Shortcuts

Zero Assumed on $t < 0$. Laplace theory assumes a given $f(t)$ is zero for $t < 0$. Therefore, a given $f(t)$ in Laplace calculations can be formally replaced by $f(t)u(t)$, where u is the unit step defined by $u(t) = 1$ for $t \ge 0$, $u(t) = 0$ for $t < 0$.

Exponential Order. Unless specifically assumed otherwise, any $f(t)$ in Laplace theory is assumed to have exponential order so that $\mathcal{L}(f(t))$ exists. Exceptions: Function $f(t) = t^\alpha$ is not of exponential order for $\alpha < 0$, but $\mathcal{L}(f(t))$ exists; Dirac impulse $\delta(t)$ is in the extended table, but it is not of exponential order, because $\delta(t)$ is not a *function.*

Unit Step and Ramp. Table entry $f(t) = 1$ is called the **unit step** and entry $f(t) = t$ is called the **unit ramp**. Entry $f(t) = 1$ is equivalent to $u(t)$, whose graph shape resembles a staircase step. Entry $f(t) = t$ is equivalent to $tu(t)$, whose graph shape resembles a wheelchair ramp.

Step and Ramp Inputs. Digital design might refer to $y''(t) + y(t) = u(t)$ as an *oscillator with step input.* Similarly, $y''(t) + y(t) = tu(t)$ is an *oscillator with unit ramp input.*

Trigonometric Shortcut. Even function $f(t) = \cos(bt)$ $\mathcal{L}$-transforms to an odd fraction $F(s) = \frac{s}{s^2+b^2}$. Similarly, odd function $f(t) = \sin(bt)$ transforms to even fraction $\frac{b}{s^2+b^2}$.

Trig table entries $\cos, \sin$ change even-odd under $\mathcal{L}$-transformation.

Gamma Function

In mathematical physics, the **Gamma function** or the **generalized factorial function** is given by the identity

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt, \quad x > 0. \tag{1}$$

This function is tabulated and available in computer languages such as `Fortran, C, C++ and C#`. It is also available in computer algebra systems and numerical laboratories, such as `maple`, `matlab`, `mathematica`.

Fundamental Properties of $\Gamma(x)$

The generalized factorial function $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt$ has the following fundamental properties.

$\boxed{1}$ $\Gamma(1) = 1$

$\boxed{2}$ $\Gamma(1+x) = x\Gamma(x) \quad \text{for } x > 0.$

$\boxed{3}$ $\Gamma(1+n) = n! \quad \text{for integers} \quad n \geq 1.$

Details for relations $\boxed{1}$, $\boxed{2}$ and $\boxed{3}$: Start with $\int_0^\infty e^{-t}dt = 1$, which gives $\Gamma(1) = 1$, hence $\boxed{1}$. Use this identity and successively relation $\boxed{2}$ to obtain relation $\boxed{3}$. To prove identity $\boxed{2}$, integration by parts is applied, as follows:

$$\begin{aligned}
\Gamma(1+x) &= \int_0^\infty e^{-t}t^x dt && \text{Definition.}\\
&= -t^x e^{-t}\big|_{t=0}^{t=\infty} + \int_0^\infty e^{-t}xt^{x-1}dt && \text{Use } u = t^x,\ dv = e^{-t}dt.\\
&= x\int_0^\infty e^{-t}t^{x-1}dt && \text{Boundary terms are zero for } x > 0.\\
&= x\Gamma(x).
\end{aligned}$$

Examples and Methods

Example 8.4 (Forward Table)
Let $f(t) = t(t-5) - \sin 2t + e^{3t}$. Compute $\mathcal{L}(f(t))$ using the forward Laplace table and transform linearity properties.

Solution:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \mathcal{L}(t^2 - 5t - \sin 2t + e^{3t}) && \text{Expand } t(t-5).\\
&= \mathcal{L}(t^2) - 5\mathcal{L}(t) - \mathcal{L}(\sin 2t) + \mathcal{L}(e^{3t}) && \text{Linearity applied.}\\
&= \frac{2}{s^3} - \frac{5}{s^2} - \frac{2}{s^2+4} + \frac{1}{s-3} && \text{Forward Table.}
\end{aligned}$$

Example 8.5 (Backward Table)

Use the backward Laplace table plus transform linearity properties to solve for $f(t)$ in the equation

$$\mathcal{L}(f(t)) = \frac{s}{s^2+16} + \frac{2}{s-3} + \frac{s+1}{s^3}.$$

Solution:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \frac{s}{s^2+16} + 2\frac{1}{s-3} + \frac{1}{s^2} + \frac{1}{2}\frac{2}{s^3} && \text{Convert to table entries.}\\
&= \mathcal{L}(\cos 4t) + 2\mathcal{L}(e^{3t}) + \mathcal{L}(t) + \tfrac{1}{2}\mathcal{L}(t^2) && \text{Backward Laplace table.}\\
&= \mathcal{L}(\cos 4t + 2e^{3t} + t + \tfrac{1}{2}t^2) && \text{Linearity applied.}\\
f(t) &= \cos 4t + 2e^{3t} + t + \tfrac{1}{2}t^2 && \text{Lerch's cancellation law.}
\end{aligned}$$

Example 8.6 (Unit Step and Pulses)

Find $\mathcal{L}(f(t))$ in Figure 1.

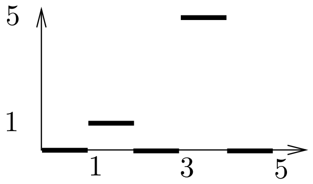


Figure 1. A piecewise defined function $f(t)$ on $0 \le t < \infty$: $f(t) = 0$ except for $1 \le t < 2$ and $3 \le t < 4$.

Solution: A **pulse** on $[a, b]$ is defined by

$$\textbf{pulse}(t, a, b) = u(t-a) - u(t-b) = \begin{cases} 1 & a \le t < b, \\ 0 & \text{otherwise.} \end{cases}$$

The formula for $f(t)$:

$$\begin{aligned}
f(t) &= \begin{cases} 1 & 1 \le t < 2, \\ 5 & 3 \le t < 4, \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} 1 & 1 \le t < 2, \\ 0 & \text{otherwise} \end{cases} + 5 \begin{cases} 1 & 3 \le t < 4, \\ 0 & \text{otherwise} \end{cases} \\
&= f_1(t) + 5f_2(t), \text{ where} \\
&\qquad f_1(t) = u(t-1) - u(t-2) = \textbf{pulse}(t, 1, 2), \\
&\qquad f_2(t) = u(t-3) - u(t-4) = \textbf{pulse}(t, 3, 4).
\end{aligned}$$

The extended Laplace table gives

$$\begin{aligned}
\mathcal{L}(f(t)) &= \mathcal{L}(f_1(t)) + 5\mathcal{L}(f_2(t)) && \text{Linearity.}\\
&= \mathcal{L}(u(t-1)) - \mathcal{L}(u(t-2)) + 5\mathcal{L}(f_2(t)) && \text{Substitute for } f_1.\\
&= \frac{e^{-s} - e^{-2s}}{s} + 5\mathcal{L}(f_2(t)) && \text{Extended table used.}\\
&= \frac{e^{-s} - e^{-2s} + 5e^{-3s} - 5e^{-4s}}{s} && \text{Similarly for } f_2.
\end{aligned}$$

Example 8.7 (Dirac Impulse)

A machine shop tool that repeatedly hammers a die is modeled by a Dirac impulse model $f(t) = \sum_{n=1}^N \delta(t-n)$. Verify the formula $\mathcal{L}(f(t)) = \sum_{n=1}^N e^{-ns}$.

Solution:

$$\begin{aligned} \mathcal{L}(f(t)) &= \mathcal{L}\left(\sum_{n=1}^N \delta(t-n)\right) & \\ &= \sum_{n=1}^N \mathcal{L}(\delta(t-n)) & \text{Linearity.} \\ &= \sum_{n=1}^N e^{-ns} & \text{Extended Laplace table.} \end{aligned}$$

Example 8.8 (Square wave)

A periodic camshaft force $f(t)$ applied to a mechanical system has the idealized graph shown in Figure 2. Verify formulas $f(t) = 1 + \textbf{sqw}(t)$ and $\mathcal{L}(f(t)) = \frac{1}{s}(1 + \tanh(s/2))$.

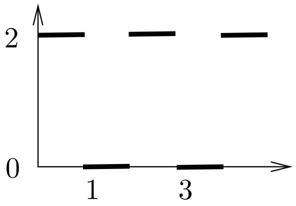


Figure 2. A periodic force $f(t)$ applied to a mechanical system.

Solution:

$$\begin{aligned} 1 + \textbf{sqw}(t) &= \begin{cases} 1+1 & 2n \le t < 2n+1, \ \ n = 0, 1, \ldots, \\ 1-1 & 2n+1 \le t < 2n+2, \ \ n = 0, 1, \ldots, \end{cases} \\ &= \begin{cases} 2 & 2n \le t < 2n+1, \ \ n = 0, 1, \ldots, \\ 0 & \text{otherwise,} \end{cases} \\ &= f(t). \end{aligned}$$

By the extended Laplace table, $\mathcal{L}(f(t)) = \mathcal{L}(1) + \mathcal{L}(\textbf{sqw}(t)) = \dfrac{1}{s} + \dfrac{\tanh(s/2)}{s}$.

Example 8.9 (Sawtooth wave)

Express the P-periodic sawtooth wave represented in Figure 3 as $f(t) = ct/P - c\,\textbf{floor}(t/P)$ and obtain the formula

$$\mathcal{L}(f(t)) = \frac{c}{Ps^2} - \frac{ce^{-Ps}}{s - se^{-Ps}}.$$

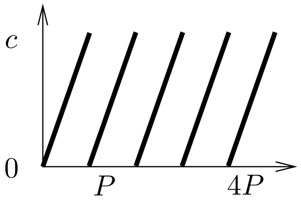


Figure 3. A P-periodic sawtooth wave $f(t)$ of height $c > 0$.

Solution: The representation originates from geometry, because the periodic function f can be viewed as derived from ct/P by subtracting the correct constant from each of intervals $[P, 2P]$, $[2P, 3P]$, etc.

The technique used to verify the identity is to define $g(t) = ct/P - c\,\mathsf{floor}(t/P)$ and then show that g is P-periodic and $f(t) = g(t)$ on $0 \le t < P$. Two P-periodic functions equal on the base interval $0 \le t < P$ have to be identical, hence the representation follows.

Periodicity: Let's show $g(u + P) - g(u) = 0$ for all u. Used below is the identity $\mathsf{floor}(1 + x) = 1 + \mathsf{floor}(x)$. Details: Let $x = u/P$, then

$$\begin{aligned} g(u+P) - g(u) &= c\tfrac{u+P}{P} - c\,\mathsf{floor}\left(\tfrac{u+P}{P}\right) - g(u) \\ &= cx + c - c\,\mathsf{floor}(1+x) - cx + c\,\mathsf{floor}(x) \\ &= 0. \end{aligned}$$

Base interval equality: On $0 \le t < P$, define $x = t/P$ so that $0 \le x < 1$. Then $\mathsf{floor}(x) = 0$ and $f(t) = ct/P = cx$. Compute $g(t) = ct/P - c\,\mathsf{floor}(t/P) = cx - c\,\mathsf{floor}(x) = cx = f(t)$.

Laplace Calculation:

$$\begin{aligned} \mathcal{L}(f(t)) &= \frac{c}{P}\mathcal{L}(t) - c\mathcal{L}(\mathsf{floor}(t/P)) && \text{Linearity.} \\ &= \frac{c}{Ps^2} - \frac{ce^{-Ps}}{s - se^{-Ps}} && \text{Basic and extended table applied.} \end{aligned}$$

Example 8.10 (Triangular wave)

Express the triangular wave f of Figure 4 in terms of the square wave **sqw** and obtain $\mathcal{L}(f(t)) = \dfrac{5}{\pi s^2}\tanh(\pi s/2)$.

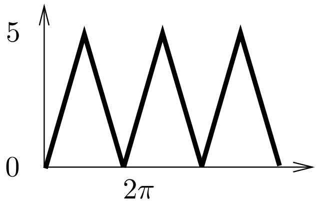


Figure 4. A 2π-periodic triangular wave $f(t)$ of height 5.

Solution: The representation of f in terms of **sqw** is $f(t) = 5\int_0^{t/\pi} \mathsf{sqw}(x)dx$.

Details: A 2-periodic triangular wave of height 1 is obtained by integrating the square wave of period 2. A wave of height c and period 2 is given by $c\,\mathsf{trw}(t) = c\int_0^t \mathsf{sqw}(x)dx$. Then $f(t) = c\,\mathsf{trw}(2t/P) = c\int_0^{2t/P} \mathsf{sqw}(x)dx$ where $c = 5$ and $P = 2\pi$.

Laplace calculation: Use the extended Laplace table as follows.

$$\mathcal{L}(f(t)) = \frac{5}{\pi}\mathcal{L}(\pi\,\mathsf{trw}(t/\pi)) = \frac{5}{\pi s^2}\tanh(\pi s/2).$$

Exercises 8.2

Laplace Transform Forward Table

Using the basic Laplace table and linearity properties of the transform, compute $\mathcal{L}(f(t))$. Do not use the direct Laplace transform!

1. $\mathcal{L}(2t)$

2. $\mathcal{L}(4t)$

3. $\mathcal{L}(1+2t+t^2)$

4. $\mathcal{L}(t^2-3t+10)$

5. $\mathcal{L}(\sin 2t)$

6. $\mathcal{L}(\cos 2t)$

7. $\mathcal{L}(e^{2t})$

8. $\mathcal{L}(e^{-2t})$

9. $\mathcal{L}(t+\sin 2t)$

10. $\mathcal{L}(t-\cos 2t)$

11. $\mathcal{L}(t+e^{2t})$

12. $\mathcal{L}(t-3e^{-2t})$

13. $\mathcal{L}((t+1)^2)$

14. $\mathcal{L}((t+2)^2)$

15. $\mathcal{L}(t(t+1))$

16. $\mathcal{L}((t+1)(t+2))$

17. $\mathcal{L}(\sum_{n=0}^{10} t^n/n!)$

18. $\mathcal{L}(\sum_{n=0}^{10} t^{n+1}/n!)$

19. $\mathcal{L}(\sum_{n=1}^{10} \sin nt)$

20. $\mathcal{L}(\sum_{n=0}^{10} \cos nt)$

Laplace Backward Table

Solve the given equation for the function $f(t)$. Use the basic table and linearity properties of the Laplace transform.

21. $\mathcal{L}(f(t)) = s^{-2}$

22. $\mathcal{L}(f(t)) = 4s^{-2}$

23. $\mathcal{L}(f(t)) = 1/s + 2/s^2 + 3/s^3$

24. $\mathcal{L}(f(t)) = 1/s^3 + 1/s$

25. $\mathcal{L}(f(t)) = 2/(s^2+4)$

26. $\mathcal{L}(f(t)) = s/(s^2+4)$

27. $\mathcal{L}(f(t)) = 1/(s-3)$

28. $\mathcal{L}(f(t)) = 1/(s+3)$

29. $\mathcal{L}(f(t)) = 1/s + s/(s^2+4)$

30. $\mathcal{L}(f(t)) = 2/s - 2/(s^2+4)$

31. $\mathcal{L}(f(t)) = 1/s + 1/(s-3)$

32. $\mathcal{L}(f(t)) = 1/s - 3/(s-2)$

33. $\mathcal{L}(f(t)) = (2+s)^2/s^3$

34. $\mathcal{L}(f(t)) = (s+1)/s^2$

35. $\mathcal{L}(f(t)) = s(1/s^2 + 2/s^3)$

36. $\mathcal{L}(f(t)) = (s+1)(s-1)/s^3$

37. $\mathcal{L}(f(t)) = \sum_{n=0}^{10} n!/s^{1+n}$

38. $\mathcal{L}(f(t)) = \sum_{n=0}^{10} n!/s^{2+n}$

39. $\mathcal{L}(f(t)) = \sum_{n=1}^{10} \frac{n}{s^2+n^2}$

40. $\mathcal{L}(f(t)) = \sum_{n=0}^{10} \frac{s}{s^2+n^2}$

Laplace Table Extension

Compute the indicated Laplace integral using the extended Laplace table, page 602.

41. $\mathcal{L}(u(t-2)+2u(t))$

42. $\mathcal{L}(u(t-3)+4u(t))$

43. $\mathcal{L}(u(t-\pi)(u(t)+u(t-1)))$

44. $\mathcal{L}(u(t-2\pi)+3u(t-1)u(t-2))$

45. $\mathcal{L}(\delta(t-2))$

46. $\mathcal{L}(5\delta(t-\pi))$

47. $\mathcal{L}(\delta(t-1)+2\delta(t-2))$

48. $\mathcal{L}(\delta(t-2)(5+u(t-1)))$

49. $\mathcal{L}(\textsf{\textbf{floor}}(3t))$

50. $\mathcal{L}(\textsf{\textbf{floor}}(2t))$

51. $\mathcal{L}(5\,\textsf{\textbf{sqw}}(3t))$

52. $\mathcal{L}(3\,\textsf{\textbf{sqw}}(t/4))$

53. $\mathcal{L}(4\,\textsf{\textbf{trw}}(2t))$

54. $\mathcal{L}(5\,\textsf{\textbf{trw}}(t/2))$

55. $\mathcal{L}(t + t^{-3/2} + t^{-1/2})$

56. $\mathcal{L}(t^3 + t^{-3/2} + 2t^{-1/2})$

Inverse Laplace, Extended Table

Solve the given equation for $f(t)$, using the extended Laplace integral table.

57. $\mathcal{L}(f(t)) = e^{-s}/s$

58. $\mathcal{L}(f(t)) = 5e^{-2s}/s$

59. $\mathcal{L}(f(t)) = e^{-2s}$

60. $\mathcal{L}(f(t)) = 5e^{-3s}$

61. $\mathcal{L}(f(t)) = \dfrac{e^{-s/3}}{s(1 - e^{-s/3})}$

62. $\mathcal{L}(f(t)) = \dfrac{e{-}2s}{s(1 - e^{-2s})}$

63. $\mathcal{L}(f(t)) = \dfrac{4\tanh(s)}{s}$

64. $\mathcal{L}(f(t)) = \dfrac{5\tanh(3s)}{2s}$

65. $\mathcal{L}(f(t)) = \dfrac{4\tanh(s)}{3s^2}$

66. $\mathcal{L}(f(t)) = \dfrac{5\tanh(2s)}{11s^2}$

67. $\mathcal{L}(f(t)) = \dfrac{1}{\sqrt{s}}$

68. $\mathcal{L}(f(t)) = \dfrac{1}{\sqrt{s^3}}$

8.3 Laplace Transform Rules

In Table 7, the basic table manipulation rules are summarized. Full statements and proofs of the rules appear in section 8.5, page 637.

The rules are applied here to several key examples using the 8 rules. Partial fraction expansions including Heaviside's coverup method will be delayed to the section on Heaviside's Method page 8.4.

Table 7. Laplace transform Rules

$\mathcal{L}(f(t)+g(t)) = \mathcal{L}(f(t)) + \mathcal{L}(g(t))$ $\mathcal{L}(cf(t)) = c\mathcal{L}(f(t))$	Linearity. The Laplace of a sum is the sum of the Laplaces. Constants move through the $\mathcal{L}$-symbol.
$\mathcal{L}(y'(t)) = s\mathcal{L}(y(t)) - y(0)$	The t-derivative or parts rule. Derivatives $\mathcal{L}(y')$ are replaced in transformed equations.
$\mathcal{L}\left(\int_0^t g(x)dx\right) = \frac{1}{s}\mathcal{L}(g(t))$	Forward t-integral rule.
$\frac{1}{s}\mathcal{L}(g(t)) = \mathcal{L}\left(\int_0^t g(x)dx\right)$	Backward t-integral rule.
$\mathcal{L}(tf(t)) = -\frac{d}{ds}\mathcal{L}(f(t))$	Forward s-differentiation rule. Each erased t-factor inserts $-\frac{d}{ds}$ in front of $\mathcal{L}$.
$\frac{d}{ds}\mathcal{L}(f(t)) = \mathcal{L}((-t)f(t))$	Backward s-differentiation rule.
$\mathcal{L}(e^{at}f(t)) = \mathcal{L}(f(t))\vert_{s\to(s-a)}$ $\mathcal{L}(f(t))\vert_{s\to(s-a)} = \mathcal{L}(e^{at}f(t))$	Forward First Shifting rule. Backward First Shifting rule. Multiplying f by e^{at} replaces s by $s-a$.
$\mathcal{L}(g(t)u(t-a)) = e^{-as}\mathcal{L}(g(t+a))$, $e^{-as}\mathcal{L}(f(t)) = \mathcal{L}(f(t-a)u(t-a))$	Forward Second Shifting rule. Backward Second Shifting rule.
$\mathcal{L}(f(t)) = \dfrac{\int_0^P f(t)e^{-st}dt}{1-e^{-Ps}}$	Rule for P-periodic functions. Assumed: $f(t+P) = f(t)$.
$\mathcal{L}(f(t))\mathcal{L}(g(t)) = \mathcal{L}((f*g)(t))$	Convolution rule. Define $(f*g)(t) = \int_0^t f(x)g(t-x)dx$.

Examples and Methods

Example 8.11 (Rutherford Decay)
Solve the radioactive chain decay problem $x' + 2x = -e^{-2t}$, $x(0) = 10$ by Laplace's method.

Solution: The solution is $x = e^{-2t} - te^{-2t}$. The details:

$\mathcal{L}(x' + 2x) = \mathcal{L}(-e^{-2t})$	$\boxed{1}$ Apply $\mathcal{L}$ across the equation.
$\mathcal{L}(x') + 2\mathcal{L}(x) = -\mathcal{L}(e^{-2t})$	Linearity of $\mathcal{L}$.
$s\mathcal{L}(x) - x(0) + 2\mathcal{L}(x) = -\mathcal{L}(e^{-2t})$	Parts rule.
$(s+2)\mathcal{L}(x) = 10 - \mathcal{L}(e^{-2t})$	Use $x(0) = 10$. Collect left on $\mathcal{L}(x)$.
$(s+2)\mathcal{L}(x) = 10 - \frac{1}{s+2}$	Forward Laplace table: $\mathcal{L}(e^{-2t}) = \frac{1}{s+2}$.

$\mathcal{L}(x) = \frac{10}{s+2} - \frac{1}{(s+2)^2}$	Divide to isolate $\mathcal{L}(x)$ left.
$\mathcal{L}(x) = \frac{10}{s+2} - \frac{1}{s^2}\big\|_{s\to s+2}$	$\boxed{2}$ First shifting rule preparation.
$\mathcal{L}(x) = \mathcal{L}(e^{-2t}) - \mathcal{L}(t)\|_{s\to s+2}$	Backward Laplace table: $\frac{1}{s-a} = \mathcal{L}(e^{at})$, $\frac{1}{s^2} = \mathcal{L}(t)$.
$\mathcal{L}(x) = \mathcal{L}(e^{-2t}) - \mathcal{L}(te^{-2t})$	$\boxed{3}$ Backwards first shifting theorem.
$\mathcal{L}(x) = \mathcal{L}(e^{-2t} - te^{-2t})$	Linearity.
$x = e^{-2t} - te^{-2t}$	Lerch's theorem: cancel $\mathcal{L}$ on each side.

Laplace's method: Multiply across by $e^{-st}dt$, then integrate across $t = 0$ to $t = \infty$. It is the same as applying $\mathcal{L}$ across the equation.

The **details** used algebraic steps and Laplace rules to obtain $\mathcal{L}(x(t))$ on the left and $\mathcal{L}$(some t-expression) on the right. In the last step Lerch's theorem applies to cancel $\mathcal{L}$ on each side, which isolates the solution $x(t) =$ some t-expression.

$\boxed{1}$: Think of $\mathcal{L}$ as a matrix and Laplace's method as matrix multiply.

$\boxed{2}$: Fraction $\frac{1}{(s+2)^2}$ is $\frac{1}{w^2}$ using substitution $w = s + 2$. Mentally replace w by s and search the Backward Table for a matching entry. The intuition comes from u-substitution in calculus, but because u is the unit step function in Laplace theory, symbol **w is used instead of u** in substitution examples.

$\boxed{3}$: The backwards first shifting theorem in words: Remove $|_{s\to s-a}$ by inserting exponential e^{at} inside the scope of $\mathcal{L}$.

Example 8.12 (Harmonic oscillator)

Solve the initial value problem $x'' + x = 0$, $x(0) = 0$, $x'(0) = 1$ by Laplace's method.

Solution: The solution is $x(t) = \sin t$. The details:

$\mathcal{L}(x'') + \mathcal{L}(x) = \mathcal{L}(0)$	Apply $\mathcal{L}$ across the equation.
$s\mathcal{L}(x') - x'(0) + \mathcal{L}(x) = 0$	The t-derivative or parts rule.
$s[s\mathcal{L}(x) - x(0)] - x'(0) + \mathcal{L}(x) = 0$	Again the parts rule.
$(s^2 + 1)\mathcal{L}(x) = 1$	Use $x(0) = 0$, $x'(0) = 1$.
$\mathcal{L}(x) = \dfrac{1}{s^2 + 1}$	Divide to isolate $\mathcal{L}(x(t))$ left.
$= \mathcal{L}(\sin t)$	Forward Laplace table.
$x(t) = \sin t$	Lerch's cancellation law.

Example 8.13 (Forward Table First Shifting Rule)

Show the steps for the identity $\mathcal{L}(t^2\, e^{-3t}) = \dfrac{2}{(s + 3)^3}$.

Solution:

$\mathcal{L}(t^2 e^{-3t}) = \mathcal{L}(t^2)\big\|_{s\to s-(-3)}$	First shifting rule.
$= \left(\dfrac{2}{s^{2+1}}\right)\Big\|_{s\to s-(-3)}$	Forward Laplace table.

$$= \frac{2}{(s+3)^3} \qquad \text{Identity verified.}$$

Example 8.14 (Backward Table First Shifting Rule I)

Solve for $f(t)$ in the equation $\mathcal{L}(f(t) = \dfrac{s+7}{s^2+4s+8}$.

Solution: The answer is $f(t) = e^{-2t}(\cos 2t + \frac{5}{2}\sin 2t)$. The details:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \frac{s+7}{(s+2)^2+4} && \text{Complete the square.}\\
&= \frac{w+5}{w^2+4} && \text{Replace } s+2 \text{ by } w.\\
&= \frac{w}{w^2+4} + \frac{5}{2}\frac{2}{w^2+4} && \text{Split into table entries.}\\
&= \left.\frac{s}{s^2+4} + \frac{5}{2}\frac{2}{s^2+4}\right|_{s\to w=s+2} && \text{Shifting rule preparation.}\\
&= \mathcal{L}\left(\cos 2t + \tfrac{5}{2}\sin 2t\right)\big|_{s\to w=s+2} && \text{Basic Laplace table.}\\
&= \mathcal{L}(e^{-2t}(\cos 2t + \tfrac{5}{2}\sin 2t)) && \text{First shifting rule.}\\
f(t) &= e^{-2t}(\cos 2t + \tfrac{5}{2}\sin 2t) && \text{Lerch's cancellation law.}
\end{aligned}$$

Example 8.15 (Backward Table First Shifting Rule II)

Solve the equation $\mathcal{L}(f(t)) = \dfrac{s+2}{2^2+2s+2}$ for $f(t)$.

Solution: The answer is $f(t) = e^{-t}\cos t + e^{-t}\sin t$. The details:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \frac{s+2}{s^2+2s+2} && \text{Signal for this method: the denominator has complex roots.}\\
&= \frac{s+2}{(s+1)^2+1} && \text{Complete the square, denominator.}\\
&= \frac{w+1}{w^2+1} && \text{Substitute } w \text{ for } s+1.\\
&= \frac{w}{w^2+1} + \frac{1}{w^2+1} && \text{Split into Laplace table entries.}\\
&= (\mathcal{L}(\cos t) + \mathcal{L}(\sin t))|_{s\to w=s+1} && \text{Basic Laplace table.}\\
&= \mathcal{L}(e^{-t}\cos t) + \mathcal{L}(e^{-t}\sin t) && \text{First shifting rule.}\\
f(t) &= e^{-t}\cos t + e^{-t}\sin t && \text{Lerch's cancellation law.}
\end{aligned}$$

Example 8.16 (Damped oscillator)

Solve by Laplace's method the initial value problem $x'' + 2x' + 2x = 0$, $x(0) = 1$, $x'(0) = -1$.

Solution: The solution is $x(t) = e^{-t}\cos t$. The details:

$\mathcal{L}(x'') + 2\mathcal{L}(x') + 2\mathcal{L}(x) = \mathcal{L}(0)$	Apply $\mathcal{L}$ across the equation.
$s\mathcal{L}(x') - x'(0) + 2\mathcal{L}(x') + 2\mathcal{L}(x) = 0$	The t-derivative rule on x'.
$s[s\mathcal{L}(x) - x(0)] - x'(0)$ $+2[\mathcal{L}(x) - x(0)] + 2\mathcal{L}(x) = 0$	The t-derivative rule on x.
$(s^2 + 2s + 2)\mathcal{L}(x) = 1 + s$	Use $x(0) = 1$, $x'(0) = -1$.
$\mathcal{L}(x) = \dfrac{s+1}{s^2+2s+2}$	Divide to isolate $\mathcal{L}(x)$.
$= \dfrac{s+1}{(s+1)^2+1}$	Complete the square.
$= \left.\dfrac{w}{w^2+1}\right\vert_{w=s+1}$	Replace $s+1$ by w.
$= \mathcal{L}(\cos t)\vert_{s\to w=s+1}$	Backward table: $\frac{s}{s^2+1} = \mathcal{L}(\cos t)$.
$= \mathcal{L}(e^{-t}\cos t)$	First shifting rule.
$x(t) = e^{-t}\cos t$	Lerch's cancellation law.

Example 8.17 (Forward Table s-Differentiation)

Show the steps for the identity $\mathcal{L}(t^2\, e^{5t}) = \dfrac{2}{(s-5)^3}$.

Solution:

$\mathcal{L}(t^2 e^{5t}) = \left(-\dfrac{d}{ds}\right)\left(-\dfrac{d}{ds}\right)\mathcal{L}(e^{5t})$	Apply s-differentiation.
$= (-1)^2 \dfrac{d}{ds}\dfrac{d}{ds}\left(\dfrac{1}{s-5}\right)$	Basic Laplace table.
$= \dfrac{d}{ds}\left(\dfrac{-1}{(s-5)^2}\right)$	Calculus power rule $(u^n)' = nu^{n-1}u'$.
$= \dfrac{2}{(s-5)^3}$	Identity verified.

Example 8.18 (Backward Table s-Differentiation)

Solve the equation $\mathcal{L}(f(t)) = \dfrac{2s}{(s^2+1)^2}$ for $f(t)$.

Solution: The solution is $f(t) = t\sin t$. The details:

$\mathcal{L}(f(t)) = \dfrac{2s}{(s^2+1)^2}$	
$= -\dfrac{d}{ds}\left(\dfrac{1}{s^2+1}\right)$	Calculus power rule $(u^n)' = nu^{n-1}u'$.
$= -\dfrac{d}{ds}\left(\mathcal{L}(\sin t)\right)$	Basic Laplace table.

$$\begin{aligned} &= \mathcal{L}(t\sin t) && \text{Apply the } s\text{-differentiation rule.} \\ f(t) &= t\sin t && \text{Lerch's cancellation law.} \end{aligned}$$

Example 8.19 (Forward Table Second shifting rule)

Show the steps for the identity $\mathcal{L}(\sin(t)\,u(t-\pi)) = \dfrac{-e^{-\pi s}}{s^2+1}$, where $u(t)$ is the **unit step function**: $u(t) = 1$ for $t \ge 0$, $u(t) = 0$ otherwise.

Solution: The second shifting rule is applied as follows, where LHS and RHS abbreviate the left and right hand side.

$$\begin{aligned} \text{LHS} &= \mathcal{L}(\sin t\, u(t-\pi)) && \text{Left side of the identity.} \\ &= \mathcal{L}(g(t)u(t-a)) && \text{Choose } g(t) = \sin t,\ a = \pi. \\ &= e^{-as}\mathcal{L}(g(t+a)) && \text{Second form, second shifting theorem.} \\ &= e^{-\pi s}\mathcal{L}(\sin(t+\pi)) && \boxed{1}\ \text{Substitute } a = \pi,\ g(t) = \sin(t). \\ &= e^{-\pi s}\mathcal{L}(-\sin t) && \text{Trig rules } \sin(a+b) = \sin a\cos b + \sin b\cos a \text{ and } \sin\pi = 0,\ \cos\pi = -1. \\ &= e^{-\pi s}\frac{-1}{s^2+1} && \text{Forward Laplace table.} \\ &= \text{RHS} && \text{Identity verified.} \end{aligned}$$

$\boxed{1}$: Easy for some readers, difficult for others. How did we change symbol $g(t+a)$ into $\sin(t+\pi)$? For $g(t) = \sin t$, the replacement process $g \to \sin$ and $a \to \pi$ can be written as $g(t+a) = g(x)|_{x=t+a} = \sin(x)|_{x=t+\pi} = \sin(t+\pi)$.

Example 8.20 (Backward Table Second Shifting Rule)

Solve the equation $\mathcal{L}(f(t)) = e^{-3s}\dfrac{s+1}{s^2+2s+2}$ for $f(t)$.

Solution: The answer is $f(t) = e^{3-t}\cos(t-3)$ for $t \ge 3$, $f(t) = 0$ otherwise. The details:

$$\begin{aligned} \mathcal{L}(f(t)) &= e^{-3s}\,\frac{s+1}{(s+1)^2+1} && \text{Complete the square.} \\ &= e^{-3w+3}\,\frac{w}{w^2+1} && \text{Let } w = s+1, \text{ like a calculus } u\text{-substitution.} \\ &= e^{-3w+3}\,\left(\mathcal{L}(\cos t)\right)|_{s\to w} && \text{Backward table: } \tfrac{s}{s^2+1} = \mathcal{L}(\cos t) \\ &= e^3\left(e^{-3s}\mathcal{L}(\cos t)\right)\Big|_{s\to w} && \text{Regroup factor } e^{-3w}. \\ &= e^3\,\left(\mathcal{L}(\cos(t-3)u(t-3))\right)|_{s\to w=s+1} && \text{Second shifting rule, 1st form.} \\ &= e^3\mathcal{L}(\cos(t-3)u(t-3)e^{-t}) && \text{First shifting rule.} \\ f(t) &= e^{3-t}\cos(t-3)u(t-3) && \text{Lerch's cancellation law.} \end{aligned}$$

Example 8.21 (Trigonometric formulas)
Show the steps used to obtain these Laplace identities:

(a) $\mathcal{L}(t\cos at) = \dfrac{s^2-a^2}{(s^2+a^2)^2}$ **(c)** $\mathcal{L}(t^2\cos at) = \dfrac{2(s^3-3sa^2)}{(s^2+a^2)^3}$

(b) $\mathcal{L}(t\sin at) = \dfrac{2sa}{(s^2+a^2)^2}$ **(d)** $\mathcal{L}(t^2\sin at) = \dfrac{6s^2a-a^3}{(s^2+a^2)^3}$

Solution: The details for **(a)**:

$$\begin{aligned}
\mathcal{L}(t\cos at) &= -(d/ds)\mathcal{L}(\cos at) && \text{Use } s\text{-differentiation.}\\
&= -\frac{d}{ds}\left(\frac{s}{s^2+a^2}\right) && \text{Basic Laplace table.}\\
&= \frac{s^2-a^2}{(s^2+a^2)^2} && \text{Calculus quotient rule.}
\end{aligned}$$

The details for **(c)**:

$$\begin{aligned}
\mathcal{L}(t^2\cos at) &= -(d/ds)\mathcal{L}((-t)\cos at) && \text{Use } s\text{-differentiation.}\\
&= \frac{d}{ds}\left(-\frac{s^2-a^2}{(s^2+a^2)^2}\right) && \text{Result of } \textbf{(a)}.\\
&= \frac{2s^3-6sa^2)}{(s^2+a^2)^3} && \text{Calculus quotient rule.}
\end{aligned}$$

The similar details for **(b)** and **(d)** are left as exercises.

Example 8.22 (Exponential Formulas)
Show the steps used to obtain these Laplace identities:

(a) $\mathcal{L}(e^{at}\cos bt) = \dfrac{s-a}{(s-a)^2+b^2}$ **(c)** $\mathcal{L}(te^{at}\cos bt) = \dfrac{(s-a)^2-b^2}{((s-a)^2+b^2)^2}$

(b) $\mathcal{L}(e^{at}\sin bt) = \dfrac{b}{(s-a)^2+b^2}$ **(d)** $\mathcal{L}(te^{at}\sin bt) = \dfrac{2b(s-a)}{((s-a)^2+b^2)^2}$

Solution: Details for **(a)**:

$$\begin{aligned}
\mathcal{L}(e^{at}\cos bt) &= \mathcal{L}(\cos bt)|_{s\to s-a} && \text{First shifting rule.}\\
&= \left.\left(\frac{s}{s^2+b^2}\right)\right|_{s\to s-a} && \text{Basic Laplace table.}\\
&= \frac{s-a}{(s-a)^2+b^2} && \text{Verified } \textbf{(a)}.
\end{aligned}$$

Details for **(c)**:

$$\begin{aligned}
\mathcal{L}(te^{at}\cos bt) &= \mathcal{L}(t\cos bt)|_{s\to s-a} && \text{First shifting rule.}\\
&= \left.\left(-\frac{d}{ds}\mathcal{L}(\cos bt)\right)\right|_{s\to s-a} && \text{Apply } s\text{-differentiation.}\\
&= \left.\left(-\frac{d}{ds}\left(\frac{s}{s^2+b^2}\right)\right)\right|_{s\to s-a} && \text{Basic Laplace table.}
\end{aligned}$$

$$= \left.\left(\frac{s^2-b^2}{(s^2+b^2)^2}\right)\right|_{s\to s-a} \qquad \text{Calculus quotient rule.}$$

$$= \frac{(s-a)^2-b^2}{((s-a)^2+b^2)^2} \qquad \text{Verified } \textbf{(c)}.$$

Left as exercises are **(b)** and **(d)**.

Example 8.23 (Hyperbolic Functions)

Establish these Laplace transform facts about $\cosh u = (e^u + e^{-u})/2$ and $\sinh u = (e^u - e^{-u})/2$.

(a) $\mathcal{L}(\cosh at) = \dfrac{s}{s^2-a^2}$

(c) $\mathcal{L}(t\cosh at) = \dfrac{s^2+a^2}{(s^2-a^2)^2}$

(b) $\mathcal{L}(\sinh at) = \dfrac{a}{s^2-a^2}$

(d) $\mathcal{L}(t\sinh at) = \dfrac{2as}{(s^2-a^2)^2}$

Solution: The details for **(a)**:

$$\mathcal{L}(\cosh at) = \tfrac{1}{2}(\mathcal{L}(e^{at}) + \mathcal{L}(e^{-at})) \qquad \text{Definition plus linearity of } \mathcal{L}.$$

$$= \frac{1}{2}\left(\frac{1}{s-a} + \frac{1}{s+a}\right) \qquad \text{Basic Laplace table.}$$

$$= \frac{s}{s^2-a^2} \qquad \text{Identity } \textbf{(a)} \text{ verified.}$$

The details for **(d)**:

$$\mathcal{L}(t\sinh at) = -\frac{d}{ds}\left(\frac{a}{s^2-a^2}\right) \qquad \text{Apply the } s\text{-differentiation rule.}$$

$$= \frac{a(2s)}{(s^2-a^2)^2} \qquad \text{Calculus power rule; } \textbf{(d)} \text{ verified.}$$

Left as exercises are **(b)** and **(c)**.

Example 8.24 (Rectified sine wave)

Compute the Laplace transform of the rectified sine wave $f(t) = |\sin\omega t|$ and show that it can be expressed in the form

$$\mathcal{L}(|\sin\omega t|) = \frac{\omega\ \coth\left(\frac{\pi s}{2\omega}\right)}{s^2+\omega^2}.$$

Solution: The periodic function formula will be applied with period $P = 2\pi/\omega$. The calculation reduces to the evaluation of $J = \int_0^P f(t)e^{-st}dt$. Because $\sin\omega t \le 0$ on $\pi/\omega \le t \le 2\pi/\omega$, integral J can be written as $J = J_1 + J_2$, where

$$J_1 = \int_0^{\pi/\omega} \sin\omega t\, e^{-st}dt, \quad J_2 = \int_{\pi/\omega}^{2\pi/\omega} -\sin\omega t\, e^{-st}dt.$$

Integral tables give the result

$$\int \sin\omega t\, e^{-st}\, dt = -\frac{\omega e^{-st}\cos(\omega t)}{s^2+\omega^2} - \frac{se^{-st}\sin(\omega t)}{s^2+\omega^2}.$$

Then

$$J_1 = \frac{\omega(e^{-\pi*s/\omega}+1)}{s^2+\omega^2}, \quad J_2 = \frac{\omega(e^{-2\pi s/\omega}+e^{-\pi s/\omega})}{s^2+\omega^2},$$

$$J = \frac{\omega(e^{-\pi s/\omega}+1)^2}{s^2+\omega^2}.$$

The remaining challenge is to write the answer for $\mathcal{L}(f(t))$ in terms of $\coth(u) = \frac{\cosh(u)}{\sinh(u)}$ where $\cosh(u) = \frac{1}{2}e^u + \frac{1}{2}e^{-u}$ and $\sinh(u) = \frac{1}{2}e^u - \frac{1}{2}e^{-u}$. The details:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \frac{J}{1-e^{-Ps}} && \text{Periodic function formula.}\\
&= \frac{J}{(1-e^{-Ps/2})(1+e^{-Ps/2})} && \text{Apply } 1-x^2=(1-x)(1+x) \text{ where } x=e^{-Ps/2}.\\
&= \frac{\omega(1+e^{-Ps/2})}{(1-e^{-Ps/2})(s^2+\omega^2)} && \text{Cancel factor } 1+e^{-Ps/2}.\\
&= \frac{e^{Ps/4}+e^{-Ps/4}}{e^{Ps/4}-e^{-Ps/4}}\frac{\omega}{s^2+\omega^2} && \text{Factor out } e^{-Ps/4}\text{, then cancel.}\\
&= \frac{2\cosh(Ps/4)}{2\sinh(Ps/4)}\frac{\omega}{s^2+\omega^2} && \text{Apply cosh, sinh identities.}\\
&= \frac{\omega\coth(Ps/4)}{s^2+\omega^2} && \text{Use } \frac{\cosh u}{\sinh u} = \coth u.\\
&= \frac{\omega\coth\left(\frac{\pi s}{2\omega}\right)}{s^2+\omega^2} && \text{Identity verified.}
\end{aligned}$$

Example 8.25 (Half–wave Rectification)
Determine the Laplace transform of the half–wave rectification $g(t)$ of $\sin\omega t$, in which the negative cycles of $\sin\omega t$ have been replaced by zero to define $g(t)$. Show in particular that

$$\mathcal{L}(g(t)) = \frac{1}{2}\frac{\omega}{s^2+\omega^2}\left(1+\coth\left(\frac{\pi s}{2\omega}\right)\right)$$

Solution: The half–wave rectification of $\sin\omega t$ is $g(t) = (\sin\omega t + |\sin\omega t|)/2$. The Forward Table plus the result of Example 8.24 gives

$$\begin{aligned}
\mathcal{L}(2g(t)) &= \mathcal{L}(\sin\omega t) + \mathcal{L}(|\sin\omega t|)\\
&= \frac{\omega}{s^2+\omega^2} + \frac{\omega\cosh(\pi s/(2\omega))}{s^2+\omega^2}\\
&= \frac{\omega}{s^2+\omega^2}(1+\cosh(\pi s/(2\omega)))
\end{aligned}$$

Dividing by 2 produces the identity.

Exercises 8.3

First Order Linear DE

Display the Laplace method details which verify the supplied answer.

The first two exercises use forward and backward Laplace tables plus the first shifting theorems. The others require a calculus background in partial fractions.

1. $x' + x = e^{-t}$, $x(0) = 1$; $x(t) = (1+t)e^{-t}$.

2. $x' + 2x = -e^{-2t}$, $x(0) = 1$; $x(t) = (1-t)e^{-2t}$.

3. $x' + x = 1$, $x(0) = 1$; $x(t) = 1$.

4. $x' + 4x = 4$, $x(0) = 1$; $x(t) = 1$.

5. $x' + x = t$, $x(0) = -1$; $x(t) = t - 1$.

6. $x' + x = t$, $x(0) = 1$; $x(t) = t - 1 + 2e^{-t}$.

Second Order Linear DE

Display the Laplace method details which verify the supplied answer.

The first 4 exercises require only forward and backward Laplace tables and the first shifting theorems. The others require methods in partial fractions beyond a calculus background.

7. $x'' + x = 0$, $x(0) = 1$, $x'(0) = 1$; $x(t) = \cos t + \sin t$.

8. $x'' + x = 0$, $x(0) = 1$, $x'(0) = 2$; $x(t) = \cos t + 2\sin t$.

9. $x'' + 2x' + x = 0$, $x(0) = 0$, $x'(0) = 1$; $x(t) = te^{-t}$.

10. $x'' + 2x' + x = 0$, $x(0) = 1$, $x'(0) = -1$; $x(t) = e^{-t}$.

11. $x'' + 3x' + 2x = 0$, $x(0) = 1$, $x'(0) = -1$; $x(t) = e^{-t}$.

12. $x'' + 3x' + 2x = 0$, $x(0) = 1$, $x'(0) = -2$; $x(t) = e^{-2t}$.

13. $x'' + 3x' = 0$, $x(0) = 5$, $x'(0) = 0$; $x(t) = 5$.

14. $x'' + 3x' = 0$, $x(0) = 1$, $x'(0) = -3$; $x(t) = e^{-3t}$.

15. $x'' + x = 1$, $x(0) = 1$, $x'(0) = 0$; $x(t) = 1$.

16. $x'' = 2$, $x(0) = 0$, $x'(0) = 0$; $x(t) = t^2$.

Forward Integral Rule

The rule is $\mathcal{L}\left(\int_0^t g(r)dr\right) = \frac{1}{s}\mathcal{L}(g(t))$

17. Relate this rule to the convolution rule with $f(t) = 1$.

18. Compute $\mathcal{L}\left(\int_0^t \sin(r)dr\right)$.

19. Compute $\mathcal{L}\left(\int_0^t (r+1)^3\,dr\right)$.

20. Compute $\mathcal{L}\left(\int_0^t \mathsf{sqw}(r)dr\right)$, where **sqw** is the square wave of period 2. Use the Extended Laplace Table.

Backward Integral Rule

Apply rule $\frac{1}{s}\mathcal{L}(g(t)) = \mathcal{L}\left(\int_0^t g(r)dr\right)$ and Lerch's theorem to solve for $f(t)$.

21. $\mathcal{L}(f(t)) = \frac{1}{s(s^2+1)}$

22. $\mathcal{L}(f(t)) = \frac{1}{s}\,\frac{s+1}{s^2+1}$

23. $\mathcal{L}(f(t)) = \frac{1}{s}\left(\frac{1}{s+1} - \frac{1}{s+2}\right)$

24. $\mathcal{L}(f(t)) = \frac{1}{s}\,\frac{e^{-s}}{s}$
Hint: $\mathcal{L}(u(t-a)) = \frac{1}{s}e^{-as}$.

The s–Integral Rule

Identity $\mathcal{L}\left(\frac{f(t)}{t}\right) = \int_s^\infty \mathcal{L}(f(t))\,ds$ requires piecewise continuous $f(t)$ of exponential order with $\lim_{t\to 0+} \frac{f(t)}{t} = L$.

25. Prove the identity.

26. Compute $\mathcal{L}\left(\frac{\sin(t)}{t}\right)$.

Forward First Shifting Rule

Apply $\mathcal{L}(f(t)e^{at}) = \mathcal{L}(f(t))|_{s\to s-a}$ to find the Laplace transform.

27. $\mathcal{L}(te^t)$

28. $\mathcal{L}(te^t + e^{2t})$

29. $\mathcal{L}(\sin(t)e^t)$

30. $\mathcal{L}(\sin(2t)e^{2t} + \cos(t)e^t)$

31. $\mathcal{L}(t\cosh(2t))$ using identity $\cosh(w) = \frac{1}{2}e^w + \frac{1}{2}e^{-w}$.

32. $\mathcal{L}((t+1)^3\,e^t)$

Backward First Shifting Rule

Apply $\mathcal{L}(f(t))|_{s\to s-a} = \mathcal{L}(f(t)e^{at})$ and Lerch's theorem to solve for $f(t)$.

33. Explain for $\mathcal{L}(t^2)\big|_{s\to s-4}$ the rule *Erase a shift $|_{s\to s-a}$ by inserting e^{at} inside the scope of $\mathcal{L}$.*

34. $\mathcal{L}(f(t)) = \left.\frac{s}{s^2+1}\right|_{s\to s-1}$

35. $\mathcal{L}(f(t)) = \frac{s-1}{(s-1)^2+4}$

36. $\mathcal{L}(f(t)) = \frac{8}{(s+1)^2+4}$

37. $\mathcal{L}(f(t)) = \frac{s+1}{s^2+2s+5}$

38. $\mathcal{L}(f(t)) = \frac{4}{s^2+8s+17}$

39. $\mathcal{L}(f(t)) = \frac{2}{(s+1)^2}$

40. $\mathcal{L}(f(t)) = \frac{1}{(s+2)^{101}}$

Forward s-Differentiation

Apply $\mathcal{L}((-t)f(t)) = \frac{d}{ds}\mathcal{L}(f(t))$ to find the Laplace transform.

41. Explain for $\mathcal{L}((-t)\cos(t))$ the rule *Multiplying by $(-t)$ differentiates the Laplace transform..*

42. $\mathcal{L}((-t)\sin(2t))$

43. $\mathcal{L}((-t)\sinh(2t))$, using identity $\sinh(w) = \frac{1}{2}e^{w} - \frac{1}{2}e^{-w}$.

44. $\mathcal{L}(te^{t}\sin(2t) + te^{2t}\cos(t))$

Backward s-Differentiation

Apply $\frac{d}{ds}\mathcal{L}(f(t)) = \mathcal{L}((-t)f(t))$ and Lerch's theorem to solve for $f(t)$.

45. Explain for $\frac{d}{ds}\mathcal{L}(\cos(t))$ the rule *Erase $\frac{d}{ds}$ by inserting factor $(-t)$ inside the scope of $\mathcal{L}$.*

46. $\mathcal{L}(f(t)) = \frac{d}{ds}\frac{s}{s^2+4}$

47. $\mathcal{L}(f(t)) = \frac{d^2}{ds^2}\frac{1}{(s+1)^5}$

48. $\mathcal{L}(f(t)) = \frac{d^3}{ds^3}\frac{s+1}{s^2+2s+5}$

Unit Step and Pulse

Define

$$\mathbf{pulse}(t,a,b) = \begin{cases} 1 & a \le t < b, \\ 0 & \text{else,} \end{cases}$$

which is a tool for encoding and decoding piecewise-defined functions.

49. Prove the identity $\mathbf{pulse}(t,a,b) = u(t-a) - u(t-b)$, where u is the **unit step**.

50. Prove the Laplace formula $\mathcal{L}(\mathbf{pulse}(t,a,b)) = \frac{e^{-at}-e^{-bt}}{s}$

51. Verify that $f(t)$ defined by $\begin{cases} 2 & 1 \le t < 2, \\ 0 & \text{else} \end{cases} + \begin{cases} 3 & 3 \le t < 4, \\ 0 & \text{else} \end{cases}$ encodes to representation $2\,\mathbf{pulse}(t,1,2) + 3\,\mathbf{pulse}(t,3,4)$.

52. Decode $f(t)$ into a piecewise–defined function and graph it by hand, no computer, given $f(t)$ is $e^{t}\,\mathbf{pulse}(t,1,3) + e^{-t}\,\mathbf{pulse}(t,4,6)$

53. Decode $f(t)$ into a piecewise–defined function and graph it, no computer, given $f(t)$ is the sum $\sum_{n=1}^{3} |\sin(n\pi t)|\,\mathbf{pulse}(t,2n,2n+1)$

54. Encode as a combination of pulses

$$f(t) = \begin{cases} 1 & 1 \le t < 2, \\ -2 & 3 \le t < 4, \\ 1 & 5 \le t < 6, \\ 0 & \text{else,} \end{cases}$$

showing all encoding details. Ans: $f(t) = \mathbf{pulse}(t,1,2) - 2\,\mathbf{pulse}(t,3,4) + \mathbf{pulse}(t,5,6)$.

Alternate Second Shifting Rule

$\mathcal{L}(g(t)u(t-a)) = e^{-as}\mathcal{L}\left(g(w)|_{w=t+a}\right)$. No Laplace here. The focus is on function notation and finding $g(t+a) = g(w)|_{w=t+a}$, which means *substitute $w = t+a$ into the $g(w)$-formula.*

55. Let $g(t) = te^{-t}$. Verify identity $g(w)|_{w=t+2} = e^{-2}(te^{-t} + 2e^{-t})$.

56. Let $g(t) = t^3$. Verify identity $g(w)|_{w=t+2} = 8 + 12t + 6t^2 + t^3$.

57. Typical polynomial $g(w) = 1 + 2w^2 + 3w^4$ upon substitution $w = t + a$ requires expansions for $(t+a)^2$ and $(t+a)^4$. Pascal's Triangle can be useful. Find the answer for $g(t+a) = g(w)|_{w=t+a}$.

58. Polynomial $1+2w^2+3w^4$ upon substitution $w=t-b$ is a Taylor polynomial expansion
$f(t)=\sum_{n=0}^{4}\frac{f^{(n)}(b)}{n!}(t-b)^n$.
Find the Maclaurin expansion
$f(t)=\sum_{n=0}^{4}\frac{f^{(n)}(0)}{n!}t^n$.

Forward Second Shifting Rule

$\mathcal{L}(g(t)u(t-a))=e^{-as}\mathcal{L}(g(t+a))$
Find $\mathcal{L}(f(t))$, where u is the unit step.

59. $f(t)=u(t-\pi)$

60. $f(t)=e^t\,u(t-1)$

61. $f(t)=t^3u(t-\pi)$

62. $f(t)=e^t\,\mathbf{pulse}(t,1,2)$, where $\mathbf{pulse}(t,a,b)=u(t-a)-u(t-b)$.

63. $f(t)=te^tu(t-2)$

64. $f(t)=t\sin(t)u(t-\pi)$

Backward Second Shifting Rule

$e^{-as}\mathcal{L}(f(t))=\mathcal{L}(f(t-a)u(t-a))$
Find $f(t)$ using the rule and Lerch's theorem, giving a piecewise–defined display and a unit step or pulse formula.

65. $\mathcal{L}(f(t))=\frac{1}{s}e^{-3s}$
Ans: $f(t)=u(t-3)=\begin{cases}1 & t\ge 3,\\ 0 & \text{else},\end{cases}$

66. $\mathcal{L}(f(t))=\dfrac{1}{s^2}e^{3-3s}$

67. $\mathcal{L}(f(t))=\dfrac{4}{s^2+8s+17}e^{-2s}$

68. $\mathcal{L}(f(t))=\dfrac{4+s}{s^2+8s+17}e^{-3s}$

69. $\mathcal{L}(f(t))=\left(\dfrac{1}{s^2}+\dfrac{2}{s^3}\right)e^{-2s}$

70. $\mathcal{L}(f(t))=\dfrac{1}{(s-4)^2}\,e^{-2s}$

Trigonometric Formulas

Supply the details in Example 8.21.

71. $\mathcal{L}(t\sin at)=\dfrac{2as}{(s^2+a^2)^2}$

72. $\mathcal{L}(t^2\sin at)=\dfrac{6s^2a-a^3}{(s^2+a^2)^3}$

Exponential Formulas

Supply the details in Example 8.22.

73. $\mathcal{L}(e^{at}\sin bt)=\dfrac{b}{(s-a)^2+b^2}$

74. $\mathcal{L}(te^{at}\sin bt)=\dfrac{2b(s-a)}{((s-a)^2+b^2)^2}$

Hyperbolic Functions

Supply the details in Example 8.23.

75. $\mathcal{L}(\sinh at)=\dfrac{a}{s^2-a^2}$

76. $\mathcal{L}(t\cosh at)=\dfrac{s^2+a^2}{(s^2-a^2)^2}$

Waves

Use Laplace ideas from Examples 8.24 and 8.25. Each $f(t)$ can be expressed as a **pulse train**, which is an expression $\sum_{n=1}^{\infty}f_n(t)\,\mathbf{pulse}(t,a_i,b_i)$ to which the second shifting theorem applies.

77. Find $\mathcal{L}(f(t))$ for the square wave
$f(t)=\sum_{n=0}^{\infty}(-1)^n\,\mathbf{pulse}(t,n,n+1)$

78. Define pulse train
$f(t)=\sum_{n=0}^{\infty}f_n(t)\,\mathbf{pulse}(t,n,n+1)$, $f_{2n}(t)=t-2n, f_{2n+1}(t)=2-t+2n$.
Show that $f(t+2)=f(t)$ and
$f(t)=\begin{cases}t & 0\le t<1,\\ 2-t & 1\le t\le 2.\end{cases}$

79. Find $\mathcal{L}(f(t))$ for
$f(t)=\begin{cases}|\sin(2t)| & 0\le t\le\pi,\\ 0 & \pi\le t\le 2\pi,\end{cases}$
and $f(t+r\pi)=f(t)$.

80. Find $\mathcal{L}(f(t))$ for
$f(t)=\begin{cases}1 & 0\le t\le\pi,\\ |\sin(t)| & \pi\le t\le 2\pi,\end{cases}$
and $f(t+2\pi)=f(t)$.

81. Given $f(t)=\frac{1}{2}(|\sin t|+\sin t)$, called the **Half–wave rectification** of the sine wave, derive $\mathcal{L}(f(t))=\frac{1}{(s^2+1)(1-e^{-\pi s})}$

82. Solve for 2–periodic function $f(t)$:
$\mathcal{L}(f(t))=\dfrac{1}{s}\tanh\left(\dfrac{s}{2}\right)$.
Use the Extended Laplace Integral Table.

8.4 Heaviside's Partial Fraction Method

This clever algebraic shortcut is used to solve an equation like

$$\mathcal{L}(f(t)) = \frac{2s}{(s+1)(s^2+1)}$$

for the time domain function $f(t) = -e^{-t} + \cos t + \sin t$. The details in Heaviside's method involve a sequence of easy-to-learn college algebra steps. The practical method was popularized by English electrical engineer Oliver Heaviside (1850–1925).

More precisely, **Heaviside's method** starts with a polynomial quotient

$$\text{(1)} \qquad \frac{a_0 + a_1 s + \cdots + a_n s^n}{b_0 + b_1 s + \cdots + b_m s^m}$$

and computes an expression $f(t)$ such that

$$\frac{a_0 + a_1 s + \cdots + a_n s^n}{b_0 + b_1 s + \cdots + b_m s^m} = \mathcal{L}(f(t)) \equiv \int_0^\infty f(t) e^{-st} dt.$$

Symbols $a_0, \ldots, a_n, b_0, \ldots, b_m$ are **real** constants. Heaviside's method assumes **limit zero at** $s = \infty$ for polynomial quotient (1). [5]

Partial Fraction Theory

It is a college algebra theorem that a rational function (1) can be expressed as the sum of **partial fractions**.

Definition 8.3 (Partial Fraction)

A partial fraction is a polynomial fraction with a constant in the numerator and a polynomial denominator having exactly one root, i.e.,

$$\text{(2)} \qquad \textbf{partial fraction} = \frac{C}{(s-s_0)^k}.$$

The numerator C in (2) is a real or complex constant. The denominator has exactly one root $s = s_0$, real or complex. We expect power $(s-s_0)^k$ to be a **divisor of the denominator** in fraction (1).

Real Root Case. If s_0 in (2) is a real number, then C is *real*.

Complex Root Case. If $s_0 = a + ib$ in (2), then $(s - \overline{s_0})^k$ also divides the denominator in (1), where $\overline{s_0} = a - ib$ is the complex conjugate of s_0. The corresponding partial fractions used in the expansion turn out to be complex

[5] Otherwise, fraction (1) equals by long division a polynomial plus a remainder. Heaviside's method applies to the remainder.

conjugates of one another, which can be paired and re-written as a fraction with numerator $Q(s)$ a *real* polynomial.

$$\text{(3)} \qquad \frac{C}{(s-s_0)^k} + \frac{\overline{C}}{(s-\overline{s_0})^k} = \frac{Q(s)}{((s-a)^2+b^2)^k}.$$

To illustrate, if $C = u + iv$, then

$$\begin{aligned} \frac{C}{(s-2i)^2} + \frac{\overline{C}}{(s+2i)^2} &= \frac{(C+\overline{C})s^2 + 4i(\overline{C}-C)s - 4(C+\overline{C})}{(s^2+4)^2} \\ &= \frac{2us^2+8vs-8u}{(s^2+4)^2}. \end{aligned}$$

The numerator $2us^2+8vs-8u$ can be expanded by the college algebra division algorithm as $Q(s) = A_1(s^2+4) + A_2 s + A_3$, with real coefficients A_1, A_2, A_3. Then the fraction can be written as

$$\frac{Q(s)}{(s^2+4)^2} = \frac{A_1}{s^2+4} + \frac{A_2 s + A_3}{(s^2+4)^2}.$$

Similarly, numerator $2us^3 - 12vs^2 - 24us + 16v$ expands as $A_1(s^2+4)^2 + A_2(s^2+4) + A_3 s + A_4$ in the following example:

$$\begin{aligned} \frac{u+iv}{(s-2i)^3} + \frac{u-iv}{(s+2i)^3} &= \frac{2us^3-12vs^2-24us+16v}{(s^2+4)^3} \\ &= \frac{A_1}{s^2+4} + \frac{A_2}{(s^2+4)^2} + \frac{A_3 s + A_4}{(s^2+4)^3} \end{aligned}$$

for some *real coefficients* A_1, A_2, A_3, A_4.

This discussion generalizes to all powers $k > 1$. Partial fractions with denominator $(s-s_0)^k$ and $(s-\overline{s_0})^k$ with s_0 = a complex number are paired and the division algorithm is employed as in the examples to replace the pair of terms by a sum of terms of the form

$$\frac{\text{linear polynomial in } s}{((s-a)^2+b^2)^j}, \quad 1 \le j \le k.$$

The numerator has the form $c_1 + c_2 s$ with *real* coefficients c_1, c_2. This **real partial fraction form** is preferred over the sum of complex fractions, because integral tables and Laplace tables typically contain only formulas with *real coefficients*. See Example 8.26, page 629.

Simple Roots

Assume that (1) has *real coefficients* and the denominator of the fraction (1) has **distinct real roots** $s_1, \ldots, s_N$ and **distinct complex roots** $\alpha_1 \pm i\beta_1, \ldots,$

$\alpha_M \pm i\beta_M$. The partial fraction expansion of (1) is a sum given in terms of *real* constants A_p, B_q, C_q by

$$(4) \qquad \frac{a_0 + a_1 s + \cdots + a_n s^n}{b_0 + b_1 s + \cdots + b_m s^m} = \sum_{p=1}^{N} \frac{A_p}{s - s_p} + \sum_{q=1}^{M} \frac{B_q + C_q(s - \alpha_q)}{(s - \alpha_q)^2 + \beta_q^2}.$$

Multiple Roots

Assume (1) has *real coefficients* and the denominator of the fraction (1) has possibly *multiple roots*. Let N_p be the multiplicity of real root s_p and let M_q be the multiplicity of complex root $\alpha_q + i\beta_q$ ($\beta_q > 0$), $1 \le p \le N$, $1 \le q \le M$. The partial fraction expansion of (1) is given in terms of *real* constants $A_{p,k}$, $B_{q,k}$, $C_{q,k}$ by

$$(5) \qquad \sum_{p=1}^{N} \sum_{1 \le k \le N_p} \frac{A_{p,k}}{(s - s_p)^k} + \sum_{q=1}^{M} \sum_{1 \le k \le M_q} \frac{B_{q,k} + C_{q,k}(s - \alpha_q)}{((s - \alpha_q)^2 + \beta_q^2)^k}.$$

Summary

A polynomial quotient p/q with limit zero at infinity has a unique expansion into partial fractions. A partial fraction is either a constant divided by a divisor of q having exactly one real root, or else a linear function divided by a real divisor of q, having exactly one complex conjugate pair of roots.

Sampling Method

Consider the expansion in partial fractions

$$(6) \qquad \frac{s-1}{s(s+1)^2(s^2+1)} = \frac{A}{s} + \frac{B}{s+1} + \frac{C}{(s+1)^2} + \frac{Ds+E}{s^2+1}.$$

The five undetermined real constants A through E are found by **clearing the fractions**, that is, multiply (6) by the denominator on the left to obtain the polynomial equation

$$(7) \qquad \begin{aligned} s - 1 \quad = \quad & A(s+1)^2(s^2+1) + Bs(s+1)(s^2+1) \\ & + Cs(s^2+1) + (Ds+E)s(s+1)^2. \end{aligned}$$

Next, five different **samples** of s are substituted into (7) to obtain equations for the five unknowns A through E.[6] Always use the **roots of the denominator**

[6] The values chosen for s are called **samples**, that is, they are cleverly chosen values. The number of s-values sampled equals the number of symbols A, B, ... to be determined, which in turn equals the degree of the denominator.

to start: $s = 0$, $s = -1$, $s = i$, $s = -i$ are the roots of $s(s+1)^2(s^2+1) = 0$. Each complex root results in two equations, by taking real and imaginary parts. The complex conjugate root $s = -i$ is not used, because it duplicates equations already obtained from $s = i$. The three roots $s = 0$, $s = -1$, $s = i$ give only four equations, so we **invent another sample** $s = 1$ to get the last equation:

$$(8)\qquad \begin{array}{rcll} -1 &=& A & (s=0) \\ -2 &=& -2C & (s=-1) \\ i-1 &=& (Di+E)i(i+1)^2 & (s=i) \\ 0 &=& 8A+4B+2C+4(D+E) & (s=1) \end{array}$$

Because D and E are real, the complex equation $(s = i)$ becomes two equations, as follows.

$i - 1 = (Di + E)i(i^2 + 2i + 1)$	Expand power $(i + 1)^2$.
$i - 1 = -2Di - 2E$	Simplify using $i^2 = -1$.
$1 = -2D$	Equate imaginary parts.
$-1 = -2E$	Equate real parts. Root i created 2 equations!

The 5×5 system of linear algebraic equations is solved for answers $A = -1$, $B = 3/2$, $C = 1$, $D = -1/2$, $E = 1/2$.

Method of Atoms

Consider the expansion in partial fractions

$$(9)\qquad \frac{2s-2}{s(s+1)^2(s^2+1)} = \frac{a}{s} + \frac{b}{s+1} + \frac{c}{(s+1)^2} + \frac{ds+e}{s^2+1}.$$

Clearing the fractions in (9) gives the polynomial equation

$$(10)\qquad \begin{array}{rcl} 2s-2 &=& a(s+1)^2(s^2+1) + bs(s+1)(s^2+1) \\ && +cs(s^2+1) + (ds+e)s(s+1)^2. \end{array}$$

The **method of atoms** expands all polynomial products and collects on powers of s. Functions $1, s, s^2, \ldots$ are by *definition* called **Euler solution atoms**, hence the terminology. The coefficients of the powers are matched to give 5 equations in the five unknowns a through e. Some details:

$$(11)\qquad \begin{array}{rcl} 2s-2 &=& (a+b+d)\,s^4 + (2a+b+c+2d+e)\,s^3 \\ && +\,(2a+b+d+2e)\,s^2 + (2a+b+c+e)\,s + a \end{array}$$

Matching powers of s implies the 5 equations

$$\begin{array}{l} a+b+d=0,\ 2a+b+c+2d+e=0,\ 2a+b+d+2e=0, \\ 2a+b+c+e=2,\ a=-2. \end{array}$$

Solving, the unique solution is $a = -2$, $b = 3$, $c = 2$, $d = -1$, $e = 1$.

Heaviside's Coverup Method

Assume distinct roots in the denominator of fraction (1). Extensions to multiple-root cases can be made; see page 625.

To illustrate Oliver Heaviside's 1890 ideas, consider the problem details

$$\begin{aligned}\frac{2s+1}{s(s-1)(s+1)} &= \frac{A}{s}+\frac{B}{s-1}+\frac{C}{s+1}\\ &= \mathcal{L}(A)+\mathcal{L}(Be^{t})+\mathcal{L}(Ce^{-t})\\ &= \mathcal{L}(A+Be^{t}+Ce^{-t})\end{aligned} \tag{12}$$

The first line in (12) uses college algebra partial fractions. The second and third lines use the basic Laplace table and linearity of $\mathcal{L}$. Missing here are the values of constants A, B, C. Heaviside's ideas provide an efficient method to evaluate $A=-1, B=\frac{3}{2}, C=-\frac{1}{2}$. Then $\mathcal{L}(y)=\frac{2s+1}{s(s-1)(s+1)}=\mathcal{L}(-1+\frac{3}{2}e^{t}-\frac{1}{2}e^{-t})$ implies $y=-1+\frac{3}{2}e^{t}-\frac{1}{2}e^{-t}$.

Mysterious Details

Oliver Heaviside proposed to find $A=-1, B=\frac{3}{2}, C=-\frac{1}{2}$ in (12) by a **cover-up method**. The method is completely mental, no writing at all. We explain in detail how Heaviside found $C=-\frac{1}{2}$.

Heaviside starts with the identity

$$\frac{2s+1}{s(s-1)(s+1)}=\frac{A}{s}+\frac{B}{s-1}+\frac{C}{s+1}. \tag{13}$$

The cover–up method finds C by **mentally** clearing the fraction $\frac{\mathbf{C}}{\mathbf{s+1}}$, that is, multiply (13) by the denominator $\boxed{\mathbf{s+1}}$ of the partial fraction $\frac{\mathbf{C}}{\mathbf{s+1}}$ to obtain the *partially-cleared fraction relation*

$$\frac{(2s+1)\boxed{(s+1)}}{s(s-1)\boxed{(s+1)}}=\frac{A\boxed{(s+1)}}{s}+\frac{B\boxed{(s+1)}}{s-1}+\frac{C\boxed{(s+1)}}{\boxed{(s+1)}}.$$

Set $\boxed{(s+1)}=0$ in the display. Cancellations left and right plus annihilation of two terms on the right give the answer for C:

$$\left.\frac{2s+1}{s(s-1)}\right|_{\boxed{s+1}=0}=0+0+C.$$

Heaviside's cryptic *instructions* are to cover–up the matching factors $(s+1)$ on the left and right in (13) with box $\boxed{(s+1)}$ (Heaviside used his fingertips), then evaluate on the left at the *root* s which causes the box contents to be zero. The other terms on the right are replaced by zero. Heaviside would find $C = -\frac{1}{2}$ by placing his fingers over the factors $(s+1)$ left and right in (13), the boxes $\boxed{(s+1)}$ below being his finger tips:

$$\left.\frac{2s+1}{s(s-1)\boxed{(s+1)}}\right|_{\boxed{s+1}=0} = \frac{C}{\boxed{(s+1)}}.$$

The factor $(s+1)$ in (13) is by no means special: the same procedure applies to find A and B. The method works for denominators with simple roots, that is, no repeated roots are allowed. Heaviside's method in words:[7]

> To determine C in partial fraction $\frac{C}{s-s_0}$, multiply the relation by $(s-s_0)$, to partially clear the fraction. Substitute root s of equation $s - s_0 = 0$ into the partially cleared relation.

Extension to Multiple Roots

Heaviside's method can be extended to the case of repeated roots. The basic idea is to *factor–out the repeats*. To illustrate, consider the partial fraction expansion details

$$R = \frac{1}{(s+1)^2(s+2)}$$

A sample rational function having repeated roots.

$$= \frac{1}{s+1}\left(\frac{1}{(s+1)(s+2)}\right)$$

Factor–out the repeats.

$$= \frac{1}{s+1}\left(\frac{1}{s+1} + \frac{-1}{s+2}\right)$$

Apply the cover–up method to the simple root fraction.

$$= \frac{1}{(s+1)^2} + \frac{-1}{(s+1)(s+2)}$$

Multiply. Observe that $\frac{1}{(s+1)^2}$ is a partial fraction!

$$= \frac{1}{(s+1)^2} + \frac{-1}{s+1} + \frac{1}{s+2}$$

Apply the cover–up method to the last fraction on the right.

Term $\frac{1}{(s+1)^2}$ has constant numerator and denominator with only one root. It is already a partial fraction.[8] Therefore the work centers on expansion of quotients in which the denominator has two or more roots.

[7] Root $s = s_0$ is called a **pole** and the answer C is called a **residue**. See page 627.

[8] Re–read the definition of partial fraction page 620.

Special Methods

Heaviside's method has a useful extension for the case of roots of multiplicity two. To illustrate, consider these details:

$$\begin{aligned}
R &= \frac{1}{(s+1)^2(s+2)} \\
&= \frac{A}{s+1} + \frac{B}{(s+1)^2} + \frac{C}{s+2} \\
&= \frac{A}{s+1} + \frac{1}{(s+1)^2} + \frac{1}{s+2} \\
&= \frac{-1}{s+1} + \frac{1}{(s+1)^2} + \frac{1}{s+2}
\end{aligned}$$

$\boxed{1}$ A fraction with multiple roots.

$\boxed{2}$ See equation (5), page 622.

$\boxed{3}$ Find B and C by Heaviside's cover–up method.

$\boxed{4}$ Details below.

Details $\boxed{4}$. Multiply the equation $\boxed{1} = \boxed{2}$ by $s+1$ to partially clear fractions, the same step as the cover-up method:

$$\frac{1}{(s+1)(s+2)} = A + \frac{B}{s+1} + \frac{C(s+1)}{s+2}.$$

Don't substitute s from $s+1=0$, because it gives infinity for the second term. Instead, set $s=\infty$ to get the equation $0 = A + C$. Because $C = 1$ from $\boxed{3}$, then $A = -1$.

The illustration works for one root of multiplicity two, because $s = \infty$ will resolve the coefficient not found by the cover–up method.

In general, if the denominator in (1) has a root s_0 of multiplicity k, then the partial fraction expansion contains terms

$$\frac{C_1}{s-s_0} + \frac{C_2}{(s-s_0)^2} + \cdots + \frac{C_k}{(s-s_0)^k}.$$

Heaviside's cover–up method directly finds C_k, but not C_1 to C_{k-1}.

Cover-up Method and Complex Numbers

Consider the partial fraction expansion

$$\frac{10}{(s+1)(s^2+9)} = \frac{A}{s+1} + \frac{Bs+C}{s^2+9}.$$

The symbols A, B, C are real. The value of A can be found directly by the cover-up method, giving $A = 1$. To find B and C, multiply the fraction expansion by s^2+9, in order to partially clear fractions, then formally set $s^2+9=0$ to obtain the two equations

$$\frac{10}{s+1} = Bs + C, \quad s^2+9=0.$$

The method applies the identical idea used for one real root. By clearing fractions in the first, the equations become

$$10 = Bs^2 + Cs + Bs + C, \quad s^2 + 9 = 0.$$

Substitute $s^2 = -9$ into the first equation to give the linear equation

$$10 = (-9B + C) + (B + C)s.$$

Because this linear equation has two complex roots $s = \pm 3i$, then real constants B, C satisfy the 2×2 system

$$\begin{array}{rcrcr} -9B & + & C & = & 10, \\ B & + & C & = & 0. \end{array}$$

Solving gives $B = -1$, $C = 1$.

The same method applies especially to fractions with 3-term denominators, like $s^2 + s + 1$. The only change made in the details is the replacement $s^2 \to -s - 1$. By repeated application of $s^2 = -s - 1$, the first equation can be distilled into one linear equation in s with two roots. As before, a 2×2 system results.

Residues, Poles and Oliver Heaviside

The language of **residues** and **poles** invaded engineering literature years ago, blamed in part on engineers who studied the foundations of complex variables. The terminology formalizes the naming of partial fraction theory constants and roots that appear in Oliver Heaviside's *cover-up method*, detailed above, which is an electrical engineering partial fraction shortcut that dates back to the year 1890.

Residues and poles do not provide any new mathematical tools for solving partial fraction problems. The service of residues and poles is to provide a new language for discussing the answers, a language that appears in current engineering and science literature. If you know how to compute coefficients in partial fractions using Heaviside's shortcut, then you already know how to find residues and poles.

A Key Example. Heaviside's shortcut finds the coefficients $c_1 = \frac{1}{2}, c_2 = -5, c_3 = \frac{5}{2}$ in the expansion

$$\frac{5 - 2(s+2)(s+3)}{(s+1)(s+2)(s+3)} = \frac{c_1}{s+1} + \frac{c_2}{s+2} + \frac{c_3}{s+3}$$

by clearing the fractions one at a time, each clearing followed by substitution of the corresponding root found in the denominator.

For instance, to clear the fraction for c_2 requires multiplication by $(s+2)$, to give the intermediate step (Heaviside did it mentally, writing nothing)

$$\frac{5 - 2(s+2)(s+3)}{(s+1)(s+3)} = \frac{c_1(s+2)}{s+1} + \frac{c_2}{1} + \frac{c_3(s+2)}{s+3}.$$

Root $s = -2$ of $s + 2 = 0$ is substituted above to give $c_2 = -5$.

Table 8. Working Definition of Pole and Residue

A **pole** is the same as a root of the denominator in a quotient $\dfrac{p(x)}{q(x)}$.

A **residue** is the same as a coefficient in the partial fraction expansion of the quotient $\dfrac{p(x)}{q(x)}$ (precise details below).

In the *key example*, the **residue** at **pole** $s = -2$ (the pole is the root of $s+2 = 0$) is **defined** by the equation

$$\lim_{s \to -2} (s + 2) \frac{5 - (2(s + 2)(s + 3)}{(s + 1)(s + 2)(s + 3)}.$$

To evaluate the limit, cancel the common factor $(s + 2)$ and substitute $s = -2$. Oliver Heaviside would be surprised by the unnecessary limit.

Definition 8.4 (Poles and Residues)

A function $f(z)$ of complex variable z has a **pole** at $z = z_0$ provided there is an integer $n \geq 0$ such that $g(z) = (z - z_0)^n f(z)$ can be written as a power series

$$g(z) = g_0 + g_1(z - z_0) + g_2(z - z_0)^2 + \cdots$$

convergent in a disk $|z - z_0| < R$ and $g_0 \neq 0$ (which means $g(z_0) \neq 0$).

The **order of the pole** is the integer n. The **residue** is g_0.

If $f(z)$ has a pole $z = z_0$ of order n, then the **residue** g_0 at the pole can be computed from the limit formula

$$g_0 = \lim_{z \to z_0} (z - z_0)^n f(z).$$

In terms of series expansion, a pole of order n means that

$$f(z) = \frac{g_0}{(z - z_0)^n} + \cdots + g_n + g_{n+1}(z - z_0) + g_{n+2}(z - z_0)^2 + \cdots,$$

which is called a **Laurent Series**.

Table 9. Pole, Residue and Applications

A real pole defines the **damping coefficient** in a transient.

A complex pole on the imaginary axis describes **frequency**.

Residues are **mode shape** information.

Examples and Methods

Example 8.26 (Partial Fractions I)

Show the details of the partial fraction expansion

$$\frac{s^3+2s^2+2s+5}{(s-1)(s^2+4)(s^2+2s+2)} = \frac{2/5}{s-1} + \frac{1/2}{s^2+4} - \frac{1}{10}\frac{7+4\,s}{s^2+2\,s+2}.$$

Solution:

Background. The problem originates as equality $\boxed{5}=\boxed{6}$ in the sequence of Example 8.28, page 632, which solves for $x(t)$ using the method of partial fractions:

$$\begin{array}{lrl} \boxed{5} & \mathcal{L}(x) = & \dfrac{s^3+2s^2+2s+5}{(s-1)(s^2+4)(s^2+2s+2)} \\ \boxed{6} & = & \dfrac{2/5}{s-1} + \dfrac{1/2}{s^2+4} - \dfrac{1}{10}\dfrac{7+4\,s}{s^2+2\,s+2} \end{array}$$

College algebra detail. College algebra partial fractions theory says that there exist real constants A, B, C, D, E satisfying the identity

$$\frac{s^3+2s^2+2s+5}{(s-1)(s^2+4)(s^2+2s+2)} = \frac{A}{s-1} + \frac{B+Cs}{s^2+4} + \frac{D+Es}{s^2+2\,s+2}.$$

As explained on page 621, the complex conjugate roots $\pm 2i$ and $-1\pm i$ are not represented as terms $c/(s-s_0)$, but in the combined real form seen in the above display, which is suited for use with Laplace tables.

The **sampling method** applies to find the constants. In this method, the fractions are cleared to obtain the polynomial relation

$$\begin{aligned} s^3+2s^2+2s+5 \;=\;& A(s^2+4)(s^2+2s+2) \\ &+(B+Cs)(s-1)(s^2+2s+2) \\ &+(D+Es)(s-1)(s^2+4). \end{aligned}$$

The roots of the denominator $(s-1)(s^2+4)(s^2+2s+2)$ to be inserted into the previous equation are $s=1$, $s=2i$, $s=-1+i$. The conjugate roots $s=-2i$ and $s=-1-i$ are not used. Each complex root generates two equations, by equating real and imaginary parts, therefore there will be 5 equations in 5 unknowns. Substitution of $s=1$, $s=2i$, $s=-1+i$ gives three equations

$$\begin{array}{lrcl} s=1 & 10 & = & 25A, \\ s=2i & -4i-3 & = & (B+2iC)(2i-1)(-4+4i+2), \\ s=-1+i & 5 & = & (D-E+Ei)(-2+i)(2-2(-1+i)). \end{array}$$

Writing each expanded complex equation in terms of its real and imaginary parts, explained in detail below, gives 5 equations

$$\begin{array}{lrcrcr} s=1 & 2 & = & 5A, & & \\ s=2i & -3 & = & -6B & + & 16C, \\ s=2i & -4 & = & -8B & - & 12C, \\ s=-1+i & 5 & = & -6D & - & 2E, \\ s=-1+i & 0 & = & 8D & - & 14E. \end{array}$$

The equations are solved to give $A = 2/5$, $B = 1/2$, $C = 0$, $D = -7/10$, $E = -2/5$ (details for B, C below).

Complex equation to two real equations. It is an algebraic mystery how exactly the complex equation

$$-4i - 3 = (B + 2iC)(2i - 1)(-4 + 4i + 2)$$

gets converted into two real equations. The process is explained here.

First, the complex equation is expanded, as though it is a polynomial in variable i, to give the steps

$$\begin{array}{rcll} -4i - 3 & = & (B + 2iC)(2i - 1)(-2 + 4i) & \\ & = & (B + 2iC)(-4i + 2 + 8i^2 - 4i) & \text{Expand.} \\ & = & (B + 2iC)(-6 - 8i) & \text{Use } i^2 = -1. \\ & = & -6B - 12iC - 8Bi + 16C & \text{Expand, use } i^2 = -1. \\ & = & (-6B + 16C) + (-8B - 12C)i & \text{Convert to form } x + yi. \end{array}$$

Next, the two sides are compared. Because B and C are real, then the real part of the right side is $(-6B + 16C)$ and the imaginary part of the right side is $(-8B - 12C)$. Equating matching parts on each side gives the equations

$$\begin{array}{rcl} -6B + 16C & = & -3, \\ -8B - 12C & = & -4, \end{array}$$

which is a 2×2 linear system for the unknowns B, C.

Solving the 2×2 system. Such a system with a unique solution can be solved by Cramer's rule, matrix inversion or elimination. The answer: $B = 1/2$, $C = 0$.

The easiest method turns out to be elimination. Multiply the first equation by 4 and the second equation by 3, then subtract to obtain $C = 0$. Then the first equation is $-6B + 0 = -3$, implying $B = 1/2$.

Example 8.27 (Partial Fractions II)

Verify the partial fraction expansion

$$\boxed{1} \quad \frac{s^5 + 8\,s^4 + 23\,s^3 + 31\,s^2 + 24\,s + 9}{(s+1)^2\,(s^2 + s + 1)^2} = \frac{4}{s+1} + \frac{5 - 3s}{s^2 + s + 1}.$$

Solution:

Basic partial fraction theory implies that there are unique real constants a, b, c, d, e, f satisfying the equation

$$\begin{array}{rcl} \dfrac{s^5 + 8\,s^4 + 23\,s^3 + 31\,s^2 + 24\,s + 9}{(s+1)^2\,(s^2 + s + 1)^2} & = & \dfrac{a}{s+1} + \dfrac{b}{(s+1)^2} \\ & & + \dfrac{c + ds}{s^2 + s + 1} + \dfrac{e + f\,s}{(s^2 + s + 1)^2} \end{array}$$

Sanity checks apply when constructing the expansion. First, the number of real constants is always the degree of the denominator, which is 6 in this example. This caused the invention of 6 symbols a, b, c, d, e, f. Only *real* polynomials appear in the fraction denominators on the right. The following checkpoints are done mentally: *we never in a partial fraction problem write out such details*:

$$\begin{array}{lll} s+1 & \text{divides} & (s+1)^2\left(s^2+s+1\right)^2 \\ (s+1)^2 & \text{divides} & (s+1)^2\left(s^2+s+1\right)^2 \\ s^2+s+1 & \text{divides} & (s+1)^2\left(s^2+s+1\right)^2 \\ (s^2+s+1)^2 & \text{divides} & (s+1)^2\left(s^2+s+1\right)^2 \end{array}$$

The **sampling** method applies to clear fractions and replace the fractional equation by the polynomial relation

$$\begin{aligned} s^5+8\,s^4+23\,s^3+31\,s^2+24\,s+9 \;&=\; a(s+1)(s^2+s+1)^2 \\ &\quad +b(s^2+s+1)^2 \\ &\quad +(c+ds)(s^2+s+1)(s+1)^2 \\ &\quad +(e+f\,s)(s+1)^2 \end{aligned}$$

However, the prognosis for the resultant algebra is grim: only three of the six required equations can be obtained by substitution of the roots ($s=-1$, $s=-1/2+i\sqrt{3}/2$) of the denominator. The sampling idea is abandoned, because of the complexity of the 6×6 system of linear equations required to solve for the six constants a through f.

Instead, the fraction R on the left of $\boxed{\mathbf{1}}$ is written with repeated factors extracted, as follows:

$$\begin{aligned} R &= \frac{1}{(s+1)(s^2+s+1)}\left(\frac{p(s)}{(s+1)(s^2+s+1)}\right), \\ p(s) &= s^5+8\,s^4+23\,s^3+31\,s^2+24\,s+9. \end{aligned}$$

Long division gives the formulas

$$\begin{aligned} \frac{p(s)}{(s+1)(s^2+s+1)} &= s^2+6s+9, \\ R=\frac{p(s)}{(s+1)^2(s^2+s+1)^2} &= \frac{(s+3)^2}{(s+1)(s^2+s+1)}. \end{aligned}$$

The simplified form of R has a partial fraction expansion

$$\frac{(s+3)^2}{(s+1)(s^2+s+1)} = \frac{a}{s+1}+\frac{b+cs}{s^2+s+1}$$

where a,b,c are real constants. Reuse of earlier symbols a,b,c,d,e,f has occurred, similar to always using symbol x in a quadratic equation. Progress: the dimension of the algebra problem went from 6×6 to 3×3.

Heaviside's cover-up method gives $a=4$. Applying Heaviside's method again to the quadratic factor implies the pair of equations

$$\frac{(s+3)^2}{s+1}=b+cs, \quad s^2+s+1=0.$$

These equations can be solved for $b=5$, $c=-3$. The details assume that s is a root of $s^2+s+1=0$, then

$$\frac{(s+3)^2}{s+1}=b+cs \qquad \text{The first equation.}$$

$\dfrac{s^2+6s+9}{s+1} = b+cs$	Expand.
$\dfrac{-s-1+6s+9}{s+1} = b+cs$	Use $s^2+s+1=0$.
$5s+8 = (s+1)(b+cs)$	Clear fractions.
$5s+8 = bs+cs+b+cs^2$	Expand again.
$5s+8 = bs+cs+b-cs-c$	Use $s^2+s+1=0$.

Conclusion $5 = b$ and $8 = b-c$ follows because the last equation is linear but has two complex roots. Solve: $b=5$, $c=-3$.

Finally: $a=4, b=5, c=-3$, which verifies $\boxed{\mathbf{1}}$.

Example 8.28 (Third Order Initial Value Problem)
Solve the third order initial value problem

$$\begin{aligned} &x''' - x'' + 4x' - 4x = 5e^{-t}\sin t,\\ &x(0)=0, \quad x'(0)=x''(0)=1. \end{aligned}$$

Solution:
The answer is

$$x(t) = \frac{2}{5}e^t + \frac{1}{4}\sin 2t - \frac{3}{10}e^{-t}\sin t - \frac{2}{5}e^{-t}\cos t.$$

Method. Apply $\mathcal{L}$ to the differential equation. In steps $\boxed{1}$ to $\boxed{3}$ the Laplace integral of $x(t)$ is isolated, by applying linearity of $\mathcal{L}$, integration by parts $\mathcal{L}(f') = s\mathcal{L}(f) - f(0)$ and the basic Laplace table.

$$\mathcal{L}(x''') - \mathcal{L}(x'') + 4\mathcal{L}(x') - 4\mathcal{L}(x) = 5\mathcal{L}(e^{-t}\sin t) \qquad \boxed{1}$$

$$(s^3\mathcal{L}(x) - s - 1) - (s^2\mathcal{L}(x) - 1) + 4(s\mathcal{L}(x)) - 4\mathcal{L}(x) = \frac{5}{(s+1)^2+1} \qquad \boxed{2}$$

$$(s^3 - s^2 + 4s - 4)\mathcal{L}(x) = 5\frac{1}{(s+1)^2+1} + s \qquad \boxed{3}$$

Steps $\boxed{5}$ and $\boxed{6}$ use the college algebra theory of partial fractions, the details of which appear in Example 8.26, page 629. Steps $\boxed{7}$ and $\boxed{8}$ write the partial fraction expansion in terms of Laplace table entries. Step $\boxed{9}$ converts the s-expressions, which are basic Laplace table entries, into Laplace integral expressions. Algebraically, we replace s-expressions by expressions in symbols $\mathcal{L}$ and t.

$$\begin{aligned}
\mathcal{L}(x) &= \frac{\frac{5}{(s+1)^2+1} + s}{s^3 - s^2 + 4s - 4} && \boxed{4}\\
&= \frac{s^3 + 2s^2 + 2s + 5}{(s-1)(s^2+4)(s^2+2s+2)} && \boxed{5}\\
&= \frac{2/5}{s-1} + \frac{1/2}{s^2+4} - 1/10\,\frac{7+4\,s}{s^2+2\,s+2} && \boxed{6}\\
&= \frac{2/5}{s-1} + \frac{1/2}{s^2+4} - 1/10\,\frac{3+4(s+1)}{(s+1)^2+1} && \boxed{7}\\
&= \frac{2/5}{s-1} + \frac{1/2}{s^2+4} - \frac{3/10}{(s+1)^2+1} - \frac{(2/5)(s+1)}{(s+1)^2+1} && \boxed{8}
\end{aligned}$$

$$= \mathcal{L}\left(\frac{2}{5}e^t + \frac{1}{4}\sin 2t - \frac{3}{10}e^{-t}\sin t - \frac{2}{5}e^{-t}\cos t\right) \quad \boxed{9}$$

The last step $\boxed{10}$ applies Lerch's cancellation theorem to $\mathcal{L}(x(t)) = \boxed{9}$.

$$x(t) = \frac{2}{5}e^t + \frac{1}{4}\sin 2t - \frac{3}{10}e^{-t}\sin t - \frac{2}{5}e^{-t}\cos t \qquad \boxed{10}$$

Example 8.29 (Second Order System)

Solve for $x(t)$ and $y(t)$ in the 2nd order system of linear differential equations

$$\begin{array}{ll} 2x'' - x' + 9x - y'' - y' - 3y = 0, & x(0) = x'(0) = 1, \\ 2'' + x' + 7x - y'' + y' - 5y = 0, & y(0) = y'(0) = 0. \end{array}$$

Solution: The answer is

$$\begin{aligned} x(t) &= \frac{1}{3}e^t + \frac{2}{3}\cos(2\,t) + \frac{1}{3}\sin(2\,t), \\ y(t) &= \frac{2}{3}e^t - \frac{2}{3}\cos(2\,t) - \frac{1}{3}\sin(2\,t). \end{aligned}$$

Transform. The intent of steps $\boxed{1}$ and $\boxed{2}$ is to transform the initial value problem into two equations in two unknowns. Used repeatedly in $\boxed{1}$ is integration by parts $\mathcal{L}(f') = s\mathcal{L}(f) - f(0)$. No Laplace tables were used. In $\boxed{2}$ the substitutions $x_1 = \mathcal{L}(x)$, $x_2 = \mathcal{L}(y)$ are made to produce two equations in the two unknowns x_1, x_2.

$$\begin{aligned} (2s^2 - s + 9)\mathcal{L}(x) + (-s^2 - s - 3)\mathcal{L}(y) &= 1 + 2s, \\ (2s^2 + s + 7)\mathcal{L}(x) + (-s^2 + s - 5)\mathcal{L}(y) &= 3 + 2s, \end{aligned} \qquad \boxed{1}$$

$$\begin{array}{lcccc} (2s^2 - s + 9)x_1 & + & (-s^2 - s - 3)x_2 & = & 1 + 2s, \\ (2s^2 + s + 7)x_1 & + & (-s^2 + s - 5)x_2 & = & 3 + 2s. \end{array} \qquad \boxed{2}$$

Step $\boxed{3}$ uses Cramer's rule. Equations $\boxed{2}$ are of the form $ax_1 + bx_2 = e$, $cx_1 + dx_2 = f$. Cramer's rule expresses answers x_1, x_2 by determinant fractions

$$x_1 = \frac{\begin{vmatrix} e & b \\ f & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}, \quad x_2 = \frac{\begin{vmatrix} a & e \\ c & f \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}.$$

The variable names x_1, x_2 stand for the Laplace integrals of the unknowns $x(t)$, $y(t)$, respectively. The answers, following a tedious calculation:

$$\left\{ \begin{aligned} x_1 &= \frac{s^2 + 2/3}{s^3 - s^2 + 4\,s - 4}, \\ x_2 &= \frac{10/3}{s^3 - s^2 + 4\,s - 4}. \end{aligned} \right. \qquad \boxed{3}$$

Step $\boxed{4}$ writes each fraction resulting from Cramer's rule as a partial fraction expansion suited for backward Laplace table look-up (details after $\boxed{6}$). Step $\boxed{5}$ does the table look-up and prepares for step $\boxed{6}$ to apply Lerch's cancellation law, in order to display the answers $x(t)$, $y(t)$.

$$\left\{\begin{aligned} x_1 &= \frac{1/3}{s-1} + \frac{2}{3}\frac{s}{s^2+4} + \frac{1}{3}\frac{2}{s^2+4}, \\ x_2 &= \frac{2/3}{s-1} - \frac{2}{3}\frac{s}{s^2+4} - \frac{1}{3}\frac{2}{s^2+4}. \end{aligned}\right. \qquad \boxed{4}$$

$$\left\{\begin{aligned} \mathcal{L}(x(t)) &= \mathcal{L}\left(\frac{1}{3}e^t + \frac{2}{3}\cos(2\,t) + \frac{1}{3}\sin(2\,t)\right), \\ \mathcal{L}(y(t)) &= \mathcal{L}\left(\frac{2}{3}e^t - \frac{2}{3}\cos(2\,t) - \frac{1}{3}\sin(2\,t)\right). \end{aligned}\right. \qquad \boxed{5}$$

$$\left\{\begin{aligned} x(t) &= \frac{1}{3}e^t + \frac{2}{3}\cos(2\,t) + \frac{1}{3}\sin(2\,t), \\ y(t) &= \frac{2}{3}e^t - \frac{2}{3}\cos(2\,t) - \frac{1}{3}\sin(2\,t). \end{aligned}\right. \qquad \boxed{6}$$

Partial fraction details. Shown below is how to obtain the expansion

$$\frac{s^2+2/3}{s^3-s^2+4\,s-4} = \frac{1/3}{s-1} + \frac{2}{3}\frac{s}{s^2+4} + \frac{1}{3}\frac{2}{s^2+4}.$$

The denominator $s^3 - s^2 + 4\,s - 4$ factors into $s-1$ times s^2+4. Partial fraction theory implies that there is an expansion with *real coefficients* A, B, C of the form

$$\frac{s^2+2/3}{(s-1)(s^2+4)} = \frac{A}{s-1} + \frac{Bs+C}{s^2+4}.$$

Let's verify $A = 1/3$, $B = 2/3$, $C = 2/3$. Clear the fractions to obtain the polynomial equation

$$s^2 + 2/3 = A(s^2+4) + (Bs+C)(s-1).$$

Instead of using $s = 1$ and $s = 2i$, which are roots of the denominator, invent samples $s = 1$, $s = 0$, $s = -1$ to get a *real* 3×3 system for A, B, C:

$$\begin{array}{lrcl} s=1: & 1+2/3 & = & A(1+4)+0, \\ s=0: & 0+2/3 & = & A(4)+C(-1), \\ s=-1: & 1+2/3 & = & A(1+4)+(-B+C)(-2). \end{array}$$

Write this system as an augmented matrix G with variables A, B, C assigned to the first three columns of G:

$$G = \left(\begin{array}{rrr|r} 5 & 0 & 0 & 5/3 \\ 4 & 0 & -1 & 2/3 \\ 5 & 2 & -2 & 5/3 \end{array}\right)$$

Using computer assist, calculate

$$\textbf{rref}(G) = \left(\begin{array}{rrr|r} 1 & 0 & 0 & 1/3 \\ 0 & 1 & 0 & 2/3 \\ 0 & 0 & 1 & 2/3 \end{array}\right)$$

Then A, B, C are the last column entries of $\textbf{rref}(G)$, which verifies the partial fraction expansion.

Heaviside cover-up detail. It is possible to rapidly check that $A = 1/3$ using the cover-up method. Less obvious is that the cover-up method also applies to the fraction with complex roots.

The idea is to multiply the fraction decomposition by s^2+4 to partially clear the fractions and then set $s^2+4=0$. This process formally sets s equal to one of the two roots $s=\pm 2i$. Complex numbers are avoided entirely by solving for B, C in the pair of equations

$$\frac{s^2+2/3}{s-1} = A(0) + (Bs+C), \quad s^2+4=0.$$

Because $s^2=-4$, the first equality is simplified to $\dfrac{-4+2/3}{s-1} = Bs+C$. Swap sides of the equation, then cross-multiply to obtain $Bs^2+Cs-Bs-C=-10/3$ and then use $s^2=-4$ again to simplify to $(-B+C)s+(-4B-C)=-10/3$. Because this linear equation in variable s has two solutions, then $-B+C=0$ and $-4B-C=-10/3$. Solve this 2×2 system by elimination to obtain $B=C=2/3$.

The algebraic method: First, find two equations in symbols s, B, C. Next, symbol s is eliminated to give two equations in symbols B, C. Finally, the 2×2 system for B, C is solved.

Exercises 8.4

Partial Fraction Mistakes

1. How many real constants appear in the partial fraction expansion of the fraction $\dfrac{s+1}{s^2(s+2)(s+3)^2}$?

2. How many real constants appear in the partial fraction expansion of $\dfrac{s+1}{s^2(s^2+4)(s^2+2s+5)^2}$?

3. Guido expanded $\dfrac{s+1}{s(s+2)(s+3)^2}$ to get $\dfrac{a}{s}+\dfrac{b}{s+2}+\dfrac{c}{(s+3)^2}$.
What is the mistake?

4. Helena made this expansion:
$$\frac{s+1}{s(s+2)} = \frac{a}{s}+\frac{b}{s+2}+\frac{c}{s+3}$$
The expansion is correct! Explain how you know that $c=0$ without computing anything.

This example explains why fractions on the right have denominators dividing the denominator on the left.

5. Marco made an expansion:
$$\frac{s+1}{s(s^2+4)} = \frac{a}{s}+\frac{b}{s+2}+\frac{c}{s-2}$$
Explain why it is a mistake.

This example explains why sanity checks have more than one item to check.

6. Violeta made an expansion
$$\frac{s+2}{s(s-2)(s+2)} = \frac{a}{s}+\frac{b}{s-2}+\frac{c}{s+2}$$
Explain why $c=0$ without computing anything.

This example explains why common factors of numerator and denominator should be removed.

7. Find the mistake in expansion
$$\frac{(s+2)^2}{s(s-2)} = \frac{a}{s}+\frac{b}{s-2}$$
This example explains why the degree of the numerator and denominator are checkpoints.

8. Is there a mistake here?
$$\frac{(s+2)^2}{s^2(s-2)} = \frac{a}{s}+\frac{b}{s^2}+\frac{c}{s-2}$$

Sampling Method

Apply the sampling method (a *failsafe method*) to verify the given equation.

9. $\dfrac{s}{s^2-1} = \dfrac{1/2}{s-1}+\dfrac{1/2}{s+1}$

10. $\dfrac{s}{s^4-1} = \dfrac{1/4}{s-1}+\dfrac{1/4}{s+1}+\dfrac{-s/2}{s^2+1}$

Method of Atoms

Apply the method of atoms to verify the given equation.

11. $\dfrac{2s}{s^2-1} = \dfrac{1}{s-1}+\dfrac{1}{s+1}$

12. $\dfrac{4s}{s^4-1}=\dfrac{1}{s-1}+\dfrac{1}{s+1}+\dfrac{-2s}{s^2+1}$

Heaviside's 1890 Shortcut

Apply Heaviside's shortcut to verify the given equation.

13. $\dfrac{2s}{s^2-4}=\dfrac{1}{s-2}+\dfrac{1}{s+2}$

14. $\dfrac{s+4}{s^3+4s}=\dfrac{1}{s}+\dfrac{-s+1}{s^2+4}$

Residues and Poles

Compute the residue for the given pole.

15. Residue at $s=2$ for $\dfrac{2s}{s^2-4}$.

16. Residue at $s=0$ for $\dfrac{s+4}{s^3+16s}$.

Scalar Differential Equations

The **transfer function** of $x''+x=f(t)$ is $H(s)=\frac{1}{s^2+1}$. A common definition is $H(s)=\mathcal{L}(f(t))$ divided by $\mathcal{L}(x(t))$, assuming $x(0)=x'(0)=0$.

17. Verify for $x''+x=e^{-t}$ with $x(0)=0$, $x'(0)=0$ that $\mathcal{L}(x)=\frac{1}{s+1}\frac{1}{s^2+1}$. Then compute $H(s)$.

18. Explain the transfer function equation
$H(s)=\frac{1}{\text{characteristic equation}}$.

19. Solve $\mathcal{L}(x(t))=\frac{1}{s+1}\frac{1}{s^2+1}$ by Heaviside cover–up for output $x(t)=\frac{1}{2}(e^{-t}-\cos t+\sin t)$.

20. Given $x''+x=te^{-t}$, $x(0)=x'(0)=0$, show all steps to find
$\mathcal{L}(x(t))=\frac{1}{(s+1)^2}\frac{1}{s^2+1}$.

First Order System

Using Example 8.29 as a guide, solve the system for $x_1(t)$ by Laplace's method.

21. $\begin{cases} x_1'=x_2, \\ x_2'=4x_1+12e^{-t}, \\ x_1(0)=x_2(0)=0. \end{cases}$
Ans: $x_1(t)=e^{2t}+3e^{-2t}-4e^{-t}$.

22. $\begin{cases} x_1'=x_2, \\ x_2'=x_3, \\ x_3'=4x_1-4x_2+x_3+10e^{-t}, \\ x_1(0)=x_2(0)=x_3(0)=0. \end{cases}$
Ans: $x_1(t)=e^{t}-e^{-t}-\sin(2t)$.

Second Order System

Using Example 8.29 as a guide, compute $x(t), y(t)$.

23. $\mathcal{L}(x(t))=\frac{3s^2+2}{(s-1)(s^2+4)}$,
$\mathcal{L}(y(t))=\frac{10}{(s-1)(s^2+4)}$.
Ans: $x=2\cos(2t)+\sin(2t)+e^{t}$,
$y=-2\cos(2t)-\sin(2t)+2e^{t}$

24. $\mathcal{L}(x(t))=\frac{2s^2+4}{(s+1)(s^2+1)}$,
$\mathcal{L}(y(t))=\frac{2}{(s+1)(s^2+1)}$.
Ans: $x=-\cos(t)+\sin(t)+3e^{-t}$,
$y=-\cos(t)+\sin(t)+e^{-t}$.

8.5 Transform Properties

Collected here are the major theorems for the manipulation of Laplace transform tables along with their derivations. Those who study in isolation are advised to dwell on the details of proof and re-read the examples of preceding sections. No exercises are appropriate and none are supplied.

Theorem 8.5 (Linearity)
The Laplace transform has these inherited integral properties:

$$\begin{array}{ll}\textbf{(a)} & \mathcal{L}(f(t)+g(t)) = \mathcal{L}(f(t)) + \mathcal{L}(g(t)),\\ \textbf{(b)} & \mathcal{L}(cf(t)) = c\mathcal{L}(f(t)).\end{array}$$

Theorem 8.6 (The t-Derivative Rule or Parts Rule)
Let $y(t)$ be continuous, of exponential order and let $y'(t)$ be piecewise continuous on $t \ge 0$. Then $\mathcal{L}(y'(t))$ exists and

$$\mathcal{L}(y'(t)) = s\mathcal{L}(y(t)) - y(0+).$$

Theorem 8.7 (The t-Integral Rule)
Let $g(t)$ be of exponential order and continuous for $t \ge 0$. Then

$$\mathcal{L}\left(\int_0^t g(x)\,dx\right) = \frac{1}{s}\mathcal{L}(g(t))$$

or equivalently

$$\mathcal{L}(g(t)) = s\mathcal{L}\left(\int_0^t g(x)\,dx\right)$$

Theorem 8.8 (The s-Differentiation Rule)
Let $f(t)$ be of exponential order. Then

$$\mathcal{L}(tf(t)) = -\frac{d}{ds}\mathcal{L}(f(t)).$$

Theorem 8.9 (First Shifting Rule)
Let $f(t)$ be of exponential order and $-\infty < a < \infty$. Then

$$\mathcal{L}(e^{at}f(t)) = \left.\mathcal{L}(f(t))\right|_{s\to(s-a)}.$$

Theorem 8.10 (Second Shifting Rule)
Let $f(t)$ and $g(t)$ be of exponential order and assume $a \ge 0$. Then

$$\begin{array}{ll}\textbf{(a)} & \mathcal{L}(f(t-a)H(t-a)) = e^{-as}\mathcal{L}(f(t)),\\ \textbf{(b)} & \mathcal{L}(g(t)H(t-a)) = e^{-as}\mathcal{L}(g(t+a)).\end{array}$$

Theorem 8.11 (Periodic Function Rule)
Let $f(t)$ be of exponential order and satisfy $f(t+P)=f(t)$. Then

$$\mathcal{L}(f(t)) = \frac{\int_0^P f(t)e^{-st}dt}{1-e^{-Ps}}.$$

Theorem 8.12 (Convolution Rule)
Let $f(t)$ and $g(t)$ be of exponential order. Then

$$\mathcal{L}(f(t))\mathcal{L}(g(t)) = \mathcal{L}\left(\int_0^t f(x)g(t-x)dx\right).$$

Theorem 8.13 (Laplace at Infinity is Zero)
Let $f(t)$ be of piecewise continuous and of exponential order. Then

$$\lim_{s\to\infty} \mathcal{L}(f(t)) = 0.$$

Theorem 8.14 (Initial and Final Value Rules)
Let $f(t)$ and $f'(t)$ be functions of exponential order. Then, when all indicated limits exist,

$$\begin{array}{llcll} \textbf{1}. & f(0+) & = & \lim_{t\to 0+} f(t) & = & \lim_{s\to\infty} s\mathcal{L}(f(t)), \\ \textbf{2}. & f(\infty) & = & \lim_{t\to\infty} f(t) & = & \lim_{s\to 0} s\mathcal{L}(f(t)). \end{array}$$

Initial and Final Value Pitfalls

In Theorem 8.14, impulses and higher order singularities at $t=0$ are disallowed, because hypotheses require $s\mathcal{L}(f(t))$ to be bounded.

Examples $f(t)=\sin t$ and $f(t)=e^t$ don't satisfy hypotheses for **2** because $f(\infty)$ is undefined, but **1** applies for both examples.

A **pole**, defined precisely on page 628, is a root of the denominator in a fraction $F(s)=\mathcal{L}(f(t))$. The location of the poles influences the possibility of using Theorem 8.14:

> If there are poles in the right s-plane, then $f(t)$ will contain exponentially growing terms, which implies $f(\infty)$ does not exist.
>
> If there are pairs of complex conjugate poles on the imaginary axis, then $f(t)$ will contain sinusoids and $f(\infty)$ is not defined.
>
> Poles in the left s-plane contribute exponentially decaying terms to $f(t)$ which do not affect the final value.
>
> Signal $f(t)$ has possibly a constant final value, the steady state of the signal, only when there are poles at the origin of the s-plane.

Proofs and Details

Proof of Theorem 8.5 (Linearity):

$$\begin{aligned}
\text{LHS} &= \mathcal{L}(f(t)+g(t)) && \text{Left side of the identity in (a).}\\
&= \textstyle\int_0^\infty (f(t)+g(t))e^{-st}dt && \text{Direct transform.}\\
&= \textstyle\int_0^\infty f(t)e^{-st}dt + \int_0^\infty g(t)e^{-st}dt && \text{Calculus integral rule.}\\
&= \mathcal{L}(f(t)) + \mathcal{L}(g(t)) && \text{Equals RHS; identity (a) verified.}\\
\text{LHS} &= \mathcal{L}(cf(t)) && \text{Left side of the identity in (b).}\\
&= \textstyle\int_0^\infty cf(t)e^{-st}dt && \text{Direct transform.}\\
&= c\textstyle\int_0^\infty f(t)e^{-st}dt && \text{Calculus integral rule.}\\
&= c\mathcal{L}(f(t)) && \text{Equals RHS; identity (b) verified.}
\end{aligned}$$

Proof of Theorem 8.6 (t-Derivative or parts rule): Already $\mathcal{L}(f(t))$ exists, because f is of exponential order and continuous. On an interval $[a,b]$ where f' is continuous, integration by parts using $u=e^{-st}$, $dv=f'(t)dt$ gives

$$\begin{aligned}
\textstyle\int_a^b f'(t)e^{-st}dt &= f(t)e^{-st}\big|_{t=a}^{t=b} - \textstyle\int_a^b f(t)(-s)e^{-st}dt\\
&= -f(a)e^{-sa} + f(b)e^{-sb} + s\textstyle\int_a^b f(t)e^{-st}dt.
\end{aligned}$$

On any interval $[0,N]$, there are finitely many intervals $[a,b]$ on each of which f' is continuous. Add the above equality across these finitely many intervals $[a,b]$. The boundary values on adjacent intervals match and the integrals add to give

$$\int_0^N f'(t)e^{-st}dt = -f(0+)e^0 + f(N)e^{-sN} + s\int_0^N f(t)e^{-st}dt.$$

Take the limit across this equality as $N\to\infty$. Then the right side has limit $-f(0)+s\mathcal{L}(f(t))$, because of the existence of $\mathcal{L}(f(t))$ and $\lim_{t\to\infty} f(t)e^{-st}=0$ for large s. Therefore, the left side has a limit, and by definition $\mathcal{L}(f'(t))$ exists and $\mathcal{L}(f'(t)) = -f(0)+s\mathcal{L}(f(t))$.

Proof of Theorem 8.7 (t-Integral rule): Let $f(t)=\int_0^t g(x)dx$. Then f is of exponential order and continuous. The details:

$$\begin{aligned}
\mathcal{L}(\textstyle\int_0^t g(x)dx) &= \mathcal{L}(f(t)) && \text{By definition.}\\
&= \frac{1}{s}\mathcal{L}(f'(t)) && \text{Because } f(0)=0 \text{ implies } \mathcal{L}(f'(t)) = s\mathcal{L}(f(t)).\\
&= \frac{1}{s}\mathcal{L}(g(t)) && \text{Because } f'=g \text{ by the Fundamental theorem of calculus.}
\end{aligned}$$

Proof of Theorem 8.8 (s-Differentiation): Let's prove $\mathcal{L}((-t)f(t)) = (d/ds)\mathcal{L}(f(t))$, an equivalent relation. If f is of exponential order, then so is $(-t)f(t)$, therefore $\mathcal{L}((-t)f(t))$ exists. It remains to show the s-derivative exists and satisfies the given equality.

The proof below is based in part upon the calculus inequality

$$\left|e^{-x}+x-1\right| \le x^2, \quad x\ge 0. \tag{1}$$

The inequality is obtained from two applications of the *mean value theorem* $g(b)-g(a) = g'(\overline{x})(b-a)$, which gives $e^{-x}+x-1 = x\overline{x}e^{-x_1}$ with $0 \le x_1 \le \overline{x} \le x$.

In addition, the existence of $\mathcal{L}(t^2|f(t)|)$ is used to define $s_0 > 0$ such that $\mathcal{L}(t^2|f(t)|) \le 1$ for $s > s_0$. This follows from the transform existence theorem for functions of exponential order, where it is shown that the transform has limit zero at $s = \infty$. See also the proof of Theorem 8.13.

Consider $h \ne 0$ and the Newton quotient $Q(s,h) = (F(s+h)-F(s))/h$ for the s-derivative of the Laplace integral. We have to show that

$$\lim_{h\to 0} |Q(s,h) - \mathcal{L}((-t)f(t))| = 0.$$

This will be accomplished by proving for $s > s_0$ and $s+h > s_0$ the inequality

$$|Q(s,h) - \mathcal{L}((-t)f(t))| \le |h|.$$

For $h \ne 0$,

$$Q(s,h) - \mathcal{L}((-t)f(t)) = \int_0^\infty f(t)\frac{e^{-st-ht} - e^{-st} + the^{-st}}{h}\,dt.$$

Assume $h > 0$. Due to the exponential rule $e^{A+B} = e^A e^B$, the quotient in the integrand simplifies to give

$$Q(s,h) - \mathcal{L}((-t)f(t)) = \int_0^\infty f(t)e^{-st}\left(\frac{e^{-ht}+th-1}{h}\right)dt.$$

Inequality (1) applies with $x = ht \ge 0$, giving

$$|Q(s,h) - \mathcal{L}((-t)f(t))| \le |h|\int_0^\infty t^2|f(t)|e^{-st}dt.$$

The right side is $|h|\mathcal{L}(t^2|f(t)|)$, which for $s > s_0$ is bounded by $|h|$, completing the proof for $h > 0$. If $h < 0$, then a similar calculation is made to obtain

$$|Q(s,h) - \mathcal{L}((-t)f(t))| \le |h|\int_0^\infty t^2|f(t)|e^{-st-ht}dt.$$

The right side is $|h|\mathcal{L}(t^2|f(t)|)$ evaluated at $s+h$ instead of s. If $s+h > s_0$, then the right side is bounded by $|h|$, completing the proof for $h < 0$.

Proof of Theorem 8.9 (First Shifting Rule): The left side LHS of the equality can be written because of the exponential rule $e^A e^B = e^{A+B}$ as

$$\text{LHS} = \int_0^\infty f(t)e^{-(s-a)t}dt.$$

This integral is $\mathcal{L}(f(t))$ with s replaced by $s-a$, which is precisely the meaning of the right side RHS of the equality. Therefore, LHS = RHS.

Proof of Theorem 8.10 (Second Shifting Rule): The details for (a) are

$$\begin{aligned} \text{LHS} &= \mathcal{L}(H(t-a)f(t-a)) \\ &= \textstyle\int_0^\infty H(t-a)f(t-a)e^{-st}dt && \text{Direct transform.} \end{aligned}$$

$$
\begin{aligned}
&= \int_a^\infty H(t-a)f(t-a)e^{-st}dt && \text{Because } a \geq 0 \text{ and } H(x)=0 \text{ for } x<0.\\
&= \int_0^\infty H(x)f(x)e^{-s(x+a)}dx && \text{Change variables } x=t-a,\ dx=dt.\\
&= e^{-sa}\int_0^\infty f(x)e^{-sx}dx && \text{Use } H(x)=1 \text{ for } x \geq 0.\\
&= e^{-sa}\mathcal{L}(f(t)) && \text{Direct transform.}\\
&= \text{RHS} && \text{Identity (a) verified.}
\end{aligned}
$$

In the details for (b), let $f(t)=g(t+a)$, then

$$
\begin{aligned}
\text{LHS} &= \mathcal{L}(H(t-a)g(t)) && \\
&= \mathcal{L}(H(t-a)f(t-a)) && \text{Use } f(t-a)=g(t-a+a)=g(t).\\
&= e^{-sa}\mathcal{L}(f(t)) && \text{Apply (a).}\\
&= e^{-sa}\mathcal{L}(g(t+a)) && \text{Because } f(t)=g(t+a).\\
&= \text{RHS} && \text{Identity (b) verified.}
\end{aligned}
$$

Proof of Theorem 8.11 (Periodic Function Rule):

$$
\begin{aligned}
\text{LHS} &= \mathcal{L}(f(t)) && \\
&= \int_0^\infty f(t)e^{-st}dt && \text{Direct transform.}\\
&= \sum_{n=0}^\infty \int_{nP}^{nP+P} f(t)e^{-st}dt && \text{Additivity of the integral.}\\
&= \sum_{n=0}^\infty \int_0^P f(x+nP)e^{-sx-nPs}dx && \text{Change variables } t=x+nP.\\
&= \sum_{n=0}^\infty e^{-nPs}\int_0^P f(x)e^{-sx}dx && \text{Because } f(x) \text{ is } P\text{–periodic and } e^Ae^B = e^{A+B}.\\
&= \int_0^P f(x)e^{-sx}dx \sum_{n=0}^\infty r^n && \text{The summation has a common factor. Define } r=e^{-Ps}.\\
&= \int_0^P f(x)e^{-sx}dx \frac{1}{1-r} && \text{Sum the geometric series.}\\
&= \frac{\int_0^P f(x)e^{-sx}dx}{1-e^{-Ps}} && \text{Substitute } r=e^{-Ps}.\\
&= \text{RHS} && \text{Periodic function identity verified.}
\end{aligned}
$$

Left unmentioned here is the convergence of the infinite series on line 3 of the proof, which follows from f of exponential order.

Proof of Theorem 8.12 (Convolution rule): The details use Fubini's integration interchange theorem for a planar unbounded region, and therefore this proof involves advanced calculus methods that may be outside the background of the reader. Modern calculus texts contain a less general version of Fubini's theorem for finite regions, usually referenced as *iterated integrals*. The unbounded planar region is written in two ways:

$$
\begin{aligned}
\mathbf{D} &= \{(r,t) : t \leq r < \infty,\ 0 \leq t < \infty\},\\
\mathcal{D} &= \{(r,t) : 0 \leq r < \infty,\ 0 \leq r \leq t\}.
\end{aligned}
$$

Readers should pause here and verify that $\mathbf{D}=\mathcal{D}$.

The change of variable $r=x+t$, $dr=dx$ is applied for fixed $t \geq 0$ to obtain the identity

$$
(2)\qquad
\begin{aligned}
e^{-st}\int_0^\infty g(x)e^{-sx}dx &= \int_0^\infty g(x)e^{-sx-st}dx\\
&= \int_t^\infty g(r-t)e^{-rs}dr.
\end{aligned}
$$

The left side of the convolution identity is expanded as follows:

$$\begin{aligned}
\mathsf{LHS} &= \mathcal{L}(f(t))\mathcal{L}(g(t)) && \\
&= \textstyle\int_0^\infty f(t)e^{-st}dt \int_0^\infty g(x)e^{-sx}dx && \text{Direct transform.} \\
&= \textstyle\int_0^\infty f(t) \int_t^\infty g(r-t)e^{-rs}drdt && \text{Apply identity (2).} \\
&= \textstyle\int_{\mathbf{D}} f(t)g(r-t)e^{-rs}drdt && \text{Fubini's theorem applied.} \\
&= \textstyle\int_{\mathcal{D}} f(t)g(r-t)e^{-rs}drdt && \text{Descriptions } \mathbf{D} \text{ and } \mathcal{D} \text{ are the same.} \\
&= \textstyle\int_0^\infty \int_0^r f(t)g(r-t)dte^{-rs}dr && \text{Fubini's theorem applied.}
\end{aligned}$$

Then

$$\begin{aligned}
\mathsf{RHS} &= \mathcal{L}\left(\textstyle\int_0^t f(u)g(t-u)du\right) && \\
&= \textstyle\int_0^\infty \int_0^t f(u)g(t-u)due^{-st}dt && \text{Direct transform.} \\
&= \textstyle\int_0^\infty \int_0^r f(u)g(r-u)due^{-sr}dr && \text{Change variable names } r \leftrightarrow t. \\
&= \textstyle\int_0^\infty \int_0^r f(t)g(r-t)dt\, e^{-sr}dr && \text{Change variable names } u \leftrightarrow t. \\
&= \mathsf{LHS} && \text{Convolution identity verified.}
\end{aligned}$$

Proof of Theorem 8.13 (Laplace at Infinity is Zero): Assumed is an inequality $|f(t)| \le Me^{kt}$ for some constants $M \ge 0$ and k. Then

$$\left|\int_0^N f(t)e^{-st}dt\right| \le \int_0^N |f(t)|e^{-st}dt \le M\int_0^N e^{-(s-k)t}dt.$$

The integral on the right is estimated for $s > k$ and all $N \ge 0$ as follows:

$$\int_0^N e^{-(s-k)t}dt = \frac{1-e^{-(s-k)N}}{s-k} \le \frac{1}{s-k}.$$

Limiting as $N \to \infty$ across the chained inequalities implies the fundamental estimate $|\mathcal{L}(f(t))| \le \frac{M}{s-k}$, $s > k$, therefore $\lim_{s\to\infty} \mathcal{L}(f(t)) = 0$.

Proof of Theorem 8.14 (Initial and Final Values):

$\boxed{1}$: Write $\mathcal{L}(f'(t))$ in two ways: (1) $\mathcal{L}(f'(t)) = s\mathcal{L}(f(t)) - f(0+)$ using the parts formula, and (2) $\mathcal{L}(f'(t)) = \int_0^\infty f'(t)e^{-st}dt$, using the direct Laplace definition.

A high-powered calculus theorem is needed to tell us that the integral on the right in (2) has limit as $s \to 0+$ equal to $\int_0^\infty f'(t)(1)dt = f(t)|_{t=0}^{t=\infty} = f(\infty) - f(0+)$. The needed result is *Lebesgue's Bounded Convergence Theorem*, which says that under certain conditions (met by the assumed hypotheses here) $\lim_{n\to\infty} \int_0^\infty f_n(t)dt = \int_0^\infty (\lim_{n\to\infty} f_n(t))\,dt$. We take $f_n(t) = f'(t)e^{-s_n t}$ where $\{s_n\}$ is any sequence of positive numbers with limit zero.

Assembling the two ways to write $\mathcal{L}(f'(t))$ implies $\lim_{s\to 0+} s\,\mathcal{L}(f(f)) - f(0) = f(\infty) - f(0+)$. Cancel $f(0+)$ from each side of this identity. Then $\lim_{s\to 0+} s\,\mathcal{L}(f(t)) = f(\infty)$.

$\boxed{2}$: Theorem 8.13 implies $\mathcal{L}(f'(t))$ has limit zero as $s \to \infty$. Limit as $s \to \infty$ across the parts formula $\mathcal{L}(f'(t)) = s\mathcal{L}(f(t)) - f(0+)$ to obtain the limit $\lim_{s\to\infty} s\mathcal{L}(f(t)) - f(0+) = 0$, which is the claimed identity.

Exercises 8.5

There are no exercises for this section. The content is exclusively statements of theorems and proofs, for the following theorems.

Linearity

The t-Derivative Rule or Parts Rule

The t-Integral Rule

The s-Differentiation Rule

First Shifting Rule

Second Shifting Rule

Periodic Function Rule

Convolution Rule

Initial and Final Value Rules

8.6 Heaviside Step and Dirac Impulse

Heaviside Function

The **unit step function** $\mathbf{u}(t)$ is distinguished from the more precise clone called the **Heaviside function** $H(t)$, which is undefined at $t = 0$. The definitions:

$$\mathbf{u}(t) = \begin{cases} 1 & \text{for} \quad t > 0, \\ 1 & \text{for} \quad t = 0, \\ 0 & \text{for} \quad t < 0, \end{cases} \qquad H(t) = \begin{cases} 1 & \text{for} \quad t > 0, \\ \text{undefined} & \text{for} \quad t = 0, \\ 0 & \text{for} \quad t < 0. \end{cases}$$

Functions $1, \mathbf{u}(t), H(t)$ agree for $t > 0$ because all functions in Laplace theory are assumed zero for $t < 0$.

An often–used formula involving the unit step function is the **characteristic function** of the interval $a \le t < b$, or **unit pulse**:

$$\text{(1)} \qquad \mathbf{pulse}(t, a, b) = \mathbf{u}(t - a) - \mathbf{u}(t - b) = \begin{cases} 1 & a \le t < b, \\ 0 & \text{otherwise.} \end{cases}$$

To illustrate, a square wave $\mathsf{sqw}(t) = (-1)^{\mathsf{floor}(t)}$ can be written in the series form $\sum_{n=0}^{\infty} (-1)^n \, \mathbf{pulse}(t, n, n+1)$ as a **pulse train**.[9]

Trouble at $t = 0$. Computer algebra systems like `maple` distinguish between the piecewise-defined unit step function and the Heaviside function. The Heaviside function $H(t)$ is left undefined at $t = 0$, whereas the unit step is defined everywhere. This seemingly minor distinction makes more sense when taking formal derivatives. On the domain $t \ne 0$ of H, the ordinary calculus derivative dH/dt is defined and equals zero. In contrast, $\mathbf{u}(t)$ on its domain $-\infty < t < \infty$ fails to have a derivative at one point: $t = 0$.

Fundamental Theorem of Calculus. Calculus rule $\int_a^b f'(t)dt = f(b) - f(a)$ fails for $f = H$, due to integrand $f'(t) = dH/dt = 0$. Riemann and Stieltjes filled the gap in the theory by providing a new definition of integral and corresponding theory of integration, these days called **Riemann–Stieltjes Integration**. In their theory, integral $\int_a^b \frac{dH}{dt}\,dt$ makes sense and $\int_a^b \frac{dH}{dt}\,dt = H(b) - H(a)$. Riemann–Stieltjes theory will be used to explain a contribution of Paul Dirac (1902-1984) to Laplace theory called the **Dirac impulse**, denoted δ in the literature.

Dirac Impulse

Following the 1932 work of Paul A. M. Dirac the definition should be

$$\delta(t) = \frac{d}{dt}\mathbf{u}(t).$$

[9] A square wave resembles a train of boxcars.

One year after Dirac introduced impulse δ, he received the Nobel Prize in physics for his quantum theory work. Laurent Schwartz later justified mathematically the use of δ.

A precise mathematical definition of the Dirac impulse is $\delta(t) = d\,\mathbf{u}(t)$, where $\mathbf{u}(t)$ is the unit step and $d\,\mathbf{u}(t)/dt$ has meaning under the integral sign in a Riemann-Stieltjes integral. This definition restrains $d\,\mathbf{u}(t)$ to have meaning only under an integral sign. It is in this sense that the Dirac impulse δ is defined.[10]

Dirac Impulse in Applications. What is the meaning of the differential equation

$$x'' + 16x = 5\delta(t - t_0)?$$

The equation $x''+16x = f(t)$ represents an undamped spring-mass system having Hooke's constant 16, subject to external force $f(t)$. In a mechanical context, the Dirac impulse term $5\delta(t - t_0)$ is an *idealization* of a hammer-hit at time $t = t_0 > 0$ with impulse 5. The hammer-hit **injects energy** into the system almost instantaneously.

Forcing term $f(t)$ in $x'' + 16x = f(t)$ can be formally written as a Riemann-Stieltjes integrator $5\,d\,\mathbf{u}(t - t_0)$ where $\mathbf{u}$ is the unit step function: $\mathbf{u}(t) = 1$ on $t \ge 0$, else $\mathbf{u}(t) = 0$.

The Dirac impulse or *derivative of the unit step*, nonsensical as it may appear, is realized in applications via the two–sided or central difference quotient $\dfrac{\mathbf{u}(t+h) - \mathbf{u}(t-h)}{2h} \approx d\,\mathbf{u}(t)$. Given t_0, let $a = t_0 - h$, $b = t_0 + h$ for $h > 0$ very small. A *simplistic approximation* for ideal impulse $5\delta(t - t_0)$ is given by the central difference approximation

$$\begin{aligned} \frac{5}{2h}\begin{cases} 1 & a \le t < b \\ 0 & \text{else} \end{cases} &= 5\frac{\mathbf{u}(t-a) - \mathbf{u}(t-b)}{b-a} \\ &= \frac{5}{b-a}\,\mathbf{pulse}(t, a, b) \end{aligned}$$

The *impulse*[11] of the actual force f is therefore approximated by

$$\int_{-\infty}^{\infty} f(t)\,dt \approx 5\int_a^b \frac{1}{b-a}\,\mathbf{pulse}(t, a, b)\,dt = 5,$$

due to the integrand being $1/(b - a)$ on $a \le t < b$ and otherwise 0.

[10] The definition of the Dirac Impulse by Laurent Schwartz uses Lebesgue integration theory. In differential equations applications, the Riemann–Stieltjes definition $\delta(t) = d\,\mathbf{u}(t)$ suffices, with an unremarkable quantity of exceptions. The presentation here requires a calculus background but no Lebesgue theory background.

[11] Momentum is defined to be mass times velocity. If the force f is given by Newton's law as $f(t) = \frac{d}{dt}(mv(t))$ and $v(t)$ is velocity, then $\int_a^b f(t)dt = mv(b) - mv(a)$ is the net momentum or impulse on $[a, b]$.

Modeling Impulses

One argument for the Dirac impulse idealization is that an infinity of choices exist for modeling an impulse. There are in addition to the central difference quotient two other popular difference quotients, the forward quotient $(\mathbf{u}(t+h)-\mathbf{u}(t))/h$ and the backward quotient $(\mathbf{u}(t)-\mathbf{u}(t-h))/h$ ($h>0$ assumed). In reality, h is unknown in any application, and the impulsive force of a hammer hit is hardly constant, as is supposed by this naive modeling.

The modeling logic often applied for the Dirac impulse is that the external force $f(t)$ will be used in the model in a limited manner, in which only the momentum $p=mv$ is important. More precisely, only the change in momentum or impulse is important, $\int_a^b f(t)dt = \Delta p = mv(b) - mv(a)$.

The precise force $f(t)$ is replaced during the modeling by a simplistic piecewise–defined force that has exactly the same impulse Δp. The replacement is justified by arguing that if only the impulse is important, and not the actual details of the force, then both models should give similar results. Most of the intuition for this modeling magic comes from investigation of two models: (1) $f(t)$ is piecewise–defined and depends on h, (2) $f(t)$ is an idealized Dirac impulse. The result of the investigation is that answers from (1) converge as $h\to 0$ to the single idealized answer from (2).

Impulses in Differential Equations. In Laplace theory, there is a natural encounter with Dirac's ideas, because $\mathcal{L}(f(t))$ routinely appears on the right of the equation after transformation. Let $a=t_0-h, b=t_0+h$ for small $h>0$. Then $2h=b-a$. Assume $t_0>0$ and $t_0-h>0$ for the purpose of illustration. If the input $f(t)$ is a simplistic impulsive force of impulse c, then representation $f(t)=\frac{c}{b-a}\,\mathbf{pulse}(t,a,b)$ permits a direct computation of the impulse:

$$\text{impulse} = \int_{-\infty}^{\infty} f(t)dt = \int_a^b \frac{c}{b-a}\,\mathbf{pulse}(t,a,b)dt = c.$$

The Laplace integral $\mathcal{L}(f(t))$ evaluates as follows:

$$\begin{aligned}
\mathcal{L}(f(t)) &= \textstyle\int_0^\infty f(t)e^{-st}dt \\
&= \textstyle\int_a^b \frac{c}{b-a}\,\mathbf{pulse}(t,a,b)e^{-st}dt \\
&= \textstyle\int_a^b \frac{c}{b-a}(1)e^{-st}dt \\
&= \textstyle\frac{c}{b-a}\left(\frac{e^{-sa}-e^{-sb}}{-s}\right) && \boxed{1} \\
&= c\,\textstyle\frac{e^{-sh}-e^{sh}}{2sh}\,e^{-s\,t_0} && \boxed{2} \\
&\approx ce^{-s\,t_0} && \boxed{3}
\end{aligned}$$

$\boxed{1}$: Factor the constant $\frac{c}{b-a}$ outside the integral, use the definition of pulse, then integrate the exponential.

$\boxed{2}$: Replace $a=t_0-h, b=t_0+h$, simplify, then collect denominator factors s

and $2h = b - a$.
$\boxed{\mathbf{3}}$: As $h \to 0$, factor $\frac{e^{sh} - e^{-sh}}{2sh}$ converges to 1 because of L'Hôspital's rule applied to $\frac{e^x - e^{-x}}{2x}$ at $x = 0$.

Mathematical Flaw. The immediate naive modeling conclusion is that the simplistic impulsive force f should be replaced by an equivalent one f^* such that

$$\mathcal{L}(f^*(t)) = c\, e^{-s\, t_0}.$$

Unfortunately, *there is no such function* f^*!

The apparent mathematical flaw in this idea was resolved by the work of Laurent Schwartz on **distributions**. In short, there is a solid foundation for introducing f^*, but unfortunately the mathematics involved is not elementary nor especially accessible to those readers whose background is just calculus.[12] The theory of distributions provides a resolution of the mathematical flaw:

$$\mathcal{L}(c\,\delta(t - t_0)) = c\, e^{-s\, t_0}.$$

It is mistake to write $f^*(t) = c\,\delta(t - t_0)$ and call f^* *a function*, because it is not. Expression f^* makes sense only under an integral sign.

Function or Operator? The work of physics Nobel prize winner Paul Dirac (1902–1984) proceeded for about 15 years before the mathematical community developed a sound mathematical theory for his impulsive force representations. A systematic theory was developed in 1936 by the Soviet mathematician S. Sobolev. The French mathematician Laurent Schwartz further developed the theory in 1945. He observed that the idealization δ is not a function but an operator or *linear functional*, in particular, δ maps or *associates* to each function $\phi(t)$ its value at $t = 0$, in short, $\delta(\phi) = \phi(0)$. This fact was observed early on by Dirac and others, during the replacement of simplistic forces by δ.

Laplace Theory and the Dirac Impulse. When Laplace theory manipulates the Dirac impulse δ, it does so by obeying the *under the integral sign rule.* The good news is that answers can be calculated formally, as though f^* was a function. What are we to do when applying the formal rules? We think of $\delta(t - t_0)$ as a simplistic impulse given on an interval $[a, b]$ that shrinks to t_0. A simplistic impulse *is a function*! Laplace theory provides a transition from simplistic impulse modeling to *idealized* Dirac impulse.

[12] Practising engineers and scientists might be able to ignore the vast literature on distributions, citing the example of physicist Paul Dirac, who succeeded in applying impulsive force ideas without the distribution theory developed by S. Sobolev and L. Schwartz. Those who wish to read current literature on partial differential equations have no such luxury, because the work on distributions has forever changed the required background for reading new literature.

Properties of the Dirac Impulse

Theorem 8.15 (Fundamental Identities for Dirac δ)
Let $\mathbf{u}(t)$ denote the unit step function. Define $\delta(t) = d\,\mathbf{u}(t)$ as a Riemann–Stieltjes integrator. Let $g(t)$ be piecewise continuous and $a \ge 0$. Then

(1) $\displaystyle\int_{-\infty}^{\infty} \delta(t)dt = 1,$ meaning $\int_{-\infty}^{\infty} d\,\mathbf{u}(t) = 1$

(2) $\displaystyle\int_{-\infty}^{\infty} g(t)\delta(t-a)dt = g(a+),$ meaning $\int_{-\infty}^{\infty} g(t)d\,\mathbf{u}(t-a) = g(a+)$

(3) $\mathcal{L}(\delta(t-a)) = e^{-s\,a},$ meaning $\int_0^{\infty} e^{-st}\,d\,\mathbf{u}(t-a) = e^{-s\,a}.$

Proof:
Symbol $g(a+)$ means $\lim_{h\to 0+} g(a+h)$, the right–hand limit at $t = a$.

Property **(1)** follows from property **(2)** by choosing $g(t) = \mathbf{u}(t)$ and $a = 0$.

Property **(3)** follows from property **(2)** by choosing $g(t) = \mathbf{u}(t)$.

Details **(2)**: The *definition* of the Dirac impulse is a formal one, in which every occurrence of $\delta(t-a)dt$ under an integrand is replaced by $d\,\mathbf{u}(t-a)$. The differential symbol $d\,\mathbf{u}(t-a)$ is taken in the sense of the Riemann-Stieltjes integral, which is defined in Rudin [Rudin] for monotonic integrators $\alpha(x)$ as

$$\int_a^b f(x)d\alpha(x) = \lim_{N\to\infty} \sum_{n=1}^{N} f(x_n)(\alpha(x_n) - \alpha(x_{n-1})).$$

Required in the definition: $x_0 = a$, $x_N = b$ and $x_0 < x_1 < \cdots < x_N$ forms a partition of $[a, b]$ whose mesh $\max\{|x_j - x_{j-1}| : 1 \le j \le N\}$ approaches zero as $N \to \infty$. Used exclusively here is nondecreasing integrator $\alpha = \mathbf{u}$, the unit step.

Steps below verify that the left and right sides in **(2)** are equal.

$\text{LHS} = \mathcal{L}(g(t)\delta(t-a))$	Left side of **(2)**.
$= \int_0^{\infty} g(t)e^{-st}\delta(t-a)dt$	Laplace integral, $a \ge 0$ assumed.
$= \int_0^{\infty} g(t)e^{-st}d\,\mathbf{u}(t-a)$	Replace $\delta(t-a)dt$ by $d\,\mathbf{u}(t-a)$.
$= \lim_{M\to\infty} \int_0^{M} g(t)e^{-st}d\,\mathbf{u}(t-a)$	Definition of improper integral.
$= g(a)e^{-sa}$	Explained below.
$= \text{RHS}$	Property **(2)** verified.

To explain the last step, apply the definition of the Riemann-Stieltjes integral to $\alpha = \mathbf{u}$ with given partition $0 = t_0 < t_1 < \cdots < t_N = M$ of $[0, M]$, M a large positive number. It is assumed that the mesh approaches zero as $N \to \infty$. Then

$$\int_0^{M} g(t)e^{-st}d\,\mathbf{u}(t-a) = \lim_{N\to\infty} \sum_{n=0}^{N-1} g(t_n)e^{-st_n}(\mathbf{u}(t_n - a) - \mathbf{u}(t_{n-1} - a))$$

Given point a satisfying $0 \le a < M$, then this point has to lie in exactly one interval: $t_{n-1} \le a < t_n$. Then $\mathbf{u}(t_n - a) - \mathbf{u}(t_{n-1} - a) = 1$ and for all other intervals this factor is zero. The sum reduces to a single term $g(t_n)e^{-s\,t_n}$. This term limits to $g(a+)\,e^{-sa}$ as $N \to \infty$, because t_n limits to a from the right. ■

Exercises 8.6

Unit Step and Heaviside

1. The unit step $\mathbf{u}(t)$ is defined on the whole real line. Is it piecewise continuous on the whole line?

2. Is there a continuous function on the real line that agrees with the Heaviside function except at $t = 0$?

3. The piecewise continuous function $\mathbf{pulse}(t, a, b)$ is defined everywhere. Redefine $\mathbf{pulse}(t, a, b)$ using $H(t)$ instead of $\mathbf{u}(t)$.

4. Write $f(t) = \textsf{\textbf{floor}}(t)\,\mathbf{u}(t)$ as a sum of terms, each of which has the form $g(t)\,\mathbf{pulse}(t, a, b)$.

Dirac Impulse

5. Verify $\int_{-\infty}^{\infty} \frac{\mathbf{pulse}(t,a,b)}{b-a}\,dt = 1$.

6. Verify by direct integration that $f(t) = 10\,\mathbf{pulse}(t, -0.001, 0.001)$ represents a simple impulse of 10 at $t = 0$ of duration 0.002. Graph it without using technology.

7. Find $\mathcal{L}(\delta(t-1) + \delta(t-2))$.

8. Find $\mathcal{L}(10\,\delta(t-1) - 5\,\delta(t-2))$.

9. Solve for $f(t)$ in terms of δ:
$\mathcal{L}(f(t)) = 10e^{-s}$

10. Solve for $f(t)$ in terms of δ:
$\mathcal{L}(f(t)) = 10e^{-s} + \frac{s}{s^2+1}\,e^{-2s}$

11. Find $\mathcal{L}\left(\sum_{n=1}^{10}(1+n)\delta(t-n)\right)$.

12. A sequence of camshaft impulses happening periodically in a finite time interval have transform $\mathcal{L}(f(t)) = \sum_{i=1}^{N} e^{-c_i\,s}$. Find the idealized impulse train f.

Riemann–Stieltjes Integral

Evaluate the integrals either directly from the definition or else by using Theorem 8.15.

13. $\int_0^2 d\,\mathbf{u}(t-1)$

14. $\int_0^\infty d\,\mathbf{u}(t-2)$

15. $\int_0^2 \tanh(t^2+1)\,d\,\mathbf{u}(t-1)$

16. $\int_0^\infty \frac{t}{1+t^2}\,d\,\mathbf{u}(t-2)$

8.7 Laplace Table Derivations

Verified here are two Laplace tables: the minimal Laplace Table 7.2-4 and its extension Table 7.2-6. This section is for reading, designed to enrich lectures and to aid those who study in isolation. Due to density of proof details, there are no exercises.

Derivation of Laplace integral formulas in Table 7.2-4, page 601.

- **Proof of $\mathcal{L}(t^n) = n!/s^{1+n}$:**

The first step is to evaluate $\mathcal{L}(t^n)$ for $n = 0$.

$$\begin{aligned}
\mathcal{L}(1) &= \textstyle\int_0^\infty (1)e^{-st}dt && \text{Laplace integral of } f(t) = 1. \\
&= -(1/s)e^{-st}\Big|_{t=0}^{t=\infty} && \text{Evaluate the integral.} \\
&= 1/s && \text{Assumed } s > 0 \text{ to evaluate } \lim_{t\to\infty} e^{-st}.
\end{aligned}$$

The value of $\mathcal{L}(t^n)$ for $n = 1$ can be obtained by s-differentiation of the relation $\mathcal{L}(1) = 1/s$, as follows.

$$\begin{aligned}
\tfrac{d}{ds}\mathcal{L}(1) &= \tfrac{d}{ds}\textstyle\int_0^\infty (1)e^{-st}dt && \text{Laplace integral for } f(t) = 1. \\
&= \textstyle\int_0^\infty \tfrac{d}{ds}\left(e^{-st}\right)dt && \text{Used } \tfrac{d}{ds}\textstyle\int_a^b F dt = \int_a^b \tfrac{dF}{ds}dt. \\
&= \textstyle\int_0^\infty (-t)e^{-st}dt && \text{Calculus rule } (e^u)' = u'e^u. \\
&= -\mathcal{L}(t) && \text{Definition of } \mathcal{L}(t).
\end{aligned}$$

Then

$$\begin{aligned}
\mathcal{L}(t) &= -\tfrac{d}{ds}\mathcal{L}(1) && \text{Rewrite last display.} \\
&= -\tfrac{d}{ds}(1/s) && \text{Use } \mathcal{L}(1) = 1/s. \\
&= 1/s^2 && \text{Differentiate.}
\end{aligned}$$

This idea can be repeated to give $\mathcal{L}(t^2) = -\frac{d}{ds}\mathcal{L}(t)$ and hence $\mathcal{L}(t^2) = 2/s^3$. The pattern is $\mathcal{L}(t^n) = -\frac{d}{ds}\mathcal{L}(t^{n-1})$ which gives $\mathcal{L}(t^n) = n!/s^{1+n}$.

- **Proof of $\mathcal{L}(e^{at}) = 1/(s-a)$:**

The result follows from $\mathcal{L}(1) = 1/s$, as follows.

$$\begin{aligned}
\mathcal{L}(e^{at}) &= \textstyle\int_0^\infty e^{at}e^{-st}dt && \text{Direct Laplace transform.} \\
&= \textstyle\int_0^\infty e^{-(s-a)t}dt && \text{Use } e^Ae^B = e^{A+B}. \\
&= \textstyle\int_0^\infty e^{-St}dt && \text{Substitute } S = s - a. \\
&= 1/S && \text{Apply } \mathcal{L}(1) = 1/s. \\
&= 1/(s-a) && \text{Back-substitute } S = s - a.
\end{aligned}$$

- **Proof of $\mathcal{L}(\cos bt) = s/(s^s + b^2)$ and $\mathcal{L}(\sin bt) = b/(s^s + b^2)$:**

Use will be made of Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, usually first introduced in trigonometry. In this formula, θ is a real number in radians and $i = \sqrt{-1}$ is the complex unit.

$e^{ibt}e^{-st} = (\cos bt)e^{-st} + i(\sin bt)e^{-st}$	Substitute $\theta = bt$ into Euler's formula and multiply by e^{-st}.
$\int_0^\infty e^{-ibt}e^{-st}dt = \int_0^\infty (\cos bt)e^{-st}dt + i\int_0^\infty (\sin bt)e^{-st}dt$	Integrate $t = 0$ to $t = \infty$. Then use properties of integrals.
$\dfrac{1}{s-ib} = \int_0^\infty (\cos bt)e^{-st}dt + i\int_0^\infty (\sin bt)e^{-st}dt$	Evaluate the left hand side using $\mathcal{L}(e^{at}) = 1/(s-a)$, $a = ib$.
$\dfrac{1}{s-ib} = \mathcal{L}(\cos bt) + i\mathcal{L}(\sin bt)$	Direct Laplace transform definition.
$\dfrac{s+ib}{s^2+b^2} = \mathcal{L}(\cos bt) + i\mathcal{L}(\sin bt)$	Use complex rule $1/z = \overline{z}/\vert z\vert^2$, $z = A + iB$, $\overline{z} = A - iB$, $\vert z\vert = \sqrt{A^2+B^2}$.
$\dfrac{s}{s^2+b^2} = \mathcal{L}(\cos bt)$	Extract the real part.
$\dfrac{b}{s^2+b^2} = \mathcal{L}(\sin bt)$	Extract the imaginary part.

Derivation of Laplace integral formulas in Table 7.2-6, page 602.

- **Proof of the Heaviside formula $\mathcal{L}(\mathbf{u}(t-a)) = e^{-as}/s$.**

$\mathcal{L}(\mathbf{u}(t-a)) = \int_0^\infty \mathbf{u}(t-a)e^{-st}dt$	Direct Laplace transform. Assume $a \ge 0$.
$= \int_a^\infty (1)e^{-st}dt$	Because $\mathbf{u}(t-a) = 0$ for $0 \le t < a$.
$= \int_0^\infty (1)e^{-s(x+a)}dx$	Change variables $t = x + a$.
$= e^{-as}\int_0^\infty (1)e^{-sx}dx$	Constant e^{-as} moves outside integral.
$= e^{-as}(1/s)$	Apply $\mathcal{L}(1) = 1/s$.

- **Proof of the Dirac impulse formula $\mathcal{L}(\delta(t-a)) = e^{-as}$.**

The *definition* of the Dirac impulse is a formal one, in which every occurrence of $\delta(t-a)dt$ under an integrand is replaced by $d\,\mathbf{u}(t-a)$. The differential symbol $d\,\mathbf{u}(t-a)$ is taken in the sense of the Riemann-Stieltjes integral. This integral is defined in Rudin [Rudin] for monotonic integrators $\alpha(x)$ as the limit

$$\int_a^b f(x)d\alpha(x) = \lim_{N\to\infty}\sum_{n=1}^{N} f(x_n)(\alpha(x_n) - \alpha(x_{n-1}))$$

where $x_0 = a$, $x_N = b$ and $x_0 < x_1 < \cdots < x_N$ forms a partition of $[a,b]$ whose mesh approaches zero as $N \to \infty$. Instance $\alpha(x) = x$ duplicates the theory of the Riemann integral in calculus.

The steps in computing the Laplace integral of the Dirac impulse appear below. Admittedly, the proof requires advanced calculus skills and a certain level of mathematical maturity. The reward is a fuller understanding of the Dirac symbol $\delta(x)$. More details and further properties of the Dirac impulse can be found in Section 8.6, page 648.

$\mathcal{L}(\delta(t-a)) = \int_0^\infty e^{-st}\delta(t-a)dt$	Laplace integral, $a \ge 0$ assumed.
$= \int_0^\infty e^{-st}d\,\mathbf{u}(t-a)$	Replace $\delta(t-a)dt$ by $d\,\mathbf{u}(t-a)$.
$= \lim_{M\to\infty}\int_0^M e^{-st}d\,\mathbf{u}(t-a)$	Definition of improper integral.
$= e^{-sa}$	Explained below.

To explain the last step, apply the definition of the Riemann-Stieltjes integral:

$$\int_0^M e^{-st} d\,\mathbf{u}(t-a) = \lim_{N\to\infty} \sum_{n=0}^{N-1} e^{-st_n}(\mathbf{u}(t_n - a) - \mathbf{u}(t_{n-1} - a))$$

where $0 = t_0 < t_1 < \cdots < t_N = M$ is a partition of $[0, M]$ whose mesh $\max_{1\le n\le N}(t_n - t_{n-1})$ approaches zero as $N \to \infty$. Given a partition, then point a in $0 \le a < M$ lies in exactly one interval: $t_{n-1} \le a < t_n$. By the definition of unit step, $\mathbf{u}(t_n - a) - \mathbf{u}(t_{n-1} - a) = 1$, while for any other interval this factor is zero. Therefore, the sum reduces to a single term $e^{-s\,t_n}$. This term approaches $e^{-s\,a}$ as $N \to \infty$, because t_n must approach a from the right.

- **Proof of** $\mathcal{L}(\textbf{floor}(t/a)) = \dfrac{e^{-as}}{s(1-e^{-as})}$:

The library function **floor** present in computer languages C and Fortran is defined by **floor**(x) = greatest whole integer $\le x$, e.g., **floor**$(5.2) = 5$ and **floor**$(-1.9) = -2$. The computation of the Laplace integral of **floor**(t) requires ideas from infinite series, as follows.

$$\begin{aligned}
F(s) &= \textstyle\int_0^\infty \textbf{floor}(t) e^{-st} dt && \text{Laplace integral definition.}\\
&= \textstyle\sum_{n=0}^\infty \int_n^{n+1} (n) e^{-st} dt && \text{On } n \le t < n+1,\ \textbf{floor}(t) = n.\\
&= \textstyle\sum_{n=0}^\infty \dfrac{n}{s}(e^{-ns} - e^{-ns-s}) && \text{Evaluate each integral.}\\
&= \dfrac{1-e^{-s}}{s} \textstyle\sum_{n=0}^\infty n e^{-sn} && \text{Common factor removed.}\\
&= \dfrac{x(1-x)}{s} \textstyle\sum_{n=0}^\infty n x^{n-1} && \text{Define } x = e^{-s}.\\
&= \dfrac{x(1-x)}{s} \dfrac{d}{dx} \textstyle\sum_{n=0}^\infty x^n && \text{Term-by-term differentiation.}\\
&= \dfrac{x(1-x)}{s} \dfrac{d}{dx} \dfrac{1}{1-x} && \text{Geometric series sum.}\\
&= \dfrac{x}{s(1-x)} && \text{Compute the derivative, simplify.}\\
&= \dfrac{e^{-s}}{s(1-e^{-s})} && \text{Substitute } x = e^{-s}.
\end{aligned}$$

To evaluate the Laplace integral of **floor**(t/a), a change of variables is made.

$$\begin{aligned}
\mathcal{L}(\textbf{floor}(t/a)) &= \textstyle\int_0^\infty \textbf{floor}(t/a) e^{-st} dt && \text{Laplace integral definition.}\\
&= a \textstyle\int_0^\infty \textbf{floor}(r) e^{-asr} dr && \text{Change variables } t = ar.\\
&= aF(as) && \text{Apply the formula for } F(s).\\
&= \dfrac{e^{-as}}{s(1-e^{-as})} && \text{Simplify.}
\end{aligned}$$

- **Proof of** $\mathcal{L}(\textbf{sqw}(t/a)) = \dfrac{1}{s}\tanh(as/2)$:

The square wave defined by **sqw**$(x) = (-1)^{\textbf{floor}(x)}$ is periodic of period 2 and piecewise-defined. Let $\mathcal{P} = \int_0^2 \textbf{sqw}(t) e^{-st} dt$.

$$\begin{aligned}
\mathcal{P} &= \textstyle\int_0^1 \mathsf{\textbf{sqw}}(t)e^{-st}dt + \int_1^2 \mathsf{\textbf{sqw}}(t)e^{-st}dt && \text{Apply } \textstyle\int_a^b = \int_a^c + \int_c^b. \\
&= \textstyle\int_0^1 e^{-st}dt - \int_1^2 e^{-st}dt && \text{Use } \mathsf{\textbf{sqw}}(x) = 1 \text{ on } 0 \le x < 1 \text{ and } \mathsf{\textbf{sqw}}(x) = -1 \text{ on } 1 \le x < 2. \\
&= \frac{1}{s}(1-e^{-s}) + \frac{1}{s}(e^{-2s} - e^{-s}) && \text{Evaluate each integral.} \\
&= \frac{1}{s}(1-e^{-s})^2 && \text{Collect terms.}
\end{aligned}$$

An intermediate step is to compute the Laplace integral of $\mathsf{\textbf{sqw}}(t)$:

$$\begin{aligned}
\mathcal{L}(\mathsf{\textbf{sqw}}(t)) &= \frac{\int_0^2 \mathsf{\textbf{sqw}}(t)e^{-st}dt}{1-e^{-2s}} && \text{Periodic function formula, page 638.} \\
&= \frac{1}{s}(1-e^{-s})^2 \frac{1}{1-e^{-2s}}. && \text{Use the computation of } \mathcal{P} \text{ above.} \\
&= \frac{1}{s}\frac{1-e^{-s}}{1+e^{-s}}. && \text{Factor } 1-e^{-2s} = (1-e^{-s})(1+e^{-s}). \\
&= \frac{1}{s}\frac{e^{s/2}-e^{-s/2}}{e^{s/2}+e^{-s/2}}. && \text{Multiply the fraction by } e^{s/2}/e^{s/2}. \\
&= \frac{1}{s}\frac{\sinh(s/2)}{\cosh(s/2)}. && \text{Use } \sinh u = (e^u - e^{-u})/2, \ \cosh u = (e^u + e^{-u})/2. \\
&= \frac{1}{s}\tanh(s/2). && \text{Use } \tanh u = \sinh u/\cosh u.
\end{aligned}$$

To complete the computation of $\mathcal{L}(\mathsf{\textbf{sqw}}(t/a))$, a change of variables is made:

$$\begin{aligned}
\mathcal{L}(\mathsf{\textbf{sqw}}(t/a)) &= \textstyle\int_0^\infty \mathsf{\textbf{sqw}}(t/a)e^{-st}dt && \text{Direct transform.} \\
&= \textstyle\int_0^\infty \mathsf{\textbf{sqw}}(r)e^{-asr}(a)dr && \text{Change variables } r = t/a. \\
&= \frac{a}{as}\tanh(as/2) && \text{See } \mathcal{L}(\mathsf{\textbf{sqw}}(t)) \text{ above.} \\
&= \frac{1}{s}\tanh(as/2)
\end{aligned}$$

- **Proof of $\mathcal{L}(a\,\mathsf{\textbf{trw}}(t/a)) = \dfrac{1}{s^2}\tanh(as/2)$:**

The triangular wave is defined by $\mathsf{\textbf{trw}}(t) = \int_0^t \mathsf{\textbf{sqw}}(x)dx$.

$$\begin{aligned}
\mathcal{L}(a\,\mathsf{\textbf{trw}}(t/a)) &= \frac{1}{s}(f(0) + \mathcal{L}(f'(t))) && \text{Let } f(t) = a\,\mathsf{\textbf{trw}}(t/a). \text{ Use } \mathcal{L}(f'(t)) = s\mathcal{L}(f(t)) - f(0)\text{, page 596.} \\
&= \frac{1}{s}\mathcal{L}(\mathsf{\textbf{sqw}}(t/a)) && \text{Use } f(0) = 0,\ (a\textstyle\int_0^{t/a}\mathsf{\textbf{sqw}}(x)dx)' = \mathsf{\textbf{sqw}}(t/a). \\
&= \frac{1}{s^2}\tanh(as/2) && \text{Table entry for } \mathsf{\textbf{sqw}}.
\end{aligned}$$

- **Proof of $\mathcal{L}(t^\alpha) = \dfrac{\Gamma(1+\alpha)}{s^{1+\alpha}}$:**

$$\begin{aligned}
\mathcal{L}(t^\alpha) &= \textstyle\int_0^\infty t^\alpha e^{-st}dt && \text{Direct Laplace transform.} \\
&= \textstyle\int_0^\infty (u/s)^\alpha e^{-u}du/s && \text{Change variables } u = st,\ du = sdt.
\end{aligned}$$

$$= \frac{1}{s^{1+\alpha}} \int_0^\infty u^\alpha e^{-u} du$$

$$= \frac{1}{s^{1+\alpha}} \Gamma(1+\alpha).$$

Where $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$, by definition.

The *generalized factorial function* $\Gamma(x)$ is defined for $x > 0$ and it agrees with the classical factorial $n! = (1)(2)\cdots(n)$ in case $x = n+1$ is an integer. In literature, $\alpha!$ means $\Gamma(1+\alpha)$. For more details about the Gamma function, see Abramowitz and Stegun [Abram-St], or `maple` documentation.

- **Proof of $\mathcal{L}(t^{-1/2}) = \sqrt{\dfrac{\pi}{s}}$:**

$$\mathcal{L}(t^{-1/2}) = \frac{\Gamma(1 + (-1/2))}{s^{1-1/2}}$$

Apply the previous formula.

$$= \frac{\sqrt{\pi}}{\sqrt{s}}$$

Use $\Gamma(1/2) = \sqrt{\pi}$.

8.8 Modeling

Laplace Modeling in Engineering

A *differential equation model* in variable t can be subjected to the Laplace transform, which produces an *algebraic model* in transform variable s.

The possibility of equivalence of models

$$mx''(t) + cx'(t) + kx(t) = 0 \quad \text{and} \quad \frac{1}{ms^2 + cs + k},$$

can be understood because of the one-to-one correspondence of the physical parameters m, c, k. Lerch's theorem provides a theoretical foundation which says that the differential equation model in the t-domain and the algebraic model in the s-domain are equivalent, that is, the solution of one model gives the solution to the other model.

$$\boxed{\begin{array}{l} mx'' + cx' + kx = f(t), \\ x(0) = x'(0) = 0 \end{array}} \qquad \longleftrightarrow \qquad \boxed{X(s) = \frac{F(s)}{ms^2 + cs + k}}$$

Figure 5. Differential Equation and Laplace Model Equivalence

In mechanical, electrical and computer engineering it is commonplace to deal *only* with the Laplace algebraic model, and to back-burner discussions of the differential equation model in the time domain.

Modeling conversations are often exclusively in terms of transforms. Differential equations are *rarely* mentioned! Terminology for such modeling is necessarily specialized, which gives rise to new contextual meanings to the terms *input* and *output*. For example, an *RLC*-circuit could be discussed with *input*

$$F(s) = \frac{\omega}{s^2 + \omega^2},$$

and the listener must know that this expression is the Laplace transform of the time input $f(t) = \sin \omega t$. The audience would then know that the *RLC*-circuit is driven by a sinusoidal input of natural frequency ω and amplitude one. The *output* could be the Laplace transform

$$X(s) = \frac{1}{s+1} + \frac{d_1 + d_2\omega}{s^2 + \omega^2}.$$

Lerch's equivalence provides extra information, deemed momentarily useless, that $X(s)$ is the Laplace transform of the time output $x(t) = e^{-t} + d_1 \cos \omega t + d_2 \sin \omega t$. It is important to know for modeling that fraction $\frac{1}{s+1}$ is the Laplace transform of the transient part of the output, while fraction $\frac{d_1+d_2\omega}{s^2+\omega^2}$ is the Laplace transform of the steady state output.

DC Gain

Background. Gain may be voltage gain (OP-amp, V/V), power gain (RF-amp, W/W) or sensor gain (light, e.g., 5 μV per photon). **Steady state gain** and **DC-gain** are synonyms for the same number. Laplace theory can compute steady-state values for a differential equation.

Signal applications might seek the signal $x(t)$ as the output of an **underdamped** model with switch at $t = 0$:

$$x'' + 2\zeta\omega x' + \omega^2 x = G_{\mathbf{DC}}\, \omega^2\, \mathbf{u}(t).$$

The three parameters $\zeta, \omega, G_{\mathbf{DC}}$ are known respectively as the **damping ratio**, **frequency** and **DC-gain**. Symbol $\mathbf{u}(t)$ is the unit step function. Under-damped for this equation means $\zeta > 1$, the case for complex roots of the characteristic equation. The Euler atoms for the homogeneous problem are exponential decay factors times sines and cosines.

On time interval $0 \le t < \infty$ the unit step $\mathbf{u}(t)$ is replaced by 1:

$$x'' + 2\zeta\omega x' + \omega^2 x = G_{\mathbf{DC}}\, \omega^2$$

Superposition implies $x = x_h + x_p$ with equilibrium solution $x_p(t) = G_{\mathbf{DC}}$ and homogeneous general solution $x_h(t) = c_1 e^{-at}\cos(bt) + c_2 e^{-at}\sin(bt)$, symbols defined by $a = \zeta\omega$ and $b = \omega\sqrt{\zeta^2 - 1}$. Because of the exponential decay of $x_h(t)$, the constant solution $x_p(t) = G_{\mathbf{DC}}$ is the signal steady state. This is why a simulator in a lab given a constant input k has DC-gain equal to the steady state of the output signal divided by k.

The DC-gain can be found mathematically from fractions in the s-domain instead of from time domain formulas. One possibility is to use the *final value rule* page 638 to find the steady state gain (DC-gain) $G_{\mathbf{DC}}$. Assume the input is the unit step function $\mathbf{u}(t)$. Then the steady state value is $\lim_{t\to\infty} x(t) = x(\infty) = \lim_{s\to 0} s\mathcal{L}(x(t))$. The s-domain product $s\mathcal{L}(x(t)) = \dfrac{\mathcal{L}(x(t))}{\mathcal{L}(\mathbf{u}(t))}$ equals the Laplace of the output divided by the Laplace of the input. Evaluation of this quotient at $s = 0$ is the system's steady state gain (DC-gain) $G_{\mathbf{DC}}$.

The **transfer function** $H(s)$ is the Laplace of the output $x(t)$ with input $\mathbf{u}(t)$ and zero initial data. For this special output $x(t)$ the equation $\mathcal{L}(x(t)) = H(s)\mathcal{L}(\mathbf{u}(t)) = \dfrac{H(s)}{s}$ holds. Therefore, the steady state gain (DC-gain) equals $H(0)$.

Illustration. An underdamped system whose transfer function is the fraction $H(s) = \dfrac{2}{s^2 + 2s + 2}$ has DC-gain $H(0) = 1$.

Engineering Inverse Problems

Linear time-invariant systems are used as building blocks to construct complex systems, in which the output of one system is the input of the next system. The systems are modeled by constant-coefficient linear differential equations. Practical applications endeavor to find a mathematical model to represent the block, technically an inverse problem.

What is the Inverse Problem? The terminology *inverse problem* applied to $x'' + px' + qx = f(t)$ means: given input $f(t)$ and output $x(t)$ as numerical data, recover the values of p and q. Imagine the experimental data is in a graph of signal $x(t)$ viewed on an oscilloscope. The graph data is imported into a numerical workbench in order to find the *system parameters* p, q in the predicted mathematical model

$$x'' + px' + qx = f(t), \quad \text{transfer function} = \frac{1}{s^2 + ps + q}.$$

Oscilloscope Experiments. It may help to think of the block as a physical device, like part of a battery charging circuit on a mobile phone. Initial states for a block are $x(0) = 0$, $x'(0) = 1$ or $x(0) = x'(0) = 0$. Possible input signals are zero input $f(t) = 0$, a step input $f(t) = k\,\mathbf{u}(t)$ or an impulse input $f(t) = \delta(t)$. Zero input means no battery. A step input can be considered a toggle which switches in a k–volt battery at $t = 0$. Impulse input $\delta(t)$ is practically a simplistic impulse $\frac{1}{h}(\mathbf{u}(t) - \mathbf{u}(t-h))$ with $h > 0$ very small; a function generator would work. A special BNC cable carries output signal $x(t)$ from the block to the oscilloscope for display. Oscilloscopes can save $x(t)$ data in text format for import into computer software.

Graphical Recognition of Block Types. A nontrivial aspect of an inverse problem is examination of graphical output in order to predict the block type.

The process can be more art than science. Multiple graphics might be produced for accurate prediction of the model type. For simplicity, there are two possibilities:

> **Non–oscillatory** $x'' + px' + qx = 0$, called *over–damped* in applications. It means characteristic equation $r^2 + pr + q = 0$ has two real distinct roots $-a, -b$. Assumed below is $a < b$.
>
> **Oscillatory** $x'' + px' + qx = 0$, called *under–damped* in applications. It means characteristic equation $r^2+pr+q = 0$ has complex conjugate roots $r = -a \pm b\,i$ with $b > 0$.

Skipped in the analysis above is the *critically–damped* case in which the characteristic equation has a double root. This case is technically non–oscillatory and physically indistinguishable from the over–damped case. In spring–mass systems, coefficient p is the damping constant, imagined as a tuning parameter adjusted by a set screw, for which the critically–damped value for p separates the two physically observable classifications **oscillatory** (small $p > 0$) and **non–oscillatory** (large $p > 0$).

The two observable cases are graphed in Figures 6 and 7. The distinction: the first curve touches the t-axis just once, while the second curve touches the t-axis infinitely often. In oscilloscope output, oscillations may be damped severely, looking non–oscillatory like Figure 6. Nonlinear blocks may have output completely different from Figures 6, 7. The choice of model is then art instead of science.

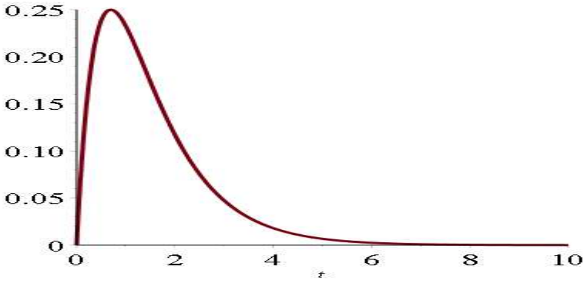


Figure 6. Block Output, Over-Damped. Non–oscillatory output $x(t)$ for $x'' + 3x' + 2x = 0$, $x(0) = 0$, $x'(0) = 1$.

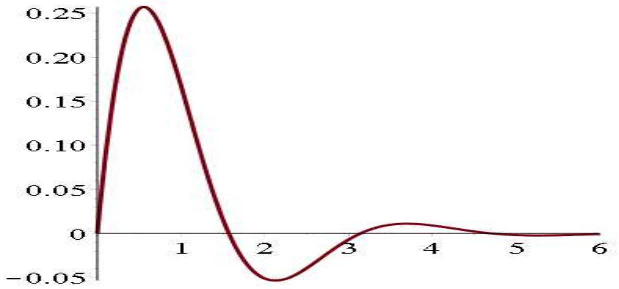


Figure 7. Block Output, Under-Damped Oscillatory output $x(t)$ for $x''+2x'+5x = 0$, $x(0) = 0$, $x'(0) = 1$.

Second Order Models

The models can be represented by

$$x''(t) + 2\zeta\omega x'(t) + \omega^2 x(t) = 0,$$

where ζ is the damping ratio and ω is the undamped natural frequency. Cases $\zeta < 1$, $\zeta = 1$ and $\zeta > 1$ are named **over-damped**, **critically-damped** and **under-damped**, respectively.

The over–damped and under–damped cases specialize respectively to

$$\begin{aligned} &x'' + (a+b)x' + abx = 0, \quad a < b, \\ &x'' + 2ax' + \left(a^2+b^2\right)x = 0, \quad b > 0. \end{aligned}$$

Because ζ, ω can be found from a, b, then the inverse problem seeks values for a, b instead of the damping ratio and undamped frequency.

Theorem 8.16 (Solution Formulas for Second Order Over–Damped)
The differential equation is $x'' + (a+b)x' + abx = f(t)$. Assume $a < b$. Formulas are for $t > 0$.

(1) Zero Input $f(t) = 0$, $x(0) = 0$, $x'(0) = 1$:
$$x(t) = \frac{1}{b-a}\left(e^{-at} - e^{-bt}\right)$$

(2) Step Input $f(t) = k\,\mathbf{u}(t)$, $x(0) = 0$, $x'(0) = 0$:
$$x(t) = \frac{k}{ab} + \frac{k}{a^2-ab}\,e^{-at} + \frac{k}{b^2-ab}\,e^{-bt}$$

(3) Dirac Input $f(t) = k\delta(t)$, $x(0) = 0$, $x'(0) = 0$:
$$x(t) = \frac{1}{2}\,\frac{1}{b-a}\left(e^{-at} - e^{-bt}\right)$$

Details for Theorem 8.16: Paper and pencil solutions use Laplace theory. The formulas can be obtained from a CAS like `maple` or `mathematica`, which use Laplace theory to solve the equation. The `maple` code:

```
deOD:=diff(z(t),t,t)+(a+b)*diff(z(t),t)+a*b*z(t);
ic1:=z(0)=0,D(z)(0)=1;ic2:=z(0)=0,D(z)(0)=0;
dsolve({deOD=0,ic1},z(t));dsolve({deOD=k,ic2},z(t));
dsolve({deOD=Dirac(t),ic2},z(t));convert(%,piecewise);
```

Theorem 8.17 (Solution Formulas for Second Order Under–Damped)
The equation is $x''(t) + 2ax'(t) + \left(a^2+b^2\right)x(t) = f(t)$. Assume $b > 0$. Formulas are for $t > 0$.

(1) Zero Input $f(t) = 0$, $x(0) = 0$, $x'(0) = 1$:
$$x(t) = e^{-a\,t}\,\frac{\sin(b\,t)}{b}$$

(2) Step Input $f(t) = k\,\mathbf{u}(t)$, $x(0) = 0$, $x'(0) = 0$:

$$x(t) = \frac{k - k\,e^{-a\,t}(\frac{a}{b}\sin(b\,t) + \cos(b\,t))}{a^2+b^2}$$

(3) Dirac Input $f(t) = \delta(t)$, $x(0) = 0$, $x'(0) = 0$:

$$x(t) = \frac{e^{-a\,t}}{2}\,\frac{\sin(b\,t)}{b}$$

Details for Theorem 8.17: Paper and pencil solutions use Laplace theory. The `maple` code:

```
deUD:=diff(z(t),t,t)+2*a*diff(z(t),t)+(a^2+b^2)*z(t);
ic1:=z(0)=0,D(z)(0)=1;ic2:=z(0)=0,D(z)(0)=0;
dsolve({deUD=0,ic1},z(t));dsolve({deUD=k,ic2},z(t));
dsolve({deUD=Dirac(t),ic2},z(t));convert(%,piecewise);
```

System Parameters for Over-Damped Problems

It is assumed that a signal $x(t)$ is known via a graphic with numerical data available. The graphic must pass the following visual test:

> The curve starts at $t = 0$, $x = 0$ and increases. The region of increase may end at a maximum and after decrease to limit zero, or else the region of increase is a half–line and $x(t)$ limits to a nonzero steady–state value.

Expected is a rich sample of the plot data, because computations use the numeric data, not the graphic. The initial data and the input are assumed to satisfy one of the following three cases.

Zero Input: $x(0) = 0$, $x'(0) = 1$, $f(t) = 0$
Step Input: $x(0) = 0$, $x'(0) = 0$, $f(t) = k\,\mathbf{u}(t)$
Impulse Input: $x(0) = 0$, $x'(0) = 0$, $f(t) = \delta(t)$

A graphic that passes the visual test predicts the model

$$x'' + (a+b)x' + abx = f(t), \text{ Transfer Function} = \frac{1}{(s+a)(s+b)},\ a < b.$$

The plan is to compute numerical values for a, b from the graphical data. The product of the computation is a mathematical model for the block represented by the graphic.

Example 8.30 (System Parameters: Over-Damped with Zero Input)

Oscilloscope data created Figure 8 from block initial state $x(0) = 0$, $x'(0) = 1$ and zero input. Explain why the graphic predicts over–damped model

$$x''(t) + (a+b)x'(t) + abx(t) = f(t), \quad \text{Transfer Function} = \frac{1}{(s+a)(s+b)}.$$

Then verify system parameters $a = 1$, $b = 2$ from graphical numeric data.

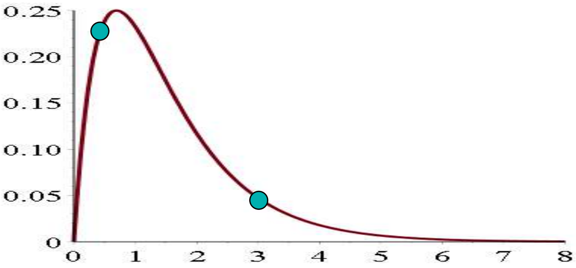


Figure 8. Oscilloscope output $x(t)$ for Example 8.30. The cyan dots are located at $(0.5, 0.2387)$ and $(3, 0.0473)$.

Solution: The curve in Figure 8 passes through $t = 0$, $x = 0$, then increases to a maximum and after decreases to zero. The curve fails to have infinity many crossings of the x-axis, therefore the system model is non–oscillatory over–damped.

We have only numeric output data for $x(t)$ and not the differential equation itself, so a, b are unknown. We discuss how to find $a = 1$ and $b = 2$ directly from the numerical data used to plot Figure 8.

Choose two points on the curve, one on the increasing section and one on the decreasing section. For example, the cyan dots in the figure, $t = 0.5, x = 0.2387$ and $t = 3.0, x = 0.0473$. Define $F(t, a, b) = \frac{1}{b-a}\left(e^{-at} - e^{-bt}\right)$, which is symbolic solution **(1)** in Theorem 8.16. Use a CAS like `maple` or `mathematica`, or a workbench like `matlab` to solve for a, b in the equations $F(0.5, a, b) = 0.2387$, $F(3.0, a, b) = 0.0473$. The answer is $b = 1.998164793$, $a = 1.000799323$. Due to $F(t, a, b) = F(t, b, a)$, there are two answers, but only one answer with requirement $a < b$. The `maple` code:

```
F:=(t,a,b)->(exp(-a*t)-exp(-b*t))/(b-a);
fsolve({F(0.5,a,b)=0.2387,F(3.0,a,b)=0.0473},{a,b});
```

Numerical computations like this might be done with algebra and shortcuts, like finding the smaller root from $x(t) \approx e^{-at}/(b-a)$ for large t.

Example 8.31 (System Parameters: Over–Damped with Step Input)

Oscilloscope data created Figure 9 from block state $x(0) = 0$, $x'(0) = 0$ and unit step input. Explain why the graphic predicts over–damped model

$$x''(t) + (a + b)x'(t) + abx(t) = f(t), \quad \text{Transfer Function} = \frac{1}{(s + a)(s + b)}.$$

Then verify system parameters $a = 1$, $b = 2$ from graphical numeric data.

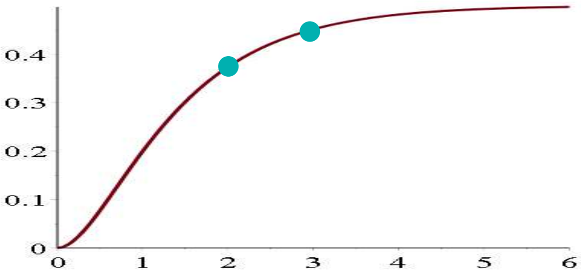


Figure 9. Oscilloscope output $x(t)$ for Example 8.31. The cyan dots are located at $(2, 0.374)$ and $(3, 0.451)$. The steady–state is $y_0 = 1/2$.

Solution:
The response curve has only one t-axis crossing, which classifies it nonoscillatory. The block type prediction is based upon seeing a response curve that starts at $(0, 0)$ and increases to nonzero steady–state, which is $y_0 = \frac{1}{2}$ in this example.

Choose two data points on the graphic, the cyan dots in the figure: $t = 2, y = 0.374$ and $t = 3, y = 0.451$. Let $F(t, a, b) = \frac{1}{ab} + \frac{1}{a^2-ab}\, e^{-at} + \frac{1}{b^2-ab}\, e^{-bt}$, which is symbolic solution **(2)** in Theorem 8.16. Solve the equations $F(2, a, b) = 0.374$, $F(3, a, b) = 0.451$ in a CAS or numerical workbench to get $b = 1.972417640$, $a = 1.017736613$. Because $F(t, a, b) = F(t, b, a)$, switching a, b values gives another solution. Only one of these meets requirement $a < b$. The `maple` code:

```
F:=(t,a,b)->1/(a*b)+exp(-a*t)/(a^2-a*b)+exp(-b*t)/(b^2-a*b);
fsolve({F(2,a,b)=0.374,F(3,a,b)=0.451},{a,b});
```

Example 8.32 (System Parameters: Over–Damped Impulse Input)

Oscilloscope data created Figure 10 from block state $x(0) = 0$, $x'(0) = 0$ and Dirac input. Explain why the graphic predicts over–damped model

$$x''(t) + (a + b)x'(t) + abx(t) = f(t), \quad \text{Transfer Function} = \frac{1}{(s + a)(s + b)}.$$

Then verify system parameters $a = 1$, $b = 2$ from graphical numeric data.

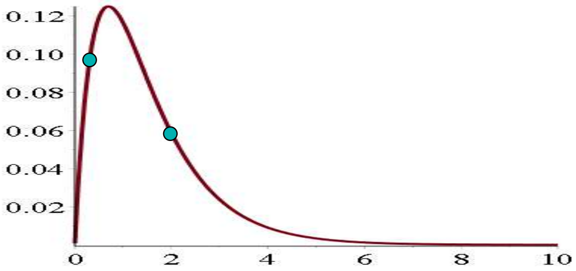


Figure 10. Oscilloscope output $x(t)$ for Example 8.32. The cyan dots are located at $(0.3, 0.0960)$ and $(2, 0.0585)$.

Solution: The graphic increases from $t = 0$, $x = 0$ to a maximum and after decreases to zero. The response curve has only one t-axis crossing, which classifies it non-oscillatory, hence over–damped.

Let $F(t,a,b) = \frac{1}{2}\frac{1}{b-a}\left(e^{-at} - e^{-bt}\right)$, which is symbolic solution **(3)** in Theorem 8.16 for $t > 0$. Choose two points on the experimental curve, for example the cyan dots in the figure $t = 0.3, x = 0.0960$ and $t = 2, x = 0.0585$. Solve for a, b in the two equations $F(0.3,a,b) = 0.0960$ and $F(2,a,b) = 0.0585$. The answer is $a = 2.000235495$, $b = 1.000004659$. The `maple` code:

```
F:=(t,a,b)->(1/2)*(exp(-a*t)-exp(-b*t))/(b-a);
fsolve({F(0.3,a,b)=0.0960,F(2,a,b)=0.0585},{a,b});
```

System Parameters for Under-Damped Problems

It is assumed that a signal $x(t)$ is known via numerical data for a graphic passing the following visual test:

> The curve starts at $t = 0$, $x = 0$ and has at least two local maxima on $t > 0$.

The initial data and the input are assumed to satisfy one of the following three cases.:

Zero Input: $x(0) = 0$, $x'(0) = 1$, $f(t) = 0$
Step Input: $x(0) = 0$, $x'(0) = 0$, $f(t) = k\,\mathbf{u}(t)$
Impulse Input: $x(0) = 0$, $x'(0) = 0$, $f(t) = \delta(t)$

A graphic that passes the above test predicts the model

$$x'' + 2ax' + \left(a^2 + b^2\right)x = f(t), \ \text{Transfer Function} = \frac{1}{(s+a)(s+b)},\ b > 0.$$

The method computes numerical values for a, b from the graphical data. The computation finds a mathematical model for the block represented by the graphic.

Theorem 8.18 (Parameters for an Under–Damped Model)
Let $x(t)$ be the response curve for $x'' + 2ax' + \left(a^2 + b^2\right)x = f(t)$ having a maximum at $t = t_1$, $x = x_1$ and next maximum at $t = t_2$, $x = x_2$.

If the input is $f(t) = 0$ or $f(t) = \delta(t)$, then the system parameters are

$$a = \frac{1}{t_2 - t_1}\ln\left|\frac{x_1}{x_2}\right| \quad \text{and} \quad b = \frac{2\pi}{t_2 - t_1}.$$

If the input is $f(t) = k\,\mathbf{u}(t)$ with steady–state solution y_0 different from x_1 and x_2, then the formulas are

$$a = \frac{1}{t_2 - t_1}\ln\left|\frac{x_1 - y_0}{x_2 - y_0}\right| \quad \text{and} \quad b = \frac{2\pi}{t_2 - t_1}.$$

Proof of Theorem 8.18, Underdamped Parameters:
For zero input or impulse input, the equation is $x'' + 2ax' + \left(a^2 + b^2\right)x = 0$ for $t > 0$ with general solution $x = c_1e^{-at}\cos(bt) + c_2e^{-at}\sin(bt)$. Convert to the form $x = Ae^{-at}\cos(bt - \phi)$. For definiteness, assume $x_1 > 0$ and $x_2 > 0$.

Verify $b = \frac{2\pi}{t_2-t_1}$. At max $t = t_1$ the cosine factor must be a max, which is 1, so $bt_1 - \phi = n\pi$. At the next maximum $t = t_2$ the cosine factor must also be 1, so $bt_2 - \phi = n\pi + m\pi$. Because the maxima are consecutive, then $m = 2$ (the period of the cosine is 2π). Subtract the two equations to obtain $2\pi = (bt_2 - \phi) - (bt_1 - \phi) = b(t_2 - t_1)$ and solve for $b = \frac{2\pi}{t_2-t_1}$.

Verify $a = \frac{1}{t_2-t_1} \ln \left|\frac{x_1}{x_2}\right|$, $x_1{=}Ae^{-at_1}\cos(bt_1{-}\phi)$. Let $x_2{=}Ae^{-at_2}\cos(bt_2{-}\phi)$. Because $\cos(bt_1{-}\phi) = \cos(bt_2{-}\phi) = 1$, as argued above, then x_1 divided by x_2 gives $\frac{x_1}{x_2} = e^{at_2-at_1}$. Take the logarithm across this equality and use $\ln(e^u) = u$, then $\ln\left|\frac{x_1}{x_2}\right| = \ln(e^{at_2-at_1}) = a(t_2 - t_1)$. Solve for $a = \frac{1}{t_2-t_1}\ln\left|\frac{x_1}{x_2}\right|$.

Case $f(t) = k\,\mathbf{u}(t)$. The equation is $x'' + 2ax' + \left(a^2 + b^2\right)x = k$ with steady–state solution $y_0 = k/(a^2 + b^2)$ and homogeneous solution $x_h = c_1e^{-at}\cos(bt) + c_2e^{-at}\sin(bt)$. The change of variables $y(t) = x(t) - y_0$ changes $x'' + 2ax' + \left(a^2 + b^2\right)x = k$ into $y'' + 2ay' + \left(a^2 + b^2\right)y = 0$. Because $x'(t) = y'(t)$, the curves $x(t)$ and $y(t)$ have the same critical points. Further, solution $y(t)$ has consecutive local maxima at $t = t_1$, $y = x(t_1) - y_0 = x_1 - y_0$ and $t = t_2$, $y = x(t_2) - y_0 = x_2 - y_0$. The zero input case applies to compute the parameters $a = \frac{1}{t_2-t_1}\ln\left|\frac{x_1-y_0}{x_2-y_0}\right|$, $b = \frac{2\pi}{t_2-t_1}$.[13] ■

Example 8.33 (Parameters: Under–Damped, Zero or Impulse Input)

Oscilloscope data created Figure 11 from block state $x(0) = 0$, $x'(0) = 1$ and zero input or block state $x(0) = 0$, $x'(0) = 0$ and impulse input. Explain why the graphic predicts under–damped model

$$x''(t) + 2ax'(t) + \left(a^2 + b^2\right)x(t) = f(t), \text{ Transfer Function } = \frac{1}{(s+a)^2 + b^2}.$$

Then verify system parameters $a = 1$, $b = 2$ from graphical numeric data.

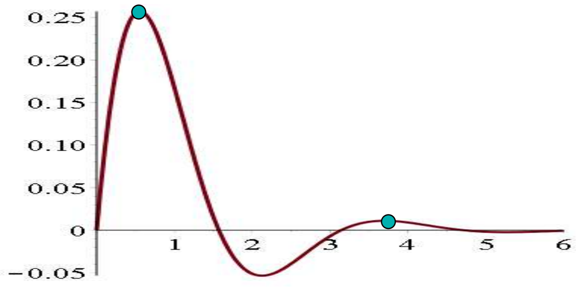


Figure 11. Oscilloscope output $x(t)$ for Example 8.33. The cyan dots are located at $(0.554, 0.257)$ and $(3.695, 0.011)$.

Solution: The curve in Figure 11 passes through $t = 0$, $x = 0$, then increases to a maximum and after decreases to a minimum. The curve crosses the t-axis twice, therefore it will have infinity many crossings of the t-axis: the system model is oscillatory under–damped.

[13] Oscilloscopes can display signal $y(t)$ directly, which simplifies external data processing to the zero input case. See Example 8.35.

Only numeric data for $x(t)$ is available and not the differential equation itself, so parameters a, b are unknown. Approximations to $a = 1$ and $b = 2$ are found directly from the numerical data used to plot Figure 11.

Choose two points on the curve, one at the first maximum and one at the very next maximum. These are the cyan dots in the figure, $t_1 = 0.554$, $x_1 = 0.257$ and $t_2 = 3.695$, $x_2 = 0.011$. Apply Theorem 8.18 to obtain $a = \dfrac{\ln(x_1/x_2)}{t_2 - t_1} = 1.003241265$ and $b = \dfrac{2\pi}{t_2 - t_1} = 2.000377366$.

Answer Check. Insert the parameters $a = 1$, $b = 2$ into the predicted model, then plot the symbolic response $x(t)$ for zero input (or $\delta(t)$ for impulse input). If the graphic matches Figure 11, then the computed parameters were likely correct.

Example 8.34 (System Parameters: Under–Damped Step Input)

Oscilloscope data created Figure 12 from block state $x(0) = 0$, $x'(0) = 0$ and unit step input. Explain why the graphic predicts under–damped model

$$x''(t) + 2ax'(t) + \left(a^2 + b^2\right) x(t) = f(t), \text{ Transfer Function} = \frac{1}{(s+a)^2 + b^2}.$$

Then verify system parameters $a = 1$, $b = 4$ from graphical numeric data.

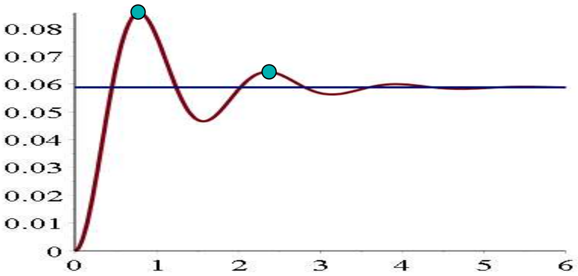


Figure 12. Oscilloscope output $x(t)$ for Example 8.34. The cyan dots are located at $(0.7854, 0.0856)$, $(2.3562, 0.0644)$. The steady–state is $y_0 = 1/17$.

Solution: The curve in Figure 12 passes through $t = 0$, $x = 0$, then increases to a local maximum and after decreases to a local minimum. The steady–state is $y_0 = 1/17$. The curve crosses the steady–state $y_0 = 1/17$ twice, therefore it will have infinity many crossings of $y_0 = 1/17$: the system model is oscillatory under–damped.

Choose two points on the curve, one at the first local maximum and one at the very next local maximum. These are the cyan dots in the figure, $t_1 = 0.7854$, $x_1 = 0.0856$, $t_2 = 2.3562$, $x_2 = 0.0644$. Let $y_0 = 1/17$. Apply Theorem 8.18 to obtain $a = \frac{1}{t_2 - t_1} \ln\left|\frac{x_1 - y_0}{x_2 - y_0}\right| = 0.9988333798$ and $b = \frac{2\pi}{t_2 - t_1} = 3.999990646$.

Answer Check. Insert the parameters $a = 1$, $b = 4$ into the predicted model, then plot the symbolic response $x(t)$ for $x(0) = x'(0) = 0$ and unit step input. If the graphic matches Figure 12, then the computed parameters are probably correct.

Example 8.35 (Special Case: Under–Damped Step Input)

Oscilloscope data created Figure 13 from block state $x(0) = 0$, $x'(0) = 1$ and unit step input. Explain why the graphic is missing a nonzero steady–state and the predicted model is

$$x''(t) + 2ax'(t) + \left(a^2 + b^2\right) x(t) = f(t), \text{ Transfer Function} = \frac{1}{(s+a)^2 + b^2}.$$

Discuss which formula to use when verifying $a = 1$, $b = 4$.

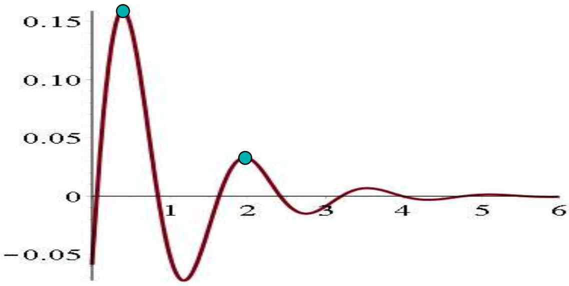


Figure 13. Oscilloscope output $x(t)$ for Example 8.35. The cyan dots are located at $(0.3927, 0.1580)$ and $(1.9635, 0.0330)$. The steady–state is $x = 0$.

Solution: The oscilloscope hardware made a change of variables $y = x - 1/17$ prior to display of the y–data. This changed the steady state from $y_0 = 1/17$ to $y_0 = 0$. The apparent oscillation of the graphic about $x = 0$ predicts the under–damped case: see the proof of Theorem 8.18 page 662.

The figure looks like the zero input case, because y satisfies the homogeneous equation $y'' + 2ay' + (a^2 + b^2)y = 0$ with steady–state zero and initial state $y(0) = -1/17$, $y'(0) = 1$. The single important distinction is that the graph fails to pass through $(0, 0)$.

Apply Theorem 8.18 with $y_0 = 0$ to obtain $a = \dfrac{1}{t_2 - t_1} \ln \left|\dfrac{x_1}{x_2}\right| = 1.0006$ and $b = \dfrac{2\pi}{t_2 - t_1} = 3.999991$.

Exercises 8.8

Oscillatory and Non–oscillatory

Assume $x'' + px' + qx = 0$ with p, q nonnegative.

Parameter p is imagined as a set screw adjustment on a screen door dashpot, larger p meaning more damping effect.

Parameter q is the Hooke's constant for the spring restoring force.

1. Let $q = 100$, $p = 99$. Verify that the equation is over–damped in two ways: (1) Graph $x(t)$; (2) Justify that $r^2 + pr + q = 0$ has real negative roots.

2. Let $q = 100$. The case which is called *critically–damped* happens at exactly one value $p = p^*$ between 0 and 99. Compute p^* numerically. Graph $x(t)$ using $q = 100$, $p = p^*$, $x(0) = 0$, $x'(0) = 1$.

3. Let $q = 100$. Verify that $p = 0$ produces the **harmonic oscillator** $x'' + \omega^2 x = 0$, $\omega = 10$.

Small set screw changes from $p = 0$ to $p > 0$ are still oscillatory. Under–damped means weak dashpot reaction.

4. Let $q = 100$, $p = 2$. Justify oscillatory under–damped from the graph of $x(t)$ and also by solving $r^2 + pr + q = 0$.

Simplistic Dirac Impulse

Define $g(t) = 7\,e^{-153800\,t}\,\mathbf{u}(t)$ and $f(t,a) = \frac{1}{a}\,(u(t) - u(t-a))$, $a > 0$. The impulse of force h is $\int_{-\infty}^{\infty} h(t)\,dt$.

5. Compute the impulse for $f(t,a)$.
Ans: 1.

6. Plot $f(t,a)$ for $a = 0.1$, 0.001, 0.0001.

7. Calculate the impulse for $g(t)$.
Ans: About 46 times 10^{-6}.

8. Try to find an **RC** discharge circuit with 10 volt *emf* and output $g(t)$.

Circuit response $g(t)$ simulates Dirac impulsive force $\frac{45.5}{1000000}\delta(t)$.

Parameters: Over–Damped

Find $a, b, \omega = \sqrt{ab}, \zeta = \frac{a+b}{2\omega}$ given the plot and two dots on the graph.

9. Step input Figure 9, dots $(1, 0.1998)$, $(4, 0.4819)$.
Ans: $a = 1.0000$, $b = 1.9997$, $\omega = 1.4141$, $\zeta = 1.0607$.

10. Impulse input Figure 10, dots $(0.5, 0.1193)$, $(2, 0.0585)$.
Ans: $a = 0.9991$, $b = 2.0021$, $\omega = 1.4143$, $\zeta = 1.0610$.

Parameters: Under–Damped

Find $a, b, \omega = \sqrt{a^2 + b^2}, \zeta = \frac{a}{\omega}$ given the plot and two dots on the graph.

11. Zero input like Figure 11, but consecutive maxima at $(2.5107, 0.0257)$, $(4.6051, 0.0032)$.
Ans: Approximately $a = 1$, $b = 3$.

12. Step input like Figure 13, but steady–state $y_0 = 1/26$ and consecutive maxima at $(0.6283, 0.0205)$, $(1.8850, 0.0058)$.
Ans: Approximately $a = 1$, $b = 5$.

Chapter 9

Eigenanalysis

Contents

9.1 Matrix Eigenanalysis

Studied here is eigenanalysis for matrix equations. The topics are *eigenanalysis*, *eigenvalue*, *eigenvector*, *eigenpair* and *diagonalization*.

What's Eigenanalysis?

The term **eigenanalysis** refers to the identification and computation of a **new coordinate system** and **scale factors**. There is one scale factor per coordinate direction. The new coordinate system has axes with measurement units defined by the scale factors. This coordinate system is employed to simplify the expression of the original mathematical model, be it a matrix model, a differential equation model, or otherwise.

Matrix eigenanalysis is a tool for a matrix equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$.

Eigenanalysis was born from ideas in the 1822 work of J. B. Fourier on heat conduction for an insulated rod, which resulted in a simple algebraic re-scaling formula for the rod temperature: Fourier's idea is explained on page 676. His ideas apply to data analysis matrix equations $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, systems of linear ordinary differential equations and partial differential equations of mathematical physics.

Larry Page and Sergey Brin in 1996 created from eigenanalysis a search algorithm which became *Google search*. Eigenanalysis is part of the mathematical toolset for research areas like machine learning and data mining.

Simplification of Linear Algebraic Equations

Consider the matrix equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, where symbol A is a square matrix of constants and symbols $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ are column vectors. The matrix equation is equivalent to simultaneous linear algebraic equations. For a 3×3 matrix $A = (a_{ij})$, $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ is equivalent to linear algebraic equations

$$\left\{\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &=& y_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &=& y_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &=& y_3. \end{array}\right.$$

Table 1. Simplification of 3×3 Linear Algebraic Equations

Matrix eigenanalysis is a tool for $\vec{\mathbf{A}}\vec{\mathbf{x}} = \vec{\mathbf{b}}$, a system of linear simultaneous algebraic equations. It invents a change of variable $\vec{\mathbf{x}} \to \vec{\mathbf{X}}$, $\vec{\mathbf{b}} \to \vec{\mathbf{B}}$ that simplifies the system of equations.

A change of variables $\vec{\mathbf{X}} = P\vec{\mathbf{x}}$, $\vec{\mathbf{B}} = P\vec{\mathbf{b}}$ with eigenanalysis vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ for the columns of P simplifies a 3×3 system of linear algebraic equations $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ into the diagonal form

$$(1) \qquad \left\{\begin{array}{lcl} \lambda_1 X_1 &=& B_1, \\ \lambda_2 X_2 &=& B_2, \\ \lambda_3 X_3 &=& B_3. \end{array}\right.$$

Scalar values $\lambda_1, \lambda_2, \lambda_3$ are scale factors (measurement units) corresponding to the directions $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$. Precise definitions are on page 669.

Coordinate Change using Eigenanalysis

Technically, matrix eigenanalysis is an **opportunistic change of coordinates**, which means the analysis must compute a set of **independent** column vectors that span $\mathcal{R}^n$. Linear algebra calls such a set of vectors a **basis**. Eigenanalysis constructs from square matrix A a *special basis*. This special basis defines a change of coordinates $\vec{\mathbf{x}} \to P\vec{\mathbf{x}}$ where P is the augmented matrix of constructed basis vectors.

Consider vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ which form a basis for $\mathcal{R}^3$. To be a *basis* means that *each possible* vector $\vec{\mathbf{x}}$ in $\mathcal{R}^3$ can be uniquely expressed as a linear combination $\vec{\mathbf{x}} = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3$. Geometrically, the triad $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ must define a parallelepiped of positive volume. For a triad basis $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, each possible $\vec{\mathbf{x}}$ in $\mathcal{R}^3$ can be constructed from the triad using solely the geometric parallelogram law for vector addition.

The claimed simplifying change of coordinates[1] is defined by:

$$\begin{array}{ll} & P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle = \text{augmented matrix}, \\ (2) & \vec{\mathbf{X}} = P\vec{\mathbf{x}}, \quad \vec{\mathbf{B}} = P\vec{\mathbf{b}}, \quad \text{a change of } A\vec{\mathbf{x}} = \vec{\mathbf{b}} \text{ into } D\vec{\mathbf{X}} = \vec{\mathbf{B}}, \\ & D = \mathbf{diag}(\lambda_1, \lambda_2, \lambda_3) \quad \text{a diagonal matrix of scale factors} \end{array}$$

Details on page 675.

Eigenvalue, Eigenvector and Eigenpair Defined

Eigenanalysis for the matrix equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ when matrix A is 3×3 is an algebraic method for discovering basis vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ and scale factors $\lambda_1, \lambda_2, \lambda_3$. The vectors are called **eigenvectors** and the scale factors are called **eigenvalues**.

A scale factor λ is thought to be a **measurement unit** along an axis $\vec{\mathbf{v}}$, therefore the eigenvectors and eigenvalues occur in pairs, called **eigenpairs**. Pairing is due to fundamental equation (3) below, which is used in references to define and/or compute an eigenpair.

Definition 9.1 (Eigenpair)

An **Eigenpair** $(\lambda, \vec{\mathbf{v}})$ is defined to be a solution of the problem

$$(3) \qquad A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}, \quad \vec{\mathbf{v}} \neq \vec{\mathbf{0}}.$$

Vector $\vec{\mathbf{v}}$ is called an **eigenvector**. The value λ is called the **eigenvalue** corresponding to the eigenvector $\vec{\mathbf{v}}$.

Important. Because $\vec{\mathbf{v}} \neq \vec{\mathbf{0}}$ in equation (3), then an eigenvector is never the zero vector: an eigenvector is a *direction*. Otherwise stated:

An eigenvector answer of zero signals an algebra error.

Motivation for the rather abstract definition of eigenpair appears below. Excuses aside, definition (3) *must be learned and memorized*, because of explicit use in computations and implicit use in literature.

[1] The triad $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ in **principal coordinate analysis** and **metric scaling** simplifies the data set to find trends and important parameters.

Why the Equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$?

The pattern is $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. However, it is **not** the problem being solved. The maddening historical event of algebraists stripping away the problem from the definition impacts everyone trying to learn eigenanalysis.

The algebraists' Definition 9.1 is a sub-problem. It is madness to try to learn eigenanalysis from it. Learning from it parallels trying to learn about trees by crawling on the ground through the forest examining tree trunks.

Assume matrix A is 3×3. The **problem to be solved** is computation of independent vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ to find an opportunistic change of variables that simplifies the linear algebraic system of equations $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$.

Algebraists were quick to discover that the problem is solved by finding a basis $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ of $\mathcal{R}^3$ satisfying the **three equations** (4) *infra*. They isolated $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ as a sub-problem to be solved many times, in order to find the basis.

History of Eigenvector and Eigenvalue Terminology

James J. Sylvester in 1883 coined the term **latent root** for what has become the term **eigenvalue**:

> ...the latent roots of a matrix – latent in a somewhat similar sense as vapour may be said to be latent in water or smoke in a tobacco-leaf.

The German term *eigenwert* was coined by David Hilbert in 1904. By 1967, Paul Halmos gave up the battle over which words to use in his new book *A Hilbert Space Problem Book*. The battle: German *eigen* means **proper**, *wert* means **value**.

> For many years I have battled for **proper values** and against the one and a half times translated German-English hybrid (Halmos means **eigenvalue**) that is often used to refer to them. I have now become convinced that the war is over, and eigenvalues have won it; in this book I use them.

No longer used are the historical terms *hidden value*, *proper value*, *characteristic value* and *latent root*. The term *hidden* arose because the vectors and scale factors are generally impossible to determine from matrix A without computation. What has persisted in literature is the *characteristic equation*, the equation which determines eigenvalues. See Theorem 9.2.

Eigenpair Equations and $AP = PD$

Eigenpair equations for a square matrix A can be written by matrix multiply as a single equation.

Theorem 9.1 (Eigenpairs and $AP = PD$)
Assume $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ independent in $\mathcal{R}^3$. Let matrix A be 3×3. Then relations

$$(4) \qquad \left\{\begin{array}{lcl} A\vec{\mathbf{v}}_1 &=& \lambda_1\vec{\mathbf{v}}_1, \\ A\vec{\mathbf{v}}_2 &=& \lambda_2\vec{\mathbf{v}}_2, \\ A\vec{\mathbf{v}}_3 &=& \lambda_3\vec{\mathbf{v}}_3. \end{array}\right. \quad \textbf{(Eigenpair Equations)}$$

hold if and only if $AP = PD$ where P and D are defined by equations

$$(5) \qquad P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

The result holds for dimension n. Proof on page 693.

Computing Eigenpairs of a Matrix

To compute an eigenpair $(\lambda, \vec{\mathbf{v}})$ of a square matrix A requires finding scalar λ and a nonzero vector $\vec{\mathbf{v}}$ satisfying the homogeneous matrix–vector equation

$$A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}.$$

Write it as $A\vec{\mathbf{x}} - \lambda\vec{\mathbf{x}} = \vec{\mathbf{0}}$, then replace $\lambda\vec{\mathbf{x}}$ by $\lambda I\vec{\mathbf{x}}$ to obtain the standard homogeneous linear algebraic system form[2]

$$(A - \lambda I)\vec{\mathbf{v}} = \vec{\mathbf{0}}, \quad \vec{\mathbf{v}} \neq \vec{\mathbf{0}}.$$

Definition 9.2 (Characteristic Equation)
Determinant equation $|A - \lambda I| = 0$ is called the **characteristic equation**. The **characteristic polynomial** is the polynomial obtained by determinant evaluation on the left, normally by cofactor expansion or the triangular rule.

Theorem 9.2 (Eigenvalues of A)
The eigenvalues of a square matrix A are exactly all the roots λ of the polynomial equation

$$\det(A - \lambda I) = 0.$$

Proof on page 694

Theorem 9.3 (Find Eigenvectors of Matrix A)
For each root λ of the characteristic equation $|A - \lambda I| = 0$, form matrix $B = A - \lambda I$. Write a toolkit sequence to $\mathbf{rref}(B)$. Solve the homogeneous equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ for $\vec{\mathbf{v}}$ in terms of invented symbols t_1, t_2, ...

A basis of eigenvectors of A for eigenvalue λ is the list of vectors $\partial_{t_1}\vec{\mathbf{v}}$, $\partial_{t_2}\vec{\mathbf{v}}$, ... They are Strang's **special solutions** of $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$, known to be independent.

These eigenvectors span the nullspace (kernel) of B: if $A\vec{\mathbf{w}} = \lambda\vec{\mathbf{w}}$, then $\vec{\mathbf{w}}$ is a linear combination of these basis vectors.

Proof on page 694.

[2] Identity I is required to factor out the matrix $A - \lambda I$. It is wrong to factor out $A - \lambda$, because A is 3×3 and λ is 1×1, incompatible sizes for matrix addition.

Characteristic Equation Illustration.

$$\begin{aligned}\det\left(\begin{pmatrix}1&3\\1&2\end{pmatrix}-\lambda\begin{pmatrix}1&0\\0&1\end{pmatrix}\right) &= \begin{vmatrix}1-\lambda & 3\\ 1 & 2-\lambda\end{vmatrix}\\ &= (1-\lambda)(2-\lambda)-6\\ &= \lambda^2-3\lambda-4\\ &= (\lambda+1)(\lambda-4).\end{aligned}$$

The characteristic equation $\lambda^2-3\lambda-4=0$ has roots $\lambda_1=-1$, $\lambda_2=4$. The characteristic polynomial is $\lambda^2-3\lambda-4$.

Table 2. Shortcut for the Characteristic Polynomial

To find the characteristic polynomial $|A-\lambda I|$, subtract symbol λ from the diagonal of A and then evaluate the determinant.

Key Examples for Finding Eigenvectors

Assume given a 3×3 matrix A. Found after at most 3 applications of Theorem 9.3 is a list of eigenpairs with independent eigenvectors.

There might not be 3 answers!

The amount of work on paper and pencil varies with the number of repeated eigenvalues. Key examples:

$$\boxed{\mathbf{1}}\quad\begin{pmatrix}1&0&1\\0&2&4\\0&0&3\end{pmatrix},\quad\boxed{\mathbf{2}}\quad\begin{pmatrix}1&0&1\\0&1&4\\0&0&1\end{pmatrix},\quad\boxed{\mathbf{3}}\quad\begin{pmatrix}1&0&1\\0&1&4\\0&0&2\end{pmatrix}$$

$\boxed{\mathbf{1}}$ Matrix A has eigenvalues 1, 2, 3. Apply Theorem 9.3 three times to write three different matrices B. Each B has a toolkit sequence to $\mathbf{rref}(B)$, a total of 3 toolkit sequences. Each sequence produces one eigenvector: *there are 3 answers.*

$\boxed{\mathbf{2}}$ Matrix A has eigenvalues 1, 1, 1. Apply Theorem 9.3 one time to write one matrix B. There is just 1 toolkit sequence to $\mathbf{rref}(B)$. Because of 2 free variables, *there are 2 answers.* In general, the number of free variables is 1, 2 or 3 with correspondingly 1,2 or 3 answers.

$\boxed{\mathbf{3}}$ Matrix A has eigenvalues 1, 1, 2. Apply Theorem 9.3 two times to write two matrices B. Each B has a toolkit sequence to $\mathbf{rref}(B)$, a total of 2 toolkit sequences. Eigenvalue 1 has a basis of 2 eigenvectors, caused by 2 free variables. Eigenvalue 2 has a basis of just one eigenvector, caused by only 1 free variable.

In general, the number of answers for a repeated eigenvalue equals the number of free variables for the toolkit sequence B to $\mathbf{rref}(B)$.

Independence of Eigenvectors

Theorem 9.4 (Independence of Eigenvectors)
If $(\lambda_1, \vec{\mathbf{v}}_1)$ and $(\lambda_2, \vec{\mathbf{v}}_2)$ are two eigenpairs of A and $\lambda_1 \neq \lambda_2$, then $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are linearly independent vectors.

More generally, if $(\lambda_1, \vec{\mathbf{v}}_1)$, ..., $(\lambda_k, \vec{\mathbf{v}}_k)$ are eigenpairs of A corresponding to distinct eigenvalues λ_1, ..., λ_k, then $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_k$ are independent.

Proof on page 694

Theorem 9.5 (Unions of Eigenvectors)
Let A be an $n \times n$ matrix A. Let variable λ denote an arbitrary eigenvalue of A. Let λ_1, ..., λ_k be a list of distinct eigenvalues of A.

Let $\mathcal{B}(\lambda)$ be some basis for the eigenpair equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. Then

(1) For $\lambda \neq \mu$, subspaces $\mathbf{span}(\mathcal{B}(\lambda))$ and $\mathbf{span}(\mathcal{B}(\mu))$ intersect in only the zero vector.

(2) The union U of bases $\mathcal{B}(\lambda_1)$, ..., $\mathcal{B}(\lambda_k)$ is a list of independent vectors in $\mathcal{C}^n$.[3]

(3) If all eigenvalues are real, then $\mathcal{C}^n$ can be replaced by $\mathcal{R}^n$ in results (1), (2).

Proof on page 695

Complete Set of Eigenvectors

Definition 9.3 (Complete Set of Eigenvectors)
A list $U = \{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_k\}$ of independent eigenvectors of an $n \times n$ matrix A is called **complete** provided $k = n$.

Lemma 9.1 (Invertible Change of Variables) Let $U = \{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_n\}$ be a list of independent eigenvectors of an $n \times n$ matrix A. Assume all eigenvalues are real. Define augmented $n \times n$ matrix $P = \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_n \right\rangle$. Then:

The eigenvectors span $\mathcal{R}^n$: $\mathbf{span}(U) = \mathcal{R}^n$.

Matrix P is invertible.

Proof: A list U of n independent vectors in $\mathcal{R}^n$ is a basis. Then U spans $\mathcal{R}^n$. An $n \times n$ matrix with independent columns is invertible. ■

[3] Symbol $\mathcal{C}^n$ is the vector space of n-vectors with complex entries.

Theorem 9.6 (Finding Independent Eigenvectors)
Let $n \times n$ matrix A be given. Solve the characteristic equation $|A - \lambda I| = 0$ for all eigenvalues λ. For each λ, let $B = A - \lambda I$ and solve $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ for general solution $\vec{\mathbf{v}}$, which contains invented symbols $t_1, t_2, \ldots$. Let $\mathcal{B}(\lambda)$ be the list of vector partial derivatives $\partial_{t_1}\vec{\mathbf{v}}, \partial_{t_2}\vec{\mathbf{v}}, \ldots$. Then the union U of all lists $\mathcal{B}(\lambda)$ is a set of independent eigenvectors. Examples exist where U is *not* a basis for $\mathcal{R}^n$.

Proof on page 696.

Eigenanalysis Facts

1. An **eigenvalue** λ of a triangular matrix A is one of the diagonal entries. If A is non-triangular, then an eigenvalue is found as a root λ of the characteristic equation $|A - \lambda I| = 0$.

2. An **eigenvalue** of a square matrix A can be zero, positive, negative or even complex. It is a pure number, with a physical meaning inherited from the model, e.g., a scale factor or measurement unit.

3. An **eigenvector** for eigenvalue λ (a scale factor) is a nonzero direction $\vec{\mathbf{v}}$ of application satisfying $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. It is found from a toolkit sequence starting at $B = A - \lambda I$ and ending at $\mathbf{rref}(B)$. Independent eigenvectors are computed from the general solution of $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ as partial derivatives $\partial\vec{\mathbf{v}}_{/}\partial t_1$, $\partial\vec{\mathbf{v}}/\partial t_2$, ...

4. If a 3×3 matrix has three independent real eigenvectors, then they collectively form a **basis** of $\mathcal{R}^3$ (a **coordinate system**).

Diagonalization and Eigenpair Packages

Definition 9.4 (Diagonalizable Matrix)
An $n \times n$ matrix A which has n independent eigenvectors is called **diagonalizable**. The eigenvalues are not required to be distinct.

Given a diagonalizable 3×3 system $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, the augmented matrix $P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle$ of eigenvectors and diagonal matrix $D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$ provide a variable change $\vec{\mathbf{X}} = P\vec{\mathbf{x}}, \vec{\mathbf{Y}} = P\vec{\mathbf{y}}$ to transform system $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ into the simplified diagonal system $\vec{\mathbf{Y}} = D\vec{\mathbf{X}}$.

Theorem 9.7 (Diagonalization and Diagonal Matrices)

A 3×3 diagonal matrix $A = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$ has eigenvalues on the diagonal. The eigenvectors are the columns of the 3×3 identity matrix:

$$\begin{array}{lll} \lambda_1 = a, & \lambda_2 = b, & \lambda_3 = c, \\ \vec{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{array}$$

The theorem extends to $n \times n$ matrices. Every $n \times n$ diagonal matrix is diagonalizable.

Definition 9.5 (Eigenpair Packages)
Let A be a diagonalizable 3×3 matrix with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, $(\lambda_3, \vec{\mathbf{v}}_3)$. Define **eigenpair packages** by:[4]

$$(6) \qquad P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Package definitions for an $n \times n$ matrix:

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \cdots | \vec{\mathbf{v}}_n \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

If all eigenvalues are real then both P and D are real. Otherwise, matrices P and D will have complex entries.

Theorem 9.8 (Diagonalization)
Let A be a diagonalizable $n \times n$ matrix with eigenpair packages P, D.

1. The matrix A is completely determined by its eigenpairs:

$$A = PDP^{-1}.$$

2. The change of variables $\vec{\mathbf{X}} = P\vec{\mathbf{x}}$, $\vec{\mathbf{Y}} = P\vec{\mathbf{y}}$ transforms the equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ into the diagonal system $\vec{\mathbf{Y}} = D\vec{\mathbf{X}}$.

3. The equation $A(c_1\vec{\mathbf{v}}_1 + \cdots + c_n\vec{\mathbf{v}}_n) = c_1\lambda_1\vec{\mathbf{v}}_1 + \cdots + c_n\lambda_n\vec{\mathbf{v}}_n$ holds for any constants $c_1, \ldots, c_n$ with matrix form

$$(7) \qquad AP\vec{\mathbf{c}} = PD\vec{\mathbf{c}}, \quad \vec{\mathbf{c}} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}.$$

See Fourier Replacement page 676.

[4] Eigenpair packages are not unique. For 3×3, there are six (6) permutations of the pairs, leading to six different packages. In addition, eigenvectors are not unique, leading to infinitely many possible eigenpair packages.

Proof on page 696.

Theorem 9.9 (Distinct Eigenvalues implies Diagonalizable)
If an $n \times n$ matrix A has n distinct eigenvalues, real or complex, then it has n eigenpairs $(\lambda_i, \vec{\mathbf{v}}_i)$, $i = 1, \ldots, n$. The eigenpair packages

$$P = \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_n \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

satisfy $AP = PD$ and matrix A is diagonalizable.

Proof on page 697.

Fourier Replacement

The subject of **eigenanalysis** was popularized by J. B. Fourier in his 1822 publication on the theory of heat, *Théorie analytique de la chaleur.* Fourier's ideas can be summarized for the $n \times n$ matrix equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$:

Vector $A\vec{\mathbf{x}}$ is obtained from $\vec{\mathbf{x}}$ and a complete set of eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, ..., $(\lambda_n, \vec{\mathbf{v}}_n)$ by replacing the eigenvectors by their scaled versions $\lambda_1\vec{\mathbf{v}}_1$, ..., $\lambda_n\vec{\mathbf{v}}_n$:

$$(8) \qquad \begin{array}{rcccccccl} \vec{\mathbf{x}} & = & c_1\vec{\mathbf{v}}_1 & + & c_2\vec{\mathbf{v}}_2 & + & \cdots & + & c_n\vec{\mathbf{v}}_n \quad \text{implies} \\ A\vec{\mathbf{x}} & = & c_1\lambda_1\vec{\mathbf{v}}_1 & + & c_2\lambda_2\vec{\mathbf{v}}_2 & + & \cdots & + & c_n\lambda_n\vec{\mathbf{v}}_n. \end{array}$$

See Example 9.10 page 690 for details about the heat problem.

For the case of $\mathcal{R}^3$, basis vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are re-scaled by invented **scale factors** $\lambda_1, \lambda_2, \lambda_3$, which we imagine as measurement units along the three directions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$. Fourier's 1822 idea: vector $\vec{\mathbf{x}}$ is replaced by a new vector $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, according to the rule

$$(9) \qquad \begin{array}{rcl} \vec{\mathbf{x}} & = & c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3 \text{ implies} \\ \vec{\mathbf{y}} & = & c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3. \end{array}$$

Table 3. Fourier's 1822 Re-Scaling Idea

Replace $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ by re-scaled vectors $\lambda_1\vec{\mathbf{v}}_1$, $\lambda_2\vec{\mathbf{v}}_2$, $\lambda_3\vec{\mathbf{v}}_3$.

Criticism: Table 3 makes no mention of a matrix A. Fourier's re-scaling idea does not need a matrix A, but it resurfaces:

Theorem 9.10 (Matrix Form of Fourier Replacement)
Let vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ be independent. Let $\lambda_1, \lambda_2, \lambda_3$ be scalars. Define

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \quad \vec{\mathbf{c}} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

Fourier replacement is defined by

$$\begin{array}{rcl} \vec{\mathbf{x}} & = & c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3 \text{ implies} \\ \vec{\mathbf{y}} & = & c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3 \end{array} \quad \text{for all scalars } c_1, c_2, c_3$$

The statement has vector-matrix forms

$$\begin{array}{l} \vec{\mathbf{x}} = P\vec{\mathbf{c}} \quad \text{implies} \quad \vec{\mathbf{y}} = PD\vec{\mathbf{c}} \\ \vec{\mathbf{y}} = A\vec{\mathbf{x}} \quad \text{where} \quad A = PDP^{-1} \\ A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\right) = c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3 \end{array}$$

The theorem extends to $n \times n$. Proof on page 697.

Theorem 9.11 (Fourier Re-scaling and Diagonalization)
Let vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ be independent. Let $\lambda_1, \lambda_2, \lambda_3$ be scalars. Define $P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle$, $D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$.

(a) Matrix $A = PDP^{-1}$ has 3 eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, $(\lambda_3, \vec{\mathbf{v}}_3)$ and A is diagonalizable.

(b) If a diagonalizable 3×3 matrix has eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, $(\lambda_3, \vec{\mathbf{v}}_3)$ with independent eigenvectors, then Fourier replacement (8) holds.

(c) Fourier replacement for matrix equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ defined in (8) is equivalent to diagonalizability of matrix A.

Proof on page 697.

Re-scaling Example: Data Conversion

Let $\vec{\mathbf{x}}$ in $\mathcal{R}^3$ be a data set variable with coordinates x_1, x_2, x_3 recorded respectively in units of meters, millimeters and centimeters. Imagine the data being recorded every few milliseconds from three different sensors.

The $\vec{\mathbf{x}}$-data set is converted into a $\vec{\mathbf{y}}$-data set with meter, kilogram, second units (MKS units) via the equations

$$(10) \qquad \left\{ \begin{array}{rcl} y_1 & = & x_1, \\ y_2 & = & 0.001x_2, \\ y_3 & = & 0.01x_3. \end{array} \right.$$

Equations (10) are an instance of Fourier's re-scaling process, Table 3. The paired scale factors and vectors are

$$\begin{array}{lll} \lambda_1 = 1, & \lambda_2 = 0.001, & \lambda_3 = 0.01, \\ \vec{\mathbf{v}}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}, & \vec{\mathbf{v}}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}, & \vec{\mathbf{v}}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}. \end{array}$$

Then equations (10) can be written as the replacement process

$$(11)\qquad \begin{array}{rcl} \vec{\mathbf{x}} & = & x_1\begin{pmatrix}1\\0\\0\end{pmatrix} + x_2\begin{pmatrix}0\\1\\0\end{pmatrix} + x_3\begin{pmatrix}0\\0\\1\end{pmatrix} \quad \text{implies} \\ \vec{\mathbf{y}} & = & x_1\lambda_1\begin{pmatrix}1\\0\\0\end{pmatrix} + x_2\lambda_2\begin{pmatrix}0\\1\\0\end{pmatrix} + x_3\lambda_3\begin{pmatrix}0\\0\\1\end{pmatrix}. \end{array}$$

Vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are the *data directions* (or *axes*) re-scaled by the measurement units λ_1, λ_2, λ_3, respectively. In particular, data direction $\vec{\mathbf{v}}_2$ is for millimeters and scale factor $\lambda_2 = 0.001$ is the measurement unit along axis $\vec{\mathbf{v}}_2$. Theorem 9.10 applied to (11) gives $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ where $A = \begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{1000} & 0\\ 0 & 0 & \frac{1}{100}\end{pmatrix}$, agreeing with conversion of (10) to matrix form.

Fourier Replacement: Matrix Example

Let

$$(12)\qquad \begin{array}{lll} A = \begin{pmatrix}1 & 3 & 0\\ 0 & 2 & -1\\ 0 & 0 & -5\end{pmatrix} & & \\ \lambda_1 = \mathbf{1}, & \lambda_2 = \mathbf{2}, & \lambda_3 = \mathbf{\text{-}5}, \\ \vec{\mathbf{v}}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}, & \vec{\mathbf{v}}_2 = \begin{pmatrix}3\\1\\0\end{pmatrix}, & \vec{\mathbf{v}}_3 = \begin{pmatrix}1\\-2\\-14\end{pmatrix}. \end{array}$$

Then Fourier's model (9) holds, details in Example 9.3:

$$\begin{array}{rcl} \vec{\mathbf{x}} & = & c_1\begin{pmatrix}1\\0\\0\end{pmatrix} + c_2\begin{pmatrix}3\\1\\0\end{pmatrix} + c_3\begin{pmatrix}1\\-2\\-14\end{pmatrix} \\ \text{implies} & & \\ A\vec{\mathbf{x}} & = & c_1(\mathbf{1})\begin{pmatrix}1\\0\\0\end{pmatrix} + c_2(\mathbf{2})\begin{pmatrix}3\\1\\0\end{pmatrix} + c_3(\mathbf{\text{-}5})\begin{pmatrix}1\\-2\\-14\end{pmatrix} \end{array}$$

Eigenanalysis and Geometry

In case the matrix A is 2×2 or 3×3, geometry can provide additional intuition about eigenanalysis.

Fourier's 2×2 replacement $A(c_1\vec{v}_1 + c_2\vec{v}_2) = c_1\lambda_1\vec{v}_1 + c_2\lambda_2\vec{v}_2$ can be interpreted as the action of the transformation $T: \ \vec{x} \to A\vec{x}$ between two copies of the plane $\mathcal{R}^2$; see Figure 1.

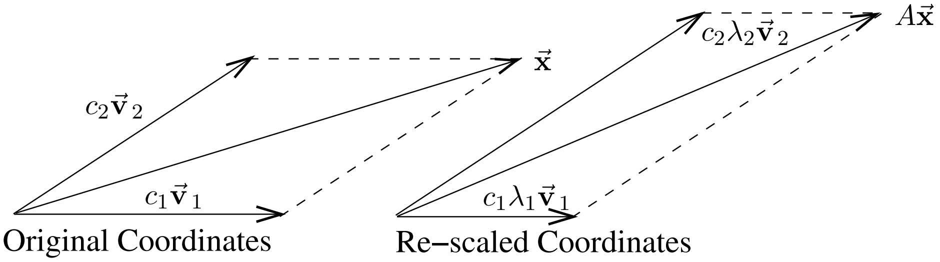


Figure 1. Transformation $T: \ \mathcal{R}^2 \to \mathcal{R}^2$.
Vector $\vec{\mathbf{x}}$ is obtained geometrically from $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$ by changing their lengths by c_1, c_2, then add with the parallelogram rule. Vector $A\vec{\mathbf{x}}$ is obtained from the two changed vectors by re-scaling by λ_1, λ_2, then apply the parallelogram rule.

Algebraically, A is replaced by the scale factors λ_1, λ_2 and the coordinate system $\vec{v}_1, \vec{v}_2$. The **eigenvalues** are the scale factors λ_1, λ_2. Vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$ used in the parallelogram rule are the **eigenvectors**.

Shear is not Equivalent to Scaling along Axes

The important geometrical operations are scaling, shears, rotations, projections, reflections and translations. Fourier replacement describes scaling along coordinate axes.

A **planar horizontal shear** $(x_1, x_2) \to (y_1, y_2)$ is a set of equations

$$\begin{array}{rcl} y_1 &=& x_1 + kx_2, \quad (k = \text{shear factor} \neq 0), \\ y_2 &=& x_2. \end{array}$$

The eigenvalues of $A = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ are $\lambda_1 = \lambda_2 = 1$. Assume it is possible to view this shear as a re-scaling. Then it must be feasible to change coordinates to new independent axes $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$ and express the shear as

$$A = PDP^{-1}, \quad D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle.$$

Then $\begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} = A = PDP^{-1} = P \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} P^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, a contradiction to the shear factor requirement $k \neq 0$.

Conclusion: A shear is not equivalent to scaling along axes. Fourier replacement fails.

Examples and Methods

Example 9.1 (Computing 2×2 Eigenpairs)

Find all eigenpairs of the 2×2 matrix $A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$.

Solution:
The method used to solve for eigenpairs in given in Theorem 9.3 page 671.

College Algebra. The eigenvalues are $\lambda_1 = 1$, $\lambda_2 = -1$. Details:

$$\begin{aligned} 0 &= \det(A - \lambda I) && \text{Characteristic equation.} \\ &= \begin{vmatrix} 1-\lambda & 0 \\ 2 & -1-\lambda \end{vmatrix} && \text{Subtract } \lambda \text{ from the diagonal.} \\ &= (1-\lambda)(-1-\lambda) && \text{Sarrus' rule.} \end{aligned}$$

Linear Algebra. The eigenpairs are $\left(1, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$, $\left(-1, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$. Details:

Eigenvector for $\lambda_1 = 1$.

$$\begin{aligned} A - \lambda_1 I &= \begin{pmatrix} 1-\lambda_1 & 0 \\ 2 & -1-\lambda_1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 2 & -2 \end{pmatrix} \\ &\approx \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} && \text{Swap and multiply rules.} \\ &= \mathbf{rref}(A - \lambda_1 I) && \text{Reduced echelon form.} \end{aligned}$$

The vector partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the scalar general solution $x = t_1$, $y = t_1$ is eigenvector $\vec{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Eigenvector for $\lambda_2 = -1$.

$$\begin{aligned} A - \lambda_2 I &= \begin{pmatrix} 1-\lambda_2 & 0 \\ 2 & -1-\lambda_2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} \\ &\approx \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} && \text{Combination and multiply.} \\ &= \mathbf{rref}(A - \lambda_2 I) && \text{Reduced echelon form.} \end{aligned}$$

The vector partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the scalar general solution $x = 0$, $y = t_1$ is eigenvector $\vec{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Example 9.2 (Computing 2×2 Complex Eigenpairs)

Find all eigenpairs of the 2×2 matrix $A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$.

Solution:

Reference: Theorem 9.3 page 671.

College Algebra. The eigenvalues are $\lambda_1 = 1+2i$, $\lambda_2 = 1-2i$. Details:

$$\begin{aligned} 0 &= \det(A-\lambda I) && \text{Characteristic equation.} \\ &= \begin{vmatrix} 1-\lambda & 2 \\ -2 & 1-\lambda \end{vmatrix} && \text{Subtract } \lambda \text{ from the diagonal.} \\ &= (1-\lambda)^2+4 && \text{Sarrus' rule.} \end{aligned}$$

The roots $\lambda = 1\pm 2i$ are found from the quadratic formula after expanding $(1-\lambda)^2+4=0$. Alternatively, use $(1-\lambda)^2=-4$ and take square roots.

Linear Algebra. The eigenpairs are $\left(1+2i, \begin{pmatrix} -i \\ 1 \end{pmatrix}\right)$, $\left(1-2i, \begin{pmatrix} i \\ 1 \end{pmatrix}\right)$.

Eigenvector for $\lambda_1 = 1+2i$.

$$\begin{aligned} A-\lambda_1 I &= \begin{pmatrix} 1-\lambda_1 & 2 \\ -2 & 1-\lambda_1 \end{pmatrix} \\ &= \begin{pmatrix} -2i & 2 \\ -2 & -2i \end{pmatrix} \\ &\approx \begin{pmatrix} i & -1 \\ 1 & i \end{pmatrix} && \text{Multiply rule.} \\ &\approx \begin{pmatrix} 0 & 0 \\ 1 & i \end{pmatrix} && \text{Combination rule, multiplier=}-i. \\ &\approx \begin{pmatrix} 1 & i \\ 0 & 0 \end{pmatrix} && \text{Swap rule.} \\ &= \mathbf{rref}(A-\lambda_1 I) && \text{Reduced echelon form.} \end{aligned}$$

The partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the general solution $x=-it_1$, $y=t_1$ is eigenvector $\vec{\mathbf{v}}_1 = \begin{pmatrix} -i \\ 1 \end{pmatrix}$.

Eigenvector for $\lambda_2 = 1-2i$.

The answer is eigenvector $\vec{\mathbf{v}} = \begin{pmatrix} i \\ 1 \end{pmatrix}$. See Lemma 9.2 page 685 for the expected shortcut, which obtains the answer from the eigenvector for $\lambda_1 = 1+2i$. The shortcut creates no matrix $B = A-\lambda I$ and no toolkit sequence B to $\mathbf{rref}(B)$.

The shortcut eliminates the following steps:

$$\begin{aligned} A-\lambda_2 I &= \begin{pmatrix} 1-\lambda_2 & 2 \\ -2 & 1-\lambda_2 \end{pmatrix} \\ &= \begin{pmatrix} 2i & 2 \\ -2 & 2i \end{pmatrix} \\ &\approx \begin{pmatrix} i & 1 \\ 1 & -i \end{pmatrix} && \text{Multiply rule.} \\ &\approx \begin{pmatrix} 0 & 0 \\ 1 & -i \end{pmatrix} && \text{Combination rule, multiplier=}-i. \\ &\approx \begin{pmatrix} 1 & -i \\ 0 & 0 \end{pmatrix} && \text{Swap rule.} \\ &= \mathbf{rref}(A-\lambda_2 I) && \text{Reduced echelon form.} \end{aligned}$$

The partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the general solution $x = it_1$, $y = t_1$ is eigenvector $\vec{\mathbf{v}}_2 = \begin{pmatrix} i \\ 1 \end{pmatrix}$.

Example 9.3 (Computing 3×3 Eigenpairs: Real Eigenvalues)

Find all eigenpairs of the 3×3 matrix

$$A = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & -5 \end{pmatrix}. \tag{13}$$

Solution:
Reference: Theorem 9.3 page 671.

The answers are

$$\begin{array}{lll} \lambda_1 = 1, & \lambda_2 = 2, & \lambda_3 = -5, \\ \vec{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_2 = \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_3 = \begin{pmatrix} 1 \\ -2 \\ -14 \end{pmatrix}. \end{array}$$

College Algebra. The eigenvalues are $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = -5$, because matrix A is triangular and the eigenvalues of a triangular matrix appear on the diagonal.

Linear Algebra. There are three toolkit sequences B to $\mathbf{rref}(B)$ to compute, one for each distinct eigenvalue λ where $B = A - \lambda I$.

Eigenvector for $\lambda_1 = 1$.

Subtract $\lambda_1 = 1$ from the diagonal of A to obtain the equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$, where

$$B = A - \lambda_1 I = \begin{pmatrix} 0 & 3 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -6 \end{pmatrix}.$$

A toolkit sequence with swap, combo, multiply will find

$$\mathbf{rref}(B) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The lead variables are v_2, v_3 and the free variable is v_1. Assign invented symbol t_1 to the free variable and back-substitute into $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ to obtain the scalar equations

$$\begin{aligned} v_1 &= t_1, \\ v_2 &= 0, \\ v_3 &= 0. \end{aligned}$$

Take the partial derivative on invented symbol t_1 across these equations to obtain the eigenvector

$$\vec{\mathbf{v}}_1 = \begin{pmatrix} \frac{\partial v_1}{\partial t_1} \\ \frac{\partial v_2}{\partial t_1} \\ \frac{\partial v_3}{\partial t_1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Eigenvector for $\lambda_2 = 2$.

Subtract $\lambda_2 = 2$ from the diagonal of A to obtain the equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$, where

$$B = A - \lambda_2 I = \begin{pmatrix} -1 & 3 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & -7 \end{pmatrix}.$$

A toolkit sequence finds

$$\mathbf{rref}(B) = \begin{pmatrix} 1 & -3 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The lead variables are v_1, v_3 and the free variable is v_2. Assign invented symbol t_1 to the free variable and back-substitute into $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ to obtain the scalar equations

$$\begin{aligned} v_1 &= 3t_1, \\ v_2 &= t_1, \\ v_3 &= 0. \end{aligned}$$

Take the partial derivative on invented symbol t_1 across these equations to obtain the eigenvector

$$\vec{\mathbf{v}}_2 = \begin{pmatrix} \frac{\partial v_1}{\partial t_1} \\ \frac{\partial v_2}{\partial t_1} \\ \frac{\partial v_3}{\partial t_1} \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}.$$

The eigenpair is $(\lambda_2, \vec{\mathbf{v}}_2) = \left(2, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}\right)$

Eigenvector for $\lambda_3 = -5$.

Subtract $\lambda_3 = -5$ from the diagonal of A to obtain the equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$, where

$$B = A - \lambda_2 I = \begin{pmatrix} 6 & 3 & 0 \\ 0 & 7 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

A toolkit sequence finds

$$\mathbf{rref}(B) = \begin{pmatrix} 1 & 0 & 1/14 \\ 0 & 1 & -1/7 \\ 0 & 0 & 0 \end{pmatrix}.$$

The lead variables are v_1, v_2 and the free variable is v_3. Assign invented symbol t_1 to the free variable and back-substitute into $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ to obtain the scalar equations

$$\begin{aligned} v_1 &= -\tfrac{1}{14}t_1, \\ v_2 &= \tfrac{1}{7}t_1, \\ v_3 &= 0. \end{aligned}$$

Take the partial derivative on invented symbol t_1 across these equations to obtain the eigenvector

$$\vec{\mathbf{v}}_3 = \begin{pmatrix} \frac{\partial v_1}{\partial t_1} \\ \frac{\partial v_2}{\partial t_1} \\ \frac{\partial v_3}{\partial t_1} \end{pmatrix} = \begin{pmatrix} -\frac{1}{14} \\ \frac{1}{7} \\ 1 \end{pmatrix}.$$

It is usual when encountering fractions in an eigenvector to replace the answer $\vec{\mathbf{v}}$ by $c\vec{\mathbf{v}}$ where $c \neq 0$ is chosen to make the answer fraction-free and the first nonzero entry positive. In this case, $c = -14$ is used, and we replace $\vec{\mathbf{v}}_3$ by $-14\vec{\mathbf{v}}_3$. The eigenpair is $(\lambda_3, \vec{\mathbf{v}}_3) = \left(-5, \begin{pmatrix} 1 \\ -2 \\ -14 \end{pmatrix}\right)$.

This completes the computation of all three eigenpairs.

Answer Check. The eigenpair equations are equivalent to the matrix identity $AP = PD$ where P is the matrix of eigenvectors and D is the diagonal matrix of corresponding eigenvalues:

$$P = \begin{pmatrix} 1 & 3 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & -14 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -5 \end{pmatrix}.$$

Eigenpairs are checked by expanding AP and PD, then compare for equality. The two calculations give

$$AP = \begin{pmatrix} 1 & 6 & -5 \\ 0 & 2 & 10 \\ 0 & 0 & 70 \end{pmatrix} = PD.$$

Fourier Replacement page 676 is explicitly

$$\begin{aligned} \vec{\mathbf{x}} &= c_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ -2 \\ -14 \end{pmatrix} \\ \text{implies} \\ A\vec{\mathbf{x}} &= c_1(1) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2(2) \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + c_3(-5) \begin{pmatrix} 1 \\ -2 \\ -14 \end{pmatrix} \end{aligned}$$

Example 9.4 (Computing 3×3 Eigenpairs: Complex Eigenvalues)

Find all eigenpairs of the 3×3 matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$.

Solution:
Reference: Theorem 9.3 page 671.

College Algebra. The eigenvalues are $\lambda_1 = 1 + 2i$, $\lambda_2 = 1 - 2i$, $\lambda_3 = 3$. Details:

$$\begin{aligned} 0 &= \det(A - \lambda I) && \text{Characteristic equation.} \\ &= \begin{vmatrix} 1-\lambda & 2 & 0 \\ -2 & 1-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{vmatrix} && \text{Subtract } \lambda \text{ from the diagonal.} \\ &= ((1-\lambda)^2 + 4)(3 - \lambda) && \text{Cofactor rule and Sarrus' rule.} \end{aligned}$$

Root $\lambda = 3$ is found from the factored form above. The roots $\lambda = 1 \pm 2i$ are found from the quadratic formula after expanding $(1 - \lambda)^2 + 4 = 0$. Alternatively, take roots across $(\lambda - 1)^2 = -4$.

Linear Algebra.

The eigenpairs are $\left(1+2i, \begin{pmatrix} -i \\ 1 \\ 0 \end{pmatrix}\right)$, $\left(1-2i, \begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}\right)$, $\left(3, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right)$.

Eigenvector for $\lambda_1 = 1+2i$.

$$\begin{aligned}
A - \lambda_1 I &= \begin{pmatrix} 1-\lambda_1 & 2 & 0 \\ -2 & 1-\lambda_1 & 0 \\ 0 & 0 & 3-\lambda_1 \end{pmatrix} & \\
&= \begin{pmatrix} -2i & 2 & 0 \\ -2 & -2i & 0 \\ 0 & 0 & 2-2i \end{pmatrix} & \text{Subtract } \lambda_1 = 1+2i \text{ from the diagonal.} \\
&\approx \begin{pmatrix} i & -1 & 0 \\ 1 & i & 0 \\ 0 & 0 & 1 \end{pmatrix} & \text{Multiply rule.} \\
&\approx \begin{pmatrix} 0 & 0 & 0 \\ 1 & i & 0 \\ 0 & 0 & 1 \end{pmatrix} & \text{Combination rule, factor}=-i. \\
&\approx \begin{pmatrix} 1 & i & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} & \text{Swap rule.} \\
&= \mathbf{rref}(A - \lambda_1 I) & \text{Reduced echelon form.}
\end{aligned}$$

The vector partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the scalar general solution $x = -it_1$, $y = t_1$, $z = 0$ is eigenvector $\vec{\mathbf{v}}_1 = \begin{pmatrix} -i \\ 1 \\ 0 \end{pmatrix}$.

Eigenvector for $\lambda_2 = 1-2i$.

There is no need for a toolkit sequence to find the eigenvector for a conjugate eigenvalue: see Lemma 9.2 *infra*. Answer: $(1-2i, \vec{\mathbf{v}}_2)$, $\vec{\mathbf{v}}_2 = \begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}$.

Details. To see why, take conjugates[5] across the equation $(A - \lambda_2 I)\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$ to give $(\overline{A} - \overline{\lambda_2} I)\overline{\vec{\mathbf{v}}_2} = \vec{\mathbf{0}}$. Then $\overline{A} = A$ (A is real) and $\lambda_1 = \overline{\lambda_2}$ gives $(A - \lambda_1 I)\overline{\vec{\mathbf{v}}_2} = \vec{\mathbf{0}}$. Then $\overline{\vec{\mathbf{v}}_2} = \vec{\mathbf{v}}_1$. Finally, $\vec{\mathbf{v}}_2 = \overline{\overline{\vec{\mathbf{v}}_2}} = \overline{\vec{\mathbf{v}}_1} = \begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}$. These details prove:

Lemma 9.2 If $(a+ib, \vec{\mathbf{v}})$ is an eigenpair of A, then formally replacing i by $-i$ in this eigenpair finds a second eigenpair for the conjugate eigenvalue.

Eigenvector for $\lambda_3 = 3$.

$$A - \lambda_3 I = \begin{pmatrix} 1-\lambda_3 & 2 & 0 \\ -2 & 1-\lambda_3 & 0 \\ 0 & 0 & 3-\lambda_3 \end{pmatrix}$$

[5] The complex conjugate is defined by $\overline{a+ib} = a - ib$ (replace i by $-i$). Two useful rules are $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ and $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$. Conjugation rules extend to vectors and matrices by applying scalar rules componentwise, e.g., $\overline{\vec{\mathbf{u}} + \vec{\mathbf{v}}} = \overline{\vec{\mathbf{u}}} + \overline{\vec{\mathbf{v}}}$ and $\overline{A\vec{\mathbf{x}}} = \overline{A}\,\overline{\vec{\mathbf{x}}}$.

$$= \begin{pmatrix} -2 & 2 & 0 \\ -2 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\approx \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{Multiply rule.}$$

$$\approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{Combination and multiply.}$$

$$= \mathbf{rref}(A - \lambda_3 I) \qquad \text{Reduced echelon form.}$$

The partial derivative $\partial_{t_1}\vec{\mathbf{v}}$ of the general solution $x = 0$, $y = 0$, $z = t_1$ is eigenvector $\vec{\mathbf{v}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Example 9.5 (Data Conversion)

The data conversion problem

$$\begin{cases} y_1 &= x_1, \\ y_2 &= 0.001x_2, \\ y_3 &= 0.01x_3. \end{cases}$$

is diagonalizable. The three eigenpairs of A are defined by

$$\begin{array}{lll} \lambda_1 = 1, & \lambda_2 = 0.001, & \lambda_3 = 0.01, \\ \vec{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, & \vec{\mathbf{v}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{array}$$

Solution: References: Theorem 9.3 page 671 and Theorem 9.7 page 674.

The example was introduced in equation (10) page 677. The equations can be written as $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, where $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.01 \end{pmatrix}$ is already a diagonal matrix, eigenpairs given by Theorem 9.7 page 674.

Answers can be verified directly from the eigenpair equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ without using theorems. For instance, when $\vec{\mathbf{v}} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $\lambda = 0.001$, then the two sides $A\vec{\mathbf{v}}$ and $\lambda\vec{\mathbf{v}}$ are computed from matrix multiply, each giving the same answer $\begin{pmatrix} 0 \\ 0.001 \\ 0 \end{pmatrix}$, therefore $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ is valid and $(\lambda, \vec{\mathbf{v}})$ is an eigenpair of A.

Example 9.6 (Decomposition $A = PDP^{-1}$)

Decompose $A = PDP^{-1}$ where P, D are eigenvector and eigenvalue packages, respectively, for the 3×3 matrix

$$A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Illustrate Fourier replacement for this matrix.

Solution: By the preceding example, the eigenpairs are

$$\left(1+2i, \begin{pmatrix} -i \\ 1 \\ 0 \end{pmatrix}\right), \quad \left(1-2i, \begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}\right), \quad \left(3, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right).$$

The packages are therefore

$$D = \textbf{diag}(1+2i, 1-2i, 3), \quad P = \begin{pmatrix} -i & i & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Fourier replacement. The model:

$$A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\right) = c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3$$

It means $A\vec{\mathbf{x}}$ changes $\vec{\mathbf{x}}$ by replacing the basis $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ by scaled vectors $\lambda_1\vec{\mathbf{v}}_1$, $\lambda_2\vec{\mathbf{v}}_2$, $\lambda_3\vec{\mathbf{v}}_3$. Explicitly,

$$\begin{aligned} \vec{\mathbf{x}} &= c_1\begin{pmatrix} -i \\ 1 \\ 0 \end{pmatrix} + c_2\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix} + c_3\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ implies} \\ A\vec{\mathbf{x}} &= c_1(1+2i)\begin{pmatrix} -i \\ 1 \\ 0 \end{pmatrix} + c_2(1-2i)\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix} + c_3(3)\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Example 9.7 (Diagonalization I)

Report **diagonalizable** or **non-diagonalizable** for the 4×4 matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

If A is diagonalizable, then report eigenvector and eigenvalue packages P, D.

Solution: Reference: page 674 for definitions and theorems.

The matrix A is **non-diagonalizable**, because it fails to have 4 eigenpairs. The details:

Eigenvalues.

$0 = \det(A - \lambda I)$ Characteristic equation.

$$= \begin{vmatrix} 1-\lambda & 2 & 0 & 0 \\ -2 & 1-\lambda & 0 & 0 \\ 0 & 0 & 3-\lambda & 1 \\ 0 & 0 & 0 & 3-\lambda \end{vmatrix}$$

$$= \begin{vmatrix} 1-\lambda & 2 \\ -2 & 1-\lambda \end{vmatrix} (3-\lambda)^2 \qquad \text{Cofactor expansion applied twice.}$$

$$= \left((1-\lambda)^2+4\right)(3-\lambda)^2 \qquad \text{Sarrus' rule.}$$

The roots are $1 \pm 2i$, 3, 3, listed according to multiplicity.

Eigenpairs. They are

$$\left(1+2i, \begin{pmatrix} -i \\ 1 \\ 0 \\ 0 \end{pmatrix}\right), \quad \left(1-2i, \begin{pmatrix} i \\ 1 \\ 0 \\ 0 \end{pmatrix}\right), \quad \left(3, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}\right).$$

Matrix A is **non-diagonalizable**, because only three eigenpairs exist, instead of four. Details:

Eigenvector for $\lambda_1 = 1 + 2i$.

$$A - \lambda_1 I = \begin{pmatrix} 1-\lambda_1 & 2 & 0 & 0 \\ -2 & 1-\lambda_1 & 0 & 0 \\ 0 & 0 & 3-\lambda_1 & 1 \\ 0 & 0 & 0 & 3-\lambda_1 \end{pmatrix}$$

$$= \begin{pmatrix} -2i & 2 & 0 & 0 \\ -2 & -2i & 0 & 0 \\ 0 & 0 & 2-2i & 1 \\ 0 & 0 & 0 & 2-2i \end{pmatrix}$$

$$\approx \begin{pmatrix} -i & 1 & 0 & 0 \\ -1 & -i & 0 & 0 \\ 0 & 0 & 2-2i & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{Multiply rule, three times.}$$

$$\approx \begin{pmatrix} -i & 1 & 0 & 0 \\ -1 & -i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{Combination and multiply rule.}$$

$$\approx \begin{pmatrix} 1 & i & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{Combination and multiply rule.}$$

$$= \mathbf{rref}(A - \lambda_1 I) \qquad \text{Reduced echelon form.}$$

The general solution is $x_1 = -it_1$, $x_2 = t_1$, $x_3 = 0$, $x_4 = 0$. Then ∂_{t_1} applied to this solution gives the reported eigenpair for $\lambda = 1 + 2i$.

Eigenvector for $\lambda_2 = 1 - 2i$.
Because λ_2 is the conjugate of λ_1 and A is real, then an eigenpair for λ_2 is found from the eigenpair for λ_1 by replacing i by $-i$ throughout. See Lemma 9.2 page 685.

Eigenvector for $\lambda_3 = 3$. In theory, there can be one or two eigenpairs to report. It turns out there is only one, because of the following details. This single toolkit sequence establishes that A is **non-diagonalizable**. The other toolkit sequences could have been skipped, if only diagonalizability was the issue and we were clever enough to examine this case first.

$$\begin{aligned}
A - \lambda_3 I &= \begin{pmatrix} 1-\lambda_3 & 2 & 0 & 0 \\ -2 & 1-\lambda_3 & 0 & 0 \\ 0 & 0 & 3-\lambda_3 & 1 \\ 0 & 0 & 0 & 3-\lambda_3 \end{pmatrix} \\
&= \begin{pmatrix} -2 & 2 & 0 & 0 \\ -2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
&\approx \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} && \text{Multiply rule, two times.} \\
&\approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} && \text{Combination and multiply rule.} \\
&= \mathbf{rref}(A - \lambda_3 I) && \text{Reduced echelon form.}
\end{aligned}$$

Apply ∂_{t_1} to the general solution $x_1 = 0$, $x_2 = 0$, $x_3 = t_1$, $x_4 = 0$ to give the eigenvector matching the eigenpair reported above for $\lambda = 3$.

Example 9.8 (Diagonalization II)

Report **diagonalizable** or **non-diagonalizable** for the 4×4 matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

If A is diagonalizable, then assemble and report eigenvalue and eigenvector packages D, P.

Solution: Reference: page 674 for definitions and theorems.

The matrix A is **diagonalizable**, because it has 4 eigenpairs

$$\left(1+2i, \begin{pmatrix} -i \\ 1 \\ 0 \\ 0 \end{pmatrix}\right), \quad \left(1-2i, \begin{pmatrix} i \\ 1 \\ 0 \\ 0 \end{pmatrix}\right), \quad \left(3, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}\right), \quad \left(3, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}\right).$$

Then the eigenpair packages are given by

$$D = \begin{pmatrix} -1+2i & 0 & 0 & 0 \\ 0 & 1-2i & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \quad P = \begin{pmatrix} -i & i & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The details parallel the previous example, except for the calculation of eigenvectors for $\lambda_3 = 3$. In this case, the reduced echelon form of $A - \lambda_3 I$ has two rows of zeros and parameters t_1, t_2 appear in the general solution. The answers given above for eigenvectors correspond to the partial derivatives ∂_{t_1}, ∂_{t_2} applied to the general solution of $(A - 3I)\vec{\mathbf{x}} = \vec{\mathbf{0}}$.

Example 9.9 (Non-diagonalizable Matrices)

Verify that the matrices

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

are all non-diagonalizable.

Solution: References: page 674 for definitions and theorems; Theorem 9.1 page 670 for $AP = PD$ and eigenpair equations.

Let A denote any one of these matrices and let n be its dimension.

Without computing eigenpairs, diagonalization will be decided. Assume, in order to reach a contradiction, that eigenpair packages D, P exist with D diagonal and P invertible such that $AP = PD$. Because A is triangular, its eigenvalues appear already on the diagonal of A. Only 0 is an eigenvalue and its multiplicity is n. Then the package D of eigenvalues is the zero matrix and an equation $AP = PD$ reduces to $AP = 0$. Multiply $AP = 0$ on the right by P^{-1} to obtain $A = 0$. But A is not the zero matrix, a contradiction. Conclusion: A is not diagonalizable.

Secondly, attack the diagonalization question directly, by solving for the eigenvectors corresponding to $\lambda = 0$. The toolkit sequence starts with $B = A - \lambda I$, but B equals $\mathbf{rref}(B)$ and no computations are required. The resulting reduced echelon system is just $x_1 = 0$, giving $n - 1$ free variables. Therefore, the eigenvectors of A corresponding to $\lambda = 0$ are the last $n - 1$ columns of the identity matrix I. Because A does not have n independent eigenvectors, then A is not diagonalizable.

Similar examples of non-diagonalizable matrices A can be constructed with A having from 1 up to $n - 1$ independent eigenvectors. The examples with ones on the super-diagonal and zeros elsewhere have exactly one eigenvector.

Example 9.10 (Fourier's 1822 Heat Model)

Fourier's 1822 treatise *Théorie analytique de la chaleur* studied dissipation of heat from a laterally insulated welding rod with ends held at 0°C (ice-packed ends). Assume the initial heat distribution along the rod at time $t = 0$ is given as a linear combination

$$f = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3.$$

Symbols $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are in the vector space V of all twice continuously differentiable functions on $0 \le x \le 1$, given explicitly as

$$\vec{\mathbf{v}}_1 = \sin \pi x, \quad \vec{\mathbf{v}}_2 = \sin 2\pi x, \quad \vec{\mathbf{v}}_3 = \sin 3\pi x.$$

Fourier's heat model re-scales[6] each of these vectors to obtain the temperature $u(t,x)$ at position x along the rod and time $t > 0$ as the model equation

$$u(t,x) = c_1 e^{-\pi^2 t}\vec{\mathbf{v}}_1 + c_2 e^{-4\pi^2 t}\vec{\mathbf{v}}_2 + c_3 e^{-9\pi^2 t}\vec{\mathbf{v}}_3.$$

Verify that $u(t,x)$ solves Fourier's partial differential equation heat model

$$\begin{array}{rcl} \dfrac{\partial u}{\partial t} & = & \dfrac{\partial^2 u}{\partial x^2}, \\ u(0,x) & = & f(x), \quad 0 \le x \le 1, \quad \text{initial temperature}, \\ u(t,0) & = & 0, \quad \text{zero Celsius at rod's left end}, \\ u(t,1) & = & 0, \quad \text{zero Celsius at rod's right end}. \end{array}$$

Solution: First, let's prove that the partial differential equation is satisfied by Fourier's solution $u(t,x)$. This is done by expanding the left side (LHS) and right side (RHS) of the differential equation separately, then comparing the two answers for equality.

Trigonometric functions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are solutions of three different linear ordinary differential equations: $u'' + \pi^2 u = 0$, $u'' + 4\pi^2 u = 0$, $u'' + 9\pi^2 u = 0$. Because of these differential equations, calculus derivatives can be computed:

$$\frac{\partial^2 u}{\partial x^2} = -\pi^2 c_1 e^{-\pi^2 t}\vec{\mathbf{v}}_1 - 4\pi^2 c_2 e^{-4\pi^2 t}\vec{\mathbf{v}}_2 - 9\pi^2 c_3 e^{-9\pi^2 t}\vec{\mathbf{v}}_3.$$

Similarly, computing $\partial_t u(t,x)$ involves just the differentiation of exponential functions, giving

$$\frac{\partial u}{\partial t} = -\pi^2 c_1 e^{-\pi^2 t}\vec{\mathbf{v}}_1 - 4\pi^2 c_2 e^{-4\pi^2 t}\vec{\mathbf{v}}_2 - 9\pi^2 c_3 e^{-9\pi^2 t}\vec{\mathbf{v}}_3.$$

Because the second display is exactly the first, then LHS = RHS, proving that the partial differential equation is satisfied.

The relation $u(0,x) = f(x)$ holds because each exponential factor becomes $e^0 = 1$ when $t = 0$.

The two relations $u(t,0) = u(t,1) = 0$ hold because each of $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ vanish at $x = 0$ and $x = 1$. The verification is complete.

Example 9.11 (Powers and Fourier Replacement)

Let 3×3 matrix A have eigenpairs $(\lambda_1, \vec{\mathbf{v}}_i)$, $i = 1, 2, 3$ and (9) holds. Find the powers $A^k\vec{\mathbf{x}}$ by Fourier's Replacement equation (8) with just the basic vector space toolkit, showing

$$A^k\vec{\mathbf{x}} = x_1\lambda_1^k\vec{\mathbf{v}}_1 + x_2\lambda_2^k\vec{\mathbf{v}}_2 + x_3\lambda_3^k\vec{\mathbf{v}}_3$$

Solution: The vector toolkit for $\mathcal{R}^3$ is used to compute powers:

$$\begin{array}{rcll} A\vec{\mathbf{x}} & = & x_1\lambda_1\vec{\mathbf{v}}_1 + x_2\lambda_2\vec{\mathbf{v}}_2 + x_3\lambda_3\vec{\mathbf{v}}_3 & \\ A^2\vec{\mathbf{x}} & = & A(x_1\lambda_1\vec{\mathbf{v}}_1 + x_2\lambda_2\vec{\mathbf{v}}_2 + x_3\lambda_3\vec{\mathbf{v}}_3) & \\ & = & x_1\lambda_1^2\vec{\mathbf{v}}_1 + x_2\lambda_2^2\vec{\mathbf{v}}_2 + x_3\lambda_3^2\vec{\mathbf{v}}_3 & \text{by (8)} \\ & \vdots & & \\ A^k\vec{\mathbf{x}} & = & x_1\lambda_1^k\vec{\mathbf{v}}_1 + x_2\lambda_2^k\vec{\mathbf{v}}_2 + x_3\lambda_3^k\vec{\mathbf{v}}_3 & \end{array}$$

[6] The scale factors are not constants nor are they eigenvalues, but rather, they are exponential functions of t for fixed t, as is the case for matrix differential equations $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$. See Example 9.13

The calculations do not use matrix multiply and the answer does not depend upon finding previous powers A^2, A^3, A^4, ...

Fourier replacement reduces computational effort. Matrix–vector multiplication to produce $\vec{\mathbf{y}}_k = A^k\vec{\mathbf{x}}$ requires $9k$ multiply operations whereas Fourier replacement gives the answer with $3k+9$ multiply operations.

Example 9.12 (Change of Variable $\vec{\mathbf{x}} = P\vec{\mathbf{u}}$ for Differential Equations)

Matrix $A = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & -5 \end{pmatrix}$ has eigenpairs $(\vec{\mathbf{v}}_1, \lambda_1), (\vec{\mathbf{v}}_2, \lambda_2), (\lambda_3, \vec{\mathbf{v}}_3)$ with three independent eigenvectors given by equation (12). Define $\vec{\mathbf{x}} = P\vec{\mathbf{u}}$, $P = \langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \rangle$, $D = \mathbf{diag}(\lambda_1, \lambda_2, \lambda_3)$. Show that $\vec{\mathbf{x}} = P\vec{\mathbf{u}}$ changes $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ into $\vec{\mathbf{u}}' = D\vec{\mathbf{u}}$, which is the diagonal system of growth-decay equations

$$\begin{cases} u_1' &= \quad u_1, \\ u_2' &= \quad 2u_2, \\ u_3' &= \quad -5u_3. \end{cases}$$

Solution: The calculus derivative of a vector function is performed componentwise. Matrix multiply as a linear combination of columns shows that equation $\vec{\mathbf{x}}(t) = P\vec{\mathbf{u}}(t)$ has derivative $\vec{\mathbf{x}}'(t) = P\vec{\mathbf{u}}'(t)$, because entries of P are constants. Then equation $\vec{\mathbf{x}}(t) = P\vec{\mathbf{u}}(t)$ can change $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ into a differential equation in variable $\vec{\mathbf{u}}$. The details:

$\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$	Given.
$P\vec{\mathbf{u}}'(t) = AP\vec{\mathbf{u}}(t)$	Use $\vec{\mathbf{x}}'(t) = P\vec{\mathbf{u}}'(t)$, $\vec{\mathbf{x}}(t) = P\vec{\mathbf{u}}(t)$.
$P\vec{\mathbf{u}}'(t) = PD\vec{\mathbf{u}}(t)$	because $AP = PD$ (A is diagonalizable).
$\vec{\mathbf{u}}'(t) = D\vec{\mathbf{u}}(t)$	because P has an inverse.

The eigenvalues of triangular matrix A are the diagonal entries: 1,2,-5. Then $D = \mathbf{diag}(1, 2, -5)$ and $\vec{\mathbf{u}}' = D\vec{\mathbf{u}}$ is the reported system of growth-decay differential equations.

Example 9.13 (Differential Equations and Fourier Replacement)

Solve by Fourier re-scaling $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ with $A = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & -5 \end{pmatrix}$. The scalar form:

$$\begin{cases} x_1' &= x_1 + 3x_2, \\ x_2' &= \phantom{x_1 +{}} 2x_2 - x_3, \\ x_3' &= - 5x_3. \end{cases}$$

The answer uses the eigenpairs $(\vec{\mathbf{v}}_1, \lambda_1), (\vec{\mathbf{v}}_2, \lambda_2), (\lambda_3, \vec{\mathbf{v}}_3)$ of matrix A in equation (12):

$$(14)\quad \begin{cases} \vec{\mathbf{x}}(t) &= c_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2 + c_3 e^{\lambda_3 t}\vec{\mathbf{v}}_3, \quad \text{realized as} \\ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= c_1 e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2 e^{2t} \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + c_3 e^{-5t} \begin{pmatrix} 1 \\ -2 \\ -14 \end{pmatrix}. \end{cases}$$

Solution: Fourier's re-scaling idea applies to linear differential equations, as follows. First, expand the initial condition $\vec{\mathbf{x}}(0)$ in terms of basis elements:

$$\vec{\mathbf{x}}(0) = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3.$$

Fourier's re-scaling replaces each $\vec{\mathbf{v}}_i$ by the re-scaled vector $e^{\lambda_i t}\vec{\mathbf{v}}_i$. The result:

$$\vec{\mathbf{y}} = c_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2 + c_3 e^{\lambda_3 t}\vec{\mathbf{v}}_3 \tag{15}$$

How is this related to Fourier re-scaling? Answer: at each fixed instant t, the basis vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are replaced by $\Lambda_1\vec{\mathbf{v}}_1$, $\Lambda_2\vec{\mathbf{v}}_2$, $\Lambda_3\vec{\mathbf{v}}_3$ where

$$\Lambda_1 = e^{\lambda_1 t}, \quad \Lambda_2 = e^{\lambda_2 t}, \quad \Lambda_3 = e^{\lambda_3 t}.$$

Why is the solution $\vec{\mathbf{x}}(t) = c_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2 + c_3 e^{\lambda_3 t}\vec{\mathbf{v}}_3$? Answer: Evaluate the LHS and RHS of the differential equation $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ and compare formulas.

$$\begin{aligned}
\text{LHS} &= \vec{\mathbf{x}}'(t) \\
&= c_1\lambda_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2\lambda_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2 + c_3\lambda_3 e^{\lambda_3 t}\vec{\mathbf{v}}_3 \\
&= c_1\lambda_1\Lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\Lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\Lambda_3\vec{\mathbf{v}}_3 \\
\text{RHS} &= A\vec{\mathbf{x}}(t) \\
&= A(c_1\Lambda_1\vec{\mathbf{v}}_1 + c_2\Lambda_2\vec{\mathbf{v}}_2 + c_3\Lambda_3\vec{\mathbf{v}}_3) \\
&= c_1\lambda_1\Lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\Lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\Lambda_3\vec{\mathbf{v}}_3 \quad \text{by Theorem 9.10.}
\end{aligned}$$

The last equality is tricky: equation

$$A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\right) = c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3$$

in Theorem 9.10 is applied with c_1, c_2, c_3 replaced by $c_1\Lambda_1, c_2\Lambda_2, c_3\Lambda_3$.

Justification of the solution is done with Example 9.12 after inserting exponential solutions for the growth-decay equations. A summary of the re-scaling method:

1. Expand $\vec{\mathbf{x}}(0)$ as a linear combination of eigenvectors.

2. Change on the left $\vec{\mathbf{x}}(0)$ to $\vec{\mathbf{x}}(t)$, then re-scale the linear combination on the right with scale factors $\Lambda_1 = e^{\lambda_1 t}$, $\Lambda_2 = e^{\lambda_2 t}$, $\Lambda_3 = e^{\lambda_3 t}$.

Proofs and Details

Proof of Theorem 9.1, Eigenpairs and $AP = PD$:

Let

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle, \quad D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Write the two matrix multiply equations AP and PD in expanded form

$$AP = \left\langle A\vec{\mathbf{v}}_1 | A\vec{\mathbf{v}}_2 | A\vec{\mathbf{v}}_3 \right\rangle, \quad PD = \left\langle \lambda_1\vec{\mathbf{v}}_1 | \lambda_2\vec{\mathbf{v}}_2 | \lambda_3\vec{\mathbf{v}}_3 \right\rangle. \tag{16}$$

***AP = PD* implies equation (4)**. Assume $AP = PD$. Because equal matrices have equal columns, the columns left and right in the equation $AP = PD$ must match, using expansion (16). Then

$$A\vec{\mathbf{v}}_1 = \lambda_1\vec{\mathbf{v}}_1, \quad A\vec{\mathbf{v}}_2 = \lambda_2\vec{\mathbf{v}}_2, \quad A\vec{\mathbf{v}}_3 = \lambda_3\vec{\mathbf{v}}_3,$$

which means equation (4) holds.

Equation (4) implies $AP = PD$. Assume eigenpair equations (4) hold. Then the two matrices AP and PD in expansion (16) have equal columns. Equality of matrices implies $AP = PD$. ■

Proof of Theorem 9.2, Eigenvalues of A:

An eigenvalue λ is a number such that equation $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$ has a nonzero solution $\vec{\mathbf{x}}$. Let $B = A - \lambda I$. Then λ is an eigenvalue means $B\vec{\mathbf{x}} = \vec{\mathbf{0}}$ has a nonzero solution $\vec{\mathbf{x}}$. Homogeneous equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ has a nonzero solution $\vec{\mathbf{v}}$ if and only if there are infinitely many solutions. Because the matrix is square, infinitely many solutions occur if and only if $\mathbf{rref}(B)$ has a row of zeros. Determinant theory gives a more concise statement: $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ has infinitely many solutions if and only if $\det(B) = 0$. ■

Proof of Theorem 9.3, Find Eigenvectors:

Question: Why does the solution of $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ have invented symbols? Isn't there just one solution?

Answer: According to the *three possibilities*, homogeneous equation $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$ should have unique solution $\vec{\mathbf{v}} = \vec{\mathbf{0}}$ or else infinitely many solutions. An eigenvector cannot be zero. To get infinitely many solutions there has to be at least one free variable, causing the *last frame algorithm* to be applied with invented symbols $t_1, t_2, \ldots$.

The equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ is equivalent to $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$. Because λ is a root of characteristic equation $|A - \lambda I| = 0$, then $\det(B) = 0$ and B has no inverse, equivalent to $\mathbf{rref}(B) \neq I$. Then square matrix $\mathbf{rref}(B)$ must have a row of zeros, which means there is at least one free variable. The last frame algorithm applies with invented symbols t_1, t_2, ... A vector basis $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots$ for the nullspace of B is obtained from the list of vector partial derivatives on symbols t_1, t_2, ... These vectors are Strang's special solutions, which are known to be collectively independent. The nullspace of B is the span of Strang's special solutions $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots$. If $A\vec{\mathbf{w}} = \lambda\vec{\mathbf{w}}$, then $B\vec{\mathbf{w}} = \vec{\mathbf{0}}$, so $\vec{\mathbf{w}}$ belongs to the nullspace of B:

$$\vec{\mathbf{w}} = \text{a linear combination of } \quad \vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots \quad ■$$

Proof of Theorem 9.4, Independence of Eigenvectors:

Let's solve $c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$ for c_1, c_2. The vectors are independent provided the only solution is $c_1 = c_2 = 0$. Apply A to this equation, obtaining $c_1 A\vec{\mathbf{v}}_1 + c_2 A\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$. Use $A\vec{\mathbf{v}}_1 = \lambda_1\vec{\mathbf{v}}_1$ and $A\vec{\mathbf{v}}_2 = \lambda_2\vec{\mathbf{v}}_2$ to obtain $c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$. Multiply $c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$ by λ_1 and subtract it from $c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$ to get $c_1(\lambda_1 - \lambda_1)\vec{\mathbf{v}}_1 + c_2(\lambda_2 - \lambda_1)\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$. Because $\lambda_2 \neq \lambda_1$, cancel $\lambda_2 - \lambda_1$ to give $c_2\vec{\mathbf{v}}_2 = \vec{\mathbf{0}}$. The assumption $\vec{\mathbf{v}}_2 \neq \vec{\mathbf{0}}$ implies $c_2 = 0$. Return to the first equation and use $c_2 = 0$ to obtain $c_1\vec{\mathbf{v}}_1 = \vec{\mathbf{0}}$. Because $\vec{\mathbf{v}}_1 \neq \vec{\mathbf{0}}$, then $c_1 = 0$. This proves $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are independent.

The general case is proved by **Mathematical Induction** on k (see the footnote in the proof of Theorem 9.5). The case $k = 1$ follows because a nonzero vector is an independent set. Assume it holds for $k-1$ and let's prove it for k, when $k > 1$. To prove independence, we must solve for c_1, ..., c_k in the test equation

$$c_1\vec{\mathbf{v}}_1 + \cdots + c_k\vec{\mathbf{v}}_k = \vec{\mathbf{0}}.$$

Create a second equation by multiplication of the test equation by A, effectively replacing each c_i by $\lambda_i c_i$, due to the eigenpair equation $A\vec{\mathbf{v}}_i = \lambda_i\vec{\mathbf{v}}_i$. Then multiply the test equation by λ_1 and subtract the two equations to get the new equation

$$c_1(\lambda_1 - \lambda_1)\vec{\mathbf{v}}_1 + c_2(\lambda_1 - \lambda_2)\vec{\mathbf{v}}_2 + \cdots + c_k(\lambda_1 - \lambda_k)\vec{\mathbf{v}}_k = \vec{\mathbf{0}}.$$

The first term is zero. Apply the induction hypothesis to the remaining $k-1$ vectors, then independence implies all coefficients $(\lambda_1 - \lambda_i)c_i$ are zero. Because $\lambda_1 - \lambda_i \neq 0$ for $i > 1$, then c_2 through c_k are zero. Substitute the zero values into the test equation to obtain $c_1\vec{\mathbf{v}}_1 = \vec{\mathbf{0}}$. Because $\vec{\mathbf{v}}_1 \neq \vec{\mathbf{0}}$, then $c_1 = 0$. Therefore all $c_i = 0$. The induction is complete. ■

Proof of Theorem 9.5, Unions of Eigenvectors:

Details (1). Assume there is a nonzero vector $\vec{\mathbf{v}}$ in the intersection, which must be an eigenvector for both λ and μ. Then two eigenpairs$(\lambda, \vec{\mathbf{v}}_1)$ and $(\mu, \vec{\mathbf{v}}_2)$ have been found, $\vec{\mathbf{v}}_1 = \vec{\mathbf{v}}_2 = \vec{\mathbf{v}}$, which violates Theorem 9.4, because $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ must be independent.

Details (2). Let's proceed by induction on the number k of eigenvalues used to construct U.[7] Let S_k be the statement that U = union of $\mathcal{B}(\lambda_1)$, …, $\mathcal{B}(\lambda_k)$ has independent elements, no matter how the k distinct eigenvalues $\{\lambda_i\}_{i=1}^k$ are selected and no matter how the bases are chosen.

Statement S_1 is true, because $\mathcal{B}(\lambda_1)$ is a list of independent elements.

Assume S_k is true. The proof that S_{k+1} is true will be deferred to the exercises. Revealed here are the fundamental ideas, by examining the cases $k = 2$ and $k = 3$.

Case $k = 2$. Then U is a list of vectors, some from $\mathcal{B}(\lambda_1)$ and some from $\mathcal{B}(\lambda_2)$. The test equation for independence of this list of vectors is a linear combination of the vectors equal to the zero vector. The objective is to prove that the coefficients in this linear combination are all zero. Rearrange the test equation in the form

$$\text{Terms using vectors from } \mathcal{B}(\lambda_1) = \text{Terms using vectors from } \mathcal{B}(\lambda_2)$$

The left side of the above equation is an eigenvector $\vec{\mathbf{v}}_1$ for eigenvalue λ_1, giving eigenpair $(\lambda_1, \vec{\mathbf{v}}_1)$. Similarly, the right side determines an eigenpair $(\lambda_2, \vec{\mathbf{v}}_2)$. The previous theorem says that $\vec{\mathbf{v}}_1$ and $\vec{\mathbf{v}}_2$ are independent, if nonzero. Analyzing cases, then both $\vec{\mathbf{v}}_1$ and $\vec{\mathbf{v}}_2$ are the zero vector. By independence of bases $\mathcal{B}(\lambda_1)$ and $\mathcal{B}(\lambda_2)$, all coefficients are zero, proving independence of the list U.

Case $k = 3$. Let U_2 be the union of bases $\mathcal{B}(\lambda_1), \mathcal{B}(\lambda_2)$, which is a list of vectors $\vec{\mathbf{v}}_1$, …, $\vec{\mathbf{v}}_q$. Given is U = the union of bases $\mathcal{B}(\lambda_1)$, $\mathcal{B}(\lambda_2)$, $\mathcal{B}(\lambda_3)$. The test equation for independence of the vectors in list U is a linear combination equal to the zero vector. This equation has a summation left and the zero vector on the right. Isolate left in this equation those terms that involve basis vectors from $\mathcal{B}(\lambda_3)$, then move the remaining terms to the right. The rearranged equation looks like

$$\text{Sum of terms from } \mathcal{B}(\lambda_3) = \text{Sum of terms from } U_2$$

The left side is an eigenvector $\vec{\mathbf{v}}$ for λ_3. The right side is a linear combination from U_2, which means $\vec{\mathbf{v}} = \sum_{j=1}^q c_j \vec{\mathbf{v}}_j$. Write two equations for $\lambda_3\vec{\mathbf{v}}$, using the eigenpair equation $A\vec{\mathbf{v}} = \lambda_3\vec{\mathbf{v}}$:

$$\lambda_3\vec{\mathbf{v}} = \sum_{j=1}^q c_j\lambda_3\vec{\mathbf{v}}_j, \quad \lambda_3\vec{\mathbf{v}} = A\vec{\mathbf{v}} = \sum_{j=1}^q c_jA\vec{\mathbf{v}}_j = \sum_{j=1}^q c_j\lambda(\vec{\mathbf{v}}_j)\vec{\mathbf{v}}_j,$$

[7] Mathematical induction is this theorem:
(1) For each counting number n, S_n is a statement that is either true or false.
(2) Statement S_1 is true.
(3) If statement S_k is true, then statement S_{k+1} is true.
Conclusion: All the statements are true.

where $\lambda(\vec{\mathbf{v}}_j)$ is the eigenvalue for eigenvector $\vec{\mathbf{v}}_j$. Put these two equations together, then move the right side to the left and collect terms:

$$\sum_{j=1}^{q} c_j(\lambda_3 - \lambda(\vec{\mathbf{v}}_j))\vec{\mathbf{v}}_j = \vec{\mathbf{0}}.$$

Because S_2 is true, then the vectors $\{\vec{\mathbf{v}}_j\}_{j=1}^{q}$ are independent. Therefore, all coefficients $c_j(\lambda_3 - \lambda(\vec{\mathbf{v}}_j)) = 0$. **Reminder**: symbols $\lambda_1, \ldots, \lambda_k$ are distinct values and list all eigenvalues of A. Then $\lambda_3 \neq \lambda(\vec{\mathbf{v}}_j)$ implies all $c_j = 0$. This implies $\vec{\mathbf{v}} = \vec{\mathbf{0}}$, which in turn implies that all coefficients in the independence test are zero. Therefore, U is a list of independent vectors. The induction proof is completed by the exercises of this section. ■

Proof of Theorem 9.6, Finding Independent Eigenvectors:

Exercises of this section show that $\partial_{t_1}\vec{\mathbf{v}}, \partial_{t_2}\vec{\mathbf{v}}, \ldots$ are independent vectors which constitute a basis $\mathcal{B}(\lambda)$ for the solution set of the eigenpair equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. These are Strang's special solutions for $B\vec{\mathbf{v}} = \vec{\mathbf{0}}$. Theorem 9.5 says that the union U of bases $\mathcal{B}(\lambda_1)$, …, $\mathcal{B}(\lambda_k)$ so constructed from the distinct eigenvalues λ_1, …, λ_k of A is an independent set. For an example where U does not span $\mathcal{R}^n$, let $n = 2$ and $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, a matrix with just one eigenpair. ■

Proof of Theorem 9.8, Diagonalization:

Details 1. To prove $A = PDP^{-1}$, multiply right across $AP = PD$ by matrix P^{-1}, which isolates A on the left. Then $A = AI = APP^{-1} = PDP^{-1}$.

Details 2. Define the change of variables $\vec{\mathbf{X}} = P\vec{\mathbf{x}}$, $\vec{\mathbf{Y}} = P\vec{\mathbf{y}}$. Substitute into the equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ as follows:

$$\vec{\mathbf{Y}} = P\vec{\mathbf{y}} = PA\vec{\mathbf{x}} = PAP^{-1}P\vec{\mathbf{x}} = D\vec{\mathbf{X}}.$$

The result is the diagonal system $\vec{\mathbf{Y}} = D\vec{\mathbf{X}}$.

Details 3. Let column vector $\vec{\mathbf{c}}$ have components $c_1, \ldots, c_n$. To be proved: the left side of $A(c_1\vec{\mathbf{v}}_1 + \cdots + c_n\vec{\mathbf{v}}_n) = c_1\lambda_1\vec{\mathbf{v}}_1 + \cdots + c_n\lambda_n\vec{\mathbf{v}}_n$ is the expansion of $AP\vec{\mathbf{c}}$, while the right side is the expansion of $PD\vec{\mathbf{c}}$. Assume these statements are proved, for the moment, details delayed. Then $AP = PD$ implies $AP\vec{\mathbf{c}} = PD\vec{\mathbf{c}}$ for all vectors $\vec{\mathbf{c}}$, which means (7) holds. It remains to expand $AP\vec{\mathbf{c}}$ and $PD\vec{\mathbf{c}}$, assuming $AP = PD$, or what is the same, the eigenpair equations hold: $A\vec{\mathbf{c}}_i = \lambda_i\vec{\mathbf{v}}_i$ for $1 \le i \le n$.

The expansion of $AP\vec{\mathbf{c}}$:

$AP\vec{\mathbf{c}} = A < \vec{\mathbf{v}}_1\vert\cdots\vert\vec{\mathbf{v}}_n > \vec{\mathbf{c}}$	Use definition $P =< \vec{\mathbf{v}}_1\vert\cdots\vert\vec{\mathbf{v}}_n >$.
$= A(c_1\vec{\mathbf{v}}_1 + \ldots + c_n\vec{\mathbf{v}}_n)$	Matrix multiply as a linear combination of the columns.
$= c_1A\vec{\mathbf{v}}_1 + \ldots + c_nA\vec{\mathbf{v}}_n$	Linearity of matrix multiply.
$= c_1\lambda_1\vec{\mathbf{v}}_1 + \ldots + c_n\lambda_n\vec{\mathbf{v}}_n$	Eigenpair equations $A\vec{\mathbf{v}}_i = \lambda_i\vec{\mathbf{v}}_i$ for $1 \le i \le n$.

The expansion of $PD\vec{\mathbf{c}}$:

$PD\vec{\mathbf{c}} = P\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}\vec{\mathbf{c}}$	Definition of D.

$$\begin{aligned}
&= P\begin{pmatrix} c_1\lambda_1 \\ \vdots \\ c_n\lambda_n \end{pmatrix} && \text{Matrix multiply as a dot product.} \\
&= < \vec{\mathbf{v}}_1|\cdots|\vec{\mathbf{v}}_n > \begin{pmatrix} c_1\lambda_1 \\ \vdots \\ c_n\lambda_n \end{pmatrix} && \text{Definition of } P. \\
&= c_1\lambda_1\vec{\mathbf{v}}_1 + \ldots + c_n\lambda_n\vec{\mathbf{v}}_n && \text{Matrix multiply as a linear combination of columns.}
\end{aligned}$$

■

Proof of Theorem 9.9, Distinct Eigenvalues:
Each eigenvalue λ has at least one eigenvector. Because there are n distinct eigenvalues, then there are n eigenvectors. The list of these eigenvectors must be independent, by Theorem 9.5. Therefore, matrix A is diagonalizable. The remaining statements in the theorem are a consequence of Theorem 9.8. ■

Proof of Theorem 9.10, Matrix Form Fourier Replacement:
$\boxed{1}$ Let's prove $\vec{\mathbf{x}} = P\vec{\mathbf{c}}$ implies $\vec{\mathbf{y}} = PD\vec{\mathbf{c}}$, assuming the Fourier replacement equation. Let $\vec{\mathbf{x}} = P\vec{\mathbf{c}}$. Expand the product $P\vec{\mathbf{c}}$ viewing matrix multiply as a linear combination of the columns. Then $\vec{\mathbf{x}} = P\vec{\mathbf{c}} = c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3$. Because Fourier replacement holds, then

$$\begin{aligned}
\vec{\mathbf{y}} &= c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3 && \text{Re-scale } \vec{\mathbf{x}}. \\
&= P\begin{pmatrix} c_1\lambda_1 \\ c_2\lambda_2 \\ c_3\lambda_3 \end{pmatrix} && \text{Matrix multiply as a linear combination of columns.} \\
&= P\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} && \text{Matrix multiply as a dot product.} \\
&= PD\vec{\mathbf{c}} && \text{Definition of } D.
\end{aligned}$$

$\boxed{2}$ Definition $A = PDP^{-1}$ was discovered by solving for A in equation $AP = PD$ ($AP = PD$ means A is diagonalizable). To prove $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, first solve $\vec{\mathbf{x}} = P\vec{\mathbf{c}}$ for $\vec{\mathbf{c}} = P^{-1}\vec{\mathbf{x}}$. Then $A\vec{\mathbf{x}} = PDP^{-1}\vec{\mathbf{x}} = PD\vec{\mathbf{c}} = \vec{\mathbf{y}}$ by $\boxed{1}$.

$\boxed{3}$ To prove $A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\right) = c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3$, replace its left side by $A\vec{\mathbf{x}}$ and right side by $\vec{\mathbf{y}}$. Then it suffices to prove $A\vec{\mathbf{x}} = \vec{\mathbf{y}}$, which has already been proved in $\boxed{2}$. ■

Proof of Theorem 9.11, Re-scaling and Diagonalization:
(a) Use relation $A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2 + c_3\vec{\mathbf{v}}_3\right) = c_1\lambda_1\vec{\mathbf{v}}_1 + c_2\lambda_2\vec{\mathbf{v}}_2 + c_3\lambda_3\vec{\mathbf{v}}_3$ from Theorem 9.10. Choose $c_1 = 1, c_2 = c_3 = 0$ to get $A\vec{\mathbf{v}}_1 = \lambda_1\vec{\mathbf{v}}_1$. Similarly, choose zeros and ones for c_1, c_2, c_3 to get $A\vec{\mathbf{v}}_2 = \lambda_2\vec{\mathbf{v}}_2$ and $A\vec{\mathbf{v}}_3 = \lambda_3\vec{\mathbf{v}}_3$. Then three eigenpair equations hold with independent eigenvectors and by definition A is diagonalizable.

(b) By Theorem 9.10 it suffices to prove $\vec{\mathbf{x}} = P\vec{\mathbf{c}}$ implies $A\vec{\mathbf{x}} = PD\vec{\mathbf{c}}$. If A is diagonalizable, then $AP = PD$, which gives $A\vec{\mathbf{x}} = AP\vec{\mathbf{c}} = PD\vec{\mathbf{c}}$ as required.

(c) If A is given and (8) holds, then **(a)** applies to prove A is diagonalizable. Conversely, if A is diagonalizable, then **(b)** applies and Fourier replacement (8) holds. ■

Exercises 9.1

Eigenanalysis

Classify as true or false. If false, then explain.

1. The purpose of eigenanalysis is to discover a new coordinate system.

2. Eigenanalysis can discover an opportunistic change of coordinates.

3. A matrix can have eigenvalue 0.

4. Eigenvalues are scale factors, imagined to be measurement units.

5. Eigenvectors are directions.

6. For each eigenvalue of a matrix A, there always exists at least one eigenpair.

7. If A^{-1} has eigenvalue λ, then A has eigenvalue $1/\lambda$.

8. Eigenvectors cannot be $\vec{\mathbf{0}}$.

9. The transpose of A has the same eigenvalues as A.

10. Eigenpairs $(\lambda, \vec{\mathbf{v}})$ of A satisfy the equation $(A-\lambda I)\vec{\mathbf{v}} = \vec{\mathbf{0}}$.

Eigenpairs of a Diagonal Matrix

Find eigenpairs of A without computation. Use Theorem 9.7.

11. $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$

12. $\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$

13. $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

14. $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

15. $\begin{pmatrix} 7 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -6 \end{pmatrix}$

16. $\begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$

Fourier Replacement

Let symbols c_1, c_2 represent arbitrary constants. Let 2×2 matrix A have Fourier replacement equation

$$A\left(c_1\begin{pmatrix}1\\1\end{pmatrix}+c_2\begin{pmatrix}1\\2\end{pmatrix}\right) = 2c_1\begin{pmatrix}1\\1\end{pmatrix}-5c_2\begin{pmatrix}1\\2\end{pmatrix}$$

17. Display the eigenpairs of A.

18. Display the replacement equation if the eigenvalues $2, -5$ are replaced by $1, 0$.

19. Display the eigenpair packages P, D such that $AP = PD$.

20. Find A.

Eigenanalysis Facts

Mark as true or false, then explain your answer.

21. If matrix A has all eigenvalues zero, then A is the zero matrix.

22. If 2×2 matrix A has all eigenvalues zero, then Fourier's replacement equation is

$$A\left(c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2\right) = \vec{\mathbf{0}}.$$

23. There are infinitely many 2×2 matrices A with complex eigenvalues $1+i, 1-i$.

24. A real 2×2 matrix A with eigenvalues $2+3i, 2-3i$ cannot have a real eigenvector.

25. A real 2×2 matrix A with real eigenvalues has only real eigenvectors.

26. A real 2×2 matrix A with complex eigenvalues has only complex eigenvectors.

Eigenpair Packages and equation $AP = PD$

27. Suppose A has eigenpair packages. Explain why there are so many different answers for P, D.

28. Suppose $AP = PD$ and $AQ = QD$ hold (same diagonal matrix D). Does $P = Q$?

29. Find one choice of P and D for $A =$ 2×2 diagonal matrix.

30. Given $A = 3 \times 3$ zero matrix, find one choice of P and D with column one of P equal to $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.

Matrix Eigenanalysis Method

31. The eigenvalues of $\begin{pmatrix} 1 & 3 \\ 1 & 4 \end{pmatrix}$ satisfy a quadratic equation. Find the equation and solve for the eigenvalues.

32. Find the eigenvalues of $\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$.

33. Find all eigenpairs of $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix}$.

34. A triangular $n \times n$ matrix with distinct diagonal entries has n eigenpairs. Provide a detailed proof for the case $n = 3$.

35. Find all eigenpairs of $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$.

36. A triangular $n \times n$ matrix may not have n eigenpairs. Provide a series of examples for dimensions $n = 2, 3, 4, 5$.

37. Prove that equations $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$ and $(A - \lambda I)\vec{\mathbf{x}} = \vec{\mathbf{0}}$ have exactly the same solutions $\vec{\mathbf{x}}$.

38. Cite basic linear algebra theorems to prove that $(A-\lambda I)\vec{\mathbf{x}} = \vec{\mathbf{0}}$ has a nonzero solution $\vec{\mathbf{x}}$ if and only if λ is a root of the characteristic equation $|A-\lambda I| = 0$.

Basis of Eigenvectors

The problem $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$ has a standard general solution $\vec{x}$ with invented symbols $t_1, t_2, t_3, \ldots$. **Strang's special solutions** are defined to be the vector partial derivatives of $\vec{x}$ with respect to the invented symbols.

39. Why are Strang's special solutions independent?

40. Prove that linear combinations of Strang's special solutions provide all possible solutions of $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$.

Independence of Eigenvectors

Eigenvectors of matrix A for eigenvalue λ are the nonzero solutions of $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$.

41. Invent a 2×2 example A with eigenpairs $\left(2, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$, $\left(2, \begin{pmatrix} 5 \\ 5 \end{pmatrix}\right)$. Then explain why an eigenvector for eigenvalue λ is never unique.

42. Explain: *For a given eigenvalue λ, there are infinitely many eigenvectors.*

43. Explain: *Each solution $\vec{\mathbf{x}}$ of $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$ is a linear combination of Strang's special solutions for $B = A - \lambda I$.*

44. Let P be an invertible 3×3 matrix. Construct a matrix A which has eigenvectors equal to the columns of P and corresponding eigenvalues $-1, 0, 0$.

Eigenspaces

Let $\mathcal{B}(\lambda)$ denote some basis of eigenvectors for the eigenpair equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. The **eigenspace** for λ is the subspace $\mathbf{span}(\mathcal{B}(\lambda))$.

45. Explain: The eigenspace of λ does not depend on the choice of basis.

46. Every nonzero vector in eigenspace $\mathbf{span}(\mathcal{B}(\lambda))$ is an eigenvector of A for eigenvalue λ. Provide details of proof.

47. Justify that $\mathbf{span}(\mathcal{B}(\lambda))$ is a vector subspace of $\mathcal{R}^n$, one possible basis being Strang's special solutions for matrix $B = A - \lambda I$.

48. Find a 4×4 matrix A with only one eigenvalue $\lambda = 1$ such that eigenspace $\mathcal{B}(\lambda)$ (defined above) has dimension two.

Independence of Unions of Eigenvectors

Denote by $\mathcal{B}(\lambda)$ some basis for the eigenpair equation $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$.

49. Define U_1 to be the union of lists $\mathcal{B}(\lambda_1)$, $\mathcal{B}(\lambda_2)$ and define U_2 to be the union of lists $\mathcal{B}(\lambda_3)$, $\mathcal{B}(\lambda_4)$, where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ is a list of distinct eigenvalues of A. Prove that subspaces $V_1 = \mathbf{span}(U_1)$ and $V_2 = \mathbf{span}(U_2)$ intersect in only the zero vector.

50. Complete the details of the induction proof of Theorem 9.5, using the textbook details for $k = 3$.

51. Let U^* be a subset of the list U of independent vectors in Theorem 9.5. Explain why U^* is an independent set.

52. Let B_i be a subset of the list of independent vectors in $\mathcal{B}(\lambda_i)$, $i = 1, \ldots, p$. Explain why the union U^* of $B_1, \ldots, B_p$ is an independent set.

Diagonalization Theory

53. Let $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 8 \end{pmatrix}$.

(a) Find Strang's special solutions for each eigenvalue.
(b) Compare to Theorem 9.7 on diagonal matrices.

54. Let $\vec{\mathbf{v}}_!, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ be independent vectors in $\mathcal{R}^3$. Explain why $(0, \vec{\mathbf{v}}_1)$, $(0, \vec{\mathbf{v}}_2)$, $(0, \vec{\mathbf{v}}_3)$ is a complete set of eigenpairs for the 3×3 zero matrix. Does this contradict Theorem 9.7?

55. Write a proof of Theorem 9.7 for $n = 3$.

56. State Theorem 9.7 for $n \times n$ diagonal matrices and outline a proof.

Non-diagonalizable Matrices

Verify that the matrix is not diagonalizable by using the equation $AP = PD$.

57. $A = \begin{pmatrix} 5 & 2 \\ 0 & 5 \end{pmatrix}$

58. $A = \begin{pmatrix} 5 & 2 & 1 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{pmatrix}$

Distinct Eigenvalues

Find the eigenvalues.

59. $A = \begin{pmatrix} 2 & 6 \\ 5 & 3 \end{pmatrix}$ Ans: $8, -3$

60. $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ Ans: $0, 5$

61. $A = \begin{pmatrix} 2 & 6 & 2 \\ 9 & 3 & 9 \\ 1 & 3 & 1 \end{pmatrix}$ Ans: $0, 12, -6$

62. $A = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 3 \end{pmatrix}$ Ans: $0, 1, 3$

63. $A = \begin{pmatrix} 7 & 12 & 6 \\ 2 & 2 & 2 \\ -7 & -12 & -6 \end{pmatrix}$ Ans: $0, 1, 2$

64. $A = \begin{pmatrix} 2 & 2 & -6 \\ -3 & -4 & 3 \\ -3 & -4 & -1 \end{pmatrix}$ Ans: $0, 1, 4$

Computing 2 × 2 Eigenpairs

65. Verify eigenpairs: $\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}$,

$\left(-1, \begin{pmatrix} -1 \\ 1 \end{pmatrix}\right)$, $\left(5, \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}\right)$

66. Verify eigenpairs: $\begin{pmatrix} 1 & 6 \\ 2 & -3 \end{pmatrix}$,

$\left(-5, \begin{pmatrix} -1 \\ 1 \end{pmatrix}\right)$, $\left(3, \begin{pmatrix} 3 \\ -1 \end{pmatrix}\right)$

67. Verify eigenpairs: $\begin{pmatrix} 1 & 6 \\ 4 & 3 \end{pmatrix}$,

$\left(7, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$, $\left(-3, \begin{pmatrix} -3 \\ 2 \end{pmatrix}\right)$

68. Verify eigenpairs: $\begin{pmatrix} 7 & 4 \\ -1 & 3 \end{pmatrix}$,

$\left(5, \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right)$, only one eigenpair

Computing 2 × 2 Complex Eigenpairs

69. Verify eigenpairs: $\begin{pmatrix} -2 & -6 \\ 3 & 4 \end{pmatrix}$,

$\left(1 + 3i, \begin{pmatrix} -1 + i \\ 1 \end{pmatrix}\right)$,

$\left(1 - 3i, \begin{pmatrix} -1 - i \\ 1 \end{pmatrix}\right)$

70. Verify eigenpairs: $\begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix}$, $\left(2+3i, \begin{pmatrix} -i \\ 1 \end{pmatrix}\right)$, $\left(2-3i, \begin{pmatrix} i \\ 1 \end{pmatrix}\right)$

71. Let a, b be real with $b \neq 0$. Assume $n \times n$ real matrix A has eigenpair $(a+ib, \vec{\mathbf{v}})$. Replace i by $-i$ throughout expression $\vec{\mathbf{v}}$ to obtain vector $\vec{\mathbf{w}}$. Prove that $(a-ib, \vec{\mathbf{w}})$ is an eigenpair.

72. Explain: Eigenpairs of a 2×2 real matrix A with complex eigenvalues are computed with just one row-reduction sequence.

Computing 3×3 Eigenpairs

73. Show algorithm steps to compute eigenpairs of $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix}$.

Answers: $\left(1, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}\right)$, $\left(3, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right)$

74. Show algorithm steps to compute eigenpairs of $A = \begin{pmatrix} 1 & -2 & 0 \\ 0 & -1 & 0 \\ 4 & -4 & -1 \end{pmatrix}$.

Answers:
$\left(1, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}\right)$, $\left(-1, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}\right)$, $\left(-1, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right)$

75. Suppose A is row-reduced to a triangular form B. Are the eigenvalues of B also the eigenvalues of A? Give a proof or a counter-example.

76. Suppose $A - \lambda I$ is row-reduced to a triangular form B. Explain: The eigenvalues of A are usually unrelated to the roots λ of $|B| = 0$.

Decomposition $A = PDP^{-1}$

Compute the eigenpairs. If diagonalizable, then display D, P and Fourier's replacement equation.

77. $A = \begin{pmatrix} 7 & 4 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

Ans: only 2 eigenpairs

.

78. $A = \begin{pmatrix} 1 & 6 & 0 \\ 2 & -3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

Ans: $\begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -5 \end{pmatrix}$, $\begin{pmatrix} 3 & 0 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

Fourier equation: $AP\vec{\mathbf{c}} = PD\vec{\mathbf{c}}$.

Diagonalization

Report **diagonalizable** or not and explain why.

79. $A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & -3 \end{pmatrix}$

Ans: diagonalizable

80. $A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}$

Ans: not diagonalizable

Non-diagonalizable Matrices

81. Verify $A = \begin{pmatrix} 1 & 2 \\ -8 & 9 \end{pmatrix}$ is not diagonalizable.

82. Verify $A = \begin{pmatrix} 1 & 2 & 0 \\ -8 & 9 & 1 \\ 0 & 0 & 5 \end{pmatrix}$ is not diagonalizable.

83. Invent a 3×3 matrix which has exactly one eigenpair.

84. Invent a 4×4 matrix which has exactly two eigenpairs.

Fourier's Heat Model

Define
$\vec{\mathbf{v}}_1 = \sin \pi x, \vec{\mathbf{v}}_2 = \sin 2\pi x, \vec{\mathbf{v}}_3 = \sin 3\pi x$
considered as vectors in the vector space V of twice continuously differentiable functions on $0 \le x \le 1$.

85. Verify that $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ are independent vectors in V.

86. Verify that $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ vanish at $x = 0$ and $x = 1$.

87. Define $u(x) = \sin \pi x$ (from $\vec{\mathbf{v}}_1$). Explain: Function u satisfies differential equation $\dfrac{d^2u}{dx^2} + \pi^2 u = 0$.

88. Write vector expression

$$c_1 e^{-\pi^2 t}\vec{\mathbf{v}}_1 + c_2 e^{-4\pi^2 t}\vec{\mathbf{v}}_2 \\ +c_3 e^{-9\pi^2 t}\vec{\mathbf{v}}_3$$

as a scalar function $u(t,x)$. Find initial heat distribution $u(0,x)$. Explain how Fourier replacement (re-scaling) constructs future state $u(t,x)$ from initial state $u(0,x)$.

9.2 Eigenanalysis Applications

Discrete Dynamical Systems

The matrix equation

$$(1) \qquad \vec{\mathbf{y}} = A\vec{\mathbf{x}}, \quad A = \frac{1}{10}\begin{pmatrix} 5 & 4 & 0 \\ 3 & 5 & 3 \\ 2 & 1 & 7 \end{pmatrix}$$

predicts the state $\vec{\mathbf{y}}$ of a system initially in state $\vec{\mathbf{x}}$ after some fixed elapsed time. The 3×3 matrix A in (1) represents the **dynamics** which changes state $\vec{\mathbf{x}}$ into state $\vec{\mathbf{y}}$.

An equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ like equation (1) is called a **discrete dynamical system**. The fixed elapsed time for changing $\vec{\mathbf{x}}$ to $\vec{\mathbf{y}}$ is called the **period** of the discrete dynamical system. Matrix A is called a **transition matrix**, provided A has nonnegative entries and column sums equal to one. See **stochastic matrices** page 704

The eigenpairs of matrix A in (1) are shown on page 713 to be $(1, \vec{\mathbf{v}}_1)$, $(1/2, \vec{\mathbf{v}}_2)$, $(1/5, \vec{\mathbf{v}}_3)$ with eigenvectors

$$(2) \qquad \vec{\mathbf{v}}_1 = \begin{pmatrix} 12 \\ 15 \\ 13 \end{pmatrix}, \quad \vec{\mathbf{v}}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{v}}_3 = \begin{pmatrix} -4 \\ 3 \\ 1 \end{pmatrix}.$$

Market Shares

A model application of discrete dynamical systems is telephone long distance company market shares x_1, x_2, x_3, which are fractions of the total market for long distance service. If three companies provide all the services, then their market fractions add to one: $x_1 + x_2 + x_3 = 1$. Equation $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ in (1) with eigenpairs (2) predicts the market shares of the three companies after a fixed time period, say one year. Market shares after one, two and three years are given by the **iterates**

$$\begin{aligned} \vec{\mathbf{y}}_1 &= A\vec{\mathbf{x}}, \\ \vec{\mathbf{y}}_2 &= A^2\vec{\mathbf{x}}, \\ \vec{\mathbf{y}}_3 &= A^3\vec{\mathbf{x}}. \end{aligned}$$

Fourier's replacement model (8) page 676 gives succinct and useful formulas for the iterates. If $\vec{\mathbf{x}} = a_1\vec{\mathbf{v}}_1 + a_2\vec{\mathbf{v}}_2 + a_3\vec{\mathbf{v}}_3$, then

$$(3) \qquad \begin{aligned} \vec{\mathbf{y}}_1 &= A\vec{\mathbf{x}} &&= a_1\lambda_1\vec{\mathbf{v}}_1 + a_2\lambda_2\vec{\mathbf{v}}_2 + a_3\lambda_3\vec{\mathbf{v}}_3, \\ \vec{\mathbf{y}}_2 &= A^2\vec{\mathbf{x}} &&= a_1\lambda_1^2\vec{\mathbf{v}}_1 + a_2\lambda_2^2\vec{\mathbf{v}}_2 + a_3\lambda_3^2\vec{\mathbf{v}}_3, \\ \vec{\mathbf{y}}_3 &= A^3\vec{\mathbf{x}} &&= a_1\lambda_1^3\vec{\mathbf{v}}_1 + a_2\lambda_2^3\vec{\mathbf{v}}_2 + a_3\lambda_3^3\vec{\mathbf{v}}_3. \end{aligned}$$

The eigenpairs of A in (2) show that $\lambda_1 = 1$ and $\lim_{n\to\infty} |\lambda_2|^n = \lim_{n\to\infty} |\lambda_3|^n = 0$. Then for large n

$$\vec{\mathbf{y}}_n \approx a_1(1)\vec{\mathbf{v}}_1 + a_2(0)\vec{\mathbf{v}}_2 + a_3(0)\vec{\mathbf{v}}_3 = \begin{pmatrix} 12a_1 \\ 15a_1 \\ 13a_1 \end{pmatrix}.$$

The numbers a_1, a_2, a_3 are related to x_1, x_2, x_3 in the expansion $\vec{\mathbf{x}} = a_1\vec{\mathbf{v}}_1 + a_2\vec{\mathbf{v}}_2 + a_3\vec{\mathbf{v}}_3$ by the equations $12a_1 - a_2 - 4a_3 = x_1$, $15a_1 + 3a_3 = x_2$, $13a_1 + a_2 + a_3 = x_3$. Because $x_1 + x_2 + x_3 = 1$, then $a_1 = 1/40$. The three market shares after a long time period are predicted to be 3/10, 3/8, 13/40. The market share identity $\frac{3}{10} + \frac{3}{8} + \frac{13}{40} = 1$ holds because approximating terms from (3) are sums of market shares adding to one.

Stochastic Matrices

The special matrix A in (1) is a **stochastic matrix**[8], defined by the properties

$$\sum_{i=1}^{n} a_{ij} = 1, \quad a_{kj} \ge 0, \quad k, j = 1, \ldots, n.$$

The definition is memorized by the phrase *each column sum is one.*

Leontief input-output models are stochastic models, popularized by 1973 Nobel Prize economist Wassily Leontief. A typical model is $A = R^T$ where

$$R = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .3 & .5 \\ .4 & .4 & .2 \end{pmatrix}.$$

The **rows** of R add to one, therefore the **columns** of A add to one. Row 1 is the bank, Row 2 is Factory 1, Row 3 is Factory 2. Matrix R tracks the money as it is being passed back and forth between the factories and the bank.

Leslie Models in population biology are similar to stochastic models. It is a discrete time model $\vec{\mathbf{v}}_{i+1} = A\vec{\mathbf{v}}_i$ of an age-structured population describing mortality, reproduction and development. The **Leslie matrix** A for $n = 4$ looks like

$$A = \begin{pmatrix} f_1 & f_2 & f_3 & f_4 \\ s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & s_3 & 0 \end{pmatrix}.$$

Neither the row sums nor the column sums are one. However, some stochastic matrix results have analogs for Leslie matrices. Population vector $\vec{\mathbf{v}}_i$ contains counts of age classes. Number $f_i \ge 0$ is the average number of female births for a mother of age class i. Number $s_i \ge 0$ is the fraction of individuals of age class i that survive to age class $i + 1$.

[8] Technically, a **right** stochastic matrix, which means columns add to one. A **left** stochastic matrix has rows adding to one. The term **transition matrix** is also used.

Theorem 9.12 (Stochastic Matrix Properties)
Let A be a stochastic matrix. Then

(a) If $\vec{\mathbf{x}}$ is a vector with $x_1 + \cdots + x_n = 1$, then $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ satisfies $y_1 + \cdots + y_n = 1$.

(b) If the components of $\vec{\mathbf{v}}$ are all 1, then $A^T\vec{\mathbf{v}} = \vec{\mathbf{v}}$. Therefore, $(1, \vec{\mathbf{v}})$ is an eigenpair of A^T.

(c) One root of the characteristic equation $\det(A - \lambda I) = 0$ is $\lambda = 1$. All other roots satisfy $|\lambda| \le 1$.

Proof on page 715.

Theorem 9.13 (Perron-Frobenius: Positive Stochastic Matrix)
Let A be a stochastic matrix all of whose entries are strictly positive. Then

(a) There exists an eigenpair $(1, \vec{\mathbf{w}})$ of A such that $\vec{\mathbf{w}}$ has nonnegative components and $\lim_{n\to\infty} A^n = \left\langle \vec{\mathbf{w}} | \vec{\mathbf{w}} | \cdots | \vec{\mathbf{w}} \right\rangle$.

(b) If $(1, \vec{\mathbf{v}})$ is an eigenpair of A, then $\vec{\mathbf{v}} = c\vec{\mathbf{w}}$ for $c = \sum_{i=1}^n v_i$. Briefly, the eigenspace for $\lambda = 1$ has dimension one.

(c) If $\lambda \ne 1$ is a real or complex eigenvalue of A, then $|\lambda| < 1$.

(d) If $(\lambda, \vec{\mathbf{v}})$ is an eigenpair of A and $\vec{\mathbf{v}}$ has nonnegative components, then all components of $\vec{\mathbf{v}}$ are strictly positive, $\lambda = 1$ and $\vec{\mathbf{v}} = c\vec{\mathbf{w}}$ for some constant c.

Proof on page 715.

Coupled and Uncoupled Systems

The linear system of differential equations

$$\begin{aligned} x_1' &= -x_1 - x_3, \\ x_2' &= 4x_1 - x_2 - 3x_3, \\ x_3' &= 2x_1 - 4x_3, \end{aligned} \tag{4}$$

is called **coupled**, whereas the linear system of growth-decay equations

$$\begin{aligned} y_1' &= -3y_1, \\ y_2' &= -y_2, \\ y_3' &= -2y_3, \end{aligned} \tag{5}$$

is called **uncoupled**. The terminology *uncoupled* means that each differential equation in system (5) depends on exactly one variable, e.g., $y_1' = -3y_1$ depends only on variable y_1. In a *coupled* system, one of the differential equations must involve two or more variables.

Matrix Formulation

Coupled system (4) and uncoupled system (5) can be written in matrix form, $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$, with coefficient matrices

$$A = \begin{pmatrix} -1 & 0 & -1 \\ 4 & -1 & -3 \\ 2 & 0 & -4 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} -3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

If the coefficient matrix is **diagonal**, then the system is **uncoupled**. If the coefficient matrix is **not diagonal**, then one of the corresponding differential equations involves two or more variables and the system is called **coupled** or **cross-coupled**.

Solving Uncoupled Systems

An uncoupled system consists of independent growth-decay equations of the form $u' = au$. The solution formula $u = ce^{at}$ then leads to the general solution of the system of equations. For instance, system (5) has general solution

$$\begin{aligned} y_1 &= c_1 e^{-3t}, \\ y_2 &= c_2 e^{-t}, \\ y_3 &= c_3 e^{-2t}, \end{aligned} \tag{6}$$

where c_1, c_2, c_3 are **arbitrary constants**. The number of constants equals the dimension of the diagonal matrix D.

Coordinates and Coordinate Systems

If vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are independent in $\mathcal{R}^3$, then augmented matrix

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \right\rangle$$

is invertible. The columns $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ of P are called a **coordinate system**. The matrix P is called a **change of coordinates**.

Independence of $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ means every vector $\vec{\mathbf{v}}$ in $\mathcal{R}^3$ can be uniquely expressed as

$$\vec{\mathbf{v}} = t_1\vec{\mathbf{v}}_1 + t_2\vec{\mathbf{v}}_2 + t_3\vec{\mathbf{v}}_3.$$

The values t_1, t_2, t_3 are called the **coordinates** of $\vec{\mathbf{v}}$ relative to the basis $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, or the coordinates of $\vec{\mathbf{v}}$ relative to P.

Viewpoint of a Driver

The physical meaning of a coordinate system $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ can be understood by considering an auto traveling up a mountain road. Choose orthogonal $\vec{\mathbf{v}}_1$ and

$\vec{\mathbf{v}}_2$ to give positions in the driver's seat and define $\vec{\mathbf{v}}_3$ be the seat-back direction. These are **local coordinates** as viewed from the driver's seat. The road map coordinates x, y and the altitude z define the **global coordinates** for the auto's position $\vec{\mathbf{p}} = x\vec{\imath} + y\vec{\jmath} + z\vec{k}$.

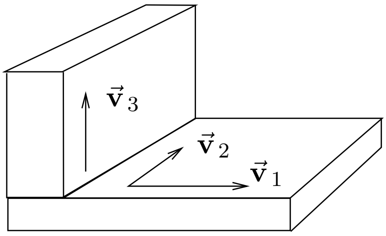


Figure 2. Driver's coordinates. The vectors $\vec{\mathbf{v}}_1(t)$, $\vec{\mathbf{v}}_2(t)$, $\vec{\mathbf{v}}_3(t)$ form an orthogonal triad which is a local coordinate system from the driver's viewpoint. The orthogonal triad changes continuously in t.

Change of Coordinates $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$

A coordinate change from $\vec{\mathbf{y}}$ to $\vec{\mathbf{x}}$ is a linear algebraic equation $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ where the $n \times n$ matrix P is required to be invertible $(\det(P) \neq 0)$. To illustrate, an instance of a change of coordinates from $\vec{\mathbf{y}}$ to $\vec{\mathbf{x}}$ is given by the linear equations

$$(7) \qquad \vec{\mathbf{x}} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & -1 \\ 2 & 0 & 1 \end{pmatrix} \vec{\mathbf{y}} \quad \text{or} \quad \begin{cases} x_1 &= y_1 + y_3, \\ x_2 &= y_1 + y_2 - y_3, \\ x_3 &= 2y_1 + y_3. \end{cases}$$

Constructing Coupled Systems

A general method exists to construct rich examples of coupled systems. The idea uses a change of variables for a given uncoupled system. Consider a diagonal system $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$, like (5), and a change of variables $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$, like (7). Differential calculus applies to give

$$(8) \qquad \begin{aligned} \vec{\mathbf{x}}' &= (P\vec{\mathbf{y}})' \\ &= P\vec{\mathbf{y}}' \\ &= PD\vec{\mathbf{y}} \\ &= PDP^{-1}\vec{\mathbf{x}}. \end{aligned}$$

The matrix $A = PDP^{-1}$ is *not triangular* in general, and therefore the change of variables produces a **cross-coupled** system.

An illustration. To give an example, substitute into uncoupled system (5) the change of variable equations (7). Use equation (8) to obtain

$$(9) \qquad \vec{\mathbf{x}}' = \begin{pmatrix} -1 & 0 & -1 \\ 4 & -1 & -3 \\ 2 & 0 & -4 \end{pmatrix} \vec{\mathbf{x}} \quad \text{or} \quad \begin{cases} x_1' = -x_1 - x_3, \\ x_2' = 4x_1 - x_2 - 3x_3, \\ x_3' = 2x_1 - 4x_3. \end{cases}$$

This **cross-coupled** system (9) can be solved using relations (7), (6) and $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ to give the general solution

$$(10) \qquad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & -1 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 e^{-3t} \\ c_2 e^{-t} \\ c_3 e^{-2t} \end{pmatrix}.$$

Changing Coupled Systems to Uncoupled

A question, motivated by the above calculations:

> Can every coupled system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$ be subjected to a change of variables $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ which converts the system into a completely uncoupled system for variable $\vec{\mathbf{y}}(t)$?

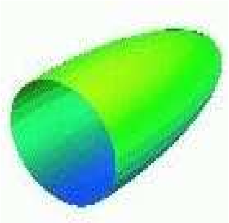
Answer: A coupled system can be so transformed if and only if matrices P and D are eigenpair packages of A. Then $AP = PD$ and A is diagonalizable. Conversely, if A is diagonalizable, then the packages P, D exist and $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ changes $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ into diagonal system $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$. The connection between $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ is like (10).

Eigenanalysis provides the opportunity to simultaneously calculate from cross-coupled system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ a change of variable $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ and a diagonal matrix D for an uncoupled system $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$. System $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$ consists of uncoupled scalar growth-decay equations like (5).

Matrices A that fail to be diagonalizable present a problem, because eigenanalysis does not apply. The demand to obtain an uncoupled system $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$ leaves no alternative, because if there is a change of variables $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ into diagonal system $\vec{\mathbf{y}}' = D\vec{\mathbf{y}}$, then $AP = PD$ and A is diagonalizable, a contradiction.

There *does exist* a change of coordinates P to change $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ into a **triangular system** $\vec{\mathbf{y}}' = T\vec{\mathbf{y}}$. This system in scalar form can be solved by the linear integrating factor method. There is again an answer $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ like (5). See page 721.

Eigenanalysis and Footballs



An ellipsoid or *football* is a geometric object described by its **semi-axes** (see Figure 3). In the vector representation, the **semi-axis directions** are unit vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ and the **semi-axis lengths** are the constants a, b, c. The vectors $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$, $c\vec{\mathbf{v}}_3$ form an **orthogonal triad**.

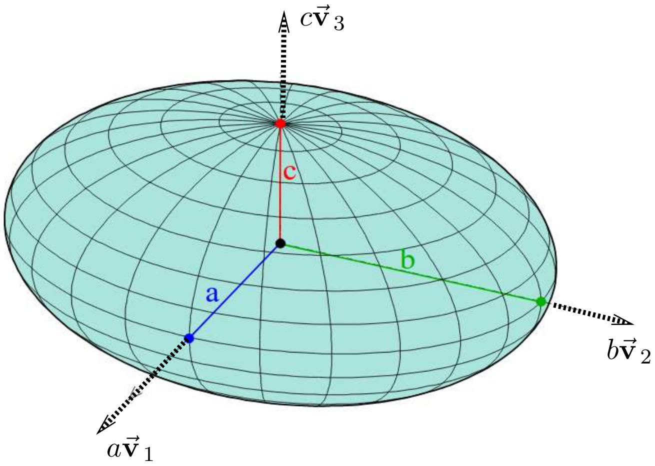


Figure 3. Ellipsoid. An ellipsoid is built from orthonormal semi-axis directions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ and the semi-axis lengths a, b, c. The semi-axis vectors are $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$, $c\vec{\mathbf{v}}_3$.

Two vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{w}}$ are *orthogonal* if both are nonzero and their dot product $\vec{\mathbf{u}} \cdot \vec{\mathbf{w}}$ is zero. Vectors are **orthonormal** if each has unit length and they are pairwise orthogonal. The orthogonal triad $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ is an **invariant** of the ellipsoid's algebraic representations. Algebra does not change the triad: the invariants $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$, $c\vec{\mathbf{v}}_3$ must somehow be **hidden** in the equations that represent the ellipsoid.

Algebraic eigenanalysis finds the hidden invariant triad $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$, $c\vec{\mathbf{v}}_3$ from the ellipsoid's algebraic equations. Suppose, for instance, that the equation of the ellipsoid is supplied as

$$x^2 + 4y^2 + xy + 4z^2 = 16.$$

A symmetric matrix A is constructed in order to write the equation in the form $\vec{\mathbf{X}}^T A \vec{\mathbf{X}} = 16$, where $\vec{\mathbf{X}}$ has components x, y, z. The replacement equation is[9]

$$(11) \qquad \begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 0 \\ 1/2 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 16.$$

It is the 3×3 symmetric matrix A in (11) that is subjected to algebraic eigenanalysis. The matrix calculation will compute the unit semi-axis directions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, called the **eigenvectors** or **hidden vectors**. The semi-axis lengths a, b, c are computed at the same time, by finding the **eigenvalues** or **hidden values**[10] λ_1, λ_2, λ_3, known to satisfy the relations

$$\lambda_1 = \frac{16}{a^2}, \quad \lambda_2 = \frac{16}{b^2}, \quad \lambda_3 = \frac{16}{c^2}.$$

For the illustration, the football dimensions are $a = 2$, $b = 1.98$, $c = 4.17$. Details of the computation are delayed until page 711.

[9] Multiply matrices to verify this statement. Halving of the entries corresponding to cross-terms generalizes to any ellipsoid.

[10] The terminology *hidden* arises because neither the semi-axis lengths nor the semi-axis directions are revealed directly by the ellipsoid equation.

Ellipse and Eigenanalysis

An ellipse equation in **standard form** is $\lambda_1 u^2 + \lambda_2 v^2 = 1$, where $\lambda_1 = 1/a^2$, $\lambda_2 = 1/b^2$ are expressed in terms of the semi-axis lengths a, b. The expression $\lambda_1 u^2 + \lambda_2 v^2$ is called a **quadratic form**. The study of the ellipse $\lambda_1 u^2 + \lambda_2 v^2 = 1$ is equivalent to the study of the quadratic form equation

$$\vec{\mathbf{r}}^T D \vec{\mathbf{r}} = 1, \quad \text{where} \quad \vec{\mathbf{r}} = \begin{pmatrix} u \\ v \end{pmatrix}, \quad D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Cross-terms. An ellipse may be represented by an equation in a xy-coordinate system having a cross-term xy, e.g., $4x^2 + 8xy + 10y^2 = 5$. The expression $4x^2 + 8xy + 10y^2$ is again called a quadratic form. Calculus courses provide methods to eliminate the cross-term and represent the equation in standard form, by a **rotation** by angle θ of the xy-system into the uv-system:

$$\begin{pmatrix} u \\ v \end{pmatrix} = R \begin{pmatrix} x \\ y \end{pmatrix}, \quad R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

Eigenanalysis computes angle θ through the columns of R, which are the unit semi-axis directions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ for the ellipse $4x^2 + 8xy + 10y^2 = 5$. If the quadratic form $4x^2 + 8xy + 10y^2$ is represented as $\vec{\mathbf{r}}^T A \vec{\mathbf{r}}$, then

$$\vec{\mathbf{r}} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad A = \begin{pmatrix} 4 & 4 \\ 4 & 10 \end{pmatrix}, \quad R = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix},$$

$$\lambda_1 = 12, \quad \vec{\mathbf{v}}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \lambda_2 = 2, \quad \vec{\mathbf{v}}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Ellipse equations. There are two coordinate systems, the xy-system and the rotated uv-system. The equations in each system, each divided by 5:

$$\begin{aligned} \tfrac{4}{5}x^2 + \tfrac{8}{5}xy + 2y^2 &= 1, \\ \tfrac{2}{5}u^2 + \tfrac{12}{5}v^2 &= 1. \end{aligned} \tag{12}$$

The rotation relation $\begin{pmatrix} u \\ v \end{pmatrix} = R \begin{pmatrix} x \\ y \end{pmatrix}$ is the set of equations

$$\begin{cases} u \;=\; = \tfrac{1}{\sqrt{5}}x - \tfrac{2}{\sqrt{5}}y, \\ v \;=\; = \tfrac{2}{\sqrt{5}}x + \tfrac{1}{\sqrt{5}}y, \end{cases} \tag{13}$$

which upon substitution into the uv-equation in (12) gives

$$\frac{2}{5}\left(\frac{1}{\sqrt{5}}x - \frac{2}{\sqrt{5}}y\right)^2 + \frac{12}{5}\left(\frac{2}{\sqrt{5}}x + \frac{1}{\sqrt{5}}y\right)^2 = 1.$$

The reader can verify that this is the first equation in (12).

Rotation matrix angle θ. The components of unit eigenvector $\vec{\mathbf{v}}_1$ can be used to determine $\theta = -63.4°$:

$$\begin{pmatrix} \cos\theta \\ -\sin\theta \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{or} \quad \left\{ \begin{array}{rcl} \cos\theta & = & \frac{1}{\sqrt{5}}, \\ -\sin\theta & = & \frac{2}{\sqrt{5}}. \end{array} \right.$$

The interpretation of angle θ: rotate the orthonormal basis $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ by angle $\theta = -63.4°$ in order to obtain the standard unit basis vectors $\vec{\imath}$, $\vec{\jmath}$. Calculus texts might discuss only the inverse rotation, where x, y are given in terms of u, v. In these references, θ is the negative of the value given here, due to a different geometric viewpoint.[11]

Semi-axis lengths. The lengths $a \approx 1.55$, $b \approx 0.63$ for the ellipse $4x^2 + 8xy + 10y^2 = 5$ are computed from the eigenvalues $\lambda_1 = 12$, $\lambda_2 = 2$ of matrix A by the equations

$$\frac{\lambda_1}{5} = \frac{1}{a^2}, \quad \frac{\lambda_2}{5} = \frac{1}{b^2}.$$

Geometry. The ellipse $4x^2 + 8xy + 10y^2 = 5$ is completely determined by the orthogonal semi-axis vectors $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$. The rotation R is a rigid motion mapping xy-plane vectors $a\vec{\mathbf{v}}_1$, $b\vec{\mathbf{v}}_2$ into uv-plane vectors $a\vec{\imath}$, $b\vec{\jmath}$.

The θ-rotation R maps $4x^2+8xy+10y^2 = 5$ into the uv-equation $\lambda_1 u^2+\lambda_2 v^2 = 5$, where λ_1, λ_2 are the eigenvalues of A. To see why, let $\vec{\mathbf{r}} = \begin{pmatrix} u \\ v \end{pmatrix}$, $\vec{\mathbf{s}} = \begin{pmatrix} x \\ y \end{pmatrix}$ in the equation $\vec{\mathbf{r}} = R\vec{\mathbf{s}}$. Then $\vec{\mathbf{r}}^T A\vec{\mathbf{r}} = \vec{\mathbf{s}}^T(R^T AR)\vec{\mathbf{s}}$. Using $R^T R = I$ gives $R^{-1} = R^T$ and $R^T AR = \mathbf{diag}(\lambda_1, \lambda_2)$. Finally, $\vec{\mathbf{r}}^T A\vec{\mathbf{r}} = \lambda_1 u^2 + \lambda_2 v^2$.

Orthogonal Triad Computation

Let's compute the semiaxis directions $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ for the ellipsoid $x^2 + 4y^2 + xy + 4z^2 = 16$. To be applied is Theorem 9.3. As explained on page 709, the starting point is to represent the ellipsoid equation as a quadratic form $\vec{\mathbf{W}}^T A\vec{\mathbf{W}} = 16$, where the symmetric matrix A and vector $\vec{\mathbf{W}}$ are defined by

$$A = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad \vec{\mathbf{W}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

College algebra. The **Characteristic Polynomial** $\det(A-\lambda I) = 0$ determines the eigenvalues or hidden values of the matrix A. By cofactor expansion, this polynomial equation is

$$(4-\lambda)((1-\lambda)(4-\lambda) - 1/4) = 0$$

with roots 4, $5/2 + \sqrt{10}/2$, $5/2 - \sqrt{10}/2$.

[11] Rod Serling, author and playwright for the SciFi series *The Twilight Zone*, enjoyed the view from the other side.

Eigenpairs. It will be shown that three eigenpairs are

$$\begin{aligned}
\lambda_1 &= 4, \quad \vec{\mathbf{x}}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \\
\lambda_2 &= \frac{5+\sqrt{10}}{2}, \quad \vec{\mathbf{x}}_2 = \begin{pmatrix} \sqrt{10}-3 \\ 1 \\ 0 \end{pmatrix}, \\
\lambda_3 &= \frac{5-\sqrt{10}}{2}, \quad \vec{\mathbf{x}}_3 = \begin{pmatrix} \sqrt{10}+3 \\ -1 \\ 0 \end{pmatrix}.
\end{aligned}$$

The vector norms of the eigenvectors are given by $\|\vec{\mathbf{x}}_1\| = 1$, $\|\vec{\mathbf{x}}_2\| = \sqrt{20+6\sqrt{10}}$, $\|\vec{\mathbf{x}}_3\| = \sqrt{20-6\sqrt{10}}$. The orthonormal semi-axis directions $\vec{\mathbf{v}}_k = \vec{\mathbf{x}}_k/\|\vec{\mathbf{x}}_k\|$, $k = 1, 2, 3$, are then given by the formulas

$$\vec{\mathbf{v}}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{v}}_2 = \begin{pmatrix} \frac{\sqrt{10}-3}{\sqrt{20-6\sqrt{10}}} \\ \frac{1}{\sqrt{20-6\sqrt{10}}} \\ 0 \end{pmatrix}, \quad \vec{\mathbf{v}}_3 = \begin{pmatrix} \frac{\sqrt{10}+3}{\sqrt{20+6\sqrt{10}}} \\ \frac{-1}{\sqrt{20+6\sqrt{10}}} \\ 0 \end{pmatrix}.$$

Eigenpair Details.

$$\left\langle A - \lambda_1 I, \vec{\mathbf{0}} \right\rangle = \left(\begin{array}{ccc|c} 1-4 & 1/2 & 0 & 0 \\ 1/2 & 4-4 & 0 & 0 \\ 0 & 0 & 4-4 & 0 \end{array}\right)$$

$$\approx \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Used Toolkit rules combination, multiply and swap. Found **rref**.

$$\left\langle A - \lambda_2 I, \vec{\mathbf{0}} \right\rangle = \left(\begin{array}{ccc|c} \frac{-3-\sqrt{10}}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{3-\sqrt{10}}{2} & 0 & 0 \\ 0 & 0 & \frac{3-\sqrt{10}}{2} & 0 \end{array}\right)$$

$$\approx \left(\begin{array}{ccc|c} 1 & 3-\sqrt{10} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Toolkit rules applied. Found **rref**.

$$\left\langle A - \lambda_3 I, \vec{\mathbf{0}} \right\rangle = \left(\begin{array}{ccc|c} \frac{-3+\sqrt{10}}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{3+\sqrt{10}}{2} & 0 & 0 \\ 0 & 0 & \frac{3+\sqrt{10}}{2} & 0 \end{array}\right)$$

$$\approx \left(\begin{array}{ccc|c} 1 & 3+\sqrt{10} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Toolkit rules applied. Found **rref**.

Solving the corresponding reduced echelon systems gives the preceding formulas for the eigenvectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, $\vec{\mathbf{x}}_3$. The equation for the ellipsoid is $\lambda_1 X^2 + \lambda_2 Y^2 + \lambda_3 Z^2 = 16$, where the multipliers of the square terms are the eigenvalues of A and X, Y, Z define the new coordinate system determined by the eigenvectors of A. This equation can be re-written in the form $\frac{X^2}{a^2} + \frac{Y^2}{b^2} + \frac{Z^2}{c^2} = 1$, provided the semi-axis lengths a, b, c are defined by the relations $a^2 = 16/\lambda_1$, $b^2 = 16/\lambda_2$, $c^2 = 16/\lambda_3$. After computation, $a = 2$, $b = 1.98$, $c = 4.17$.

Proofs, Methods and Details

Eigenpairs of (1), Telephone Carriers:

To be computed are the eigenvalues λ and eigenvectors $\vec{\mathbf{v}}$ for the 3×3 matrix

$$A = \frac{1}{10}\begin{pmatrix} 5 & 4 & 0 \\ 3 & 5 & 3 \\ 2 & 1 & 7 \end{pmatrix}.$$

The eigenpairs are $(1, \vec{\mathbf{v}}_1), \left(\frac{1}{2}, \vec{\mathbf{v}}_2\right), \left(\frac{1}{5}, \vec{\mathbf{v}}_3\right)$ where

$$(14) \qquad \vec{\mathbf{v}}_1 = \begin{pmatrix} 12 \\ 15 \\ 13 \end{pmatrix}, \quad \vec{\mathbf{v}}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{v}}_3 = \begin{pmatrix} -4 \\ 3 \\ 1 \end{pmatrix}.$$

Eigenvalues. The roots $\lambda = 1, 1/2, 1/5$ of the characteristic equation $\det(A - \lambda I) = 0$ are found by these details:

$$\begin{aligned} 0 &= \det(A - \lambda I) \\ &= \begin{vmatrix} .5 - \lambda & .4 & 0 \\ .3 & .5 - \lambda & .3 \\ .2 & .1 & .7 - \lambda \end{vmatrix} \\ &= \frac{1}{10} - \frac{8}{10}\lambda + \frac{17}{10}\lambda^2 - \lambda^3 && \text{Expand by cofactors.} \\ &= -\frac{1}{10}(\lambda - 1)(2\lambda - 1)(5\lambda - 1) && \text{Factor the cubic.} \end{aligned}$$

The factorization was found by long division of the cubic by $\lambda - 1$, the idea born from the fact that 1 is a root and therefore $\lambda - 1$ is a factor, by the Factor Theorem of college algebra. The root $\lambda = 1$ was discovered from the Rational Root theorem of college algebra.[12]

Eigenpairs. To each eigenvalue $\lambda = 1, 1/2, 1/5$ corresponds one **rref** calculation, to find the eigenvectors paired to λ. The three eigenvectors are given by (2). The details:

Eigenvalue $\lambda = 1$.

$$\begin{aligned} A - (1)I &= \begin{pmatrix} .5 - 1 & .4 & 0 \\ .3 & .5 - 1 & .3 \\ .2 & .1 & .7 - 1 \end{pmatrix} \\ &\approx \begin{pmatrix} -5 & 4 & 0 \\ 3 & -5 & 3 \\ 2 & 1 & -3 \end{pmatrix} && \text{Multiply rule, multiplier=10.} \end{aligned}$$

[12] A rational root x of $a_n x^n + \cdots + a_0 = 0$ is a rational factor of a_0/a_n.

$$
\begin{aligned}
&\approx \begin{pmatrix} 0 & 0 & 0 \\ 3 & -5 & 3 \\ 2 & 1 & -3 \end{pmatrix} && \text{Combination rule twice.} \\
&\approx \begin{pmatrix} 0 & 0 & 0 \\ 1 & -6 & 6 \\ 2 & 1 & -3 \end{pmatrix} && \text{Combination rule.} \\
&\approx \begin{pmatrix} 0 & 0 & 0 \\ 1 & -6 & 6 \\ 0 & 13 & -15 \end{pmatrix} && \text{Combination rule.} \\
&\approx \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -\frac{12}{13} \\ 0 & 1 & -\frac{15}{13} \end{pmatrix} && \text{Multiply rule and combination rule.} \\
&\approx \begin{pmatrix} 1 & 0 & -\frac{12}{13} \\ 0 & 1 & -\frac{15}{13} \\ 0 & 0 & 0 \end{pmatrix} && \text{Swap rule.} \\
&= \mathbf{rref}(A - (1)I)
\end{aligned}
$$

An equivalent reduced echelon system is $x - 12z/13 = 0$, $y - 15z/13 = 0$. The free variable assignment is $z = t_1$ and then $x = 12t_1/13$, $y = 15t_1/13$.

An eigenvector can be selected as the partial derivative on variable t_1 across the general solution $x = 12t_1/13$, $y = 15t_1/13$, $z = t_1$ (equivalent here to setting $t_1 = 1$). This computation gives eigenvector $x = 12/13, y = 15/13, z = 1$.

An eigenvector can be multiplied by a constant $c \neq 0$ to obtain another eigenvector. To eliminate fractions in the answer, the practice is to multiply by an integer c to eliminate all fractions. Choose constant $c = 13$ to obtain eigenvector $x = 12, y = 15, z = 13$.

Eigenvalue $\lambda = 1/2$.

$$
\begin{aligned}
A - (1/2)I &= \begin{pmatrix} .5 - .5 & .4 & 0 \\ .3 & .5 - .5 & .3 \\ .2 & .1 & .7 - .5 \end{pmatrix} \\
&\approx \begin{pmatrix} 0 & 4 & 0 \\ 3 & 0 & 3 \\ 2 & 1 & 2 \end{pmatrix} && \text{Multiply rule, factor=10.} \\
&\approx \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} && \text{Combination and multiply rules.} \\
&= \mathbf{rref}(A - .5I)
\end{aligned}
$$

An eigenvector is found from the equivalent reduced echelon system $y = 0$, $x + z = 0$ to be $x = -1$, $y = 0$, $z = 1$.

Eigenvalue $\lambda = 1/5$.

$$
\begin{aligned}
A - (1/5)I &= \begin{pmatrix} .5 - .2 & .4 & 0 \\ .3 & .5 - .2 & .3 \\ .2 & .1 & .7 - .2 \end{pmatrix} \\
&\approx \begin{pmatrix} 3 & 4 & 0 \\ 1 & 1 & 1 \\ 2 & 1 & 5 \end{pmatrix} && \text{Multiply rule.}
\end{aligned}
$$

$$\approx \begin{pmatrix} 1 & 0 & 4 \\ 0 & 1 & -3 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{Combination rule.}$$
$$= \mathbf{rref}(A - (1/5)I)$$

An eigenvector is found from the equivalent reduced echelon system $x+4z = 0$, $y-3z = 0$ to be $x = -4$, $y = 3$, $z = 1$.

An answer check in `maple`:

```
with(LinearAlgebra):
A:=(1/10)*Matrix([[5,4,0],[3,5,3],[2,1,7]]);
B:=A-lambda*IdentityMatrix(3);
DD,P:=Eigenvectors(A);
factor(Determinant(B));
```

Proof of Theorem 9.12, Stochastic Matrix Properties:

(a) $\sum_{i=1}^n y_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_j = \sum_{j=1}^n \left(\sum_{i=1}^n a_{ij}\right) x_j = \sum_{j=1}^n (1)x_j = 1.$

(b) Entry j of $A^T\vec{\mathbf{v}}$ is given by $\sum_{i=1}^n (a_{ij})(1) = \text{column sum} = 1$.

(c) The determinant rule $\det(B^T) = \det(B)$ applied to $B = A - \lambda I$ implies A and A^T have the same eigenvalues. Apply **(b)** to verify that A has eigenvalue 1. Any other root λ of $|A - \lambda I| = 0$ is also a root of $|A^T - \lambda I| = 0$ with corresponding eigenvector $\vec{\mathbf{x}}$ satisfying $A^T\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$. Because $\vec{\mathbf{x}} \neq \vec{\mathbf{0}}$, then $\vec{\mathbf{x}}$ has a component x_j with largest magnitude $|x_j| > 0$. Isolate index j across equation $\lambda\vec{\mathbf{x}} = A^T\vec{\mathbf{x}}$, then divide by $|x_j|$, to obtain $\lambda = \sum_{i=1}^n a_{ij}\frac{x_i}{x_j}$. Because $a_{ji} \geq 0$ and $0 \leq \left|\frac{x_i}{x_j}\right| \leq 1$, then $|\lambda| \leq 1$, because

$$|\lambda| \leq \sum_{i=1}^n a_{ij}\left|\frac{x_i}{x_j}\right| \leq \sum_{i=1}^n (a_{ij})(1) = \text{column sum} = 1.$$

Proof of Theorem 9.13, Perron-Frobenius:[13]

Proof of (a)

Definition 9.6 (Positive Matrix)

Notation $A > 0$ means all $a_{ij} > 0$. Notation $A \leq B$ means $a_{ij} \leq b_{ij}$, also written $B \geq A$.

Definition 9.7 (Max, Min and Ones Matrices)

Matrix $\mathbf{max_r}(A)$ (resp. $\mathbf{min_r}(A)$) is obtained from A by replacing each entry a_{ij} by the maximum (resp. minimum) element of row i. Symbol $\delta = \min_{i,j} a_{ij}$ is the least element in matrix A. Matrix $\mathcal{O}$ is the $n \times n$ matrix of all ones.

The proof is organized as five lemmas. Assume throughout that $A > 0$ is stochastic with least element δ, $B \geq 0$ and $\mathcal{O}$ is the matrix of all ones.

Lemma 1a. If A, B are stochastic, then BA is stochastic.

Lemma 2a. $\mathbf{min_r}(B) \leq \mathbf{min_r}(BA) \leq BA \leq \mathbf{max_r}(BA) \leq \mathbf{max_r}(B)$.

[13] Perron-Frobenius theory is a basis for the Google Search **PageRank algorithm**.

Proof: The maximum along row i of $C = BA$ is some $c_{ij} = \sum_{k=1}^n b_{ik}a_{kj}$. Let M denote the maximum along row i of B. Because columns of A sum to 1, then $c_{ij} = \sum_{k=1}^n b_{ik}a_{kj} \le \sum_{k=1}^n M a_{kj} = M$. Then $BA \le \mathbf{max_r}(BA) \le \mathbf{max_r}(B)$. Details for inequality $\mathbf{min_r}(B) \le \mathbf{min_r}(BA) \le BA$ are similar.

Lemma 3a. $\mathbf{max_r}(BA) - \mathbf{min_r}(BA) \le (1-\delta)\,(\mathbf{max_r}(B) - \mathbf{min_r}(B))$.

Proof: Let $C = BA$ have row i maximum at c_{ij} and row minimum at c_{ik}. Then all elements in row i of matrix $\mathbf{max_r}(BA) - \mathbf{min_r}(BA)$ have value $S = c_{ij} - c_{ik}$. Let M (resp. m) be the common entry along row i of $\mathbf{max_r}(B)$ (resp. $\mathbf{min_r}(B)$). We'll verify $S \le (1-\delta)\,(M-m)$, which proves the lemma.

Re-write $S = c_{ij} - c_{ik} = \sum_{p=1}^n b_{ip}a_{pj} - \sum_{p=1}^n b_{ip}a_{pk} = \sum_{p=1}^n b_{ip}(a_{pj} - a_{pk})$. Let $p_1, \ldots, p_r$ be the set of indices p such that $a_{pj} - a_{pk} > 0$ and let $q_1, \ldots, q_s$ be the set of indices q such that $a_{qj} - a_{qk} < 0$. Indices p that satisfy $a_{pj} - a_{pk} = 0$ contribute zero to S. In cases $r = 0$ and/or $s = 0$ we have $S \le 0$, so the conclusion follows. Henceforth, assume $r \ge 1$ and $s \ge 1$. The column sums of A are 1, which implies for instance $\sum_{\ell=1}^r a_{p_\ell j} + \sum_{\ell=1}^s a_{q_\ell j} = 1$. We estimate:

$$\begin{array}{rcl}
S & = & \sum_{p=1}^n b_{ip}(a_{pj} - a_{pk}) \\
 & = & \sum_{\ell=1}^r b_{ip}(a_{p_\ell j} - a_{p_\ell k}) + \sum_{\ell=1}^s b_{ip}(a_{q_\ell j} - a_{q_\ell k}) \\
 & \le & M\sum_{\ell=1}^r (a_{p_\ell j} - a_{p_\ell k}) + m\sum_{\ell=1}^s (a_{q_\ell j} - a_{q_\ell k}) \\
 & = & M\,(1 - \sum_{\ell=1}^s a_{q_\ell j} - 1 + \sum_{\ell=1}^s a_{q_\ell k}) + m\sum_{\ell=1}^s (a_{q_\ell j} - a_{q_\ell k}) \\
 & = & (M-m)\,(-\sum_{\ell=1}^s a_{q_\ell j} + \sum_{\ell=1}^s a_{q_\ell k}) \\
 & \le & (M-m)\,(-s\delta + 1) \\
 & \le & (M-m)\,(-\delta + 1)\,.
\end{array}$$

Lemma 4a. $\mathbf{max_r}(A^{k+1}) - \mathbf{min_r}(A^{k+1}) \le (1-\delta)^k \mathcal{O}$.

Proof: Let $B = A^k$ and apply Lemmas 1a and 3a. Then $\mathbf{max_r}(A^{k+1}) - \mathbf{min_r}(A^{k+1}) \le (1-\delta)\big(\mathbf{max_r}(A^k) - \mathbf{min_r}(A^k)\big)$. Induction on k implies the result, because $\mathbf{max_r}(A) - \mathbf{min_r}(A) \le \mathcal{O}$.

Lemma 5a. There exists a vector $\vec{\mathbf{w}}$ with all positive components such that $\lim_{k\to\infty} A^k = \Big\langle \vec{\mathbf{w}}\,|\vec{\mathbf{w}}\,|\cdots|\vec{\mathbf{w}}\Big\rangle$. Then $A\vec{\mathbf{w}} = \vec{\mathbf{w}}$ and $(1, \vec{\mathbf{w}})$ is an eigenpair.[14]

Proof: The preceding lemmas and the calculus squeeze theorem for limits imply that $\mathbf{max_r}(A^k)$ and $\mathbf{min_r}(A^k)$ converge as $k \to \infty$ to some matrix P. Because $\mathbf{max_r}(A^k)$ has identical elements in each row, then so does P. Therefore, the columns of P are the same vector $\vec{\mathbf{w}}$. Take limits across inequality $\mathbf{min_r}(A^k) \ge \delta\mathcal{O}$ to prove $\vec{\mathbf{w}} > \vec{\mathbf{0}}$. Vector $\vec{\mathbf{w}}$ equals $P\vec{\mathbf{u}}$, where $\vec{\mathbf{u}}$ = column 1 of the identity matrix. Then $\vec{\mathbf{w}} = P\vec{\mathbf{u}} = \lim_{k\to\infty} A^{k+1}\vec{\mathbf{u}} = A\left(\lim_{k\to\infty} A^k\vec{\mathbf{u}}\right) = A\vec{\mathbf{w}}$, which is the eigenpair equation $\vec{\mathbf{w}} = A\vec{\mathbf{w}}$.

Proof of (b)

Eigenpair equation $\vec{\mathbf{v}} = A\vec{\mathbf{v}}$ is multiplied repeatedly by A to give $\vec{\mathbf{v}} = A^{k+1}\vec{\mathbf{v}}$. Take the limit using part (a): $\vec{\mathbf{v}} = P\vec{\mathbf{v}}$, where $P = \Big\langle \vec{\mathbf{w}}\,|\vec{\mathbf{w}}\,|\cdots|\vec{\mathbf{w}}\Big\rangle$. Then $\vec{\mathbf{v}} = P\vec{\mathbf{v}} = (\sum_{i=1}^n v_i)\,\vec{\mathbf{w}}$.

Proof of (c)

Consider an eigenpair $(\lambda, \vec{\mathbf{v}})$. Apply A across $\lambda\vec{\mathbf{v}} = A\vec{\mathbf{v}}$ to obtain $\lambda^k\vec{\mathbf{v}} = A^k\vec{\mathbf{v}}$. Use part (a) to take the limit as $k \to \infty$. Then, as in part (b), $\lim_{k\to\infty} \lambda^k\vec{\mathbf{v}} = (\sum_{i=1}^n v_i)\,\vec{\mathbf{w}}$. This limit exists only in case $|\lambda| \le 1$. If $|\lambda| = 1$, then $\lambda = e^{i\theta}$ for some angle θ. The limit fails to exist unless $\theta = 0$ modulo 2π. Therefore, $\lambda = 1$ and $\vec{\mathbf{v}} = (\sum_{i=1}^n v_i)\,\vec{\mathbf{w}}$.

Proof of (d)

Let's suppose some $v_j = 0$, in order to reach a contradiction. Component j of the identity

[14] The numerical **power method** can be used to approximate eigenvector $\vec{\mathbf{w}}$.

$A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ says that $\sum_{k=1}^n a_{jk}v_k = 0$. Because $\vec{\mathbf{v}} \neq \vec{\mathbf{0}}$, then at least one $v_k \neq 0$. Because $a_{jk} > 0$, then $\sum_{k=1}^n a_{jk}v_k > 0$, a contradiction.

Perron-Frobenius proof completed. ■

Exercises 9.2

Discrete Dynamical Systems

Define matrix A via equation

$$\vec{\mathbf{y}} = \frac{1}{10}\begin{pmatrix} 5 & 1 & 0 \\ 3 & 4 & 3 \\ 2 & 5 & 7 \end{pmatrix} \vec{\mathbf{x}} \tag{15}$$

1. Find eigenpair packages of A.
Answers:

$$D = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$P = \begin{pmatrix} -1 & 1 & 1 \\ 0 & -4 & 5 \\ 1 & 3 & 9 \end{pmatrix}$$

2. Explain: A is a **transition matrix**.[15]

3. Assume $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$ has period one year. Find the system state after two years.

4. Explain: $A^n\vec{\mathbf{x}}$ is the system state after n periods.

Market Shares

Define matrix A via equation

$$\vec{\mathbf{y}} = \frac{1}{10}\begin{pmatrix} 5 & 4 & 0 \\ 3 & 5 & 3 \\ 2 & 1 & 7 \end{pmatrix} \vec{\mathbf{x}} \tag{16}$$

5. Find with software the eigenpairs of A given by equation 2.

6. Compute A^2, A^3, A^4 using software. Predict the limit of A^n as n approaches infinity.

7. Compute with software (rounded)

$$A^{10} = \begin{pmatrix} .30 & .30 & .30 \\ .37 & .38 & .37 \\ .32 & .32 & .33 \end{pmatrix} \tag{17}$$

8. Let $\vec{\mathbf{x}} = \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$. Compute

$$A^{10}\vec{\mathbf{x}} = \begin{pmatrix} 0.30 \\ 0.37 \\ 0.33 \end{pmatrix} \text{ (rounded)}$$

in two ways by calculator:
(1) Fourier replacement (3).
(2) Matrix multiply using (17).

Stochastic Matrices

Reference: Perron-Frobenius proof on page 715.

9. Establish the identity $|A - \lambda I| = |A^T - \lambda I|$.

10. Explain why A and A^T have the same eigenvalues but not necessarily the same eigenvectors.

11. Verify $\mathbf{max_r}(A) = \langle \vec{\mathbf{w}} | \vec{\mathbf{w}} | \cdots | \vec{\mathbf{w}} \rangle$, where $\vec{\mathbf{w}}$ has components $w_i = \max\{a_{ij}, 1 \leq j \leq n\}$.

12. Verify $\mathbf{max_r}(A) = D\mathcal{O}$, where D is the diagonal matrix of row maxima and $\mathcal{O}$ is the matrix of all ones.

Perron-Frobenius Theorem

Let $A > 0$ be $n \times n$ stochastic with unique eigenpair $(1, \vec{\mathbf{w}})$, all $w_i > 0$ and $\sum_{i=1}^n w_i = 1$. Assume $\vec{\mathbf{v}} \geq \vec{\mathbf{0}}$, $\sum_{i=1}^n v_i = 1$ and $\delta = \min_{i,j} a_{ij}$.

13. Apply inequality $\mathbf{min_r}(A^n)\vec{\mathbf{v}} \leq A^n\vec{\mathbf{v}} \leq \mathbf{max_r}(A^n)\vec{\mathbf{v}}$ to prove $\lim_{n\to\infty} A^n\vec{\mathbf{v}} = \left(\sum_{i=1}^n v_i\right)\vec{\mathbf{w}} = \vec{\mathbf{w}}$.

14. Verify Euclidean norm inequality
$\|A^{k+1}\vec{\mathbf{v}} - \vec{\mathbf{w}}\| \leq \sqrt{n}\,(1-\delta)^k$

Weierstrass Proof

These exercises establish existence of an eigenpair $(1, \vec{\mathbf{v}})$ for stochastic matrix A having only nonnegative entries.

[15] Perron-Frobenius theory extensions in the literature apply to transition matrices. See the Weierstrass Proof exercises.

Weierstrass Compactness Theorem
A sequence of vectors $\{\vec{\mathbf{v}}_i\}_{i=1}^{\infty}$ contained in a closed, bounded set K in $\mathcal{R}^n$ has a subsequence converging in the vector norm of $\mathcal{R}^n$ to some vector $\vec{\mathbf{v}}$ in K.

Define set K to be all vectors $\vec{\mathbf{v}}$ with non-negative components adding to 1. Let $\vec{\mathbf{v}}_0$ be any element of K. Assume stochastic A with $a_{ij} \ge 0$ and define $\vec{\mathbf{v}}_N = \frac{1}{N}\sum_{j=0}^{N-1} A^j\vec{\mathbf{v}}_0$.

15. Verify K is closed and bounded in $\mathcal{R}^n$. Then prove $\lambda\vec{\mathbf{x}} + (1-\lambda)\vec{\mathbf{y}}$ is in K for $0 \le \lambda \le 1$ and $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ in K.

16. Prove identity
$\vec{\mathbf{v}}_{N+1} = \lambda\vec{\mathbf{v}}_N + (1-\lambda)A^N\vec{\mathbf{v}}_0$
where $\lambda = \frac{N}{N+1}$ and then prove by induction that $\vec{\mathbf{v}}_N$ is in K.

17. Verify all hypotheses in the Weierstrass theorem applied to $\{\vec{\mathbf{v}}_N\}_{N=0}^{\infty}$. Applying the theorem produces a subsequence $\{\vec{\mathbf{v}}_{N_p}\}_{p=1}^{\infty}$ limiting to some $\vec{\mathbf{v}}$ in K.

18. Verify identity
$\vec{\mathbf{v}}_N - A\vec{\mathbf{v}}_N = \frac{1}{N}(\vec{\mathbf{v}}_0 - A^N\vec{\mathbf{v}}_0)$.

19. Explain why $A\vec{\mathbf{v}} = \lim_{p\to\infty} A\vec{\mathbf{v}}_{N_p}$. Then prove $\vec{\mathbf{v}} = A\vec{\mathbf{v}}$.

20. The claimed eigenpair $(1, \vec{\mathbf{v}})$ has been found, provided $\vec{\mathbf{v}} \ne \vec{\mathbf{0}}$. Explain why $\vec{\mathbf{v}} \ne \vec{\mathbf{0}}$.

Coupled Systems

Find the coefficient matrix A. Identify as coupled or uncoupled and explain why.

21. $x' = 2x + 3y$, $y' = x + y$

22. $x' = 3y$, $y' = x$

23. $x' = 3x$, $y' = 2y$

24. $x' = 3x$, $y' = 2y$, $z' = z$

Solving Uncoupled Systems

Solve for the general solution.

25. $x' = 3x$, $y' = 2y$

26. $x' = 3x$, $y' = 2y$, $z' = z$

Change of Coordinates

Given the change of coordinates $\vec{\mathbf{y}} = A\vec{\mathbf{x}}$, find the matrix B for the inverse change $\vec{\mathbf{x}} = B\vec{\mathbf{y}}$.

27. $\vec{\mathbf{y}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \vec{\mathbf{x}}$

28. $\vec{\mathbf{y}} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \vec{\mathbf{x}}$

Constructing Coupled Systems

Given the uncoupled system and change of coordinates $\vec{\mathbf{y}} = P\vec{\mathbf{x}}$, find the coupled system.

29. $x_1' = 2x_1$, $x_2' = 3x_2$, $P = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}$

30. $x_1' = x_1$, $x_2' = -x_2$, $P = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$

Uncoupling a System

Change the given coupled system into an uncoupled system using the eigenanalysis change of variables $\vec{\mathbf{y}} = P\vec{\mathbf{x}}$.

31. $x_1' = 2x_1$, $x_2' = x_1 + x_2$, $x_3' = x_3$
Ans: $P = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, $y_1' = 2y_1$, $y_2' = y_2$, $y_3' = y_3$

32. $x_1' = x_1 + x_2$, $x_2' = x_1 + x_2$, $x_3' = x_3$
Ans: $P = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $y_1' = 0$, $y_2' = 2y_2$, $y_3' = y_3$

Solving Coupled Systems

Report the answers for $x(t), y(t)$.

33. $x' = -x - 2y$, $y' = -4x + y$

34. $x' = 8x - y$, $y' = -2x + 7y$

Eigenanalysis and Footballs

The exercises study the ellipsoid
$17x^2 + 8y^2 - 12xy + 80z^2 = 80$.

35. Let $A = \begin{pmatrix} 17 & -6 & 0 \\ -6 & 8 & 0 \\ 0 & 0 & 80 \end{pmatrix}$. Expand equation $\vec{\mathbf{W}}^T A\vec{\mathbf{W}} = 80$, where $\vec{\mathbf{W}}$ has components x, y, z.

36. Find the eigenpairs of
$$A = \begin{pmatrix} 17 & -6 & 0 \\ -6 & 8 & 0 \\ 0 & 0 & 80 \end{pmatrix}.$$

37. Verify the semi-axis lengths $4, 2, 1$.

38. Verify that the ellipsoid has semi-axis unit directions
$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{5}}\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}$$

The Ellipse and Eigenanalysis

The exercises study the ellipse $2x^2 + 4xy + 5y^2 = 24$.

39. Let $A = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix}$. Expand equation $\vec{\mathbf{W}}^T A\vec{\mathbf{W}} = 24$, where $\vec{\mathbf{W}} = \begin{pmatrix} x \\ y \end{pmatrix}$.

40. Find the eigenpairs of $A = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix}$.

41. Verify the semi-axis lengths $2, 2\sqrt{6}$.

42. Verify that the ellipse has semi-axis unit directions
$$\frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \frac{1}{\sqrt{5}}\begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Orthogonal Triad Computation

The exercises fill in details from page 711. The ellipsoid equation: $x^2{+}4y^2{+}xy{+}4z^2{=}16$ or $\vec{\mathbf{x}}^T A\vec{\mathbf{x}}{=}16$,
$$A = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

43. Find the characteristic equation of A. Then verify the roots are 4, $5/2 + \sqrt{10}/2$, $5/2 - \sqrt{10}/2$.

44. Show the steps from **rref** to second eigenvector $\vec{\mathbf{x}}_2$:
$$\mathbf{rref} = \begin{pmatrix} 1 & 3-\sqrt{10} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$
$$\vec{\mathbf{x}}_2 = \begin{pmatrix} \sqrt{10}-3 \\ 1 \\ 0 \end{pmatrix}$$

9.3 Advanced Topics in Linear Algebra

Diagonalization and Jordan's Theorem

A system of differential equations $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ can be transformed to an uncoupled system $\vec{\mathbf{y}}' = \textbf{diag}(\lambda_1, \ldots, \lambda_n)\vec{\mathbf{y}}$ by a change of variables $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ if and only if A is **diagonalizable** and P is an invertible matrix of independent eigenvectors of A from eigenpairs $(\lambda_k, \vec{\mathbf{v}}_k)$, $1 \le k \le n$.

If A fails to be diagonalizable, then eigenanalysis does not help. Jordan's theorem 9.14 is a possible generator of a change of coordinates $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$. System $\vec{\mathbf{y}}' = T\vec{\mathbf{y}}$ is not uncoupled, but triangular: the linear integrating factor method applies to solve the triangular system, details forthcoming.

The sad truth about Jordan's theorem: matrix P has no algorithm for construction. The matrix P used as replacement is a matrix of **generalized eigenvectors** constructed from an algorithm for the Jordan normal form page 894. See page 898 for a `maple` example.

Theoretical existence of P for a change of variables may be enough for proofs. Computation requires a formula for P. What has emerged historically are mathematical algorithms to solve system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ independent of both Jordan's theorem and the Jordan normal form page 894. The foundation for computer algebra algorithms and low dimensional hand algorithms is the Cayley-Hamilton theorem 9.16 on page 721.

Theorem 9.14 (Jordan's theorem)
Any $n \times n$ matrix A can be represented in the form

$$A = PTP^{-1}$$

where P is invertible and T is upper triangular. The diagonal entries of T are eigenvalues of A.

Proof on page 740.

Theorem 9.15 (Jordan's Extension)
Any $n \times n$ matrix A can be represented in the block triangular form

$$A = PTP^{-1}, \quad T = \textbf{diag}(T_1, \ldots, T_k),$$

where P is invertible and each matrix T_i is upper triangular with diagonal entries equal to a single eigenvalue of A.

Remarks. An induction proof of the theorem can be based upon Jordan's Theorem 9.14. No proof is supplied. The theorem is presented again in Proposition 9.15 page 720 as a special case of the Jordan decomposition $A = PJP^{-1}$, in which J is the Jordan Form of $n \times n$ matrix A. Because the Jordan form is a triangular matrix, then $T = J$ gives an algorithm for generation of columns in matrix P. Jordan form is largely used in proofs and theoretical investigations and rarely in computation.
Computer algebra systems can find matrices J and P in the Jordan form of matrix A. With limitations, there is a constructible matrix P for Jordan's two theorems 9.14 and 9.15. See page 898 for a `maple` example.

Cayley-Hamilton Identity

A celebrated and deep result for powers of matrices was discovered by Cayley and Hamilton (see Birkhoff–MacLane [Birkhoff]), which says that an $n \times n$ matrix A satisfies its own characteristic equation. More precisely:

Theorem 9.16 (Cayley-Hamilton)
Let $\det(A - \lambda I)$ be expanded as the nth degree polynomial

$$p(\lambda) = \sum_{j=0}^{n} c_j \lambda^j,$$

for some coefficients $c_0, \ldots, c_{n-1}$ and $c_n = (-1)^n$. Then A satisfies the equation $p(\lambda) = 0$, that is,

$$p(A) \equiv \sum_{j=0}^{n} c_j A^j = \mathbf{0}.$$

In factored form in terms of the eigenvalues $\{\lambda_j\}_{j=1}^{n}$ (duplicates possible), the matrix equation $p(A) = \mathbf{0}$ can be written as

$$(-1)^n (A - \lambda_1 I)(A - \lambda_2 I) \cdots (A - \lambda_n I) = \mathbf{0}.$$

Proof on page 741.

Solving Block Triangular Differential Systems

A matrix differential system $\vec{\mathbf{y}}\,'(t) = T\vec{\mathbf{y}}\,(t)$ with T block upper triangular splits into scalar equations which can be solved by elementary methods for first order scalar differential equations. To illustrate, consider the system

$$\begin{aligned} y_1' &= 3y_1 + x_2 + y_3, \\ y_2' &= 3y_2 + y_3, \\ y_3' &= 2y_3. \end{aligned}$$

The techniques that apply are the growth-decay formula for $u' = ku$ and the integrating factor method for $u' = ku + p(t)$. Working backwards from the last equation with back-substitution gives

$$\begin{array}{rcl} y_3 &=& c_3e^{2t}, \\ y_2 &=& c_2e^{3t} - c_3e^{2t}, \\ y_1 &=& (c_1 + c_2t)e^{3t}. \end{array}$$

What has been said here applies to any triangular system $\vec{\mathbf{y}}\,'(t) = T\vec{\mathbf{y}}\,(t)$, in order to write an exact formula for the solution $\vec{\mathbf{y}}\,(t)$.

If A is an $n \times n$ matrix, then Jordan's theorem gives $A = PTP^{-1}$ with T block upper triangular and P invertible. The change of variable $\vec{\mathbf{x}}\,(t) = P\vec{\mathbf{y}}\,(t)$ changes $\vec{\mathbf{x}}\,'(t) = A\vec{\mathbf{x}}\,(t)$ into the block triangular system $\vec{\mathbf{y}}\,'(t) = T\vec{\mathbf{y}}\,(t)$.

There is no special condition on A, to effect the change of variable $\vec{\mathbf{x}}\,(t) = P\vec{\mathbf{y}}\,(t)$. The solution $\vec{\mathbf{x}}\,(t)$ of $\vec{\mathbf{x}}\,'(t) = A\vec{\mathbf{x}}\,(t)$ is a product of the invertible matrix P and a column vector $\vec{\mathbf{y}}\,(t)$; the latter is the solution of the block triangular system $\vec{\mathbf{y}}\,'(t) = T\vec{\mathbf{y}}\,(t)$, obtained by growth-decay and integrating factor methods.

The *importance of this idea* is to provide a theoretical method for solving any system $\vec{\mathbf{x}}\,'(t) = A\vec{\mathbf{x}}\,(t)$.

Matrices P and T in Jordan's extension $A = PTP^{-1}$ can be found using computer algebra systems. See page 898 for a `maple` example in which T is the Jordan normal form of A and P is the matrix of generalized eigenvectors.

Symmetric Matrices and Orthogonality

A **symmetric matrix** A is defined by the identity $A^T = A$. In applications the symmetric matrix A might be obtained as $A = B^TB$ for some non-square matrix B. Studied here is the eigenanalysis of symmetric matrices, which reproduces $AP = PD$ from classical eigenanalysis with a difference: the eigenvectors in columns of P are of **unit length**, meaning $\|\vec{\mathbf{x}}\,\| = 1$, and also **orthogonal**, meaning dot product zero or 90 degrees apart. See Chapter 5 Section 1.

Definition 9.8 (Unitize)
A vector $\vec{\mathbf{x}}$ is said to be **unitized** into vector $\vec{\mathbf{y}}$ if $\vec{\mathbf{y}} = c\vec{\mathbf{x}}$ for some scalar c and $\|\vec{\mathbf{y}}\,\| = 1$.

An eigenpair $(\lambda, \vec{\mathbf{x}}\,)$ of A can always be selected so that $\|\vec{\mathbf{x}}\,\| = 1$: replace eigenvector $\vec{\mathbf{x}}$ by $\frac{1}{\|\vec{\mathbf{x}}\|}\vec{\mathbf{x}}$.

Theorem 9.17 (Orthogonality of Eigenvectors)
Assume that $n \times n$ matrix A is **symmetric**, $A^T = A$. If $(\alpha, \vec{\mathbf{x}}\,)$ and $(\beta, \vec{\mathbf{y}}\,)$ are eigenpairs of A with $\alpha \ne \beta$, then $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ are orthogonal: $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = 0$. Proof on page 741.

Theorem 9.18 (Real Eigenvalues)
If $A^T = A$, then all eigenvalues of A are real. Consequently, matrix A has n real eigenvalues counted according to multiplicity. Proof on page 741.

Proposition 9.1 (Independence of Orthogonal Sets) Let $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_k$ be a set of nonzero orthogonal vectors. Then this set is independent.

Duplicated by the orthogonal vector test Chapter 5 Section 3, Theorem 5.33.

The Gram-Schmidt process

The eigenvectors of a symmetric matrix A may be constructed to be orthogonal. First of all, observe that eigenvectors corresponding to distinct eigenvalues are orthogonal by Theorem 9.17. It remains to construct from k independent eigenvectors $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k$, corresponding to a single eigenvalue λ, another set of independent eigenvectors $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_k$ for λ which are pairwise orthogonal. The idea, due to Gram-Schmidt, applies to any set of k independent vectors.

Theorem 9.19 (Gram-Schmidt)
Let $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k$ be independent n-vectors. The set of vectors $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_k$ constructed below as linear combinations of $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k$ are pairwise orthogonal, independent and $\mathbf{span}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k) = \mathbf{span}(\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_k)$.

$$\begin{aligned}
\vec{\mathbf{y}}_1 &= \vec{\mathbf{x}}_1 \\
\vec{\mathbf{y}}_2 &= \vec{\mathbf{x}}_2 - \frac{\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}}_1}{\vec{\mathbf{y}}_1 \cdot \vec{\mathbf{y}}_1}\vec{\mathbf{y}}_1 \\
\vec{\mathbf{y}}_3 &= \vec{\mathbf{x}}_3 - \frac{\vec{\mathbf{x}}_3 \cdot \vec{\mathbf{y}}_1}{\vec{\mathbf{y}}_1 \cdot \vec{\mathbf{y}}_1}\vec{\mathbf{y}}_1 - \frac{\vec{\mathbf{x}}_3 \cdot \vec{\mathbf{y}}_2}{\vec{\mathbf{y}}_2 \cdot \vec{\mathbf{y}}_2}\vec{\mathbf{y}}_2 \\
&\vdots \\
\vec{\mathbf{y}}_k &= \vec{\mathbf{x}}_k - \frac{\vec{\mathbf{x}}_k \cdot \vec{\mathbf{y}}_1}{\vec{\mathbf{y}}_1 \cdot \vec{\mathbf{y}}_1}\vec{\mathbf{y}}_1 - \cdots - \frac{\vec{\mathbf{x}}_k \cdot \vec{\mathbf{y}}_{k-1}}{\vec{\mathbf{y}}_{k-1} \cdot \vec{\mathbf{y}}_{k-1}}\vec{\mathbf{y}}_{k-1}
\end{aligned}$$

Proof on page 742.

Example 9.14 (Gram-Schmidt on Four Eigenvectors)

Let $(-1, \vec{\mathbf{v}}_1)$, $(2, \vec{\mathbf{v}}_2)$, $(2, \vec{\mathbf{v}}_3)$, $(2, \vec{\mathbf{v}}_4)$ be eigenpairs of a 4×4 symmetric matrix A. Apply the Gram-Schmidt process to find 4 pairwise orthogonal eigenvectors of A.

Solution: Because eigenvector $\vec{\mathbf{v}}_1$ is for eigenvalue 1 and the others are for eigenvalue 2, then Theorem 9.17 implies that $\vec{\mathbf{v}}_1$ is orthogonal to $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$. Eigenvectors $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$ belong to eigenvalue $\lambda = 2$, but they are not assumed orthogonal. The Gram-Schmidt process applied to eigenvectors $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$ finds pairwise orthogonal vectors $\vec{\mathbf{y}}_2$, $\vec{\mathbf{y}}_3$, $\vec{\mathbf{y}}_4$ that are linear combinations of eigenvectors $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$. Then $\vec{\mathbf{y}}_2$, $\vec{\mathbf{y}}_3$, $\vec{\mathbf{y}}_4$ are also eigenvectors for $\lambda = 2$. The four eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{y}}_2$, $\vec{\mathbf{y}}_3$, $\vec{\mathbf{y}}_4$ are pairwise orthogonal, as required.

Orthogonal Projection

Reproduced here for reference is the basic material on shadow projection. The ideas are then extended to obtain the orthogonal projection onto a subspace V of $\mathcal{R}^n$. Finally, the orthogonal projection formula will be related to the Gram-Schmidt equations.

The **shadow projection** of vector $\vec{X}$ onto the direction of vector $\vec{Y}$ is the number d defined by

$$d = \frac{\vec{X}\cdot\vec{Y}}{|\vec{Y}|}.$$

The triangle determined by $\vec{X}$ and $d\dfrac{\vec{Y}}{|\vec{Y}|}$ is a right triangle.

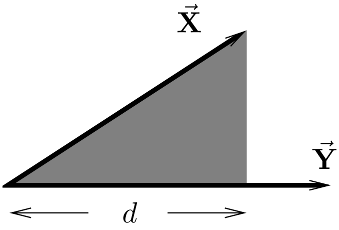


Figure 4. Shadow projection d of vector $\vec{\mathbf{X}}$ onto the direction of vector $\vec{\mathbf{Y}}$.

The **vector shadow projection** of $\vec{X}$ onto the line L through the origin in the direction of $\vec{Y}$ is the vector representing the shadow, direction $\vec{Y}$ and length d, defined by

$$\mathbf{proj}_{\vec{Y}}(\vec{X}) = d\frac{\vec{Y}}{|\vec{Y}|} = \frac{\vec{X}\cdot\vec{Y}}{\vec{Y}\cdot\vec{Y}}\vec{Y}.$$

Definition 9.9 (One-Dimensional Orthogonal Projection)
Let V be the line through the origin in the direction of nonzero vector $\vec{\mathbf{Y}}$. Then $V = \mathbf{span}\{\vec{\mathbf{Y}}\}$. Define the **orthogonal projection**:

$$\mathbf{Proj}_V(\vec{\mathbf{x}}) = (\vec{\mathbf{u}}\cdot\vec{\mathbf{x}})\vec{\mathbf{u}}, \quad \vec{\mathbf{u}} = \vec{\mathbf{Y}}/\|\vec{\mathbf{Y}}\|$$

Is Definition 9.9 the same as vector shadow projection? Yes. Does the definition depend on $\vec{\mathbf{Y}}$? No, because of Theorem 9.20 below.

Definition 9.10 (Orthogonal Projection onto a Subspace)
Let subspace V of $\mathcal{R}^n$ be spanned by orthonormal vectors $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_k$. Define the **orthogonal projection** of vector $\vec{\mathbf{x}}$ in $\mathcal{R}^n$ onto subspace V by the formula (justified in Theorem 9.20):

$$(1) \qquad \begin{aligned} \mathbf{Proj}_V(\vec{\mathbf{x}}) &= \textstyle\sum_{j=1}^k (\vec{\mathbf{u}}_j\cdot\vec{\mathbf{x}})\vec{\mathbf{u}}_j, \\ &= \textstyle\sum_{j=1}^k \text{vector shadow projection } \vec{\mathbf{x}} \text{ onto } \vec{\mathbf{u}}_j \end{aligned}$$

Theorem 9.20 (Formula $\mathbf{Proj}_V(\vec{\mathbf{x}})$ is Well-Defined)
Orthogonal projection formula $\mathbf{Proj}_V(\vec{\mathbf{x}}) = \sum_{j=1}^{k}(\vec{\mathbf{u}}_j \cdot \vec{\mathbf{x}})\vec{\mathbf{u}}_j$ is independent of the choice of orthonormal vectors $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_k$ that span V.

Proof on page 742

Important: Formula $\mathbf{Proj}_V(\vec{\mathbf{x}}) = \sum_{j=1}^{k}(\vec{\mathbf{u}}_j \cdot \vec{\mathbf{x}})\vec{\mathbf{u}}_j$ requires a basis which is **orthonormal**. An **orthogonal** basis suffices with the shadow projection summation in (1). Applications might use either formula.

Orthogonal Projection and Gram-Schmidt. Define $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_k$ by the Gram-Schmidt relations on page 723. Define

$$\vec{\mathbf{u}}_j = \vec{\mathbf{y}}_j / \|\vec{\mathbf{y}}_j\|$$

for $j = 1, \ldots, k$. Then $V_{j-1} = \mathbf{span}\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_{j-1}\}$ is a subspace of $\mathcal{R}^n$ of dimension $j-1$ with orthonormal basis $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_{j-1}$ and

$$\begin{aligned} \vec{\mathbf{y}}_j &= \vec{\mathbf{x}}_j - \left(\frac{\vec{\mathbf{x}}_j \cdot \vec{\mathbf{y}}_1}{\vec{\mathbf{y}}_1 \cdot \vec{\mathbf{y}}_1}\vec{\mathbf{y}}_1 + \cdots + \frac{\vec{\mathbf{x}}_k \cdot \vec{\mathbf{y}}_{j-1}}{\vec{\mathbf{y}}_{j-1} \cdot \vec{\mathbf{y}}_{j-1}}\vec{\mathbf{y}}_{j-1}\right) \\ &= \vec{\mathbf{x}}_j - \mathbf{Proj}_{V_{j-1}}(\vec{\mathbf{x}}_j) \end{aligned}$$

The Gram-Schmidt relations are memorized by the formula

$$\vec{\mathbf{y}}_j = \vec{\mathbf{x}}_j - \sum_{k<j} (\text{vector shadow projection of } \vec{\mathbf{x}}_j \text{ onto } \vec{\mathbf{y}}_k)$$

Near Point Theorem

Developed here is the characterization of the orthogonal projection of a vector $\vec{\mathbf{x}}$ onto a subspace V as the unique point $\vec{\mathbf{v}}$ in V which minimizes $\|\vec{\mathbf{x}} - \vec{\mathbf{v}}\|$, that is, the point in V which is nearest to $\vec{\mathbf{x}}$.

Theorem 9.21 (Orthogonal Projection Properties)
Let subspace V be the span of orthonormal vectors $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_k$.

(a) Each vector $\vec{\mathbf{v}}$ in V has an orthogonal expansion $\vec{\mathbf{v}} = \sum_{j=1}^{k}(\vec{\mathbf{u}}_j \cdot \vec{\mathbf{v}})\vec{\mathbf{u}}_j$.

(b) The orthogonal projection $\mathbf{Proj}_V(\vec{\mathbf{x}})$ is a vector in V.

(c) Vector $\vec{\mathbf{w}} = \vec{\mathbf{x}} - \mathbf{Proj}_V(\vec{\mathbf{x}})$ is orthogonal to every vector in V.

(d) Among all vectors $\vec{\mathbf{v}}$ in V, the minimum value of $\|\vec{\mathbf{x}} - \vec{\mathbf{v}}\|$ is uniquely obtained by the orthogonal projection $\vec{\mathbf{v}} = \mathbf{Proj}_V(\vec{\mathbf{x}})$.

(e) Let $n \times k$ matrix A have independent columns that span V. If vector $\vec{\mathbf{w}}$ is orthogonal to every vector in V, then $A^T\vec{\mathbf{w}} = \vec{\mathbf{0}}$.

Proof on page 743.

Theorem 9.22 (Near Point to a Subspace)

Let V be a subspace of $\mathcal{R}^n$ and $\vec{\mathbf{x}}$ a vector not in V. The **near point** to $\vec{\mathbf{x}}$ in V is the orthogonal projection of $\vec{\mathbf{x}}$ onto V. This point is characterized as the minimum of $\|\vec{\mathbf{x}} - \vec{\mathbf{v}}\|$ over all vectors $\vec{\mathbf{v}}$ in the subspace V.

Proof by part **(d)** of Theorem 9.21.

Theorem 9.23 (Cross Product and Projections)

The cross product $\vec{\mathbf{a}} \times \vec{\mathbf{b}}$ is a constant multiple of $\vec{\mathbf{c}} - \mathbf{Proj}_V(\vec{\mathbf{c}})$, where vector $\vec{\mathbf{c}}$ is not in $V = \mathbf{span}\{\vec{\mathbf{a}}, \vec{\mathbf{b}}\}$.

Proof: The cross product makes sense only in $\mathcal{R}^3$. Subspace V is two dimensional when $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$ are independent, and Gram-Schmidt applies to find an orthonormal basis $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$. By (c) of Theorem 9.21, the vector $\vec{\mathbf{c}} - \mathbf{Proj}_V(\vec{\mathbf{c}})$ has the same or opposite direction to the cross product. ■

Linear Least Squares

A primary application of linear least squares is fitting of large data sets to an equation. Desired is a simple equation which can be used to interpolate or extrapolate missing data items or to find trends in the data.

Example 9.15 (Height-Weight Data)

Verify that slope $m = 61.27$ and intercept $b = -39, 05$ best fit equation $y = mx + b$ to the 15 data items in Table 4, where x=height, y=weight. Graphic in Figure 5.

The solution is on page 744.

Table 4. Height-Weight Data for 15 women ages $30 - 39$ years.
Source: The World Almanac and Book of Facts, 1975.

Height (m)	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65
Weight (kg)	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29
Height (m)	1.68	1.70	1.73	1.75	1.78	1.80	1.83	
Weight (kg)	63.11	64.47	66.28	68.10	69.92	72.19	74.46	

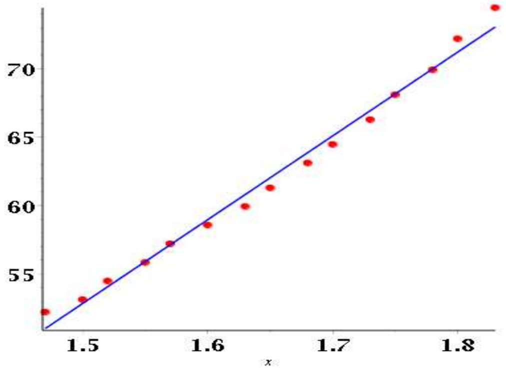


Figure 5. Best fit
The least squares fit straight line in blue $y = 61.27x - 39.06$.
Red dots are the 15 data points from Table 4.

Least Squares Normal Equation

Let $m \times n$ matrix A and vector $\vec{\mathbf{b}}$ be given. Assume hereafter that the problem $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ has *no solution*. The discussion will be guided by the unsolvable system

$$(2) \qquad \begin{cases} x + y = 1, \\ x + y = 0. \end{cases}$$

System (2) is the special case $b_1^* = 1$, $b_2^* = 0$ for system

$$(3) \qquad \begin{cases} x + y = b_1^*, \\ x + y = b_2^*. \end{cases}$$

Least squares chooses two values b_1^*, b_2^* such that system (3) is solvable for x, y. The choice requires that substitution of x, y into the original unsolvable equation (2) gives the **least error**, in some well-defined sense.

Mathematically, the least error for a trial solution x, y in (5) might be realized [16] by choosing x, y to minimize the vector norm $\|\vec{\mathbf{E}}\|$ for error vector

$$(4) \qquad \vec{\mathbf{E}} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Minimization leads to the geometry problem solved in Figure 6. By geometry, points on the line $x + y = \frac{1}{2}$ are half way between the two lines of the original system. Point $x = \frac{1}{2}$, $y = 0$ is isolated as a proper candidate for a *best solution* to original unsolvable problem (2).

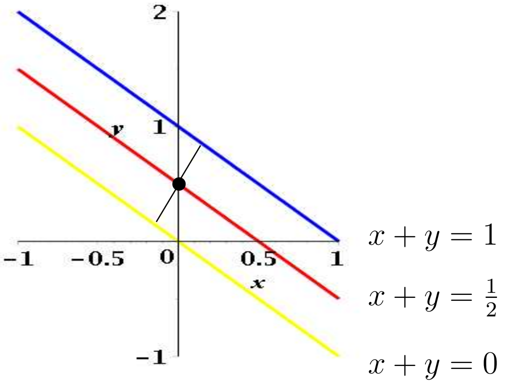


Figure 6. Black dot $x^* = \frac{1}{2}$, $y^* = 0$ is one best solution to system $x + y = 1$, $x + y = 0$. Any point along the red line $x + y = \frac{1}{2}$ makes minimum vector norm error between the two lines $x + y = 1$, $x + y = 0$.

Warning: The isolated point $x = \frac{1}{2}$, $y = 0$ does not actually work in the original equations! The geometrical solution invents one possible solvable replacement

[16] The expression to minimize is controversial: at the very least, it depends on the intended application.

system (3) with best fit to the original unsolvable equations (2):

$$\left\{\begin{array}{l} x+y=\frac{1}{2}, \\ x+y=\frac{1}{2}. \end{array}\right. \tag{5}$$

System (5) has a name:

Definition 9.11 (Normal Equation for Linear Least Squares)
The **normal equation** for unsolvable problem $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ is the solvable system

$$A^T A\vec{\mathbf{x}} = A^T\vec{\mathbf{b}} \tag{6}$$

It is **not** implied that a solution $\vec{\mathbf{x}}$ of (6) is also a solution of $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$: the original equation is assumed to have *no solution*.

System (3) has matrix form $A\vec{\mathbf{x}} = \vec{\mathbf{b}}^*$. If vector $\vec{\mathbf{x}}$ solves $A\vec{\mathbf{x}} = \vec{\mathbf{b}}^*$, then $\vec{\mathbf{b}}^*$ equals $A\vec{\mathbf{x}}$, which means $\vec{\mathbf{b}}^*$ is a linear combination of the columns of A, or $\vec{\mathbf{b}}^*$ belongs to subspace $S = \mathbf{colspace}(A)$. Overloaded symbol $\vec{\mathbf{x}}$ is not the same as in equation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$: the latter has no solution.

Geometrically, $\vec{\mathbf{b}}^*$ is a specific given vector in S and equation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}^*$ can have infinitely many solutions $\vec{\mathbf{x}}$, or just one. Important: the no solution case has been eliminated from the *three possibilities*.

Error minimization seeks a *best solution* $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ to the unsolvable problem $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$. Applied literature suggests to find $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ as a minimizer for a function which measures the error between $A\vec{\mathbf{x}}$ and $\vec{\mathbf{b}}$.

Proposition 9.2 $A\vec{\mathbf{x}} = \vec{\mathbf{b}}^*$ has a solution $\vec{\mathbf{x}}$ if and only if $\vec{\mathbf{b}}^*$ belongs to subspace $S = \mathbf{colspace}(A)$.

Proposition 9.3 Let $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ achieve the minimum for vector norm $\|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|$, taken over all $\vec{\mathbf{x}}$ in $\mathcal{R}^n$. Then $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ is a *best possible* solution of unsolvable equation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$, because it minimizes the vector norm error $\|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|$ over all possible $\vec{\mathbf{x}}$.

Theorem 9.24 (Least Squares Solution of Unsolvable $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$)
Let $\vec{\mathbf{x}}^*$ satisfy

$$\|A\vec{\mathbf{x}}^* - \vec{\mathbf{b}}\| = \min_{\vec{\mathbf{x}}} \|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|$$

Then $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ is a solution of **Normal Equation** $A^T A\vec{\mathbf{x}} = A^T\vec{\mathbf{b}}$. Vector $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ is a *best possible solution* for unsolvable equation $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$.
Proof on page 743.

Data Fitting

Assume given experimentally measured values $y_1, y_2, \ldots, y_m$ taken at independent variable values $x_1, x_2, \ldots, x_m$. Data fitting invents a model equation of the form[17]

$$y = \sum_{j=1}^{n} c_j f_j(x).$$

The invented functions f_j will have additional requirements, for example they could be polynomials $1, x, x^2, \ldots$ or trigonometric functions, e.g., a model motivated by truncation of Taylor series or Fourier series. The problem: find values for the constants $c_1, \ldots, c_n$.

Ideally, the model equation fits the data exactly. What actually holds is an exact equation with error terms $E_1, \ldots, E_m$:

$$y_i = \sum_{j=1}^{n} c_j f_j(x_i) + E_j$$

Linear least squares minimizes the sum of squares of the errors:

$$\min \sum_{j=1}^{m} |E_j|^2 \quad \text{over all choices of } c_1, \ldots, c_n$$

Minimization is assumed to return special values $c_1^*, \ldots, c_n^*$ giving the **best fit**. The predicted model for the data set is then:

$$y = \sum_{j=1}^{n} c_j^* f_j(x).$$

The QR Decomposition

Matrix multiply can express Gram-Schmidt formulas as $A = QR$, where A has independent columns $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_n$ and the columns of Q are the unitized Gram-Schmidt orthonormal vectors $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n$.

Definition 9.12 (Orthogonal Matrix)
A matrix Q having pairwise orthogonal columns of unit length is called **orthogonal**. Alternatively, $Q^TQ = I$. If Q is square, then $QQ^T = I$.[18]

[17] Statistical experiments might use vector variables. For instance a 3-vector $\vec{\mathbf{x}}$ with components of sex, age and height replaces scalar variable x. Scalars y_j could be vectors. Symbol c_j is replaced by symbol β_j, these parameters called **regressors**.

[18] Non-square matrices with *orthonormal columns* certainly exist. A warning: terminology **orthonormal matrix** usually means the matrix A is **square** and has orthonormal columns: $A^TA = AA^T = I$.

Theorem 9.25 (The QR-Decomposition)
Let the $m \times n$ matrix A have independent columns $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_n$. Then there is an upper triangular matrix R with positive diagonal entries and an orthogonal matrix Q such that

$$A = QR.$$

Proof: Let $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n$ be the Gram-Schmidt orthogonal vectors given by relations on page 723. Define $\vec{\mathbf{u}}_k = \vec{\mathbf{y}}_k/\|\vec{\mathbf{y}}_k\|$ and $r_{kk} = \|\vec{\mathbf{y}}_k\|$ for $k = 1, \ldots, n$, and otherwise $r_{ij} = \vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}}_j$. Let $Q = \left\langle \vec{\mathbf{u}}_1 | \cdots | \vec{\mathbf{u}}_n \right\rangle$. Then

$$(7) \qquad \begin{array}{rcl} \vec{\mathbf{x}}_1 & = & r_{11}\vec{\mathbf{u}}_1, \\ \vec{\mathbf{x}}_2 & = & r_{22}\vec{\mathbf{u}}_2 + r_{21}\vec{\mathbf{u}}_1, \\ \vec{\mathbf{x}}_3 & = & r_{33}\vec{\mathbf{u}}_3 + r_{31}\vec{\mathbf{u}}_1 + r_{32}\vec{\mathbf{u}}_2, \\ & \vdots & \\ \vec{\mathbf{x}}_n & = & r_{nn}\vec{\mathbf{u}}_n + r_{n1}\vec{\mathbf{u}}_1 + \cdots + r_{nn-1}\vec{\mathbf{u}}_{n-1}. \end{array}$$

It follows from (7) and matrix multiplication that $A = QR$. The columns of Q have unit length and they are pairwise orthogonal: Q is orthogonal. ■

Theorem 9.26 (Matrices Q and R in $A = QR$)
Let $m \times n$ matrix A have independent columns $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_n$. Let $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n$ be the Gram-Schmidt orthogonal vectors from page 723. Define $\vec{\mathbf{u}}_k = \vec{\mathbf{y}}_k/\|\vec{\mathbf{y}}_k\|$. Then $AQ = QR$ is satisfied by $Q = \left\langle \vec{\mathbf{u}}_1 | \cdots | \vec{\mathbf{u}}_n \right\rangle$ and

$$R = \begin{pmatrix} \|y_1\| & \vec{\mathbf{u}}_1 \cdot \vec{\mathbf{x}}_2 & \vec{\mathbf{u}}_1 \cdot \vec{\mathbf{x}}_3 & \cdots & \vec{\mathbf{u}}_1 \cdot \vec{\mathbf{x}}_n \\ 0 & \|y_2\| & \vec{\mathbf{u}}_2 \cdot \vec{\mathbf{x}}_3 & \cdots & \vec{\mathbf{u}}_2 \cdot \vec{\mathbf{x}}_n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \|y_n\| \end{pmatrix}.$$

Proof: Details are contained in the proof of Theorem 9.25 above. ■

Some references cite the diagonal entries as $\|\vec{\mathbf{x}}_1\|, \|\vec{\mathbf{x}}_2^{\perp}\|, \ldots, \|\vec{\mathbf{x}}_n^{\perp}\|$, where $\vec{\mathbf{x}}_j^{\perp} = \vec{\mathbf{x}}_j - \mathbf{Proj}_{V_{j-1}}(\vec{\mathbf{x}}_j)$, $V_{j-1} = \mathbf{span}\{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_{j-1}\}$. Because $\vec{\mathbf{y}}_1 = \vec{\mathbf{x}}_1$ and $\vec{\mathbf{y}}_j = \vec{\mathbf{x}}_j - \mathbf{Proj}_{V_{j-1}}(\vec{\mathbf{x}}_j)$, the formulas for the entries of R are identical.

Theorem 9.27 (Uniqueness of Q and R)
Let $m \times n$ matrix A have independent columns and satisfy the decomposition $A = QR$. If Q is $m \times n$ orthogonal and R is $n \times n$ upper triangular with positive diagonal elements, then Q and R are uniquely determined.

Proof: The problem is to show that $A = Q_1R_1 = Q_2R_2$ implies $R_2R_1^{-1} = I$ and $Q_1 = Q_2$. We start with $Q_1 = Q_2R_2R_1^{-1}$. Define $P = R_2R_1^{-1}$. Then $Q_1 = Q_2P$. Because $I = Q_1^TQ_1 = P^TQ_2^TQ_2P = P^TP$, then P is orthogonal. Matrix P is the product of square upper triangular matrices with positive diagonal elements, which implies P itself is square upper triangular with positive diagonal elements. The only orthogonal matrix with these properties is the identity matrix I. Then $R_2R_1^{-1} = P = I$, which implies $R_1 = R_2$ and $Q_1 = Q_2$. ■

Theorem 9.28 (The QR Decomposition and Least Squares)
Let $m \times n$ matrix A have independent columns and satisfy the decomposition $A = QR$ with Q orthogonal and R invertible. Then the normal equation

$$A^T A \vec{\mathbf{x}} = A^T \vec{\mathbf{b}}$$

in the theory of least squares can be represented as

$$R\vec{\mathbf{x}} = Q^T \vec{\mathbf{b}}.$$

Proof: Because Q is orthogonal, then $Q^TQ = I$. Let's use the identity $(CD)^T = D^TC^T$, the equation $A = QR$, and assumed R^T invertible to obtain

$$\begin{aligned}
&A^T A\vec{\mathbf{x}} = A^T\vec{\mathbf{b}} && \text{Normal equation} \\
&R^TQ^TQR\vec{\mathbf{x}} = R^TQ^T\vec{\mathbf{x}} && \text{Substitute } A = QR. \\
&R\vec{\mathbf{x}} = Q^T\vec{\mathbf{x}} && \text{Multiply by the inverse of } R^T.
\end{aligned}$$

■

The formula $R\vec{\mathbf{x}} = Q^T\vec{\mathbf{b}}$ can be solved by back-substitution, which accounts for its popularity in numerical solution of least squares problems.

Theorem 9.29 (Spectral Theorem)
Let A be a given $n \times n$ real matrix. Then $A = QDQ^{-1}$ with Q orthogonal and D diagonal if and only if $A^T = A$.

Proof: Requirement Q *is orthogonal* means that the columns of Q are **orthonormal** and $n \times n$. The equation $A = A^T$ means A is **symmetric**.

Assume first that $A = QDQ^{-1}$ with $Q = Q^T$ orthogonal ($Q^TQ = I$) and D diagonal. Then $Q^T = Q = Q^{-1}$. This implies $A^T = (QDQ^{-1})^T = (Q^{-1})^T D^T Q^T = QDQ^{-1} = A$.

Conversely, assume $A^T = A$. Then the eigenvalues of A are real and eigenvectors corresponding to distinct eigenvalues are orthogonal. The proof proceeds by induction on the dimension n of the $n \times n$ matrix A.

For $n = 1$, let Q be the 1×1 identity matrix. Then Q is orthogonal and $AQ = QD$ where D is 1×1 diagonal.

Assume the decomposition $AQ = QD$ for dimension n. Let's prove it for A of dimension $n + 1$. Choose a real eigenvalue λ of A and eigenvector $\vec{\mathbf{v}}_1$ with $\|\vec{\mathbf{v}}_1\| = 1$. Complete a basis $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_{n+1}$ of $\mathcal{R}^{n+1}$. By Gram-Schmidt, we assume as well that this basis is orthonormal. Define $P = \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_{n+1} \right\rangle$. Then P is square, orthogonal and satisfies $P^T = P^{-1}$. Define $B = P^{-1}AP$. Then B is symmetric ($B^T = B$) and $\mathbf{col}(B, 1) = \lambda\, \mathbf{col}(I, 1)$. These facts imply that B is a block matrix

$$B = \left(\begin{array}{c|c} \lambda & 0 \\ \hline 0 & C \end{array} \right)$$

where C is symmetric ($C^T = C$). The induction hypothesis applies to C to obtain the existence of an orthogonal matrix Q_1 such that $CQ_1 = Q_1D_1$ for some diagonal matrix

D_1. Define block diagonal matrix D, block matrix W and square matrix Q as follows:

$$\begin{aligned} D &= \left(\begin{array}{c|c} \lambda & 0 \\ \hline 0 & D_1 \end{array}\right), \\ W &= \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1 \end{array}\right), \\ Q &= PW. \end{aligned}$$

Then Q is the product of two orthogonal matrices, which makes Q orthogonal. Compute

$$W^{-1}BW = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1^{-1} \end{array}\right)\left(\begin{array}{c|c} \lambda & 0 \\ \hline 0 & C \end{array}\right)\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1 \end{array}\right) = \left(\begin{array}{c|c} \lambda & 0 \\ \hline 0 & D_1 \end{array}\right).$$

Then $Q^{-1}AQ = W^{-1}P^{-1}APW = W^{-1}BW = D$. This completes the induction, ending the proof of the theorem. ■

Spectral Theorem Consequence: The eigenpair equation $AP = PD$ with $A \neq A^T$ (A not symmetric) cannot be converted to $AQ = QD$ with Q orthogonal.

Theorem 9.30 (Schur's Theorem)
Given any real $n \times n$ matrix A, possibly non-symmetric, there is an upper triangular matrix T, whose diagonal entries are the eigenvalues of A, and a complex matrix Q satisfying $\overline{Q}^T = Q^{-1}$ (Q is unitary), such that

$$AQ = QT.$$

If $A = A^T$, then Q is real orthogonal ($Q^T = Q$).

Schur's theorem can be proved by induction, following the induction proof of Jordan's theorem, or the induction proof of the Spectral Theorem. The result can be used to prove the Spectral Theorem in two steps. Indeed, Schur's Theorem implies Q is real, T equals its transpose, and T is triangular. Then T must equal a diagonal matrix D.

Theorem 9.31 (Eigenpairs of a Symmetric A)
Let A be a symmetric $n \times n$ real matrix. Then A has n eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, …, $(\lambda_n, \vec{\mathbf{v}}_n)$, with independent eigenvectors $\vec{\mathbf{v}}_1$, …, $\vec{\mathbf{v}}_n$.

Proof: Apply the Spectral Theorem 9.29, page 731, to prove the existence of an orthogonal matrix Q and a diagonal matrix D such that $AQ = QD$. The diagonal entries of D are the eigenvalues of A, in some order. For a diagonal entry λ of D appearing in row j, the relation $A\,\mathbf{col}(Q, j) = \lambda\,\mathbf{col}(Q, j)$ holds, which implies that A has n eigenpairs. The eigenvectors are the columns of Q, which are orthogonal and hence independent. ■

Theorem 9.32 (Diagonalization of Symmetric A)
Let A be a symmetric $n \times n$ real matrix. Then A has n eigenpairs $(\lambda_i, \vec{\mathbf{x}}_i)$. Assume the eigenvalues are listed with duplicates grouped together. For each distinct eigenvalue

λ, replace its eigenvectors by orthonormal eigenvectors, using the Gram-Schmidt process. Let $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n$ be the orthonormal vectors so obtained and define

$$Q = \left\langle \vec{\mathbf{u}}_1 | \cdots | \vec{\mathbf{u}}_n \right\rangle \quad D = \mathbf{diag}(\lambda_1, \ldots, \lambda_n).$$

Then Q is an orthogonal matrix and $AQ = QD$.

Proof: Theorem 9.31 justifies the eigenanalysis result. Already, eigenpairs corresponding to distinct eigenvalues are orthogonal. Within the set of eigenpairs with the same eigenvalue λ, the Gram-Schmidt process produces a replacement basis of orthonormal eigenvectors. Then the union of all the eigenvectors is orthonormal. The process described here does not disturb the ordering of eigenpairs, because it only replaces an eigenvector. ■

The Singular Value Decomposition

Coined the **SVD** in literature, the singular value decomposition $A = U\Sigma V^T$ has some interesting algebraic properties and it conveys important geometrical and theoretical insights about linear transformations.

Data science uses the SVD as a compression algorithm. Machine vision uses the SVD to find the nearest orthogonal matrix to A. Linear regression modeling uses the SVD to find the pseudo-inverse. Signal processing noise reduction and image processing size reduction use the SVD. Latent semantic indexing in natural-language text processing uses the SVD to identify patterns in unstructured text. Geometric interpretations of the SVD appear in a later subsection.

Theorem 9.33 (Positive Eigenvalues of A^TA)
Given an $m \times n$ real matrix A, then A^TA is a real symmetric matrix whose eigenpairs $(\lambda, \vec{\mathbf{v}})$ satisfy[19]

$$\lambda = \frac{\|A\vec{\mathbf{v}}\|^2}{\|\vec{\mathbf{v}}\|^2} \geq 0. \tag{8}$$

Proof: Symmetry follows from $(A^TA)^T = A^T(A^T)^T = A^TA$. An eigenpair $(\lambda, \vec{\mathbf{v}})$ satisfies $\lambda\vec{\mathbf{v}}^T\vec{\mathbf{v}} = \vec{\mathbf{v}}^TA^TA\vec{\mathbf{v}} = (A\vec{\mathbf{v}})^T(A\vec{\mathbf{v}}) = \|A\vec{\mathbf{v}}\|^2$, hence (8). ■

Definition 9.13 (Singular Values of A)
Let the real symmetric matrix A^TA have real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$. The numbers

$$\sigma_k = \sqrt{\lambda_k}, \quad 1 \leq k \leq n,$$

are called the **singular values** of the matrix A. The ordering of the singular values is always with decreasing magnitude.

[19] Can a real symmetric matrix have negative or complex eigenvalues?
The answer is **NO**.

Theorem 9.34 (Orthonormal Set $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_m$)
Let the real symmetric matrix A^TA have real eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$ and corresponding orthonormal eigenvectors $\vec{\mathbf{v}}_1,\ldots,\vec{\mathbf{v}}_n$, obtained by the Gram-Schmidt process. Define the vectors

$$\vec{\mathbf{u}}_1 = \frac{1}{\sigma_1} A\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{u}}_r = \frac{1}{\sigma_r} A\vec{\mathbf{v}}_r.$$

Because $\|A\vec{\mathbf{v}}_k\| = \sigma_k$, then $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_r\}$ is orthonormal. Gram-Schmidt can extend this set to an orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_m\}$ of $\mathcal{R}^m$.

Proof of Theorem 9.34: Compute $\|\vec{\mathbf{u}}_k\|^2 = \vec{\mathbf{v}}_k \cdot (A^TA\vec{\mathbf{v}}_k)/\lambda_k = \|\vec{\mathbf{v}}_k\|^2 = 1$, because $\{\vec{\mathbf{v}}_k\}_{k=1}^n$ is an orthonormal set. Then the vectors $\vec{\mathbf{u}}_k$ are nonzero. Given $i \ne j$, then $\sigma_i\sigma_j\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j = (A\vec{\mathbf{v}}_i)^T(A\vec{\mathbf{v}}_j) = \lambda_j\vec{\mathbf{v}}_i^T\vec{\mathbf{v}}_j = 0$, showing that the vectors $\vec{\mathbf{u}}_k$ are orthogonal.

The extension of the $\vec{\mathbf{u}}_k$ to an orthonormal basis of $\mathcal{R}^m$ is not unique, because it depends upon a choice of independent spanning vectors $\vec{\mathbf{y}}_{r+1}, \ldots, \vec{\mathbf{y}}_m$ for the set $\{\vec{\mathbf{x}} : \vec{\mathbf{x}} \cdot \vec{\mathbf{u}}_k = 0, \quad 1 \le k \le r\}$. Once selected, Gram-Schmidt is applied to $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_r, \vec{\mathbf{y}}_{r+1}, \ldots, \vec{\mathbf{y}}_m$ to obtain the desired orthonormal basis.

Computer algebra systems can compute the orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_m\}$ of $\mathcal{R}^m$ by appending all columns of the identity matrix to columns $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_r\}$ to define an augmented matrix Z. Then the reduced row echelon form of Z identifies the pivot columns of Z. The first r pivot columns are $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_r$. The remaining pivot columns are columns of the identity. Apply Gram-Schmidt to the pivot columns to obtain the orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_m\}$. ■

Theorem 9.35 (The Singular Value Decomposition (svd))
Let A be a given real $m \times n$ matrix. Let $(\lambda_1, \vec{\mathbf{v}}_1),\ldots,(\lambda_n, \vec{\mathbf{v}}_n)$ be a set of orthonormal eigenpairs for A^TA such that $\sigma_k = \sqrt{\lambda_k}$ $(1 \le k \le r)$ defines the positive singular values of A and $\lambda_k = 0$ for $r < k \le n$. Complete $\vec{\mathbf{u}}_1 = (1/\sigma_1)A\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{u}}_r = (1/\sigma_r)A\vec{\mathbf{v}}_r$ to an orthonormal basis $\{\vec{\mathbf{u}}_k\}_{k=1}^m$for $\mathcal{R}^m$. Define

$$\begin{aligned} U &= \left\langle \vec{\mathbf{u}}_1 | \cdots | \vec{\mathbf{u}}_m \right\rangle, \quad \Sigma = \left(\begin{array}{c|c} \mathbf{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ \hline 0 & 0 \end{array} \right), \\ V &= \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_n \right\rangle. \end{aligned}$$

Then the columns of U and V are orthonormal and

$$\begin{aligned} A &= U\Sigma V^T \\ &= \sigma_1 \vec{\mathbf{u}}_1 \vec{\mathbf{v}}_1^T + \cdots + \sigma_r \vec{\mathbf{u}}_r \vec{\mathbf{v}}_r^T \\ &= A(\vec{\mathbf{v}}_1)\vec{\mathbf{v}}_1^T + \cdots + A(\vec{\mathbf{v}}_r)\vec{\mathbf{v}}_r^T \end{aligned}$$

Proof of Theorem 9.35: The product of U and Σ is the $m \times n$ matrix

$$\begin{aligned} U\Sigma &= \left\langle \sigma_1\vec{\mathbf{u}}_1 | \cdots | \sigma_r \vec{\mathbf{u}}_r | \vec{\mathbf{0}} | \cdots | \vec{\mathbf{0}} \right\rangle \\ &= \left\langle A(\vec{\mathbf{v}}_1) | \cdots | A(\vec{\mathbf{v}}_r) | \vec{\mathbf{0}} | \cdots | \vec{\mathbf{0}} \right\rangle. \end{aligned}$$

Let $\vec{\mathbf{v}}$ be any vector in $\mathcal{R}^n$. It will be shown that $U\Sigma V^T\vec{\mathbf{v}}$, $\sum_{k=1}^r A(\vec{\mathbf{v}}_k)(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})$ and $A\vec{\mathbf{v}}$ are the same column vector. We have the equalities

$$\begin{aligned} U\Sigma V^T\vec{\mathbf{v}} &= U\Sigma\begin{pmatrix}\vec{\mathbf{v}}_1^T\vec{\mathbf{v}}\\ \vdots\\ \vec{\mathbf{v}}_n^T\vec{\mathbf{v}}\end{pmatrix}\\ &= \left\langle A(\vec{\mathbf{v}}_1)|\cdots|A(\vec{\mathbf{v}}_r)|\vec{\mathbf{0}}|\cdots|\vec{\mathbf{0}}\right\rangle\begin{pmatrix}\vec{\mathbf{v}}_1^T\vec{\mathbf{v}}\\ \vdots\\ \vec{\mathbf{v}}_n^T\vec{\mathbf{v}}\end{pmatrix}\\ &= \sum_{k=1}^r(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})A(\vec{\mathbf{v}}_k).\end{aligned}$$

Because $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_n$ is an orthonormal basis of $\mathcal{R}^n$, then $\vec{\mathbf{v}} = \sum_{k=1}^n(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})\vec{\mathbf{v}}_k$. Additionally, $A(\vec{\mathbf{v}}_k) = \vec{\mathbf{0}}$ for $r < k \le n$ implies

$$\begin{aligned} A\vec{\mathbf{v}} &= A\left(\sum_{k=1}^n(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})\vec{\mathbf{v}}_k\right)\\ &= \sum_{k=1}^r(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})A(\vec{\mathbf{v}}_k)\end{aligned}$$

Then $A\vec{\mathbf{v}} = U\Sigma V^T\vec{\mathbf{v}} = \sum_{k=1}^r A(\vec{\mathbf{v}}_k)(\vec{\mathbf{v}}_k^T\vec{\mathbf{v}})$, which proves the theorem. ■

Singular Values and Geometry

Discussed here is how to interpret singular values geometrically, especially in low dimensions 2 and 3. Conics will be reviewed, adopting the viewpoint of eigenanalysis.

Standard Equation of an Ellipse

Calculus courses consider ellipse equations like

$$85x^2 - 60xy + 40y^2 = 2500$$

and discuss removal of the cross term $-60xy$. The objective is to obtain a **standard** ellipse equation

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1.$$

We re-visit this old problem from a different point of view, and in the derivation establish a connection between the ellipse equation, the symmetric matrix A^TA, and the singular values of A.

Example 9.16 (Image of the Unit Circle)
Let $A = \begin{pmatrix}-2 & 6\\ 6 & 7\end{pmatrix}$.
Verify that the invertible matrix A maps the unit circle into the ellipse

$$85x^2 - 60xy + 40y^2 = 2500.$$

Solution: The unit circle has parameterization $\theta \to (\cos\theta, \sin\theta)$, $0 \le \theta \le 2\pi$.

The unit circle is mapped by matrix A via the set of dual relations

$$\begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} = A^{-1} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The Pythagorean identity $\cos^2\theta + \sin^2\theta = 1$ used on the vector norm of second relation implies

$$85x^2 - 60xy + 40y^2 = 2500.$$

Example 9.17 (Removing the xy-Term in an Ellipse Equation)
After a rotation $(x, y) \to (X, Y)$ to remove the xy-term in

$$85x^2 - 60xy + 40y^2 = 2500,$$

verify that the ellipse equation in the new XY-coordinates is

$$\frac{X^2}{25} + \frac{Y^2}{100} = 1.$$

Solution: The xy-term removal is accomplished by a change of variables $(x, y) \to (X, Y)$ which transforms the ellipse equation $85x^2 - 60xy + 40y^2 = 2500$ into the ellipse equation $100X^2 + 25Y^2 = 2500$, details below. It's standard form is obtained by dividing by 2500, to give

$$\frac{X^2}{25} + \frac{Y^2}{100} = 1.$$

Analytic geometry says that the semi-axis lengths are $\sqrt{25} = 5$ and $\sqrt{100} = 10$.

In previous discussions of the ellipse, the equation $85x^2 - 60xy + 40y^2 = 2500$ was represented by the vector-matrix identity

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 85 & -30 \\ -30 & 40 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 2500.$$

The program used earlier to remove the xy-term was to diagonalize the coefficient matrix $B = \begin{pmatrix} 85 & -30 \\ -30 & 40 \end{pmatrix}$ by calculating the eigenpairs of B:

$$\left(100, \begin{pmatrix} -2 \\ 1 \end{pmatrix}\right), \quad \left(25, \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right).$$

Because B is symmetric, then the eigenvectors are orthogonal. The eigenpairs above are replaced by unitized pairs:

$$\left(100, \frac{1}{\sqrt{5}}\begin{pmatrix} -2 \\ 1 \end{pmatrix}\right), \quad \left(25, \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right).$$

Then the diagonalization theory for B can be written as

$$BQ = QD, \quad Q = \frac{1}{\sqrt{5}}\begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}, \quad D = \begin{pmatrix} 100 & 0 \\ 0 & 25 \end{pmatrix}.$$

The single change of variables

$$\begin{pmatrix} x \\ y \end{pmatrix} = Q \begin{pmatrix} X \\ Y \end{pmatrix}$$

then transforms the ellipse equation $85x^2 - 60xy + 40y^2 = 2500$ into $100X^2 + 25Y^2 = 2500$ as follows:

$$\begin{array}{ll}
85x^2 - 60xy + 40y^2 = 2500 & \text{Ellipse equation.} \\
\vec{\mathbf{u}}^T B \vec{\mathbf{u}} = 2500 & \text{Where } B = \begin{pmatrix} 85 & -30 \\ -30 & 40 \end{pmatrix} \text{ and } \vec{\mathbf{u}} = \begin{pmatrix} x \\ y \end{pmatrix}. \\
(Q\vec{\mathbf{w}})^T B (Q\vec{\mathbf{w}}) = 2500 & \text{Change } \vec{\mathbf{u}} = Q\vec{\mathbf{w}}\text{, where } \vec{\mathbf{w}} = \begin{pmatrix} X \\ Y \end{pmatrix}. \\
\vec{\mathbf{w}}^T (Q^T B Q) \vec{\mathbf{w}} = 2500 & \text{Expand, ready to use } BQ = QD. \\
\vec{\mathbf{w}}^T (D\vec{\mathbf{w}}) = 2500 & \text{Because } D = Q^{-1}BQ \text{ and } Q^{-1} = Q^T. \\
100X^2 + 25Y^2 = 2500 & \text{Expand } \vec{\mathbf{w}}^T D \vec{\mathbf{w}}.
\end{array}$$

Rotations, Reflections and Scaling

The 2×2 singular value decomposition $A = U\Sigma V^T$ can be used to decompose the change of variables $(x, y) \to (X, Y)$ into three distinct changes of variables, each with a geometrical meaning:

$$(x, y) \longrightarrow (x_1, y_1) \longrightarrow (x_2, y_2) \longrightarrow (X, Y).$$

Table 5. Three Changes of Variable

Domain	Equation	Image	Meaning
Circle 1	$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = V^T \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$	Circle 2	Proper Rotation
Circle 2	$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \Sigma \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$	Ellipse 1	Scale axes
Ellipse 1	$\begin{pmatrix} X \\ Y \end{pmatrix} = U \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$	Ellipse 2	Improper Rotation

Proper Rotation. Matrix $R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ satisfies $R^T R = I$ and $|R| = 1$, called a **proper rotation**. The rotation is clockwise about the origin, following use in computer graphics. Replace θ by $-\theta$ for a counterclockwise rotation about the origin.

Improper Rotation. Matrix $R = \begin{pmatrix} 0.936 & 0.352 \\ 0.352 & -0.936 \end{pmatrix}$ is orthogonal with $|R| = -1$, called an **improper rotation**. It represents a reflection, which inverts orientation. Reference:

https://en.wikipedia.org/wiki/Rotation_matrix

Geometry

Figure 7 provides a geometrical interpretation for the singular value decomposition

$$A = U\Sigma V^T.$$

For illustration, the matrix A is assumed 2 × 2 and invertible.

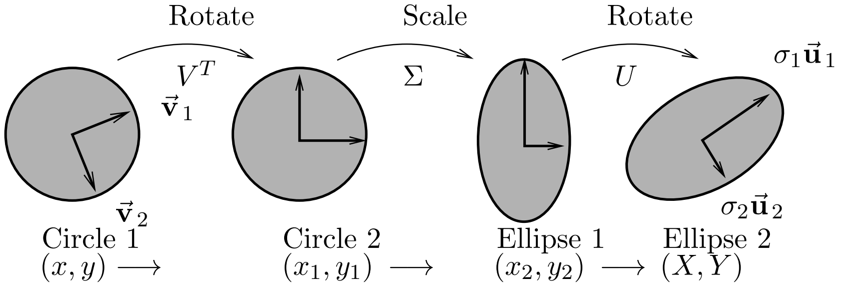


Figure 7. Mapping the unit circle.

- Invertible matrix A maps Circle 1 into Ellipse 2.
- Orthonormal vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are mapped by matrix $A = U\Sigma V^T$ into orthogonal vectors $A\vec{\mathbf{v}}_1 = \sigma_1\vec{\mathbf{u}}_1$, $A\vec{\mathbf{v}}_2 = \sigma_2\vec{\mathbf{u}}_2$, which are exactly the semi-axes vectors of Ellipse 2.
- The semi-axis lengths of Ellipse 2 equal the singular values σ_1, σ_2 of matrix A.
- The semi-axis directions of Ellipse 2 are equal to the basis vectors $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$.
- The process is a rotation $(x, y) \to (x_1, y_1)$, followed by an axis-scaling $(x_1, y_1) \to (x_2, y_2)$, followed by $(x_2, y_2) \to (X, Y)$, a rotation.

Example 9.18 (Mapping and the SVD)
The singular value decomposition $A = U\Sigma V^T$ for $A = \begin{pmatrix} -2 & 6 \\ 6 & 7 \end{pmatrix}$ is given by

$$U = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix}, \quad V = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}.$$

- Invertible matrix $A = \begin{pmatrix} -2 & 6 \\ 6 & 7 \end{pmatrix}$ maps the unit circle into an ellipse.
- The columns of V are orthonormal vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, computed as eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$ of A^TA, ordered by $\lambda_1 \geq \lambda_2$.

$$\left(100, \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right), \quad \left(25, \frac{1}{\sqrt{5}}\begin{pmatrix} -2 \\ 1 \end{pmatrix}\right).$$

- The singular values are $\sigma_1 = \sqrt{\lambda_1} = 10$, $\sigma_2 = \sqrt{\lambda_2} = 5$.
- The image of $\vec{\mathbf{v}}_1$ is $A\vec{\mathbf{v}}_1 = U\Sigma V^T \vec{\mathbf{v}}_1 = U \begin{pmatrix} \sigma_1 \\ 0 \end{pmatrix} = \sigma_1 \vec{\mathbf{u}}_1$.
- The image of $\vec{\mathbf{v}}_2$ is $A\vec{\mathbf{v}}_2 = U\Sigma V^T \vec{\mathbf{v}}_2 = U \begin{pmatrix} 0 \\ \sigma_2 \end{pmatrix} = \sigma_2 \vec{\mathbf{u}}_2$.

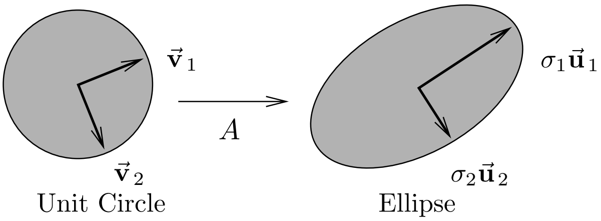


Figure 8.
Mapping the unit circle into an ellipse.

The Four Fundamental Subspaces

The subspaces appearing in the Fundamental Theorem of Linear Algebra are called the **Four Fundamental Subspaces**. They are:

Subspace	**Notation**
Row Space of A	**Image** (A^T)
Nullspace of A	**kernel**(A)
Column Space of A	**Image**(A)
Nullspace of A^T	**kernel** (A^T)

The singular value decomposition $A = U\Sigma V^T$ computes orthonormal bases for the row and column spaces of of A. In the table below, symbol $r =$ **rank**(A). Matrix A is assumed $m \times n$, which implies A maps $\mathcal{R}^n$ into $\mathcal{R}^m$.

Table 6. Four Fundamental Subspaces and the SVD

Orthonormal Basis	**Subspace**	**Name**
First r columns of U $(m \times n)$	**Image**(A)	Column Space of A
Last $n - r$ columns of U	**kernel** (A^T)	Nullspace of A^T
First r columns of V $(n \times m)$	**Image** (A^T)	Row Space of A
Last $m - r$ columns of V	**kernel**(A)	Nullspace of A

Table 7. Fundamental Subspaces by Columns of U and V

$$m \times n \quad A = U\,\Sigma\,V^T \quad \text{Singular Value Decomposition}$$

$$m \times m \quad U = \left\|\begin{array}{c|c} \textbf{colspace}(A) & \textbf{nullspace}(A^T) \\ r \;\; \text{columns} & m-r \;\; \text{columns} \end{array}\right\|$$

$$m \times n \quad \Sigma = \left\|\begin{array}{c|c} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & \sigma_r \end{pmatrix} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right\|$$

$$n \times n \quad V = \left\|\begin{array}{c|c} \textbf{rowspace}(A) & \textbf{nullspace}(A) \\ r \;\; \text{columns} & n-r \;\; \text{columns} \end{array}\right\|$$

A Change of Basis Interpretation of the SVD

The singular value decomposition can be described as follows:

> For every $m \times n$ matrix A of rank r, orthonormal bases
>
> $$\{\vec{\mathbf{v}}_i\}_{i=1}^n \text{ and } \{\vec{\mathbf{u}}_j\}_{j=1}^m$$
>
> can be constructed such that
>
> - Matrix A maps basis vectors $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_r$ to nonnegative multiples of basis vectors $\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_r$, respectively.
> - The $n - r$ left-over basis vectors $\vec{\mathbf{v}}_{r+1}, \ldots \vec{\mathbf{v}}_n$ map by A into the zero vector.
> - With respect to these bases, matrix A is represented by a real diagonal matrix Σ with non-negative entries.

Proofs, Methods and Details

Proof of Theorem 9.14, Jordan's Theorem:
Proceed by induction on the dimension n of A. For $n = 1$ there is nothing to prove. Assume the result for dimension n. Assume A is $(n+1) \times (n+1)$. To prove the induction step, choose an eigenpair $(\lambda_1, \vec{\mathbf{v}}_1)$ of A with $\vec{\mathbf{v}}_1 \neq \vec{\mathbf{0}}$. Complete a basis $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_{n+1}$ for

$\mathcal{R}^{n+1}$ and define $V = \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_{n+1} \right\rangle$. Then $V^{-1}AV = \left(\begin{array}{c|c} \lambda_1 & B \\ \hline \vec{\mathbf{0}} & A_1 \end{array}\right)$ for some matrices B and A_1. The induction hypothesis implies there is an invertible $n \times n$ matrix P_1 and an upper triangular matrix T_1 such that $A_1 = P_1 T_1 P_1^{-1}$. Let $R = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & P_1 \end{array}\right)$ and $T = \left(\begin{array}{c|c} \lambda_1 & BT_1 \\ \hline 0 & T_1 \end{array}\right)$. Then T is upper triangular and $(V^{-1}AV)R = RT$, which implies $A = PTP^{-1}$ for $P = VR$. The induction is complete. ■

Proof of Theorem 9.16, Cayley-Hamilton:
An algebraic proof was given in Chapter 5 Section 3. It depended on the adjugate identity $\mathbf{adj}(A)\,A = A\,\mathbf{adj}(A) = |A|I$. Below is a different proof which suggests how the theorem might have been discovered.

If A is diagonalizable, $AP = P\,\mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, then the proof is obtained from the expansion

$$A^j = P\,\mathbf{diag}(\lambda_1^j, \ldots, \lambda_n^j)P^{-1},$$

because summing across this identity leads to

$$\begin{array}{rcl} p(A) & = & \sum_{j=0}^{n} c_j A^j \\ & = & P\left(\sum_{j=0}^{n} c_j\,\mathbf{diag}(\lambda_1^j, \ldots, \lambda_n^j)\right)P^{-1} \\ & = & P\,\mathbf{diag}(p(\lambda_1), \ldots, p(\lambda_n))P^{-1} \\ & = & P\,\mathbf{diag}(0, \ldots, 0)P^{-1} \\ & = & \mathbf{0}. \end{array}$$

If A is not diagonalizable, then this proof fails. To handle the general case, apply **Jordan's theorem** 9.14 to write $A = PTP^{-1}$ where T is *upper triangular* (instead of *diagonal*) and the not necessarily distinct eigenvalues λ_1, $\ldots$, λ_n of A appear on the diagonal of T. Define

$$A_\epsilon = P(T + \epsilon\,\mathbf{diag}(1, 2, \ldots, n))P^{-1}.$$

For small $\epsilon > 0$, the matrix A_ϵ has distinct eigenvalues $\lambda_j + j\epsilon$, $1 \le j \le n$. Then the diagonalizable case implies that A_ϵ satisfies its characteristic equation. Let $p_\epsilon(\lambda) = \det(A_\epsilon - \lambda I)$. Use $\mathbf{0} = \lim_{\epsilon \to 0} p_\epsilon(A_\epsilon) = p(A)$ to complete the proof.

Proof of Theorem 9.17, orthogonality: Compute $\alpha\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = (A\vec{\mathbf{x}})^T\vec{\mathbf{y}} = \vec{\mathbf{x}}^T A^T \vec{\mathbf{y}} = \vec{\mathbf{x}}^T A\vec{\mathbf{y}}$. Analogously, $\beta\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = \vec{\mathbf{x}}^T A\vec{\mathbf{y}}$. Subtract the relations, then $(\alpha - \beta)\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = 0$. Because $\alpha \ne \beta$, then $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = 0$. ■

Proof of Theorem 9.18, real eigenvalues: The second statement is due to the fundamental theorem of algebra. To prove the eigenvalues are real, it suffices to prove $\lambda = \overline{\lambda}$ when $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ with $\vec{\mathbf{v}} \ne \vec{\mathbf{0}}$. A complex conjugate is computed by replacing i by $-i$. Conjugates of vectors and matrices are found componentwise. Assume that $\vec{\mathbf{v}}$ may have complex entries. Because A is real, then $\overline{A} = A$. Take the complex conjugate across $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ to obtain $A\overline{\vec{\mathbf{v}}} = \overline{\lambda}\overline{\vec{\mathbf{v}}}$. Transpose $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ to obtain $\vec{\mathbf{v}}^T A^T = \lambda\vec{\mathbf{v}}^T$ and then conclude $\vec{\mathbf{v}}^T A = \lambda\vec{\mathbf{v}}^T$ from $A^T = A$. Multiply this equation by $\overline{\vec{\mathbf{v}}}$ on the right to obtain $\vec{\mathbf{v}}^T A\overline{\vec{\mathbf{v}}} = \lambda\vec{\mathbf{v}}^T\overline{\vec{\mathbf{v}}}$. Then multiply $A\overline{\vec{\mathbf{v}}} = \overline{\lambda}\overline{\vec{\mathbf{v}}}$ by $\vec{\mathbf{v}}^T$ on the left to obtain $\vec{\mathbf{v}}^T A\overline{\vec{\mathbf{v}}} = \overline{\lambda}\vec{\mathbf{v}}^T\overline{\vec{\mathbf{v}}}$. The result:

$$\lambda\vec{\mathbf{v}}^T\overline{\vec{\mathbf{v}}} = \overline{\lambda}\vec{\mathbf{v}}^T\overline{\vec{\mathbf{v}}}.$$

Because $\vec{\mathbf{v}}^T\overline{\vec{\mathbf{v}}} = \sum_{j=1}^{n} |v_j|^2 > 0$, then it cancels: $\lambda = \overline{\lambda}$ and λ is real. ■

Proof of Theorem 9.19, Gram-Schmidt relations: Induction will be applied on k to show that $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_k$ are nonzero and orthogonal. If $k = 1$, then there is just one nonzero vector constructed $\vec{\mathbf{y}}_1 = \vec{\mathbf{x}}_1$. Orthogonality for $k = 1$ is not discussed because there are no pairs to test. Assume the result holds for $k - 1$ vectors. Let's verify that it holds for k vectors, $k > 1$. Assume orthogonality $\vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_j = 0$ for $i \neq j$ and $\vec{\mathbf{y}}_i \neq \vec{\mathbf{0}}$ for $1 \leq i, j \leq k - 1$. It remains to test $\vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_k = 0$ for $1 \leq i \leq k - 1$ and $\vec{\mathbf{y}}_k \neq \vec{\mathbf{0}}$. The test depends upon the identity

$$\vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_k = \vec{\mathbf{y}}_i \cdot \vec{\mathbf{x}}_k - \sum_{j=1}^{k-1} \frac{\vec{\mathbf{x}}_k \cdot \vec{\mathbf{y}}_j}{\vec{\mathbf{y}}_j \cdot \vec{\mathbf{y}}_j} \vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_j,$$

which is obtained from the formula for $\vec{\mathbf{y}}_k$ by taking the dot product with $\vec{\mathbf{y}}_i$. In the identity, $\vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_j = 0$ by the induction hypothesis for $1 \leq j \leq k - 1$ and $j \neq i$. Therefore, the summation in the identity contains just the term for index $j = i$, and the contribution is $\vec{\mathbf{y}}_i \cdot \vec{\mathbf{x}}_k$. This contribution cancels the leading term on the right in the identity, resulting in the orthogonality relation $\vec{\mathbf{y}}_i \cdot \vec{\mathbf{y}}_k = 0$. If $\vec{\mathbf{y}}_k = \vec{\mathbf{0}}$, then $\vec{\mathbf{x}}_k$ is a linear combination of $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_{k-1}$. But each $\vec{\mathbf{y}}_j$ is a linear combination of $\{\vec{\mathbf{x}}_i\}_{i=1}^{j}$, therefore $\vec{\mathbf{y}}_k = \vec{\mathbf{0}}$ implies $\vec{\mathbf{x}}_k$ is a linear combination of $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{k-1}$, a contradiction to the independence of $\{\vec{\mathbf{x}}_i\}_{i=1}^{k}$. ■

Proof of Theorem 9.20, Formula $\mathbf{Proj}_V(\vec{\mathbf{x}})$ is Well-Defined:

Suppose that $\{\vec{\mathbf{w}}_j\}_{j=1}^k$ is another orthonormal basis of V. Define $\vec{\mathbf{u}} = \sum_{i=1}^k (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}})\vec{\mathbf{u}}_j$ and $\vec{\mathbf{w}} = \sum_{j=1}^k (\vec{\mathbf{w}}_j \cdot \vec{\mathbf{x}})\vec{\mathbf{w}}_j$. It will be established that $\vec{\mathbf{u}} = \vec{\mathbf{w}}$, which justifies that the projection formula is independent of basis. First, two lemmas.

Lemma 9.3 (Orthonormal Basis Expansion)
Let $\{\vec{\mathbf{v}}_j\}_{j=1}^k$ be an orthonormal basis of a subspace V in $\mathcal{R}^n$. Then each vector $\vec{\mathbf{v}}$ in V is represented as

$$\vec{\mathbf{v}} = \sum_{j=1}^{k} (\vec{\mathbf{v}}_j \cdot \vec{\mathbf{v}})\vec{\mathbf{v}}_j.$$

Proof: First, $\vec{\mathbf{v}}$ has a basis expansion $\vec{\mathbf{v}} = \sum_{j=1}^k c_j \vec{\mathbf{v}}_j$ for some constants $c_1, \ldots, c_k$. Take the inner product of this equation with vector $\vec{\mathbf{v}}_i$ to prove that $c_i = \vec{\mathbf{v}}_i \cdot \vec{\mathbf{v}}$, hence the claimed expansion is proved.

Lemma 9.4 (Orthogonality) Let $\{\vec{\mathbf{u}}_i\}_{i=1}^k$ be an orthonormal basis of a subspace V in $\mathcal{R}^n$. Let $\vec{\mathbf{x}}$ be any vector in $\mathcal{R}^n$ and define $\vec{\mathbf{u}} = \sum_{i=1}^k (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}})\vec{\mathbf{u}}_i$. Then $\vec{\mathbf{y}} \cdot (\vec{\mathbf{x}} - \vec{\mathbf{u}}) = 0$ for all vectors $\vec{\mathbf{y}}$ in V.

Proof: The first lemma implies $\vec{\mathbf{u}}$ can be written a second way as a linear combination of $\vec{\mathbf{u}}_1, \ldots.\ \vec{\mathbf{u}}_k$. Independence implies equal basis coefficients, which gives $\vec{\mathbf{u}}_j \cdot \vec{\mathbf{u}} = \vec{\mathbf{u}}_j \cdot \vec{\mathbf{x}}$. Then $\vec{\mathbf{u}}_j \cdot (\vec{\mathbf{x}} - \vec{\mathbf{u}}) = 0$. Because $\vec{\mathbf{y}}$ is in V, then $\vec{\mathbf{y}} = \sum_{j=1}^k c_j \vec{\mathbf{u}}_j$, which implies $\vec{\mathbf{y}} \cdot (\vec{\mathbf{x}} - \vec{\mathbf{u}}) = \sum_{j=1}^k c_j\, \vec{\mathbf{u}}_j \cdot (\vec{\mathbf{x}} - \vec{\mathbf{u}}) = 0$. ■

Justification of $\vec{\mathbf{w}} = \vec{\mathbf{u}}$

The justification of Formula (1) is concluded here, showing that $\vec{\mathbf{w}} = \vec{\mathbf{u}}$.

$\vec{\mathbf{w}} = \sum_{j=1}^k (\vec{\mathbf{w}}_j \cdot \vec{\mathbf{x}})\vec{\mathbf{w}}_j$

$\quad = \sum_{j=1}^k (\vec{\mathbf{w}}_j \cdot \vec{\mathbf{u}})\vec{\mathbf{w}}_j$ — Because $\vec{\mathbf{w}}_j \cdot (\vec{\mathbf{x}} - \vec{\mathbf{u}}) = 0$ by the second lemma.

$$\begin{aligned}
&= \textstyle\sum_{j=1}^{k} \left(\vec{\mathbf{w}}_j \cdot \sum_{i=1}^{k}(\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}})\vec{\mathbf{u}}_i\right) \vec{\mathbf{w}}_j && \text{Definition of } \vec{\mathbf{u}}. \\
&= \textstyle\sum_{j=1}^{k} \sum_{i=1}^{k}(\vec{\mathbf{w}}_j \cdot \vec{\mathbf{u}}_i)(\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}})\vec{\mathbf{w}}_j && \text{Dot product properties.} \\
&= \textstyle\sum_{i=1}^{k} \left(\sum_{j=1}^{k}(\vec{\mathbf{w}}_j \cdot \vec{\mathbf{u}}_i)\vec{\mathbf{w}}_j\right) (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}}) && \text{Switch summations.} \\
&= \textstyle\sum_{i=1}^{k} \vec{\mathbf{u}}_i(\vec{\mathbf{u}}_i \cdot \vec{\mathbf{x}}) && \text{First lemma with } \vec{\mathbf{v}} = \vec{\mathbf{u}}_i. \\
&= \vec{\mathbf{u}} && \text{Definition of } \vec{\mathbf{u}}.
\end{aligned}$$

Proof of Theorem 9.21, Projection properties: Properties (a), (b) and (c) were proved in preceding lemmas. Details are outlined here, in case the lemmas were skipped.

(a): Write a basis expansion $\vec{\mathbf{v}} = \sum_{j=1}^{k} c_j \vec{\mathbf{u}}_j$ for some constants $c_1, \ldots, c_k$. Take the inner product of this equation with vector $\vec{\mathbf{u}}_i$ to prove that $c_i = \vec{\mathbf{u}}_i \cdot \vec{\mathbf{v}}$.

(b): Vector $\mathbf{Proj}_V(\vec{\mathbf{x}})$ is a linear combination of basis elements of V.

(c): Represent a given vector $\vec{\mathbf{v}}$ in V by the orthogonal expansion of$\vec{\mathbf{v}}$ from (a). Let's compute the dot product of $\vec{\mathbf{w}}$ and $\vec{\mathbf{v}}$:

$$\begin{aligned}
\vec{\mathbf{w}} \cdot \vec{\mathbf{v}} &= (\vec{\mathbf{x}} - \mathbf{Proj}_V(\vec{\mathbf{x}})) \cdot \vec{\mathbf{v}} \\
&= \vec{\mathbf{x}} \cdot \vec{\mathbf{v}} - \left(\sum_{j=1}^{k}(\vec{\mathbf{x}} \cdot \vec{\mathbf{u}}_j)\vec{\mathbf{u}}_j\right) \cdot \vec{\mathbf{v}} \\
&= \sum_{j=1}^{k}(\vec{\mathbf{u}}_j \cdot \vec{\mathbf{v}})(\vec{\mathbf{x}} \cdot \vec{\mathbf{u}}_j) - \sum_{j=1}^{k}(\vec{\mathbf{x}} \cdot \vec{\mathbf{u}}_j)(\vec{\mathbf{u}}_j \cdot \vec{\mathbf{v}}) \\
&= 0.
\end{aligned}$$

(d): Begin with the Pythagorean identity

$$\|\vec{\mathbf{a}}\|^2 + \|\vec{\mathbf{b}}\|^2 = \|\vec{\mathbf{a}} + \vec{\mathbf{b}}\|^2$$

valid exactly when $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}} = 0$ (a right triangle, $\theta = 90°$). Using an arbitrary $\vec{\mathbf{v}}$ in V, define $\vec{\mathbf{a}} = \mathbf{Proj}_V(\vec{\mathbf{x}}) - \vec{\mathbf{v}}$ and $\vec{\mathbf{b}} = \vec{\mathbf{x}} - \mathbf{Proj}_V(\vec{\mathbf{x}})$. By (b), vector $\vec{\mathbf{a}}$ is in V. Because of (c), then $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}} = 0$. This gives the identity

$$\|\,\mathbf{Proj}_V(\vec{\mathbf{x}}) - \vec{\mathbf{v}}\,\|^2 + \|\vec{\mathbf{x}} - \mathbf{Proj}_V(\vec{\mathbf{x}})\|^2 = \|\vec{\mathbf{x}} - \vec{\mathbf{v}}\,\|^2,$$

which establishes $\|\vec{\mathbf{x}} - \mathbf{Proj}_V(\vec{\mathbf{x}})\| < \|\vec{\mathbf{x}} - \vec{\mathbf{v}}\,\|$ except for the unique $\vec{\mathbf{v}}$ such that $\|\,\mathbf{Proj}_V(\vec{\mathbf{x}}) - \vec{\mathbf{v}}\,\| = 0$.

(e): Let $\vec{\mathbf{w}}$ be orthogonal to all vectors in V. Because the columns of A are in V, then $\vec{\mathbf{w}}$ is orthogonal to the columns of A, which are rows of A^T. Equation $A^T\vec{\mathbf{w}} = \vec{\mathbf{0}}$ means $\vec{\mathbf{w}}$ is orthogonal to the rows of A^T. ■

Proof of Theorem 9.24, Least Squares Solution:
Let $V = \mathbf{colspace}(A)$. Let $\vec{\mathbf{y}} = \mathbf{proj}_V(\vec{\mathbf{b}})$. Let $\vec{\mathbf{w}} = \vec{\mathbf{b}} - \vec{\mathbf{y}}$. Because $\vec{\mathbf{y}}$ is in the column space of A, then $\vec{\mathbf{y}} = A\vec{\mathbf{x}}^*$ for some $\vec{\mathbf{x}}^*$. By Theorem 9.21 (c), $\vec{\mathbf{w}} \cdot \vec{\mathbf{u}} = 0$ for every vector $\vec{\mathbf{u}}$ in V. This means $\vec{\mathbf{w}}\,\vec{\mathbf{u}}^T = 0$ for every column $\vec{\mathbf{u}}$ of A, which in turn means $A^T\vec{\mathbf{w}} = \vec{\mathbf{0}}$. Then $A^T(\vec{\mathbf{b}} - \vec{\mathbf{y}}) = \vec{\mathbf{0}}$ or equivalently $A^T\vec{\mathbf{b}} = A^TA\vec{\mathbf{x}}^*$. The Normal Equation has been verified for any $\vec{\mathbf{x}}^*$ such that $A\vec{\mathbf{x}}^* = \vec{\mathbf{y}} = \mathbf{proj}_V(\vec{\mathbf{b}})$. Theorem 9.21 (d) says that $\vec{\mathbf{x}} = \vec{\mathbf{x}}^*$ is a minimizer for $\|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|$ over all $\vec{\mathbf{x}}$. Then

$$\|A\vec{\mathbf{x}}^* - \vec{\mathbf{b}}\| = \min_{\vec{\mathbf{x}}} \|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|$$

■

Solution to Example 9.15, Height-Weight Best Fit:

The answer with 10 digits is $y = 61.2721865421106x - 39.0619559188439$, using assumptions made below for what is meant by **best fit**. A plot of the data and this straight line are in Figure 5.

Literature for the example might use the statistical term **simple linear regression**. The **regressors** in the example are unknowns m, b. The regression line $y = mx + b$ gives the expected weight y for height x, so y is the average or mean weight for a given height. Historically, **regression** abbreviates **regress back to the mean**, attributed to Sir Francis Galton (1822-1911) in work on genetics.

Linear algebraic equations in the unknowns m, b are discovered by inserting Table 4 data into $y = mx + b$, x=height, y=weight:

$$\begin{array}{llll}
52.21 = 1.47m + b & 53.12 = 1.50m + b & 54.48 = 1.52m + b & 55.84 = 1.55m + b \\
57.20 = 1.57m + b & 58.57 = 1.60m + b & 59.93 = 1.63m + b & 61.29 = 1.65m + b \\
63.11 = 1.68m + b & 64.47 = 1.70m + b & 66.28 = 1.73m + b & 68.10 = 1.75m + b \\
69.92 = 1.78m + b & 72.19 = 1.80m + b & 74.46 = 1.83m + b &
\end{array}$$

Define height vector $\vec{\mathbf{H}}$ and weight vector $\vec{\mathbf{W}}$ from Table 4, both vectors in $\mathcal{R}^{15}$. Let vector $\vec{\mathbf{O}}$ in $\mathcal{R}^{15}$ have all entries 1. Define augmented matrix $A = \left\langle \vec{\mathbf{H}} | \vec{\mathbf{O}} \right\rangle$. The fifteen linear algebraic equations become:

$$A \begin{pmatrix} m \\ b \end{pmatrix} = \vec{\mathbf{W}} \tag{9}$$

Among the *three possibilities* for a system of linear algebraic equations (Chapter 3 Section 1), system (9) has **no solution**. Terminology **best fit** has multiple possibilities, from which a single interpretation is isolated:

Best Fit

Find m, b to minimize the error between vectors $A \begin{pmatrix} m \\ b \end{pmatrix}$ and $\vec{\mathbf{W}}$.

The two vectors are the LHS and RHS of equation (9).

The answers $m = 61.2721865421106$, $b = -39.0619559188439$ are found by solving 2×2 matrix equation (12) on page 745. Details follow.

The idea for *solving* the unsolvable equation (9) is geometric: replace it with a solvable equation:

$$A \begin{pmatrix} m \\ b \end{pmatrix} = \vec{\mathbf{Z}} \tag{10}$$

Mystery vector $\vec{\mathbf{Z}}$ in (10) is the *unique* near point in $V = \mathbf{span}(\vec{\mathbf{H}}, \vec{\mathbf{O}})$ to $\vec{\mathbf{W}}$ given by near point Theorem 9.22.

Uniqueness of $\vec{\mathbf{Z}}$ means that the new equation $A \begin{pmatrix} m \\ b \end{pmatrix} = \vec{\mathbf{Z}}$ has a unique solution for m, b. The solution is efficiently found by multiplication of equation (10) by A^T:

$$A^T A \begin{pmatrix} m \\ b \end{pmatrix} = A^T \vec{\mathbf{Z}} = A^T \vec{\mathbf{W}}. \tag{11}$$

Equality $A^T \vec{\mathbf{Z}} = A^T \vec{\mathbf{W}}$ results from Theorem 9.21 (c): vector $\vec{\mathbf{w}} = \vec{\mathbf{W}} - \mathbf{Proj}_V(\vec{\mathbf{W}})$ is orthogonal to the columns of A and by Theorem 9.21 (e) $A^T \vec{\mathbf{w}} = \vec{\mathbf{0}}$. The simplified equation

$$A^TA\begin{pmatrix} m \\ b\end{pmatrix} = A^T\vec{\mathbf{W}} \tag{12}$$

is called the **normal equation** for unsolvable system (9).

The new system is a 2×2 system with a unique solution m, b given by matrix inversion:

$$\begin{aligned}
\begin{pmatrix} m \\ b\end{pmatrix} &= \left(A^TA\right)^{-1} A^T\vec{\mathbf{W}} \\
&= \begin{pmatrix} 41.0532 & 24.76 \\ 24.76 & 15 \end{pmatrix}^{-1} \begin{pmatrix} 1548.245 \\ 931.17 \end{pmatrix} \\
&= \begin{pmatrix} 61.2721865421106 \\ -39.0619559188439 \end{pmatrix}.
\end{aligned}$$

```
with(LinearAlgebra):# Maple check
H:=Vector([ 1.47,1.50,1.52,1.55,1.57,1.60,1.63,1.65,
            1.68,1.70,1.73,1.75,1.78,1.80,1.83]);
W:=Vector([ 52.21,53.12,54.48,55.84,57.20,58.57,59.93,61.29,
            63.11,64.47,66.28,68.10,69.92,72.19,74.46]);
ONE:=Vector([1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]);
A:=<H|ONE>;
LinearSolve(A,W);Rank(A);# fail expected
B:=A^+ . A; Z:=A^+ . W; (1/B) . Z;
```

Exercises 9.3

Diagonalization

Find the eigenpair packages P and D in the relation $AP = PD$.

1. $A = \begin{pmatrix} -4 & 2 \\ 0 & -1 \end{pmatrix}$

2. $A = \begin{pmatrix} 7 & 5 \\ 10 & -7 \end{pmatrix}$

3. $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$

4. $A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$

5. $A = \begin{pmatrix} -1 & 0 & 3 \\ 3 & 4 & -9 \\ -1 & 0 & 3 \end{pmatrix}$

6. $A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix}$

7. $A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$

8. $A = \begin{pmatrix} 4 & 0 & 0 & 1 \\ 12 & -2 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 21 & -6 & 1 & 0 \end{pmatrix}$

Jordan's Theorem

Given matrices P and T, verify Jordan's relation $AP = PT$.

9. $A = \begin{pmatrix} -4 & 2 \\ 0 & -1 \end{pmatrix}$, $P = I$, $T = A$.

10. $A = \begin{pmatrix} 0 & 1 \\ -2 & 3 \end{pmatrix}$, $P = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $T = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$

Cayley-Hamilton Theorem

11. Verify that $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ satisfies
$$A^2=(a+d)A-(ad-bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

12. Verify $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}^{20} = \begin{pmatrix} 1 & 0 \\ 40 & 1 \end{pmatrix}$ by induction using Cayley-Hamilton.

Gram-Schmidt Process

Find the Gram–Schmidt orthonormal basis from the given independent set.

13. $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$

Ans: Columns of I.

14. $\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}.$

15. $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}.$

16. $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$

Ans: Columns of I.

Gram-Schmidt on Polynomials

Define $V = \mathbf{span}(1, x, x^2)$ with inner product $\int_0^1 f(x)g(x)dx$. Find a Gram–Schmidt orthonormal basis.

17. 1, $1+x$, x^2

18. $1-x$, $1+x$, $1+x^2$

Gram-Schmidt: Coordinate Map

Define $V = \mathbf{span}(1, x, x^2)$ with inner product $\int_0^1 f(x)g(x)dx$. The coordinate map is
$$T: \ c_1 + c_2x + c_3x^2 \to \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

19. Find the images of $1-x$, $1+x$, $1+x^2$ under T.

20. Assume column vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, $\vec{\mathbf{x}}_3$ in $\mathcal{R}^3$ orthonormalize under Gram-Schmidt to $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$, $\vec{\mathbf{u}}_3$. Are the pre-images $T^{-1}(\vec{\mathbf{u}}_1)$, $T^{-1}(\vec{\mathbf{u}}_2)$, $T^{-1}(\vec{\mathbf{u}}_3)$ orthonormal in V?

Shadow Projection

Compute shadow vector $(\vec{\mathbf{x}} \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}$ for direction $\vec{\mathbf{u}} = \frac{\vec{\mathbf{v}}}{|\vec{\mathbf{v}}|}$. Illustrate with a hand–drawn figure.

21. $\vec{\mathbf{x}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

Ans: $-\frac{1}{5}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$

22. $\vec{\mathbf{x}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

23. $\vec{\mathbf{x}} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$, $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$

Ans: $\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$

24. $\vec{\mathbf{x}} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}$, $\vec{\mathbf{v}} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 1 \end{pmatrix}$

Orthogonal Projection

Find an orthonormal basis $\{\vec{\mathbf{u}}_k\}_{k=1}^n$ for $V = \mathbf{span}(1+x, x, x+x^2)$, inner product $\int_0^1 f(x)g(x)dx$. Then compute the orthogonal projection $\vec{\mathbf{p}} = \sum_{k=1}^n (\vec{\mathbf{x}} \cdot \vec{\mathbf{u}}_k)\vec{\mathbf{u}}_k$.

25. $\vec{\mathbf{x}} = 1 + x + x^2$

26. $\vec{\mathbf{x}} = 1 + 2x + x^2 + x^3$

Orthogonal Projection: Theory

27. **P**rove that the orthogonal projection $\mathbf{Proj}_V(\vec{\mathbf{x}})$ on $V = \mathbf{span}\{\vec{\mathbf{Y}}\}$ is the vector shadow projection $\mathbf{proj}_{\vec{\mathbf{Y}}}(\vec{\mathbf{x}})$.

28. **(Gram-Schmidt Construction)** Define $\vec{\mathbf{x}}_j^{\perp} = \vec{\mathbf{x}}_j - \mathbf{Proj}_{W_{j-1}}(\vec{\mathbf{x}}_j)$, and $W_{j-1} = \mathbf{span}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{j-1})$. Prove these properties.

(a) Subspace W_{j-1} is equal to the Gram-Schmidt $V_{j-1} = \mathbf{span}(\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_j)$.

(b) Vector $\vec{\mathbf{x}}_j^{\perp}$ is orthogonal to all vectors in W_{j-1}.

(c) The vector $\vec{\mathbf{x}}_j^{\perp}$ is not zero.

(d) The Gram-Schmidt vector is

$$\vec{\mathbf{u}}_j = \frac{\vec{\mathbf{x}}_j^{\perp}}{\|\vec{\mathbf{x}}_j^{\perp}\|}.$$

Near Point Theorem

Find the near point to the subspace V.

29. $\vec{\mathbf{x}} = \begin{pmatrix}1\\1\end{pmatrix}$, $V = \mathbf{span}\left(\begin{pmatrix}1\\2\end{pmatrix}\right)$

30. $\vec{\mathbf{x}} = \begin{pmatrix}1\\1\end{pmatrix}$, $V = \mathbf{span}\left(\begin{pmatrix}0\\1\end{pmatrix}\right)$

31. $\vec{\mathbf{x}}=\begin{pmatrix}1\\1\\0\end{pmatrix}$,$V=\mathbf{span}\left(\begin{pmatrix}1\\2\\0\end{pmatrix}, \begin{pmatrix}1\\0\\1\end{pmatrix}\right)$

32. $\vec{\mathbf{x}}=\begin{pmatrix}1\\0\\1\end{pmatrix}$,$V=\mathbf{span}\left(\begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}\right)$

QR-Decomposition

Give A, find an orthonormal matrix Q and an upper triangular matrix R such that $A = QR$.

33. $A=\begin{pmatrix}5&9\\1&7\\1&5\\3&5\end{pmatrix}$, Ans: $R = \begin{pmatrix}6&12\\0&6\end{pmatrix}$

34. $A=\begin{pmatrix}2&1\\2&0\\2&0\\2&1\end{pmatrix}$, Ans: $R = \begin{pmatrix}4&1\\0&1\end{pmatrix}$

35. $A=\begin{pmatrix}1&0&0\\1&1&0\\1&1&0\\1&0&0\end{pmatrix}$, Ans: $R=\begin{pmatrix}2&1&0\\0&1&0\end{pmatrix}$

36. $A=\begin{pmatrix}1&0&0\\1&1&1\\1&1&1\\1&0&0\end{pmatrix}$, Ans: $R=\begin{pmatrix}2&1&1\\0&1&1\end{pmatrix}$

Linear Least Squares: 3×2

Let $A=\begin{pmatrix}2&0\\0&2\\1&1\end{pmatrix}$, $\vec{\mathbf{b}}=\begin{pmatrix}1\\0\\5\end{pmatrix}$.

37. Find the normal equations for $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$.

38. Solve $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ by least squares.

Linear Least Squares: 4×3

Let $A=\begin{pmatrix}4&0&1\\1&0&1\\0&1&0\\1&1&1\end{pmatrix}$, $\vec{\mathbf{b}}=\begin{pmatrix}3\\0\\0\\0\end{pmatrix}$.

39. Find the normal equations for $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$.

40. Solve $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ by least squares.

Orthonormal Diagonal Form

Let $A = A^T$. The **spectral theorem** implies $AQ = QD$ where D is diagonal and Q has orthonormal columns. Find Q and D.

41. $A=\begin{pmatrix}7&2\\2&4\end{pmatrix}$

42. $A=\begin{pmatrix}1&5\\5&1\end{pmatrix}$

43. $A=\begin{pmatrix}1&5&0\\5&1&0\\0&0&2\end{pmatrix}$

Ans: Eigenvalues $-4, 2, 6$, orthonormal eigenvectors $\frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\\0\end{pmatrix}$,

$minicolvectorC001$, $\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix}$

44. $A=\begin{pmatrix}1&5&0\\5&1&1\\0&1&1\end{pmatrix}$

Eigenpairs of Symmetric Matrices: **Spectral Theorem**.

45. Let $A=\begin{pmatrix}3&-1&1\\-1&3&-1\\1&-1&3\end{pmatrix}$. Eigenvalues are $2, 2, 5$. Find three orthonormal eigenpairs.

46. Let $A=\begin{pmatrix} 5 & -1 & 1 \\ -1 & 5 & -1 \\ 1 & -1 & 5 \end{pmatrix}$. Then $|A-\lambda I|=(4-\lambda)^2(7-\lambda)$. Find three orthonormal eigenpairs.

47. Let $A=\begin{pmatrix} 6 & -1 & 1 \\ -1 & 6 & -1 \\ 1 & -1 & 6 \end{pmatrix}$. Eigenvectors $\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ are for $\lambda = 5, 5, 8$. Illustrate $AQ = QD$ with D diagonal and Q orthogonal.

48. Matrix A for $\lambda = 1, 1, 4$ has orthogonal eigenvectors
$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.
Find A and directly verify $A = A^T$.

Singular Value Decomposition

Find the SVD $A = U\Sigma V^T$.

49. $A=\begin{pmatrix} -1 & 1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix}$.
Ans: U=3×3, V=2×2. Matrix $\Sigma=\begin{pmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$=$3 \times 2$, the size of A.

50. $A=\begin{pmatrix} -1 & 1 \\ -2 & 2 \\ 1 & 1 \end{pmatrix}$.
Ans: $\sigma_1 = \sqrt{10}, \sigma_2 = \sqrt{2}$.

51. $A=\begin{pmatrix} -3 & 3 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}$.

52. $A=\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & -1 \end{pmatrix}$.

Ellipse and the SVD

Repeat Example 9.17, page 736 for the given ellipse equation.

53. $50x^2 - 30xy + 10y^2 = 2500$

54. $40x^2 - 16xy + 10y^2 = 2500$

Mapping and the SVD

Reference: Example 9.18, page 738.

Let $\vec{\mathbf{w}}=\begin{pmatrix} x \\ y \end{pmatrix}=c_1\vec{\mathbf{v}}_1+c_2\vec{\mathbf{v}}_2$,
$U=\frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}$, $\Sigma=\begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix}$, $V=\frac{1}{\sqrt{5}}\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$,
$A=\begin{pmatrix} -2 & 6 \\ 6 & 7 \end{pmatrix}$. Then $A=U\Sigma V^T$.

55. Verify $\|\vec{\mathbf{w}}\|^2 = \vec{\mathbf{w}} \cdot \vec{\mathbf{w}} = c_1^2 + c_2^2$.

56. Verify $V^T\vec{\mathbf{w}}=\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ from the general identity $V^TV = I$. Then show that $\Sigma V^T\vec{\mathbf{w}}=\begin{pmatrix} 10c_1 \\ 5c_2 \end{pmatrix}$.

Therefore, coordinate map $\vec{\mathbf{w}} \to \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ undergoes re-scaling by 10 in direction $\vec{\mathbf{v}}_1$ and 5 in direction $\vec{\mathbf{v}}_2$.

57. Find the angle θ of rotation for V^T and the reflection axis for U.

58. Assume $|\vec{\mathbf{w}}\| = 1$, a point on the unit circle. Is $A\vec{\mathbf{w}}$ on an ellipse with semi-axes 10 and 5? Justify your answer geometrically, no proof expected. Check your answer with a computer plot.

Four Fundamental Subspaces

Compute matrices S_1, S_2 such that the column spaces of S_1, S_2 are the nullspaces of A and A^T. Verify the two orthogonality relations of the four subspaces page 739 from the matrix identities $AS_1 = 0$, $A^TS_2 = 0$.

59. $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 0 \end{pmatrix}$. Answer:
$S_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $S_2 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$.

60. $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix}$. Answer:
$S_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $S_2 = \begin{pmatrix} -1 & -1 \\ -2 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$

61. $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 2 & 2 & 0 & 2 \end{pmatrix}$ Answer:

$S_1 = \begin{pmatrix} 0 & 0 \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$, $S_2 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$

62. $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$ Answer:

$S_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}$, $S_2 = \begin{pmatrix} 2 & 0 \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$,

Fundamental Theorem of Linear Algebra

Strang's Theorem says that the four subspaces built from $n \times m$ matrix A and $m \times n$ matrix A^T satisfy

$$\begin{array}{l} \mathbf{colspace}(A^T) \perp \mathbf{nullspace}(A), \\ \mathbf{colspace}(A) \perp \mathbf{nullspace}(A^T). \end{array}$$

Let $r = \mathbf{rank}(A) = \mathbf{rank}(A^T)$. The four subspace dimensions are:

$$\begin{array}{l} \dim(\mathbf{colspace}(A)) = r, \\ \dim(\mathbf{nullspace}(A)) = n - r, \\ \dim(\mathbf{colspace}(A^T)) = r, \\ \dim(\mathbf{nullspace}(A^T)) = m - r. \end{array}$$

63. Explain why $\dim(\mathbf{colspace}(A)) = \dim(\mathbf{colspace}(A^T)) = r$ from the Pivot Theorem.

64. Suppose A is 10×4. What are the dimensions of the four subspaces?

65. Invent a 4×4 matrix A where one of the four subspaces is the zero vector alone.

66. Prove that the only vector in common with $\mathbf{rowspace}(A)$ and $\mathbf{nullspace}(A)$ is the zero vector.

67. Prove that each vector $\vec{\mathbf{x}}$ in $\mathcal{R}^n$ can be uniquely written as $\vec{\mathbf{x}} = \vec{\mathbf{x}}_1 + \vec{\mathbf{x}}_2$ where $\vec{\mathbf{x}}_1$ is in $\mathbf{colspace}(A^T)$ and $\vec{\mathbf{x}}_2$ is in $\mathbf{nullspace}(A)$. See **direct sum** in exercise page 428.

68. Prove that each vector $\vec{\mathbf{y}}$ in $\mathcal{R}^m$ can be uniquely written as $\vec{\mathbf{y}} = \vec{\mathbf{y}}_1 + \vec{\mathbf{y}}_2$ where $\vec{\mathbf{y}}_1$ is in $\mathbf{colspace}(A)$ and $\vec{\mathbf{y}}_2$ is in $\mathbf{nullspace}(A^T)$.

Chapter 10

Phase Plane Methods

Contents

Studied here are planar autonomous systems of differential equations. The topics:

1. Planar Autonomous Systems: Phase Portraits, Stability.

2. Planar Constant Linear Systems: Classification of isolated equilibria, Phase portraits.

3. Planar Almost Linear Systems: Phase portraits, Nonlinear classifications of equilibria.

4. Biological Models: Predator-prey models, Competition models, Survival of one species, Co-existence, Alligators, doomsday and extinction.

5. Mechanical Models: Nonlinear spring-mass system, Soft and hard springs, Energy conservation, Phase plane and scenes.

10.1 Planar Autonomous Systems

A set of two scalar differential equations of the form

$$\begin{aligned} x'(t) &= f(x(t), y(t)), \\ y'(t) &= g(x(t), y(t)). \end{aligned} \tag{1}$$

is called a **planar autonomous system**. The term **Autonomous** means **Self-Governing**, justified by the absence of the time variable t in the functions $f(x, y)$, $g(x, y)$.

To obtain the vector form, let $\vec{u}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, $\vec{F}(x, y) = \begin{pmatrix} f(x, y) \\ g(x, y) \end{pmatrix}$ and write (1) as the first order vector-matrix system

$$\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t)). \tag{2}$$

It is assumed that f, g are continuously differentiable in some region $\mathcal{D}$ in the xy-plane. This assumption makes $\vec{F}$ continuously differentiable in $\mathcal{D}$ and guarantees that Picard's existence-uniqueness theorem for initial value problems applies to the initial value problem $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$, $\vec{u}(0) = \vec{u}_0$. Accordingly, to each $\vec{u}_0 = (x_0, y_0)$ in $\mathcal{D}$ there corresponds a unique solution $\vec{u}(t) = (x(t), y(t))$, represented as a planar curve in the xy-plane, which passes through $\vec{u}_0$ at $t = 0$.

Such a planar curve is called a **Trajectory** or **Orbit** of the system and its parameter interval is some maximal interval of existence $T_1 < t < T_2$, where T_1 and T_2 might be infinite. A graphic of trajectories drawn as parametric curves in the xy-plane is called a **Phase Portrait** and the xy-plane in which it is drawn is called the **Phase Plane**.

Trajectories Don't Cross

Autonomy of the planar system plus uniqueness of initial value problems implies that trajectories $(x_1(t), y_1(t))$ and $(x_2(t), y_2(t))$ cannot touch or cross. Hand-drawn phase portraits are accordingly limited: *you cannot draw a solution trajectory that touches another solution curve*!

Theorem 10.1 (Identical Trajectories)
Assume that Picard's existence-uniqueness theorem applies to initial value problems in $\mathcal{D}$ for the planar system

$$\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t)), \quad \vec{u}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

Let $(x_1(t), y_1(t))$ and $(x_2(t), y_2(t))$ be two trajectories of the system. If times t_1, t_2 exist such that

$$x_1(t_1) = x_2(t_2), \quad y_1(t_1) = y_2(t_2), \tag{3}$$

then for the value $c = t_1 - t_2$ the equations $x_1(t+c) = x_2(t)$ and $y_1(t+c) = y_2(t)$ are valid for all allowed values of t. This means that the two trajectories are on one and the same planar curve, or in the contrapositive, two different trajectories cannot touch or cross in the phase plane.

Proof: Define $x(t) = x_1(t+c)$, $y(t) = y_1(t+c)$. By the chain rule, $(x(t), y(t))$ is a solution of the planar system, because $x'(t) = x_1'(t+c) = f(x_1(t+c), y_1(t+c)) = f(x(t), y(t))$, and similarly for the second differential equation. Further, (3) implies $x(t_2) = x_2(t_2)$ and $y(t_2) = y_2(t_2)$, therefore Picard's uniqueness theorem implies that $x(t) = x_2(t)$ and $y(t) = y_2(t)$ for all allowed values of t. ■

Equilibria

A trajectory that reduces to a point, or a constant solution $x(t) = x_0$, $y(t) = y_0$, is called an **Equilibrium Solution**. The equilibrium solutions or **Equilibria** are found by solving the nonlinear equations

$$f(x_0, y_0) = 0, \quad g(x_0, y_0) = 0.$$

Each such (x_0, y_0) in $\mathcal{D}$ is a trajectory whose graphic in the phase plane is a single point, called an **Equilibrium Point**. In applied literature, it may be called a **Critical Point**, **Stationary Point** or **Rest Point**. Theorem 10.1 has the following geometrical interpretation.

> Assuming uniqueness, no other trajectory $(x(t), y(t))$ in the phase plane can touch an equilibrium point (x_0, y_0).

Equilibria (x_0, y_0) are often found from linear equations

$$ax_0 + by_0 = e, \quad cx_0 + dy_0 = f,$$

which are solved by linear algebra methods. They constitute an important subclass of algebraic equations which can be solved symbolically. In this special case, symbolic solutions exist for the equilibria.

It is interesting to report that in a practical sense the equilibria may be reported incorrectly, due to the limitations of computer software, even in the case when exact symbolic solutions are available. An example is $x' = x + y$, $y' = \epsilon y - \epsilon$ for small $\epsilon > 0$. The root of the problem is translation of ϵ to a machine constant, which is zero for small enough ϵ. The result is that computer software detects infinitely many equilibria when in fact there is exactly one equilibrium point. This example suggests that symbolic computation be used by default.

Practical Methods for Computing Equilibria

There exists no supporting theory to find equilibria for all choices of F and G. However, there is a rich library of special methods for solving nonlinear algebraic equations, including numerical methods based on celebrated univariate methods, such as **Newton's method** and the **Bisection method**.

Computer algebra systems like `maple`, `maxima` and `mathematica` offer convenient codes to solve the equations, when possible, including symbolic solutions. Applied mathematics depends on the dynamically expanding library of special methods, which grows due to new mathematical discoveries. See the exercises for examples.

Population Biology

Planar autonomous systems have been applied to two-species populations like two species of trout, who compete for food from the same supply, and foxes and rabbits, who compete in a predator-prey situation.

Certain equilibria are significant, because they represent the population sizes for **Cohabitation**. A point in the phase space that is not an equilibrium point corresponds to population sizes that cannot coexist, they must change with time. Some equilibria are consequently **Observable** or **average** population sizes while non-equilibria correspond to snapshot population sizes that are subject to flux. Biologists expect population sizes of such two-species competition models to undergo change until they reach approximately the observable values, on the average.

Rabbit-Fox System

This example is a **Predator-Prey** system, in which the expected observable population sizes are averages, about which the actual populations size oscillate about, periodically over time. Certain equilibria for these systems represent **ideal cohabitation**. Biological experiments suggest that initial population sizes close to the equilibrium values cause populations to stay near the initial sizes, even though the populations oscillate periodically. Observations by field biologists of large population variations seem to verify that individual populations oscillate periodically around the ideal cohabitation sizes.

A typical planar system for predator-prey dynamics of $x(t)$ rabbits and $y(t)$ foxes is the system

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{200}x(40-y), \\ \frac{dy}{dt} &= \frac{1}{100}y(x-50). \end{aligned}$$

Time variable t is in months. The equilibria are $(0,0),(50,40)$. With initial populations $x(0)=60$ rabbits and $y(0)=30$ foxes, both x' and y' are positive near $t=0$, which implies the populations initially increase in size.

After time, the signs of x' and y' are alternately positive and negative, which reflects the oscillating behavior of the populations about the ideal equilibrium values $x=50$, $y=40$. The period of oscillation is about 20 months. This predator-prey model predicts coexistence with average populations of 50 rabbits and 40 foxes.

Trout System

Consider a population of two species of trout who compete for the same food supply. A typical autonomous planar system for the species x and y is

$$\begin{aligned}\frac{dx}{dt} &= x(-2x - y + 180),\\ \frac{dy}{dt} &= y(-x - 2y + 120).\end{aligned}$$

Equilibria. The equilibrium solutions for the trout system are

$$(0,0), \quad (90,0), \quad (0,60), \quad (80,20).$$

Only nonnegative population sizes are physically significant. Units for the population sizes might be in hundreds or thousands of fish. The equilibrium $(0,0)$ corresponds to **Extinction** of both species, while $(0,60)$ and $(90,0)$ correspond to the unusual situation of extinction for one species. The last equilibrium $(80,20)$ corresponds to **Co-Existence** of the two trout species with observable population sizes of 80 and 20.

Phase Portraits

A graphic which contains some equilibria and typical trajectories of a planar autonomous system (1) is called a **Phase Portrait**.

While graphing equilibria is not a challenge, graphing typical trajectories, also called **orbits**, seems to imply that we are going to solve the differential system. This is not the case. Approximations will be used that do not require solution of the differential system.

Equilibria Plot in the xy-plane all equilibria of (1). See Figure 3.

Window Select an x-range and a y-range for the graph window which includes all significant equilibria (Figure 3).

Grid Plot a uniform grid of N grid points ($N \approx 50$ for hand work) within the graph window, to populate the graphical white space (Figure 4). The isocline method might also be used to select grid points.

Field Draw at each grid point a short tangent vector, a **replacement curve** for a solution curve through a grid point on a small time interval (Figure 5).

Orbits Draw additional threaded trajectories on long time intervals into the remaining white space of the graphic (Figure 6). This is guesswork, based upon tangents to threaded trajectories matching nearby field tangents drawn in the previous step. See Figures 1 and 2 for details.

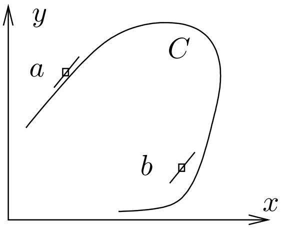


Figure 1. Badly threaded orbit.
Threaded solution curve C correctly matches its tangent to the tangent at nearby grid point a, but it fails to match at grid point b.

Why does a threaded solution curve tangent $\vec{T}_1$ have to *match* [1] a tangent $\vec{T}_2$ at a nearby grid point (see Figure 2)? A tangent vector is given by $\vec{T} = \frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$. Then $\vec{T}_1 = \vec{F}(\vec{u}_1)$, $\vec{T}_2 = \vec{F}(\vec{u}_2)$. However, $\vec{u}_1 \approx \vec{u}_2$ in the graphic, hence by continuity of $\vec{F}$ it follows that $\vec{F}(\vec{u}_1) \approx \vec{F}(\vec{u}_2)$, which implies $\vec{T}_1 \approx \vec{T}_2$.

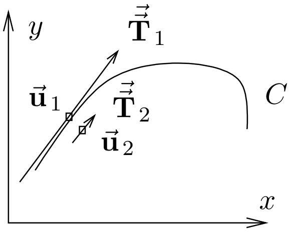


Figure 2. Tangent matching.
Threaded solution curve C matches its tangent $\vec{T}_1$ at $\vec{u}_1$ to direction field tangent $\vec{T}_2$ at nearby grid point $\vec{u}_2$.

It is important to emphasize that solution curves starting at a grid point are defined for a small t-interval about $t = 0$, and therefore their graphics extend on both sides of the grid point. We intend to shorten these curves until they appear to be straight line segments, graphically atop the tangent line, to pixel resolution. Adding an arrowhead pointing in the tangent vector direction is usual. After all this construction, *the shaft of the arrow is graphically atop a short solution curve segment.* In fact, if 50 grid points were used, then 50 short solution curve segments have already been entered onto the graphic! Threaded orbits are added to show what happens to solutions that are plotted on longer and longer t-intervals.

Phase Portrait Illustration

The method outlined above will be applied to the illustration

$$\begin{aligned} x'(t) &= x(t) + y(t), \\ y'(t) &= 1 - x^2(t). \end{aligned} \tag{4}$$

The equilibria are $(1, -1)$ and $(-1, 1)$. The graph window is selected as $|x| \le 2$, $|y| \le 2$, in order to include both equilibria. The uniform grid will be 11×11,

[1] *Match* means nearly identical, in an approximate sense: graphics of the two tangents are identical to pixel resolution.

although for hand work 5 × 5 is normal. Tangents at the grid points are short line segments which do not touch each another – they are graphically the same as short solution curves.

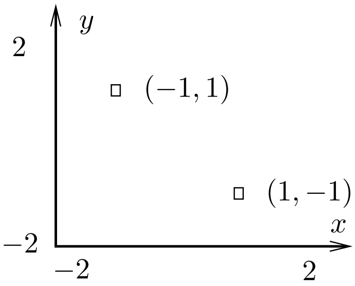


Figure 3. Equilibria $(1,-1)$**,** $(-1,1)$ **with Invented Graph Window.**
The equilibria (x,y) are calculated from equations $0 = x + y$, $0 = 1 - x^2$. The graph window $|x| \le 2$, $|y| \le 2$ is invented initially, then updated until Figure 5 reveals sufficiently rich field details.

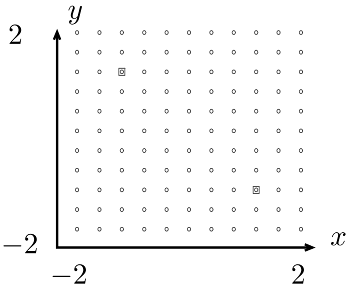


Figure 4. Equilibria $(1,-1)$**,** $(-1,1)$ **and Invented** 11×11 **Uniform Grid.**
The equilibria (squares) happen to cover up two grid points. The invented size 11×11 should fill the white space in the graphic.

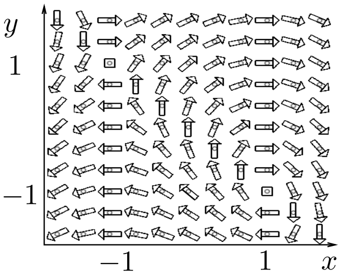


Figure 5. Equilibria, Uniform Grid and Direction Field.
An arrow shaft at a grid point represents a solution curve over a small time interval. Threaded solution curves on long time intervals have tangents matching nearby arrow shafts directions.

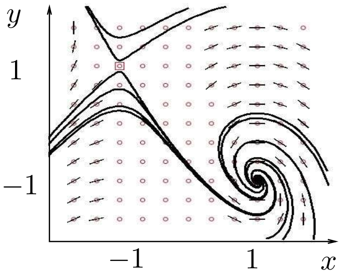


Figure 6. Initial Phase Portrait.
Equilibria $(1,-1)$, $(-1,1)$ and 11×11 uniform grid with threaded solution curves. Arrow shafts included from some direction field arrows. Threaded solution curve tangents are to match nearby direction field arrow shafts. See Figures 1 and 2 for how to match tangents.

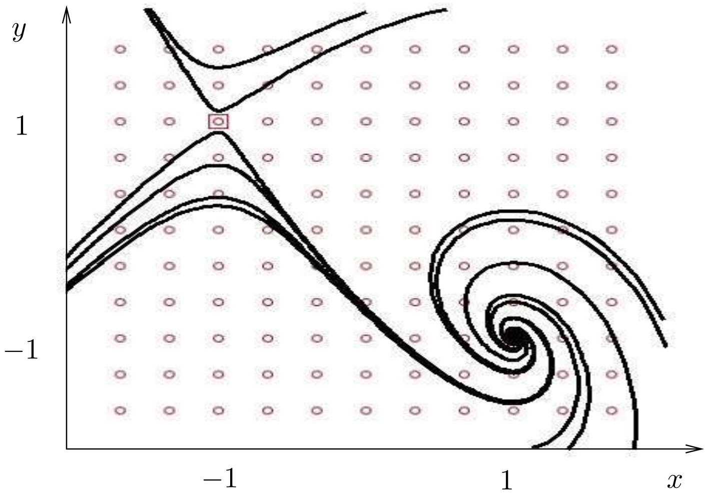


Figure 7. Final Phase portrait.
Shown are some threaded solution curves and an 11 × 11 grid. The direction field has been removed for clarity. Threaded solution curves do not actually cross, even though graphical resolution might suggest otherwise.

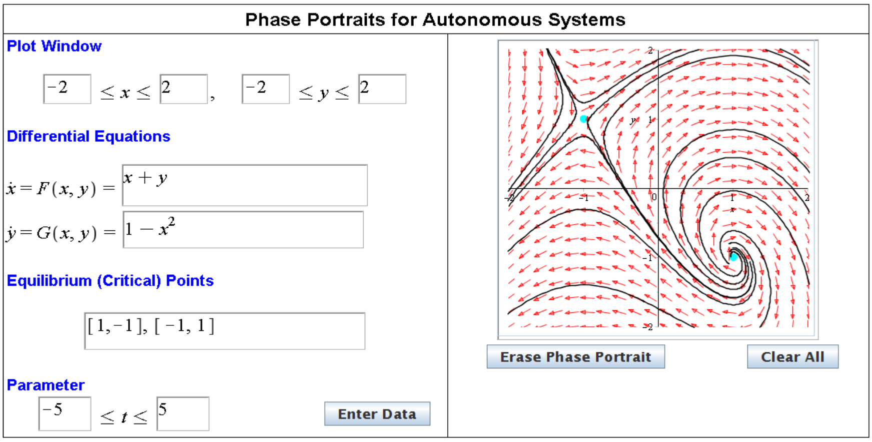
Phase Plot by Computer

Illustrated here is how to make a phase plot like Figure 8 or Figure 9, *infra*, with computer algebra system `maple`, for the system of differential equations

$$\begin{aligned} x'(t) &= x(t) + y(t), \\ y'(t) &= 1 - x^2(t). \end{aligned} \tag{5}$$

Before the computer work begins, the differential equation is defined and the equilibria are computed. Defaults supplied by `maple` allow an initial phase portrait to be plotted, from which the graph window is invented.

Phase plot tools can simplify initial plot production. To illustrate, `maple` task **Phase Portrait** has this interface:



Phase Portraits for Autonomous Systems

Plot Window

$-2 \le x \le 2$, $-2 \le y \le 2$

Differential Equations

$\dot{x} = F(x, y) =$ $x + y$

$\dot{y} = G(x, y) =$ $1 - x^2$

Equilibrium (Critical) Points

[1, −1], [−1, 1]

Parameter

$-5 \le t \le 5$

Figure 8. `PhasePortrait` **task in computer algebra system** `Maple` **for equations (5).**

Minimal input requires two differential equations, equilibria, a graph window and time interval for threaded curves. Clicking on the graphic produces threaded solution curves.

The Phase Portrait Task is unlikely to be able to produce a final, production figure. Other tools are normally used afterwards, to make the final figure.

The initial plot code:

```
des:=diff(x(t),t)=x(t)+y(t),diff(y(t),t)=1-x(t)^2:
wind:=x=-2..2,y=-2..2:Times:=t=-20..20:
DEtools[DEplot]([des],[x(t),y(t)],Times,wind);
```

The initial plot suggests which initial conditions near the equilibria should be selected in order to create typical orbits on the graphic. The final code with initial data and options:

```
des:=diff(x(t),t)=x(t)+y(t),diff(y(t),t)=1-x(t)^2:
wind:=x=-2..2,y=-2..2:Times:=t=-20..20:
opts:=stepsize=0.05,dirgrid=[13,13],
axes=none,thickness=3,arrows=small:
ics:=[[x(0)=-1,y(0)=1.1],[x(0)=-1,y(0)=1.5],
[x(0)=-1,y(0)=.9],[x(0)=-1,y(0)=.6],[x(0)=-1,y(0)=.3],
[x(0)=1,y(0)=-0.9],[x(0)=1,y(0)=-0.6],[x(0)=1,y(0)=-0.6],
[x(0)=1,y(0)=-0.3],[x(0)=1,y(0)=-1.6],[x(0)=1,y(0)=-1.3],
[x(0)=1,y(0)=-1.1]]:
DEtools[DEplot]([des],[x(t),y(t)],Times,wind,ics,opts);
```

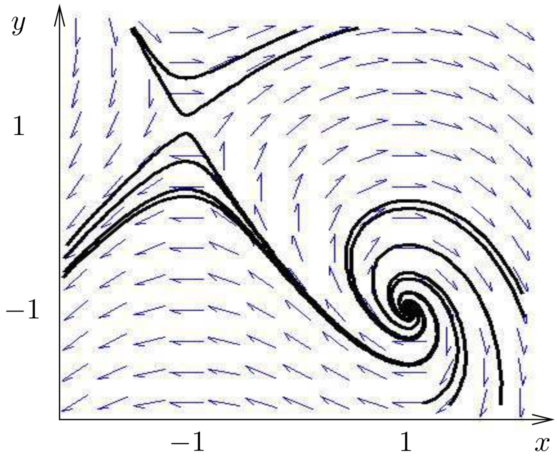


Figure 9. Phase Portrait for (5).
The graphic shows typical solution curves and a direction field. The graphic was produced in `maple` using a 13 × 13 grid.

Stability

Consider an autonomous system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ with $\vec{F}$ continuously differentiable in a region $\mathcal{D}$ in the plane.

Stable equilibrium. An equilibrium point $\vec{u}_0$ in $\mathcal{D}$ is said to be **Stable** provided for each $\epsilon > 0$ there corresponds $\delta > 0$ such that

(a) given $\vec{u}(0)$ in $\mathcal{D}$ with $\|\vec{u}(0) - \vec{u}_0\| < \delta$, then the solution $\vec{u}(t)$ exists on $0 \le t < \infty$ and

(b) $\|\vec{u}(t) - \vec{u}_0\| < \epsilon$ for $0 \le t < \infty$.

Unstable equilibrium. The equilibrium point $\vec{u}_0$ is called **Unstable** provided it is **not stable**, meaning at least one of **(a)** or **(b)** fails.

Asymptotically stable equilibrium. The equilibrium point $\vec{u}_0$ is said to be **Asymptotically Stable** provided **(a)** and **(b)** hold (it is **stable**), and additionally

(c) $\lim_{t\to\infty} \|\vec{u}(t) - \vec{u}_0\| = 0$ for $\|\vec{u}(0) - \vec{u}_0\| < \delta$.

Applied accounts of stability tend to emphasize item **(b)**. Careful application of stability theory requires attention to **(a)**, which is the question of extension of solutions of initial value problems to the half-axis.

Basic extension theory for solutions of autonomous equations says that **(a)** will be satisfied provided **(b)** holds for those values of t for which $\vec{u}(t)$ is already defined. Stability verifications in mathematical and applied literature often implicitly use extension theory, in order to present details compactly. The reader is advised to adopt the same predisposition as researchers, who assume the reader to be equally clever as they.

Physical stability. In the model $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$, physical stability addresses changes in $\vec{F}$ as well as changes in $\vec{u}(0)$. The meaning is this: physical parameters of the model, e.g., the mass $m > 0$, damping constant $c > 0$ and Hooke's constant $k > 0$ in a damped spring-mass system

$$\begin{aligned} x' &= y, \\ y' &= -\frac{c}{m}\,y - \frac{k}{m}\,x, \end{aligned}$$

may undergo small changes without significantly affecting the solution.

In physical stability, stable equilibria correspond to **Physically Observed** data whereas other solutions correspond to **Transient Observations** that disappear over time.

A typical instance is the trout system

$$\begin{aligned} x'(t) &= x(-2x - y + 180), \\ y'(t) &= y(-x - 2y + 120). \end{aligned} \tag{6}$$

Physically observed data in the trout system (6) corresponds to the **carrying capacity**, represented by the **Stable Equilibrium** point $(80, 20)$, whereas transient observations are snapshot population sizes that are subject to change over time. The strange extinction equilibria $(90, 0)$ and $(0, 60)$ are **unstable equilibria**, which disagrees with intuition about zero births for less than two individuals, but agrees with graphical representations of the trout system in Figure 10. Changing $\vec{F}(\vec{u})$ for a trout system adjusts the physical constants which describe the birth and death rates, whereas changing $\vec{u}(0)$ alters the initial population sizes of the two trout species.

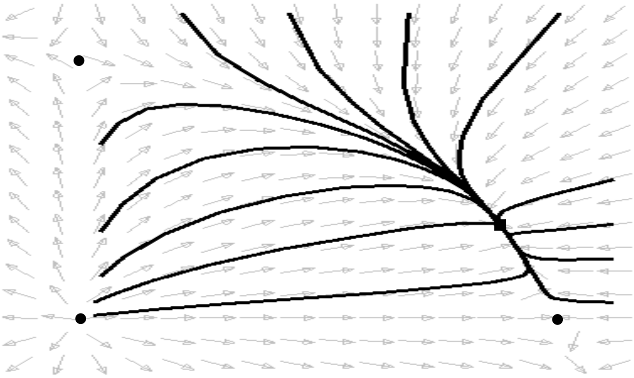
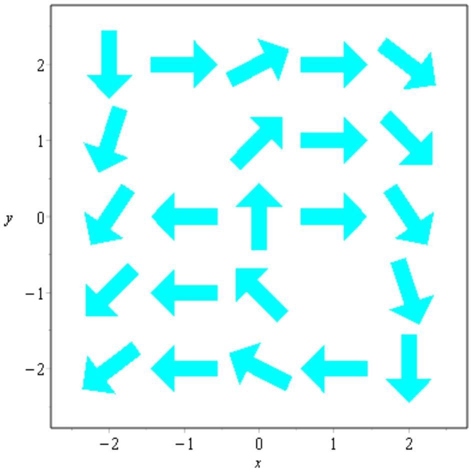


Figure 10. Phase Portrait for Trout System (6).
Shown are typical solution curves and a direction field. Equilibrium $(80, 20)$ is asymptotically stable (a square). Equilibria $(0, 0)$, $(90, 0)$, $(0, 60)$ are unstable (circles).

Direction Fields by Computer

Direction fields are produced by `Maple` with tool `DEtools[dfieldplot]` or with interactive graphical task `PhasePortrait`. Basic code that produces a direction field can be written with minimal effort:



Maple code:

```
de1:=diff(x(t),t)=x(t)+y(t);
de2:=diff(y(t),t)=1-x(t)*x(t);vars:=[x(t),y(t)];
trange:=t=-10..10:xrange:=x=-2..2:yrange:=y=-2..2:
opts1:=trange,xrange,yrange:
opts2:=arrows=large,color=cyan,dirfield=[5,5]:
DEtools[dfieldplot]([de1,de2],vars,opts1,opts2);
```

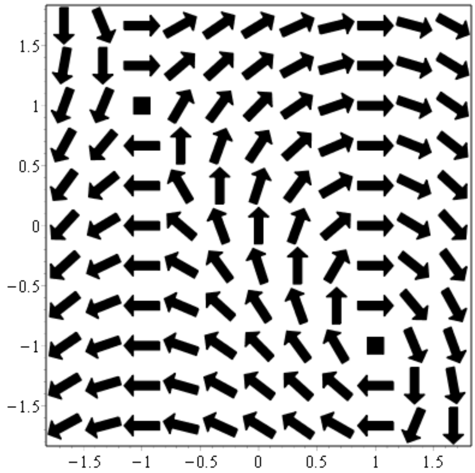
A Direction Field Procedure

The ideas discussed below for `maple` apply to other programming languages, such as `Maxima, Mathematica, Ruby, Python` and Microsoft developer languages. `Maple` code below considers the system

$$x' = F_1(x,y), \quad y' = F_2(x,y)$$

with example $x' = F_1 = x + y, y' = F_2 = 1 - x^2$, which was treated above.

```
F1:=(x,y)->evalf(x+y):F2:=(x,y)->evalf(1-x^2):
P:=directionField(F1,F2):plots[display](P);# proc below
```



Maple function plottools[rectangle] requires two arguments ul, lr, which are the upper left (ul) and lower right (lr) vertices of the rectangle.

Maple function plottools[arrow] requires five arguments P, Q, sw, aw, af: the two points P, Q which define the arrow shaft and direction, plus the shaft width sw, arrowhead width aw and arrowhead length fraction af (fraction of the shaft length).

The two functions rectangle, arrow plot a polygon from its vertices. Function rectangle computes four vertices and function arrow computes seven vertices. Maple function plots[display] plots the vertices.

```
 # 2D phase plane direction field with uniform nxm grid.
 # Tangent length is 9/10 the grid box width W0.
directionField:=
 proc(F1,F2,a:=-2,b:=2,c:=-2,d:=2,n:=11,m:=11)
description "Custom direction field for F1,F2\
Window: a <= x <= b, c <= y <= d, Grid: n by m\
Tangent length = 9/10 grid box width W0.";
local x,y,X,Y,V,H,K,i,j,M1,M2,W0,h,p1,p2,q1,q2; global P;
H:=evalf((b-a)/(n+1)):K:=evalf((d-c)/(m+1)):W0:=min(H,K):
X:=t->a+H*(t):Y:=t->c+K*(t):P:=[]:
for i from 1 to n do
for j from 1 to m do
x:=X(i):y:=Y(j):M1:=F1(x,y): M2:=F2(x,y):
if (M1 =0 and M2 =0) then # no tangent, make a box
h:=W0/5:V:=plottools[rectangle]([x-h,y+h],[x+h,y-h]):
else
h:=evalf(((1/2)*9*W0/10)/sqrt(M1^2+M2^2)):
p1:=x-h*M1:p2:=y-h*M2:q1:=x+h*M1:q2:=y+h*M2:
V:=plottools[arrow]([p1,p2],[q1,q2],0.2*W0,0.5*W0,1/4):
fi:if (P = []) then P:=V: else P:=P,V: fi:
od:od:
```

```
RETURN (P);
end proc:
```

Exercises 10.1

Autonomous Planar Systems.

Consider

$$(7) \quad \begin{array}{rcl} x'(t) &=& x(t)+y(t), \\ y'(t) &=& 1-x^2(t). \end{array}$$

1. (Vector-Matrix Form) System (7) can be written in vector-matrix form

$$\frac{d}{dt}\vec{u} = \vec{F}(\vec{u}(t)).$$

Display formulas for $\vec{u}$ and $\vec{F}$.

2. (Picard's Theorem) Picard's vector existence-uniqueness theorem applies to system (7) with initial data $x(0)=x_0$, $y(0)=y_0$. Show the details.

Trajectories Don't Cross.

3. (Theorem 10.1 Details) Show $\frac{dy}{dt} = g(x_1(t+c), y_1(t+c))$, then show that $y'(t) = g(x(t), y(t))$ in the proof of Theorem 10.1.

4. (Orbits Can Cross) The example

$$\frac{dx}{dt} = 1, \quad \frac{dy}{dt} = 3y^{2/3}$$

has infinitely many orbits crossing at $x=y=0$. Exhibit two distinct orbits which cross at $x=y=0$. Does this example contradict Theorem 10.1?

Equilibria.

A point (x_0, y_0) is called an **Equilibrium** provided $x(t)=x_0$, $y(t)=y_0$ is a solution of the dynamical system.

5. Justify that $(1,-1), (-1,1)$ are the only equilibria for the system $x' = x+y$, $y' = 1-x^2$.

6. Display the details which justify that $(0,0), (90,0), (0,60), (80,20)$ are all equilibria for the system $x'(t) = x(-2x-y+180)$, $y'(t) = y(-x-2y+120)$.

Practical Methods for Computing Equilibria.

7. (Murray System) The biological system

$$x' = x(6-2x-y), y' = y(4-x-y)$$

has equilibria $(0,0)$, $(3,0)$, $(0,4)$, $(2,2)$. Justify the four answers.

8. (Nullclines) Curves along which either $x'=0$ or $y'=0$ are called **nullclines**. The biological system

$$x' = x(6-2x-y), y' = y(4-x-y)$$

has nullclines $x=0$, $y=0$, $6-2x-y=0$, $4-x-y=0$. Justify the four answers.

9. (Nullclines by Computer) Produce a graphical display of the nullclines of the Murray System above. Maple code below makes a plot from equations $x(6-2x-y)=0$, $y(4-x-y)=0$.

```
eqns:={x*(6-2*x-y),y*(4-x-y)};
wind:=x=-5..5,y=-10..10;
opts:=wind,contours=[0];
plots[contourplot](eqns,opts);
```

10. (Isoclines by Computer) Level curves $f(x,y)=c$ are called **Isoclines**.

`Maple` will plot level curves $f(x,y)=-2$, $f(x,y)=0$, $f(x,y)=2$ using the nullcline code above, with replacement `contours=[-2,0,2]`. Produce an isocline plot for the Murray System above with these same contours.

11. (Implicit Plot) Equilibria can be found graphically by an implicit plot.

```
# MAPLE implicit plot
eqns:={x*(6-2*x-y),y*(4-x-y)};
wind:=x=-5..5,y=-10..10;
plots[implicitplot](eqns,wind);
```

Produce the implicit plot. Is it the same as the nullcline plot?

12. (Implicit Plot) Find the equilibria graphically by an implicit plot. Then find the equilibria exactly.

$$\left\{ \begin{array}{rcl} x'(t) &=& x(t)+y(t), \\ y'(t) &=& 4-x^2(t). \end{array} \right.$$

Rabbit-Fox System.

13. (Predator-Prey) Consider a rabbit and fox system

$$\begin{aligned} x' &= \frac{1}{200}x(30-y), \\ y' &= \frac{1}{100}y(x-40). \end{aligned}$$

Argue why extinction of the rabbits ($x=0$) implies extinction of the foxes ($y=0$).

14. (Predator-Prey) The rabbit and fox system

$$\begin{aligned} x' &= \frac{1}{200}x(40-y), \\ y' &= \frac{1}{100}y(x-40), \end{aligned}$$

has extinction of the foxes ($y=0$) implying Malthusian population explosion of the rabbits ($\lim_{t=\infty} x(t)=\infty$). Explain.

Trout System. Consider

$$\begin{aligned} x'(t) &= x(-2x-y+180), \\ y'(t) &= y(-x-2y+120). \end{aligned}$$

15. (Carrying Capacity) Show details for calculation of the equilibrium $x=80$, $y=20$, which is **co-existence**.

16. (Stability) Equilibrium point $x=80$, $y=20$ is stable. Explain this statement using geometry from Figure 10 and the definition of stability.

Phase Portraits. Consider

$$\begin{aligned} x'(t) &= x(t)+y(t), \\ y'(t) &= 1-x^2(t). \end{aligned}$$

17. (Equilibria) Solve for x,y in the system

$$\begin{aligned} 0 &= x+y, \\ 0 &= 1-x^2, \end{aligned}$$

for equilibria $(1,-1)$, $(-1,1)$. Explain why $|x|\le 2$, $|y|\le 2$ is a suitable graph window.

18. (Grid Points) Draw a 5×5 grid on the graph window $|x|\le 2, |y|\le 2$. Label the equilibria.

19. (Direction Field) Draw direction field arrows on the 5×5 grid of the previous exercise. They coincide with the tangent direction $\vec{v}=x'\vec{\imath}+y'\vec{\jmath}=(x+y)\vec{\imath}+(1-x^2)\vec{\jmath}$, where (x,y) is the grid point. The arrows may not touch.

20. (Threaded Orbits) On the direction field of the previous exercise, draw orbits (*threaded solution curves*), using the rules:

1. Orbits don't cross.
2. Orbits pass direction field arrows with nearly matching tangent.

Phase Plot by Computer. Use a computer algebra system or a numerical workbench to produce phase portraits for the given dynamical system. A graph window should contain all equilibria.

21. (Rabbit-Fox System I)

$$\begin{aligned} x' &= \frac{1}{200}x(30-y), \\ y' &= \frac{1}{100}y(x-40). \end{aligned}$$

22. (Rabbit-Fox System II)

$$\begin{aligned} x' &= \frac{1}{100}x(50-y), \\ y' &= \frac{1}{200}y(x-40). \end{aligned}$$

23. (Trout System I)

$$\begin{aligned} x'(t) &= x(-2x-y+180), \\ y'(t) &= y(-x-2y+120). \end{aligned}$$

24. (Trout System II)

$$\begin{aligned} x'(t) &= x(-2x-y+200), \\ y'(t) &= y(-x-2y+120). \end{aligned}$$

Stability Conditions. Consider equilibrium point $(0,0)$ and nearby solution curves $x(t), y(t)$ with $(x(0),y(0))$ near $(0,0)$.

25. (Instability: Repeller) Prove: If for every $\delta > 0$ there is one solution with $|x(0)^2 + y(0)^2| < \delta^2$ such that $\lim_{t\to\infty} |x(t)| + |y(t)| = \infty$ then equilibrium $(0, 0)$ is unstable.

26. (Stability: Attractor) Prove that $x'(t) < 0$ and $y'(t) < 0$ for all nearby solutions implies stability at $(0, 0)$, but not asymptotic stability.

27. (Instability in x) Prove that $\lim_{t\to\infty} |x(t)| = \infty$ implies instability at $(0, 0)$.

28. (Instability in y) Prove that $\lim_{t\to\infty} |y(t)| = \infty$ implies instability at $(0, 0)$.

Geometric Stability.

29. (Attractor) Imagine a dust particle in a fluid draining down a funnel, whose trace is a space curve. Assume fluid drains at $x = 0$, $y = 0$ and the funnel centerline is along the z-axis. Project the space curve onto the xy-plane. Is this planar orbit stable at $(0, 0)$ in the sense of the definition?

30. (Repeller) Imagine a paint droplet from a paint spray can, pointed downward, which traces a space curve. Project the space curve onto the xy-plane orthogonal to the spray nozzle direction, centerline along the z-axis. Is this planar orbit stable at $(0, 0)$ in the sense of the definition?

Geometric Stability: Phase Portrait.

31. (Rabbit–Fox I Stability) Plot a phase portrait for system

$$\begin{aligned} x' &= \frac{1}{200}x(30 - y), \\ y' &= \frac{1}{100}y(x - 40). \end{aligned}$$

Provide geometric evidence for stability of equilibrium $x = 40$, $y = 30$.

32. (Rabbit–Fox II Instability) Plot a phase portrait for system

$$\begin{aligned} x' &= \frac{1}{100}x(50 - y), \\ y' &= \frac{1}{200}y(x - 40). \end{aligned}$$

Provide geometric evidence for instability of equilibrium $x = 0$, $y = 0$ and stability of equilibrium $x = 40$, $y = 50$.

10.2 Planar Constant Linear Systems

A **constant linear** planar system is a set of two scalar differential equations of the form

$$(1)\qquad \begin{aligned} x'(t) &= ax(t) + by(t)),\\ y'(t) &= cx(t) + dy(t)), \end{aligned}$$

where a, b, c and d are constants. In matrix form,

$$\frac{d}{dt}\vec{u}(t) = A\vec{u}(t), \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \vec{u}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

Solutions drawn in phase portraits don't cross, because of Picard's theorem. The system is autonomous. The origin is always an equilibrium solution. There can be infinitely many equilibria, found by solving $A\vec{u} = \vec{0}$ for the constant vector $\vec{u}$, when A is not invertible.

Formula. System (1) can be solved by a formula which parallels the theorem for second order constant coefficient equations $Ay'' + By' + Cy = 0$. You are invited to learn Putzer's spectral method, page 866, which is used to derive the formulas. For now, let's accept the formulas displayed in the next theorem. Putzer's result depends only on the Cayley-Hamilton theorem, which says that a matrix A satisfies the characteristic equation $|A - \lambda I| = 0$ under substitution $\lambda = A$.

Theorem 10.2 (Planar Constant Linear System: Putzer's Formula)
Consider the real planar system $\frac{d}{dt}\vec{u}(t) = A\vec{u}(t)$. Let λ_1, λ_2 be the roots of the characteristic equation $\det(A - \lambda I) = 0$. The real general solution $\vec{u}(t)$ is given by the formula

$$\vec{u}(t) = \Phi(t)\vec{u}(0)$$

where the 2×2 real invertible matrix $\Phi(t)$ is defined as follows.

Real $\lambda_1 \ne \lambda_2$	$\Phi(t) = e^{\lambda_1 t}\, I + \dfrac{e^{\lambda_2 t} - e^{\lambda_1 t}}{\lambda_2 - \lambda_1}\,(A - \lambda_1 I).$
Real $\lambda_1 = \lambda_2$	$\Phi(t) = e^{\lambda_1 t}\, I + te^{\lambda_1 t}\,(A - \lambda_1 I).$
Complex $\lambda_1 = \overline{\lambda}_2$, $\lambda_1 = a + bi$, $b > 0$	$\Phi(t) = e^{at}\left(\cos(bt)\, I + (A - aI)\dfrac{\sin(bt)}{b}\right).$

Continuity and Redundancy

The formulas are continuous in the sense that limiting $\lambda_1 \to \lambda_2$ in the first formula or $b \to 0$ in the last formula produces the middle formula for real equal roots. The first formula is also valid for complex conjugate roots λ_1, $\lambda_2 = \overline{\lambda}_1$ and it reduces to the third when $\lambda_1 = a + ib$, therefore the third formula is technically redundant, but nevertheless useful, because it contains no complex numbers.

Recommended: Memorize the first formula, derive the other two.

About the Newton Quotient. The Newton quotient $\frac{g(x)-g(x_0)}{x-x_0}$ in the first formula of the theorem uses $g(x) = e^{xt}$, $x = \lambda_2$, $x_0 = \lambda_1$, $x - x_0 = \lambda_2 - \lambda_1$. Calculus defines $g'(x_0)$ as the Newton quotient limit as $x \to x_0$.

Illustrations

Typical cases are represented by the following 2×2 matrices A. The two roots λ_1, λ_2 of the characteristic equation must fall into one of the three possibilities: real distinct, real equal or complex conjugate.

$\lambda_1 = 5$, $\lambda_2 = 2$ Real distinct roots.

$$A = \begin{pmatrix} -1 & 3 \\ -6 & 8 \end{pmatrix} \qquad \vec{u}(t) = \left(e^{5t} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{e^{2t} - e^{5t}}{2-5} \begin{pmatrix} -6 & 3 \\ -6 & 3 \end{pmatrix} \right) \vec{u}(0).$$

$\lambda_1 = \lambda_2 = 3$ Real equal roots.

$$A = \begin{pmatrix} 2 & 1 \\ -1 & 4 \end{pmatrix} \qquad \vec{u}(t) = e^{3t} \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \vec{u}(0).$$

$\lambda_1 = \overline{\lambda_2} = 2 + 3i$ Complex conjugate roots.

$$A = \begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix} \qquad \vec{u}(t) = e^{2t} \begin{pmatrix} \cos 3t & \sin 3t \\ -\sin 3t & \cos 3t \end{pmatrix} \vec{u}(0).$$

Isolated Equilibria

An autonomous system is said to have an **isolated equilibrium** at $\vec{u} = \vec{u}_0$ provided $\vec{u}_0$ is the only constant solution of the system in $|\vec{u} - \vec{u}_0| < r$, for $r > 0$ sufficiently small.

Theorem 10.3 (Isolated Equilibrium)
The following are equivalent for a constant planar system $\frac{d}{dt}\vec{u}(t) = A\vec{u}(t)$:

1. The system has an isolated equilibrium at $\vec{u} = \vec{0}$.

2. $\det(A) \neq 0$.

3. The roots λ_1, λ_2 of $\det(A - \lambda I) = 0$ satisfy $\lambda_1\lambda_2 \neq 0$.

Proof: The expansion $\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2$ shows that $\det(A) = \lambda_1\lambda_2$. Hence **2** $\equiv$ **3**. We prove now **1** $\equiv$ **2**. If $\det(A) = 0$, then $A\vec{u} = \vec{0}$ has infinitely many solutions $\vec{u}$ on a line through $\vec{0}$, therefore $\vec{u} = \vec{0}$ is not an isolated equilibrium. If $\det(A) \neq 0$, then $A\vec{u} = \vec{0}$ has exactly one solution $\vec{u} = \vec{0}$, so the system has an isolated equilibrium at $\vec{u} = \vec{0}$.

Classification of Isolated Equilibria

For linear equations

$$\frac{d}{dt}\vec{u}(t) = A\vec{u}(t),$$

we explain the phase portrait classifications

spiral, center, saddle, node

near the isolated equilibrium point $\vec{u} = \vec{0}$, and how to detect them when they occur. Below, λ_1, λ_2 are the roots of $\det(A - \lambda I) = 0$.

Figures 13–12 illustrate the classifications. See also duplicate Figures 16–19, which are organized by geometry.

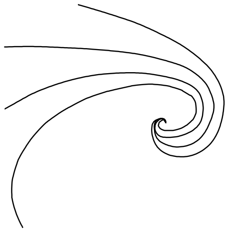


Figure 11. Spiral

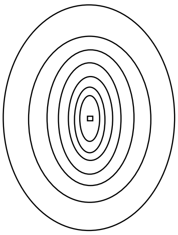


Figure 12. Center

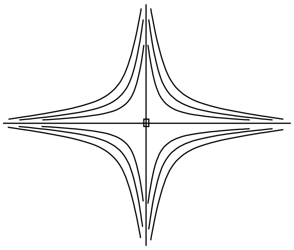


Figure 13. Saddle

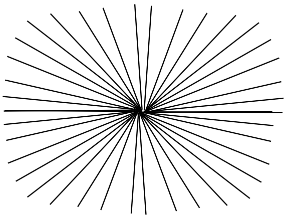


Figure 14. Proper node

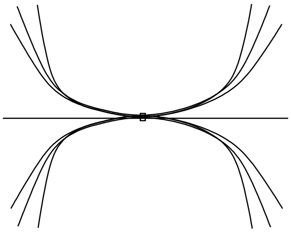


Figure 15. Improper node

Spiral $\lambda_1 = \overline{\lambda}_2 = a + ib$ complex, $a \neq 0$, $b > 0$.

A **Spiral** has solution formula

$$\begin{aligned} \vec{u}(t) &= e^{at}\cos(bt)\,\vec{c}_1 + e^{at}\sin(bt)\,\vec{c}_2, \\ \vec{c}_1 &= \vec{u}(0), \quad \vec{c}_2 = \frac{A - aI}{b}\,\vec{u}(0). \end{aligned}$$

All solutions are bounded harmonic oscillations of natural frequency b times an exponential amplitude which grows if $a > 0$ and decays if $a < 0$. An orbit in the phase plane **spirals out** if $a > 0$ and **spirals in** if $a < 0$.

Center $\lambda_1 = \overline{\lambda}_2 = a + ib$ complex, $a = 0$, $b > 0$

A **center** has solution formula

$$\begin{aligned} \vec{u}(t) &= \cos(bt)\,\vec{c}_1 + \sin(bt)\,\vec{c}_2, \\ \vec{c}_1 &= \vec{u}(0), \quad \vec{c}_2 = \frac{1}{b}\,A\vec{u}(0). \end{aligned}$$

All solutions are bounded harmonic oscillations of natural frequency b. Orbits in the phase plane are periodic closed curves of period $2\pi/b$ which encircle the origin.

Saddle λ_1, λ_2 real, $\lambda_1\lambda_2 < 0$

A **saddle** has solution formula

$$\begin{aligned} \vec{u}(t) &= e^{\lambda_1 t}\vec{c}_1 + e^{\lambda_2 t}\vec{c}_2, \\ \vec{c}_1 &= \frac{A - \lambda_2 I}{\lambda_1 - \lambda_2}\,\vec{u}(0), \quad \vec{c}_2 = \frac{A - \lambda_1 I}{\lambda_2 - \lambda_1}\,\vec{u}(0). \end{aligned}$$

The phase portrait shows two lines through the origin which are tangents at $t = \pm\infty$ for all orbits.
The line directions are given by the eigenvectors of matrix A. See Figure 13.

Node λ_1, λ_2 real, $\lambda_1\lambda_2 > 0$

The solution formulas are

$$\begin{aligned} \vec{u}(t) &= e^{\lambda_1 t}\left(\vec{a}_1 + t\vec{a}_2\right), \quad \text{when} \quad \lambda_1 = \lambda_2, \\ \vec{a}_1 &= \vec{u}(0), \quad \vec{a}_2 = (A - \lambda_1 I)\vec{u}(0), \end{aligned}$$

$$\begin{aligned} \vec{u}(t) &= e^{\lambda_1 t}\vec{b}_1 + e^{\lambda_2 t}\vec{b}_2, \quad \text{when} \quad \lambda_1 \neq \lambda_2, \\ \vec{b}_1 &= \frac{A - \lambda_2 I}{\lambda_1 - \lambda_2}\,\vec{u}(0), \quad \vec{b}_2 = \frac{A - \lambda_1 I}{\lambda_2 - \lambda_1}\,\vec{u}(0). \end{aligned}$$

Node subclassifications **proper** and **improper** are discussed below.

Definition 10.1 (Node)
A **node** is defined to be an equilibrium point (x_0, y_0) such that

1. Either $\lim_{t\to\infty}(x(t), y(t)) = (x_0, y_0)$ or else $\lim_{t\to-\infty}(x(t), y(t)) = (x_0, y_0)$, for all initial conditions $(x(0), y(0)$ close to (x_0, y_0).
2. For each initial condition $(x(0), y(0))$ near (x_0, y_0), there exists a straight line L through (x_0, y_0) such that $(x(t), y(t))$ is **tangent** at $t = \infty$ to L. More precisely, line L has a tangent vector $\vec{v}$ and $\lim_{t\to\infty}(x'(t), y'(t)) = c\vec{v}$ for some constant c.

Proper Node. Also called a **Star Node**. Matrix A is required to have two eigenpairs $(\lambda_1, \vec{v}_1), (\lambda_2, \vec{v}_2)$ with $\lambda_1 = \lambda_2$. Then $\vec{u}(0)$ in $\mathcal{R}^2 = \mathbf{span}(\vec{v}_1, \vec{v}_2)$ implies $\vec{u}(0) = c_1\vec{v}_1 + c_2\vec{v}_2$ and $\vec{a}_2 = (A - \lambda_1 I)\vec{u}(0) = \vec{0}$. Therefore, $\vec{u}(t) = e^{\lambda_1 t}\vec{a}_1$ implies trajectories are tangent to the line through $(0, 0)$ in direction $\vec{v} = \vec{a}_1/|\vec{a}_1|$. Because $\vec{u}(0) = \vec{a}_1$ is arbitrary, $\vec{v}$ can be any direction, which explains the star-like phase portrait in Figure 14.

Improper Node with One Eigenpair. The non-diagonalizable case is also called a **Degenerate Node**. Matrix A is required to have just one eigenpair $(\lambda_1, \vec{v}_1)$ and $\lambda_1 = \lambda_2$. Then $\vec{u}'(t) = (\vec{a}_2 + \lambda_1\vec{a}_1 + t\lambda_1\vec{a}_2)e^{\lambda_1 t}$ implies $\vec{u}'(t)/|\vec{u}'(t)| \approx \vec{a}_2/|\vec{a}_2|$ at $|t| = \infty$. Matrix $A - \lambda_1 I$ has rank 1, hence $\mathbf{Image}(A - \lambda_1 I) = \mathbf{span}(\vec{v})$ for some nonzero vector $\vec{v}$. Then $\vec{a}_2 = (A - \lambda_1 I)\vec{u}(0)$ is a multiple of $\vec{v}$. Trajectory $\vec{u}(t)$ is tangent to the line through $(0, 0)$ with direction $\vec{v}$, as in Figure 15.

Improper Node with Distinct Eigenvalues. Discussed here is the first possibility when matrix A has real eigenvalues with $\lambda_2 < \lambda_1 < 0$. Not discussed is the second possibility $\lambda_2 > \lambda_1 > 0$, which has similar details. Then $\vec{u}'(t) = \lambda_1\vec{b}_1 e^{\lambda_1 t} + \lambda_2\vec{b}_2 e^{\lambda_2 t}$ implies $\vec{u}'(t)/|\vec{u}'(t)| \approx \vec{b}_1/|\vec{b}_1|$ at $t = \infty$. In terms of eigenpairs $(\lambda_1, \vec{v}_1), (\lambda_2, \vec{v}_2)$, we compute $\vec{b}_1 = c_1\vec{v}_1$ and $\vec{b}_2 = c_2\vec{v}_2$ where $\vec{u}(0) = c_1\vec{v}_1 + c_2\vec{v}_2$. Trajectory $\vec{u}(t)$ is tangent to the line through $(0, 0)$ with direction $\vec{v}_1$. See Figure 15.

Attractor and Repeller

An equilibrium point is called an **Attractor** provided orbits starting nearby limit to the point as $t \to \infty$. A **Repeller** is an equilibrium point such that orbits starting nearby limit to the point as $t \to -\infty$. Terms like **Attracting node** and **Repelling spiral** are defined analogously.

Linear Classification Shortcut for $\frac{d}{dt}\vec{u} = A\vec{u}$

Presented here is a practical method for deciding the classification of center, spiral, saddle or node for a linear system $\frac{d}{dt}\vec{u} = A\vec{u}$. The method uses just the eigenvalues of A and the corresponding Euler atoms.

Cayley-Hamilton Basis.

A system $\frac{d}{dt}\vec{u} = A\vec{u}$ will have general solution

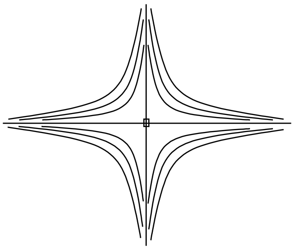
$$\vec{u} = \vec{d_1}(\text{Euler Atom 1}) + \vec{d_2}(\text{Euler Atom 2}).$$

The vectors $\vec{d_1}, \vec{d_2}$ depend on A and $\vec{u}(0)$. They are never explicitly used in the shortcut, hence never computed.

The two Euler solution atoms are found from roots λ of the characteristic equation $|A - \lambda I| = 0$. There are two kinds of atoms:

> No sine or cosine appear in the atoms, making a **non-rotating** phase portrait, which is either a node or a saddle.
>
> Sine and cosine appear in the atoms, which make a **rotating** phase portrait, which is either a center or a spiral.

Table 1. Non-Rotating Phase Portraits



Euler solution atoms for a saddle or node have form e^{at}, e^{bt} or else e^{at}, te^{at}. There are no sine or cosine terms.

Figure 16. Saddle

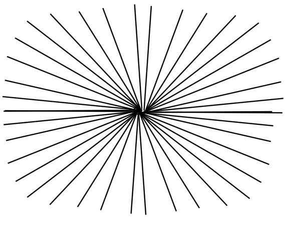


Figure 17. Proper node

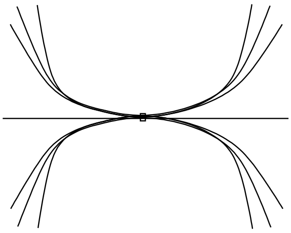


Figure 18. Improper node

Table 2. Rotating Phase Portraits

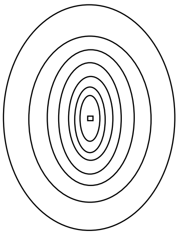


Figure 19. Center

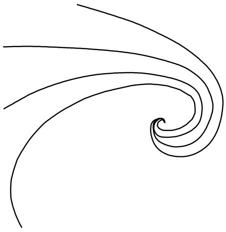


Figure 20. Spiral

Divide and Conquer. Given 2×2 matrix A with $|A| \neq 0$, find the roots of the characteristic equation $|A - \lambda I| = 0$ and construct the two Euler solution atoms. The classification figure, selected from center, spiral, node, saddle, depends only on the atoms. Examine the atoms for sines and cosines. If present, then it will be a rotating figure (center, spiral), otherwise it will be a non-rotating figure (node, saddle). One more divide and conquer decides the figure, because within each figure group, rotating or non-rotating, there is only one possible choice.

> **Rotation Test**. Suppose sines and cosines appear in the Euler atoms. If the Euler atoms are pure sine and cosine, then $(0, 0)$ is a **center**, otherwise $(0, 0)$ is a **spiral**.
>
> **Non-Rotation Test**. Suppose no sines or cosines appear in the Euler atoms. If at $t = \infty$ one Euler atom limits to zero and the other Euler atom limits to infinity, then $(0, 0)$ is a **saddle**, otherwise it is a **node**.

Stability Classification by Euler Atoms.

> A center is always stable, characterized by Euler atoms being pure sine and cosine.
>
> If $(0, 0)$ is not a center, then $(0, 0)$ is stable at $t = \infty$ if and only if both Euler atoms limit to zero at $t = \infty$.

Divide and conquer via Euler atoms requires no table to decide upon the basic phase portrait classification: spiral, center, saddle, node. Stability is likewise decided by Euler atoms.

Node Sub-classifications

If finer geometric sub-classifications of a node are useful to you, then eigenanalysis is required. Assumed below are λ_1, λ_2 real and $\lambda_1\lambda_2 > 0$. *Diagonalizable* means there are two eigenpairs $(\lambda_1, \vec{v}_1), (\lambda_2, \vec{v}_2)$.

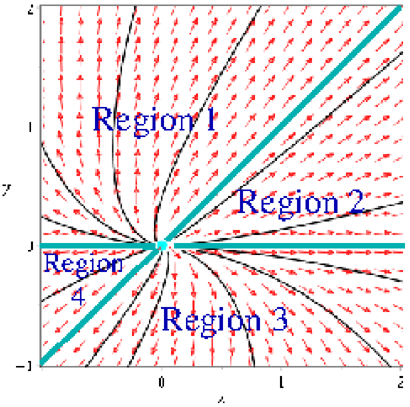
Let $(x_0, y_0) \neq (0, 0)$ denote an arbitrary initial point. Start at this point a trajectory $(x(t), y(t))$. Think of (x_0, y_0) as click point on the graphic in a computer phase portrait plotter: the threaded curve goes through (x_0, y_0).

Separatrix

A **separatrix** is a union S of equilibria and special trajectories. Separatrices are graphing tools. The possible separatrices include every solution curve, so there is art involved to construct a useful separatrix.

Literature may try to describe the phase portrait geometry of linear system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ using **eigenvector directions**. The terminology assumes you know how to construct a separatrix S from the eigenvectors. A separatrix for a nonlinear system $\vec{\mathbf{u}}' = \vec{\mathbf{F}}(\vec{\mathbf{u}})$ is not constructed from eigenvectors but from experimentally found trajectories in a phase portrait plotter.

For nodes, a separatrix S is constructed which divides the plane into two regions or four regions. A trajectory from (x_0, y_0) stays in the region where it starts: **trajectories do not cross** S. If (x_0, y_0) is in S then the trajectory remains in S: crossing means the trajectory changed regions.



Four regions are separated by four cyan lines each of which is a trajectory, their union a separatrix S. The linear system is

$$x' = 2x + y, \quad y' = 3y$$

with eigenpairs

$$\left(2, \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right), \quad \left(3, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$$

The construction for nodes uses eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$ with real nonzero eigenvalues λ_1, λ_2. Let $\vec{\mathbf{v}}_2 = \vec{\mathbf{v}}_1$ if $\lambda_1 = \lambda_2$ and there is only one eigenpair.

Lemma 10.1 A separatrix for a node is $S = \mathbf{span}(\vec{\mathbf{v}}_1) \cup \mathbf{span}(\vec{\mathbf{v}}_2)$.

Proof. Euler's method provides trajectories of $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$:

$$\vec{\mathbf{u}}_1(t) = e^{\lambda_1 t}\vec{\mathbf{v}}_1, \;\; \vec{\mathbf{u}}_2(t) = -e^{\lambda_1 t}\vec{\mathbf{v}}_1, \vec{\mathbf{u}}_3(t) = e^{\lambda_2 t}\vec{\mathbf{v}}_2, \vec{\mathbf{u}}_4(t) = -e^{\lambda_2 t}\vec{\mathbf{v}}_2$$

The separatrix is constructed as the union of equilibrium $(0, 0)$ and the four trajectories, it being understood that $\vec{\mathbf{v}}_1 = \vec{\mathbf{v}}_2$ causes there to be only two trajectories. Then

$$S = (0, 0) \cup \vec{\mathbf{u}}_1 \cup \vec{\mathbf{u}}_2 \cup \vec{\mathbf{u}}_3 \cup \vec{\mathbf{u}}_4 = \mathbf{span}(\vec{\mathbf{v}}_1) \cup \mathbf{span}(\vec{\mathbf{v}}_2)$$

■

The **exceptional case** where the Lemma is not used as a graphing tool is equal eigenvalues $\lambda_1 = \lambda_2$ and independent eigenvectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$. The general solution is $\vec{\mathbf{u}}(t) = (c_1\vec{\mathbf{v}}_1 + c_2\vec{\mathbf{v}}_2)\, e^{\lambda_1 t} = \vec{\mathbf{u}}(0)e^{\lambda_1 t}$. Geometrically, a trajectory starting at (x_0, y_0) traverses for $-\infty < t < \infty$ the ray determined by vector $\vec{\mathbf{u}}(0)$, which is the vector joining $(0, 0)$ to (x_0, y_0). Each such ray is a separatrix in the sense that trajectories cannot cross it. The Lemma is correct: S is a separatrix, but it is not useful for phase plotting. The phase portrait is a **star node**.

Node with Equal Eigenvalues

There are two sub-classifications for a matrix A with real equal eigenvalues $\lambda_1 = \lambda_2$.

> **Star Node**: Matrix A is diagonalizable with $\lambda_1 = \lambda_2 \neq 0$. Trajectories are rays from the origin. Equilibrium $(0, 0)$ is an attractor (or a repeller) from all points (x_0, y_0). Separatrix not used.
>
> **Degenerate Node**: Matrix A is not diagonalizable with $\lambda_1 = \lambda_2 \neq 0$ and one eigenpair $(\lambda_1, \vec{v}_1)$. Equilibrium $(0, 0)$ is an attractor (or a repeller) from all points (x_0, y_0). A threaded trajectory from (x_0, y_0) does not cross separatrix $S = \mathbf{span}(\vec{\mathbf{v}}_1)$, which is the union of $(0, 0)$ and two trajectories.

Node with Unequal Eigenvalues

Matrix A has two eigenpairs $(\lambda_1, \vec{v}_1), (\lambda_2, \vec{v}_2)$, because $\lambda_1 \neq \lambda_2$. Define separatrix $S = \mathbf{span}(\vec{\mathbf{v}}_1) \cup \mathbf{span}(\vec{\mathbf{v}}_2)$, which is a union of two lines through the origin separating the plane into four regions. Equilibrium $(0, 0)$ is an attractor (or a repeller) from all (x_0, y_0), the trajectory not crossing separatrix S.

Proper Node and Improper Node Classifications

The classifications **proper** and **improper** organize the possible node phase portraits. This terminology may appear in dynamical system literature.

> **Proper Node**: The equilibrium is an attractor (or repeller) from all (x_0, y_0)s. Phase portrait: *star node*. Separatrix not used.
>
> **Improper Node**: The equilibrium is an attractor (or repeller) from all (x_0, y_0). Separatrix: $S = \mathbf{span}(\vec{\mathbf{v}}_1)$ for one eigenpair $(\lambda_1, \vec{\mathbf{v}}_1)$ and $S = \mathbf{span}(\vec{\mathbf{v}}_1) \cup \mathbf{span}(\vec{\mathbf{v}}_2)$ for two eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$. Trajectories do not cross S. Phase portraits: *degenerate node* and *node with unequal eigenvalues*.

How to sort out the terminology? The rule is: **proper** = **star**. Every non-star node is **improper**. It may help to associate the terminology with phase portrait plots in Figures 17 and 18 on page 772.

Examples and Methods

Example 10.1 (Spiral)

Show the classification details for the spirals represented by the matrices

$$\begin{pmatrix} 5 & 2 \\ -2 & 5 \end{pmatrix}, \quad \begin{pmatrix} -1 & 3 \\ -3 & -1 \end{pmatrix}.$$

Solution: Matrix $\begin{pmatrix} 5 & 2 \\ -2 & 5 \end{pmatrix}$ has characteristic equation $(\lambda - 5)^2 + 4 = 0$. Then $\lambda = 5 \pm 2i$ and the Euler atoms are $e^{5t}\cos(2t), e^{5t}\sin(2t)$. The atoms have sines and cosines, which limits the classification to a center or a spiral. The presence of the exponential factor e^{5t} implies it is not a center, therefore it is a spiral. Because the atoms limit to zero at $t = -\infty$, then $(0, 0)$ is a repeller. Classification: unstable spiral.

Matrix $\begin{pmatrix} -1 & 3 \\ -3 & -1 \end{pmatrix}$ has characteristic equation $(\lambda+1)^2+9 = 0$. Then $\lambda = -1 \pm 3i$ and the Euler atoms are $e^{-t}\cos(3t), e^{-t}\sin(3t)$. The atoms have sines and cosines, which implies rotation, either a center or a spiral. The presence of the exponential factor e^{-t} implies it is not a center, therefore it is a spiral. Because the atoms limit to zero at $t = \infty$, then $(0, 0)$ is an attractor. Classification: stable spiral.

Example 10.2 (Center)

Show the classification details for matrix $\begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}$, which represents a center.

Solution: The characteristic equation $\lambda^2+4 = 0$ has complex roots $\lambda = \pm 2i$. The Euler atoms are $\cos(2t), \sin(2t)$, therefore a rotating figure is expected. Because of pure sines and cosines and no exponentials, the initial classification of spiral or center reduces to a center. Always a center is stable. Classification: stable center.

Example 10.3 (Saddle)

Show the classification details for the saddles represented by the matrices

$$\begin{pmatrix} 5 & 4 \\ 10 & 1 \end{pmatrix}, \quad \begin{pmatrix} -5 & 4 \\ 2 & 1 \end{pmatrix}$$

Solution: We'll use the theorem $|A - \lambda I| = \lambda^2 + \textbf{trace}(A)(-\lambda) + |A|$ to find the characteristic equation. Symbol $\textbf{trace}(A)$ is the sum of the diagonal elements of A and symbol $|A|$ is the determinant of A, evaluated by Sarrus's rule.

The characteristic equations are

$$\lambda^2 - 6\lambda - 35 = 0, \quad \lambda^2 + 4\lambda - 13 = 0.$$

The roots are $3 \pm 2\sqrt{11}$ $(9.6, -3.6)$ and $-2 \pm \sqrt{17}$ $(2.1, -6.1)$, respectively. Therefore, the roots a, b are real with $a > 0$ and $b < 0$. Euler atoms are e^{at}, e^{bt}. The absence of sines and cosines implies the equilibrium $(0, 0)$ is non-rotating, either a saddle or a node. Because one atom limits to ∞ and the other to zero, at $t = \pm\infty$, then $(0, 0)$ is a saddle. A saddle is always unstable. Classifications: $(0, 0)$ is an unstable saddle for both matrices.

Saddles have a separatrix $S = \textbf{span}(\vec{\mathbf{v}}_1) \cup \textbf{span}(\vec{\mathbf{v}}_2)$ that divides the plane into four regions. The analysis follows the node case, $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$ being the eigenvectors. Calculus uses the terminology **asymptotes** to describe S and the limit of a point (x, y) on a saddle graphic as $x^2 + y^2 \to \infty$. For instance, the second matrix has separatrix $S = \textbf{span}\left(\begin{pmatrix} 0.56 \\ 1 \end{pmatrix}\right) \cup \textbf{span}\left(\begin{pmatrix} -3.56 \\ 1 \end{pmatrix}\right)$, the column vectors defining the calculus asymptotes.

Example 10.4 (Node Sub-Classification: Equal Eigenvalues)

Show the node classification details for the matrices $\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}$.

Solution: A 2×2 matrix is called **diagonalizable** provided it has 2 eigenpairs. Then $\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$ is diagonalizable whereas $\begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}$ is not diagonalizable.

The eigenvalues of both matrices are $5, 5$. Euler atoms are the same for both matrices: e^{5t}, te^{5t}. The absence of sines and cosines limits the classification to saddle or node. Because these atoms limit to zero at $t = -\infty$, then $(0,0)$ is a node. For both, $(0,0)$ is a repeller.

Classifications: $\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$ is an **unstable proper node** (*star node*) and $\begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}$ is an **unstable improper node** (*degenerate node*). See page 773. The star node does not use a separatrix as a graphing tool. A separatrix S for the degenerate node is the line through $(0,0)$ with direction $\vec{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, making for two regions separated by S: the upper half-plane and the lower half-plane. Expect orbits to be tangent to S at $t = -\infty$.

Example 10.5 (Node Sub-Classification: Unequal Eigenvalues)

Show the node classification details for the matrices $\begin{pmatrix} -5 & 0 \\ 0 & -7 \end{pmatrix}$, $\begin{pmatrix} 5 & 0 \\ 0 & 7 \end{pmatrix}$.

Solution: Both matrices are diagonal. Each has two independent eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, the columns of the identity matrix. Eigenvalues are the diagonal elements.

Matrix $\begin{pmatrix} -5 & 0 \\ 0 & -7 \end{pmatrix}$ has unequal eigenvalues $-5, -7$ with Euler atoms e^{-5t}, e^{-7t}. Absence of sines and cosines limits the classification to saddle or node. The atoms have limit zero at $t = \infty$, which eliminates the saddle classification and classifies $(0,0)$ as an attractor, a stable improper node. Orbits are tangent at $t = \infty$ to $\pm\vec{v}_1$, eigenvector for $\lambda_1 = -5$. A separatrix S constructed from eigenvectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2$ has four regions: the usual 4 quadrants in the plane.

Matrix $\begin{pmatrix} 5 & 0 \\ 0 & 7 \end{pmatrix}$ has unequal eigenvalues $5, 7$ with Euler atoms e^{5t}, e^{7t}. Absence of sines and cosines limits the classification to saddle or node. The atoms have limit zero at $t = -\infty$, which eliminates the saddle classification. Therefore, $(0,0)$ is a repeller, an unstable improper node. Orbits are tangent to eigenvector $\pm\vec{v}_1$ at $t = -\infty$. A separatrix S is identical to the separatrix for the first matrix, because of identical eigenvectors. ■

Computer Phase Portraits. In computer **node** plots for unequal eigenvalues, an eigenvector direction can be detected from orbit limits at $t = \pm\infty$. Attractors will have the eigenvector direction for eigenvalue λ with $|\lambda|$ smallest. Repellers will have the eigenvector direction for eigenvalue λ with $|\lambda|$ largest.

Exercises 10.2

Planar Constant Linear Systems

1. (Picard's Theorem) Explain why planar solutions don't cross, by appeal to Picard's existence-uniqueness theorem for $\frac{d}{dt}\vec{u} = A\vec{u}$.

2. (Equilibria) System $\frac{d\vec{u}}{dt} = A\vec{u}$ always has solution $\vec{u}(t) = \vec{0}$, so there is always one equilibrium point. Give an example of a matrix A for which there are infinitely many equilibria.

Putzer's Formula

3. (Cayley-Hamilton) Define matrices $\vec{\mathbf{I}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\vec{\mathbf{0}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. Given matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, expand left and right sides to verify the **Cayley-Hamilton identity**
$A^2 - (a+d)A + (ad-bc)\vec{\mathbf{I}} = \vec{\mathbf{0}}$.

4. (Complex Roots) Verify the Putzer solution $\vec{u} = \Phi(t)\vec{u}(0)$ of $\vec{u}' = A\vec{u}$ for complex roots $\lambda_1 = \overline{\lambda_2} = a+bi$, $b > 0$, where $\Phi(t)$ is

$$e^{at}\left(\cos(bt)\, I + (A - aI)\frac{\sin(bt)}{b}\right).$$

5. (Distinct Eigenvalues) Solve

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix}\vec{u}.$$

6. (Real Equal Eigenvalues) Solve

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} 6 & -4 \\ 4 & -2 \end{pmatrix}\vec{u}.$$

7. (Complex Eigenvalues) Solve

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix}\vec{u}.$$

8. (Purely Complex Eigenvalues) Solve

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} 0 & 3 \\ -3 & 0 \end{pmatrix}\vec{u}.$$

Continuity and Redundancy

9. (Real Equal Eigenvalues) Show that limiting $\lambda_2 \to \lambda_1$ in the Putzer formula for distinct eigenvalues gives Putzer's formula for real equal eigenvalues.

10. (Complex Eigenvalues) Assume $\lambda_1 = \overline{\lambda_2} = a + ib$ with $b > 0$. Then Putzer's first formula holds. Show the third formula details for $\Phi(t)$:

$$e^{at}\left(\cos(bt)\, I + (A - aI)\frac{\sin(bt)}{b}\right).$$

Illustrations

11. (Distinct Eigenvalues) Show the details for the solution of

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} -1 & 3 \\ -6 & 8 \end{pmatrix}\vec{u}.$$

12. (Complex Eigenvalues) Show the details for the solution of

$$\frac{d\vec{u}}{dt} = \begin{pmatrix} 2 & 5 \\ -5 & 2 \end{pmatrix}\vec{u}.$$

Isolated Equilibria

13. (Determinant Expansion) Verify that $|A - \lambda I|$ equals

$$\lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2.$$

14. (Infinitely Many Equilibria) Explain why $A\vec{u} = \vec{0}$ has infinitely many solutions when $\det(A) = 0$.

Classification of Equilibria

15. (Rotating Figures) When sines and cosines appear in the Euler atoms, the phase portrait at $(0,0)$ rotates around the origin. Explain precisely why this is true.

16. (Non-Rotating Figures) When sines and cosines do not appear in the Euler atoms, the phase portrait at $(0,0)$ has no rotation. Give a precise explanation.

Attractor and Repeller

17. (Classification) Which of spiral, center, saddle, node can be an attractor or a repeller?

18. (Attractor) Prove that $(0,0)$ is an attractor if and only if the Euler atoms have limit zero at $t=\infty$.

19. (Repeller) Prove that $(0,0)$ is a repeller if and only if the Euler atoms have limit zero at $t=-\infty$.

20. (Center) A center is neither an attractor nor a repeller. Explain, using Euler atoms.

Phase Portrait Linear

Show the classification details for spiral, center, saddle, proper node, improper node. Include for saddle and node a drawing which shows eigenvector directions. Notation: $' = \frac{d}{dt}$.

21. (Spiral)

$$\begin{aligned} x' &= 2x+3y, \\ y' &= -3x+2y. \end{aligned}$$

22. (Center)

$$\begin{aligned} x' &= 3y, \\ y' &= -3x. \end{aligned}$$

23. (Saddle)

$$\begin{aligned} x' &= 3x, \\ y' &= -5y. \end{aligned}$$

24. (Proper Node)

$$\begin{aligned} x' &= 2x, \\ y' &= 2y. \end{aligned}$$

25. (Improper Node: Degenerate)

$$\begin{aligned} x' &= 2x+y, \\ y' &= 2y. \end{aligned}$$

26. (Improper Node: $\lambda_1 \ne \lambda_2$)

$$\begin{aligned} x' &= 2x+y, \\ y' &= 3y. \end{aligned}$$

10.3 Planar Almost Linear Systems

A nonlinear planar autonomous system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ is called **almost linear** at equilibrium point $\vec{u} = \vec{u}_0$ if

$$\vec{F}(\vec{u}) = A(\vec{u} - \vec{u}_0) + \vec{G}(\vec{u}),$$
$$\lim_{\|\vec{u}-\vec{u}_0\| \to 0} \frac{\|\vec{G}(\vec{u})\|}{\|\vec{u} - \vec{u}_0\|} = 0.$$

The function $\vec{G}$ has the same smoothness as $\vec{F}$. We investigate the possibility that a local phase portrait at $\vec{u} = \vec{u}_0$ for the nonlinear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ is graphically identical to the one for the linear system $\vec{v}'(t) = A\vec{v}(t)$ at $\vec{v} = 0$.

The results will apply to **all isolated equilibria** of $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$. This is accomplished by expanding F in a Taylor series about each equilibrium point, which implies that the ideas are applicable to different choices of A and G, depending upon which equilibrium point $\vec{u}_0$ was considered.

Define the **Jacobian matrix** of $\vec{F} = \begin{pmatrix} f \\ g \end{pmatrix}$ at equilibrium point $\vec{u}_0$ by the formula

$$J = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}.$$

Taylor's theorem for functions of two variables says that

$$\vec{F}(\vec{u}) = J(\vec{u} - \vec{u}_0) + \vec{G}(\vec{u})$$

where $\vec{G}(\vec{u})/\|\vec{u} - \vec{u}_0\| \to 0$ as $\|\vec{u} - \vec{u}_0\| \to 0$. Therefore, for $\vec{F}$ continuously differentiable, we may always take $A = J$ to obtain from the almost linear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ its **linearization** $\frac{d}{dt}\vec{v}(t) = A\vec{v}(t)$.

Phase Portrait of an Almost Linear System

For planar almost linear systems $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$, phase portraits have been studied extensively, by Poincaré-Bendixson and a long list of researchers. It is known that only a finite number of local phase portraits are possible near each isolated equilibrium point of the nonlinear system, the library of figures being identical to those possibilities for a linear system $\vec{v}'(t) = A\vec{v}(t)$. A precise statement without proof appears below, followed by a summary that is easier to remember.

Theorem 10.4 (Paste Theorem: Almost Linear Phase Portrait)
Let the planar almost linear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ be given with $\vec{F}(\vec{u}) = A(\vec{u} - \vec{u}_0) + \vec{G}(\vec{u})$ near the isolated equilibrium point $\vec{u}_0$ (an isolated root of $\vec{F}(\vec{u}_0) = \vec{0}$ with $|A| \neq 0$). Let λ_1, λ_2 be the roots of $\det(A - \lambda I) = 0$. Then:

1. If $\lambda_1 = \lambda_2$, then the equilibrium $\vec{u}_0$ of the nonlinear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ is either a node or a spiral. The equilibrium $\vec{u}_0$ is an asymptotically stable attractor if $\lambda_1 < 0$ and it is a repeller if $\lambda_1 > 0$. In short, the nonlinear system inherits stability from the linear system.

2. If $\lambda_1 = \overline{\lambda_2} = ib$ with $b > 0$, then the equilibrium $\vec{u}_0$ of the nonlinear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ is either a center or a spiral. The stability of the equilibrium $\vec{u}_0$ cannot be predicted from properties of A.

3. In all other cases, the isolated equilibrium $\vec{u}_0$ has graphically the same local phase portrait as the associated linear system $\frac{d}{dt}\vec{v}(t) = A\vec{v}(t)$ at $\vec{v} = \vec{0}$. In particular, local phase portraits of a saddle, spiral or node can be graphed from the linear system. The nonlinear system inherits locally the linearized system properties of stability and instability.

Paste Theorem Summary: The linearized phase portrait **locally pastes** onto the nonlinear phase portrait **with two exceptions**:
(1) Nodes from equal roots cause pasting of either a node or spiral.
(2) Centers (complex roots $\pm ib$) cause pasting a center or spiral.
Local **stability** and **instability** are inherited except for a center.

Classification of Almost Linear Equilibria

A system $\frac{d}{dt}\vec{u}(t) = A(\vec{u}(t) - \vec{u}_0) + \vec{G}(\vec{u}(t))$ has a local phase portrait determined by the linear system $\vec{v}\,'(t) = A\vec{v}(t)$, except in the case when the roots λ_1, λ_2 of the characteristic equation $\det(A - \lambda I) = 0$ are equal or purely imaginary (see Theorem 10.4). To summarize:

Table 3. Equilibria classification for almost linear systems

Eigenvalues of A	Nonlinear Classification
$\lambda_1 < 0 < \lambda_2$	Unstable saddle
$\lambda_1 < \lambda_2 < 0$	Stable improper node
$\lambda_1 > \lambda_2 > 0$	Unstable improper node
$\lambda_1 = \lambda_2 < 0$	Stable node or spiral
$\lambda_1 = \lambda_2 > 0$	Unstable node or spiral
$\lambda_1 = \overline{\lambda_2} = a + ib$, $a < 0$, $b > 0$	Stable spiral
$\lambda_1 = \overline{\lambda_2} = a + ib$, $a > 0$, $b > 0$	Unstable spiral
$\lambda_1 = \overline{\lambda_2} = ib$, $b > 0$	Stable or unstable, center or spiral

Almost Linear Equilibria Geometry

Applied literature may refer to an equilibrium point $\vec{u}_0$ of a nonlinear system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ as a spiral, center, saddle or node. The geometry of these classifications is explained below.

Spiral. To describe a **nonlinear spiral**, we require that an orbit starting on a given ray emanating from the equilibrium point must intersect that ray in infinitely many distinct points on $(-\infty, \infty)$.

Intuition. Basic understanding of a **nonlinear spiral** is obtained from a linear example, e.g.,

$$\frac{d}{dt}\vec{u}(t) = \begin{pmatrix} -1 & 2 \\ -2 & -1 \end{pmatrix} \vec{u}(t).$$

An orbit has component solutions

$$x(t) = e^{-t}(A\cos 2t + B\sin 2t), \quad y(t) = e^{-t}(-A\sin 2t + B\cos 2t)$$

which oscillate infinity often on $(-\infty, \infty)$, rotating around equilibrium point $(0, 0)$ with amplitude Ce^{-t}, for some constant $C > 0$.

Center. Local orbits are periodic solutions. Each local orbit is a closed curve which forms a planar region with boundary, having the equilibrium point interior. As the periodic orbits shrink, the planar region also shrinks, limiting as a planar set to the equilibrium point. Drawings often portray the periodic orbit as a convex figure, but this is not correct, in general, because the periodic orbit can have any shape. In particular, the linearized system may have phase portrait consisting of concentric circles, but the nonlinear phase portrait has no such exact geometric structure.

Saddle. The term implies that *locally* the phase portrait looks like a linear saddle. In nonlinear phase portraits, the straight lines to which orbits are asymptotic appear to be curves instead. These curves are called **separatrices**, which are generally unions of certain orbits and equilibria.

Node. Each orbit starting near the equilibrium is expected to limit to the equilibrium at either $t = \infty$ (stable attractor) or $t = -\infty$ (unstable repeller), in a fashion asymptotic to a direction $\vec{v}$. The terminology applies when the linearized system is a **proper node** (a.k.a. *star node*), in which case there is an orbit asymptotic to $\vec{v}$ for every direction $\vec{v}$. If there is only one direction $\vec{v}$ possible, or all orbits are asymptotic to just one separatrix, then the equilibrium is classified as an **improper node**. The term *degenerate node* applies to a subclass of improper nodes – see Example 10.4 page 776.

Pasting Figures to make a Nonlinear Phase Portrait

The plan provided by the theorem is to paste a library source figure, one of spiral, center, saddle or node, overlaying $(0, 0)$ in the source figure atop equilibrium point $\vec{u} = \vec{u}_0$ in the nonlinear phase portrait. Some observations follow, about what works and what fails.

1. The local paste is valid to graphical resolution near $\vec{u} = \vec{u}_0$, and invalid far away from the equilibrium point.

2. The pasted figure can mutate into a spiral, if the source figure is either a center, or else a node with $\lambda_1 = \lambda_2$. Otherwise, saddle, spiral and node locally paste into saddle, spiral, node.

3. Stability of the source figure is inherited by the nonlinear portrait, except when the source is a center. In this one exceptional case, no stability conclusion can be drawn. However, an attractor or repeller source figure always pastes into an attractor or a repeller.

Examples and Methods

Example 10.6 (Compute Isolated Equilibria)
Find all equilibria for the nonlinear system

$$x'(t) = x(t) + y(t), \quad y'(t) = 1 - x^2(t).$$

Solution: Equilibria are constant solutions, obtained formally by setting $x' = y' = 0$ in the two differential equations $x' = x + y, y' = 1 - x^2$. Then solve for constants x, y. The details:

Set $x' = 0$	$0 = x + y$
Set $y' = 0$	$0 = 1 - x^2$
Solve for x, y	$x = \pm 1$, $y = -x$.
Equilibria	$(1, -1)$ and $(-1, 1)$

Example 10.7 (Linearization at Equilibria)
Find the two linearizations at equilibria $(1, -1), (-1, 1)$ for the nonlinear system

$$x'(t) = x(t) + y(t), \quad y'(t) = 1 - x^2(t).$$

Solution: The system of differential equations is written with function notation in the form $x' = f(x, y), y' = g(x, y)$. Then

$$f(x, y) = x + y, \quad g(x, y) = 1 - x^2.$$

The Jacobian matrix

$$J(x, y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}$$

is computed with symbols x, y, f, g as follows.

Partial derivative $f_x(x, y)$:	$f_x = \partial_x(x + y) = 1 + 0 = 1$
Partial derivative $g_x(x, y)$:	$g_x = \partial_x(1 - x^2) = 0 - 2x = -2x$
Partial derivative $f_y(x, y)$:	$f_y = \partial_y(x + y) = 0 + 1 = 1$
Partial derivative $g_y(x, y)$:	$g_y = \partial_y(1 - x^2) = 0 - 0 = 0$

Then

$$J(x,y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -2x & 0 \end{pmatrix}.$$

The symbols x, y are used for the two substitutions: $x = 1, y = -1$ and $x = -1, y = 1$.

$$J(1,-1) = \begin{pmatrix} 1 & 1 \\ -2 & 0 \end{pmatrix}, \quad J(-1,1) = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}.$$

The two linearized problems are

$$\frac{d}{dt}\vec{u} = \begin{pmatrix} 1 & 1 \\ -2 & 0 \end{pmatrix}\vec{u}, \quad \frac{d}{dt}\vec{u} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}\vec{u}.$$

Example 10.8 (Classification of Linearized Problems)
Classify the two linear problems

$$\frac{d}{dt}\vec{u} = \begin{pmatrix} 1 & 1 \\ -2 & 0 \end{pmatrix}\vec{u}, \quad \frac{d}{dt}\vec{u} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}\vec{u}.$$

Solution:

The answers: $\begin{pmatrix} 1 & 1 \\ -2 & 0 \end{pmatrix}$ is an unstable spiral; $\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}$ is an unstable saddle.

The two characteristic equations are $\lambda^2 - \lambda + 2 = 0$ and $\lambda^2 + \lambda + 2 = 0$ with roots, respectively, $\frac{1}{2} \pm i\frac{\sqrt{7}}{2}$ and $2, -1$. According to the classification theory, page 769, the equilibrium $(0,0)$ is respectively an unstable spiral or an unstable saddle.

Example 10.9 (Pasting Linear Portraits onto Nonlinear Portraits)
Classify equilibria $(1,-1), (-1,1)$ for the nonlinear system

$$x'(t) = x(t) + y(t), \quad y'(t) = 1 - x^2(t),$$

as *nonlinear* spiral, center, saddle or node. Paste the linear portraits onto the nonlinear direction field portrait for Jacobians $J(-1,1), J(1,-1)$, if possible.

Solution: **Classifications**: $(-1,1)$ is a nonlinear unstable saddle; $(1,-1)$ is a nonlinear unstable spiral.

Previous examples show that for the linearized problems, $(-1,1)$ is an unstable saddle and $(-1,1)$ is an unstable spiral. Theorem 10.4 applies to conclude that the two linear phase portraits directly transfer onto the nonlinear phase portrait. This means that $(0,0)$ in each source figure can be pasted atop the corresponding equilibrium point in the nonlinear system, the pasted figure valid locally.

Computer phase portraits show the two pasted library figures with automatic fine tuning. Especially, the saddle will be tuned, because a library source figure usually has asymptotes parallel to the coordinate axes, whereas the computer graphic will show tuned asymptotes in eigenvector directions.

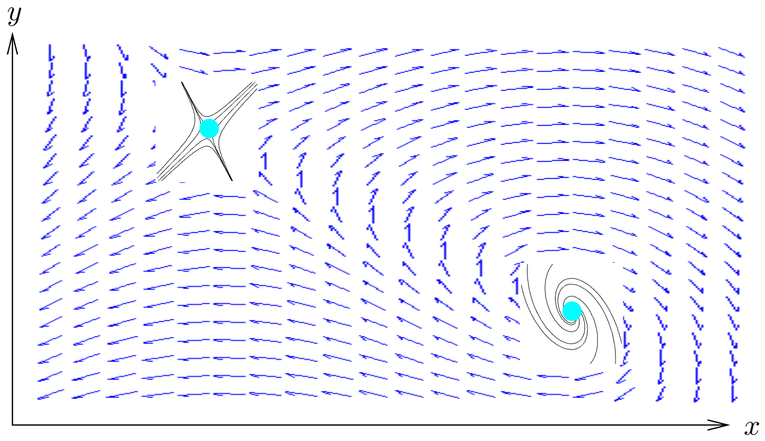


Figure 21. Pasting Source Figures onto a Nonlinear Phase portrait. Saddle at $(-1,1)$, spiral at $(1,-1)$. The saddle source uses a linear phase portrait for $\frac{d}{dt}\vec{v} = J(-1,1)\vec{v}$. The standard saddle source can be rotated to match the nonlinear direction field, with a similar result.

Example 10.10 (Trout System)

Consider a trout model for two species x, y:

$$\begin{aligned} x'(t) &= x(-2x - y + 180), \\ y'(t) &= y(-x - 2y + 120). \end{aligned}$$

The equilibria are $(0,0), (90,0), (0,60), (80,20)$. Find the linearized problem for each equilibrium, then make a tuned computer plot.

Solution:

System Form. Let $f(x,y) = x(-2x - y + 180), g(x,y) = y(-x - 2y + 120)$ to convert to system form $x' = f(x,y), y' = g(x,y)$.

Jacobian Matrix. Use symbols f, g, x, y to compute the Jacobian $J(x,y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}$.

$f_x = \frac{\partial}{\partial x}\left(-2x^2 - xy + 180x\right) = -4x - y - 180$

$f_y = \frac{\partial}{\partial y}\left(-2x^2 - xy + 180x\right) = -x$

$g_x = \frac{\partial}{\partial x}\left(-xy - 2y^2 + 120y\right) = -y$

$g_y = \frac{\partial}{\partial y}\left(-xy - 2y^2 + 120y\right) = -x - 4y + 120$

$J(x,y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} = \begin{pmatrix} -4x - y - 180 & -x \\ -y & -x - 4y + 120 \end{pmatrix}$

Equilibria. To find the equilibria, formally set $x' = y' = 0$. Details:

$x' = 0 = f(x,y)$ becomes $x(-2x - y + 180) = 0$

$y' = 0 = g(x,y)$ becomes $y(-x - 2y + 120) = 0$

Set the factors to zero, in four possible ways, to obtain the solutions

$$x = y = 0, \quad x = 0, y = 60, \quad x = 90, y = 0, \quad x = 80, y = 20.$$

Linearized Differential Equations. The linear problems $\frac{d}{dt}\vec{u} = J(x_0, y_0)\vec{u}$ at equilibria $(0,0), (0,60), (90,0), (80,20)$ are created from the four Jacobian matrices

$$J(0,0) = \begin{pmatrix} -180 & 0 \\ 0 & 120 \end{pmatrix}, \qquad J(0,60) = \begin{pmatrix} 120 & 0 \\ -60 & -120 \end{pmatrix},$$
$$J(90,0) = \begin{pmatrix} -180 & -90 \\ 0 & 30 \end{pmatrix}, \qquad J(80,20) = \begin{pmatrix} -160 & -80 \\ -20 & -40 \end{pmatrix}.$$

Eigenvalues. Answers for the four matrices are respectively:

$$120, 180; \quad 120, -120; \quad 30, -180; \quad -27.89, -172.11$$

Linear Classifications. Because there are no complex eigenvalues, then the possible linear phase portraits are either saddle or node. Checking limits of Euler atoms at $t = \infty$ reveals the classifications unstable node, saddle, saddle, stable node. No equal eigenvalues implies both nodes are **improper**.

Paste Theorem. All linear source figures paste directly onto the nonlinear phase portrait with stability properties inherited. See Theorem 10.4.

Eigenvectors help understanding of the phase portrait. In all four figures, asymptote directions are along an eigenvector. For instance, at $(80,20)$ the two eigenvector directions are $\vec{v}_1 = \begin{pmatrix} -0.6 \\ 1 \end{pmatrix}$, $\vec{v}_2 = \begin{pmatrix} 6.6 \\ 1 \end{pmatrix}$.

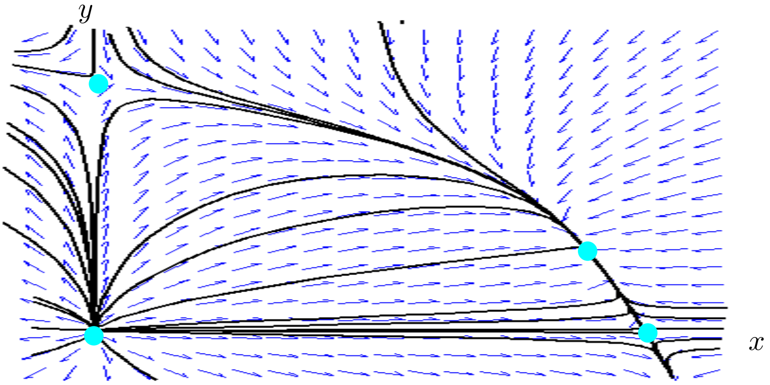


Figure 22. Trout System Phase portrait.
Saddles at $(0,60)$ and $(90,0)$. Improper nodes with unequal eigenvalues at $(0,0)$ and $(80,20)$. A separatrix can be visualized, which connects $(90,0)$ to $(0,0)$ to $(60,0)$ along the coordinate axes, and then to $(80,20)$.

Example 10.11 (Rabbit-Fox System)

Consider a predator-prey model for rabbits $x(t)$ and foxes $y(t)$:

$$\begin{aligned} x' &= \frac{1}{200}x(40-y), \\ y' &= \frac{1}{100}y(x-50). \end{aligned}$$

The equilibria are $(0,0), (50,40)$. Find the linearized problem for each equilibrium, then make a tuned computer plot.

Solution:

System Form. Let $f(x,y) = \frac{1}{200}x(40-y), g(x,y) = \frac{1}{100}y(x-50)$ to convert to system form $x' = f(x,y), y' = g(x,y)$.

Jacobian Matrix. Symbols f, g, x, y are used in the Jacobian $J(x,y) = \begin{pmatrix} f_x \; f_y \\ g_x \; g_y \end{pmatrix}$.

$$f_x = \frac{\partial}{\partial x}\left(x/5 - xy/200\right) = 1/5 - y/200$$

$$f_y = \frac{\partial}{\partial y}\left(x/5 - xy/200\right) = -x/200$$

$$g_x = \frac{\partial}{\partial x}\left(-y/2 + xy/100\right) = y/100$$

$$g_y = \frac{\partial}{\partial y}\left(-y/2 + xy/100\right) = -x - 4y + 120$$

$$J(x,y) = \begin{pmatrix} f_x \; f_y \\ g_x \; g_y \end{pmatrix} = \begin{pmatrix} -4x - y - 180 & -x \\ -y & -x - 4y + 120 \end{pmatrix}$$

Equilibria. To find the equilibria $(0,0), (50,40)$, formally set $x' = y' = 0$. Details:

$0 = f(x,y)$ becomes $\frac{1}{200}x(40-y) = 0$

$0 = g(x,y)$ becomes $\frac{1}{100}y(x-50) = 0$

The solutions are $x = y = 0$ or else $x = 50, y = 40$.

Linearized Differential Equations. The linear problems $\frac{d}{dt}\vec{u} = J(x_0, y_0)\vec{u}$ at equilibria $(0,0), (50,40)$ are created from the two Jacobian matrices

$$J(0,0) = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}, \quad J(50,40) = \begin{pmatrix} 0 & -\frac{1}{4} \\ \frac{2}{5} & 0 \end{pmatrix}.$$

Eigenvalues. The answers are $\frac{1}{5}, -\frac{1}{2}$ and $\pm i/\sqrt{10}$, respectively.

Linear Classifications. Complex eigenvalues imply linear phase portraits of either center or node. Checking Euler atoms reveals the classification **center** at $(50,40)$. Real unequal eigenvalues at $(0,0)$ implies a saddle or node. Checking limits of the Euler atoms at $t = \infty$ implies $(0,0)$ is a **saddle**. Both linear source figures are **stable**.

Paste Theorem. The linear saddle source figure for $(0,0)$ pastes directly onto the nonlinear phase portrait at $(0,0)$ with stability properties inherited. The linear center source figure for $(50,40)$ pastes into a center or a spiral at $(50,40)$. The paste stability or instability is not decided. See Theorem 10.4.

The easiest path to deciding the nonlinear portrait at $(50,40)$ is a computer phase portrait, which shows a center structure.

Eigenvectors help understanding of the phase portrait. At $(0,0)$ the two eigenvector directions are $\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

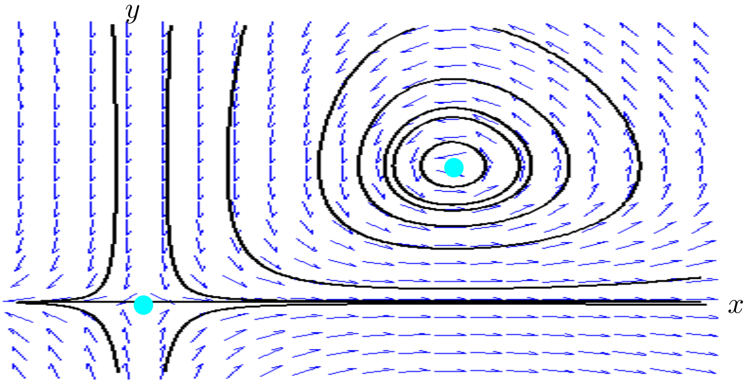


Figure 23. Rabbit-Fox System Phase portrait.
Eigenvector directions for the saddle at $(0, 0)$ are parallel to the coordinate axes. The linear center from $J(50, 40)$ happens to transfer to a nonlinear center at $(50, 40)$.

Exercises 10.3

Almost Linear Systems. Find all equilibria (x_0, y_0) of the given nonlinear system. Then compute the Jacobian matrix $A = J(x_0, y_0)$ for each equilibria.

1. (Spiral and Saddle)

$$\begin{aligned} \tfrac{d}{dt}x &= x + 2y, \\ \tfrac{d}{dt}y &= 1 - x^2. \end{aligned}$$

2. (Two Improper Nodes, Spiral)

$$\begin{aligned} \tfrac{d}{dt}x &= x - 3y + 2xy, \\ \tfrac{d}{dt}y &= 4x - 6y - xy - x^2. \end{aligned}$$

3. (Proper Node, Saddle)

$$\begin{aligned} \tfrac{d}{dt}x &= 3x - 2y - x^2 - y^2, \\ \tfrac{d}{dt}y &= 2x - y. \end{aligned}$$

4. (Center and Three Saddles)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^2 - y^2, \\ \tfrac{d}{dt}y &= 2x - y - xy. \end{aligned}$$

5. (Proper Node and Three Saddles)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^2 - y^2, \\ \tfrac{d}{dt}y &= y - xy. \end{aligned}$$

6. (Degenerate Node, Spiral and Two Saddles)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^3 + y^3, \\ \tfrac{d}{dt}y &= y + 3xy. \end{aligned}$$

7. (Improper Node, Saddle)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^3, \\ \tfrac{d}{dt}y &= 2y + 3xy. \end{aligned}$$

8. (Proper Node and a Saddle)

$$\begin{aligned} \tfrac{d}{dt}x &= 2x + y^3, \\ \tfrac{d}{dt}y &= 2y + 3xy. \end{aligned}$$

Phase Portrait Almost Linear. Linear library phase portraits can be locally pasted atop the equilibria of an almost linear system, with limitations. Apply the theory for the following examples. Complete the phase diagram by computer, thereby resolving the possible mutation of a center or node into a spiral. Label eigenvector directions where it makes sense.

9. (Center and Three Saddles)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^2 - y^2, \\ \tfrac{d}{dt}y &= 2x - y - xy. \end{aligned}$$

10. (Degenerate Node, Three Saddles)

$$\begin{aligned} \tfrac{d}{dt}x &= x - y + x^2 - y^2, \\ \tfrac{d}{dt}y &= y - xy. \end{aligned}$$

11. (Degenerate Node, Spiral, Two Saddles)

$$\begin{aligned}\tfrac{d}{dt}x &= x - y + x^3 + y^3,\\ \tfrac{d}{dt}y &= y + 3xy.\end{aligned}$$

12. (Improper Node, Saddle)

$$\begin{aligned}\tfrac{d}{dt}x &= x - y + x^3,\\ \tfrac{d}{dt}y &= 2y + 3xy.\end{aligned}$$

13. (Proper Node, Saddle)

$$\begin{aligned}\tfrac{d}{dt}x &= 2x + y^3,\\ \tfrac{d}{dt}y &= 2y + 3xy.\end{aligned}$$

14. (Two Improper Nodes and Two Saddles)

$$\begin{aligned}\tfrac{d}{dt}x &= (120 - 4x - 2y)x,\\ \tfrac{d}{dt}y &= (60 - x - 2y)y\end{aligned}$$

Classification of Almost Linear Equilibria. With computer assist, find and classify the nonlinear equilibria.

15. (Co-existing Species)

$$\begin{aligned}x'(t) &= x(t)(24 - 2x(t) - y(t)),\\ y'(t) &= y(t)(30 - 2y(t) - x(t)).\end{aligned}$$

16. (Doomsday-Extinction)

$$\begin{aligned}x'(t) &= x(t)(x(t) - y(t) - 4),\\ y'(t) &= y(t)(x(t) + y(t) - 8).\end{aligned}$$

Almost Linear Geometry. A separatrix S is a union of curves and equilibria. Ideally, orbits limit to S. With computer assist, make a plot of threaded curves which identify one or more separatrices near the equilibrium.

17. (Saddle $(-1, 1)$**)**

$$\begin{aligned}\tfrac{d}{dt}x &= x + y,\\ \tfrac{d}{dt}y &= 1 - x^2.\end{aligned}$$

18. (Saddle $(-1/5, -2/5)$**)**

$$\begin{aligned}\tfrac{d}{dt}x &= 3x - 2y - x^2 - y^2,\\ \tfrac{d}{dt}y &= 2x - y.\end{aligned}$$

19. (Saddle $(-2/3, \sqrt[3]{4/3})$**)**

$$\begin{aligned}\tfrac{d}{dt}x &= 2x + y^3,\\ \tfrac{d}{dt}y &= 2y + 3xy.\end{aligned}$$

20. (Degenerate Improper Node $(0, 0)$**)**

$$\begin{aligned}\tfrac{d}{dt}x &= x - y + x^3 + y^3,\\ \tfrac{d}{dt}y &= y + 3xy.\end{aligned}$$

Rayleigh and van der Pol. Each example below has a unique periodic orbit surrounding an equilibrium point that is the limit at $t = \infty$ of any other orbit. Discuss the spiral repeller at $(0, 0)$ in the attached figure, from the linearized problem at $(0, 0)$ and **Paste Theorem** 10.4. Create a phase portrait with computer assist for the nonlinear problem.

21. (Lord Rayleigh 1877, Clarinet Reed Model)

$$\begin{aligned}\tfrac{d}{dt}x &= y,\\ \tfrac{d}{dt}y &= -x + y - y^3.\end{aligned}$$

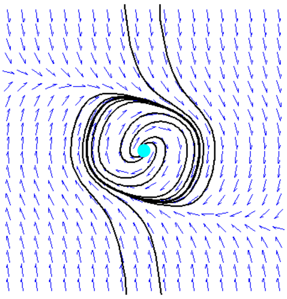


Figure 24. Clarinet Reed.

22. (van der Pol 1924, Radio Oscillator Circuit Model)

$$\begin{aligned}\tfrac{d}{dt}x &= y,\\ \tfrac{d}{dt}y &= -x + (1 - x^2)y.\end{aligned}$$

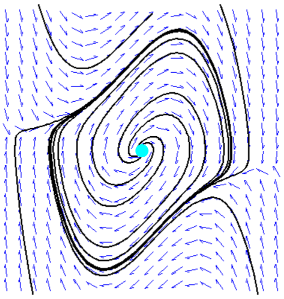


Figure 25. Oscillator Circuit.

10.4 Biological Models

Studied here are **predator-prey models** and **competition models** for two populations. Assumed as background from Malthus' Law (Chapter 1 Section 1) are the one-dimensional Malthusian model $\frac{d}{dt}P = kP$ and the one-dimensional Verhulst model $\frac{d}{dt}P = (a - bP)P$.

Predator-Prey Models

One species called the **Predator** feeds on the other species called the **Prey**. The prey feeds on some constantly available food supply, e.g., rabbits eat plants and foxes eat rabbits.

Credited with the classical predator-prey model is the Italian mathematician **Vito Volterra** (1860-1940), who worked on cyclic variations in shark and prey-fish populations in the Adriatic sea. The following biological assumptions apply to model a predator-prey system.

Malthusian Growth	The prey population grows according to the growth equation $x'(t) = a\,x(t)$, $a > 0$, in the absence of predators.
Malthusian Decay	The predator population decays according to the decay equation $y'(t) = -b\,y(t)$, $b > 0$, in the absence of prey.
Chance Encounters	The prey decrease population at a rate $-pxy$, $p > 0$, due to chance encounters of predators y with prey x. Predators increase population due to these chance interactions at a rate qxy, $q > 0$.

The interaction terms qxy and $-pxy$ are justified by arguing that the frequency of chance encounters is proportional to the product xy. Biologists explain the proportionality by saying that doubling either population should double the frequency of chance encounters. Adding the Malthusian rates and the chance encounter rates gives the **Volterra predator-prey system**[2]

$$(1) \qquad \begin{aligned} x'(t) &= (a - p\,y(t))x(t), \\ y'(t) &= (q\,x(t) - b)y(t). \end{aligned}$$

The differential equations are displayed in this form in order to emphasize that each of $x(t)$ and $y(t)$ satisfy a scalar first order differential equation $u'(t) = r(t)u(t)$ in which the rate function $r(t)$ depends on time. For initial population sizes near zero, the two differential equations behave very much like the Malthusian growth model $u'(t) = a\,u(t)$ and the Malthusian decay model $u'(t) = -b\,u(t)$. This basic growth/decay property allows us to identify the predator variable y, or the prey variable x, regardless of the order in which the differential equations are

[2] The system is written with prey x and predator y. Alphabetical order **predator-prey** would suggest the variables to be reversed, y and then x. History is otherwise.

written. As viewed from Malthus' law $u' = ru$, the prey population has growth rate $r = a - py$ which gets smaller as the number y of predators grows, resulting in fewer prey. Likewise, the predator population has decay rate $r = -b + qx$, which gets larger as the number x of prey grows, causing increased predation. These are the basic ideas of Verhulst, applied to the individual populations x and y.

System Variables

The system of two differential equations (1) can be written as a planar vector autonomous system

$$\frac{d}{dt}\vec{u} = \vec{F}(\vec{u})$$

where vector functions $\vec{F}$ and $\vec{u}$ are defined by

$$(2) \qquad \vec{F}(\vec{u}) = \begin{pmatrix} (a-py)x \\ (qx-b)y) \end{pmatrix}, \quad \vec{u} = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

The vector function $\vec{F}$ is everywhere defined and continuously differentiable. The Picard–Lindelöf theorem provides existence-uniqueness.

A planar vector autonomous system $\frac{d}{dt}\vec{u} = \vec{F}(\vec{u})$ can be written in standard scalar system form

$$x' = f(x,y), \quad y' = g(x,y)$$

by providing definitions for $f(x,y)$ and $g(x,y)$. For predator-prey system (1), the definitions are

$$f(x,y) = (a - p\,y)x, \quad g(x,y) = (q\,x - b)y.$$

Equilibria

The equilibrium points $\vec{u} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ satisfy $\vec{F}(\vec{u}) = \vec{0}$. For predator-prey system (1), the equilibria are $(0,0)$ and $(b/q, a/p)$, found by solving for x_0, y_0 in the equations $(a - p\,y_0)x_0 = 0$, $(q\,x_0 - b)y_0 = 0$.

Linearized Predator-Prey System

The linearized system at equilibrium (x_0, y_0) is the vector-matrix system $\frac{d}{dt}\vec{v}(t) = A\vec{v}(t)$, where A is the Jacobian matrix $J(x,y)$ evaluated at point $x = x_0, y = y_0$, briefly $A = J(x_0, y_0)$. In terms of system variables[3],

$$J(x_0,y_0) = \begin{pmatrix} f_x(x_0,y_0) & f_y(x_0,y_0) \\ g_x(x_0,y_0) & g_y(x_0,y_0) \end{pmatrix}.$$

[3] Notation f_x means $\partial f/\partial x$, the calculus x-derivative with all other variables held constant.

For the predator-prey system, we start by computing

$$\begin{aligned} f_x &= \frac{\partial}{\partial x}(a\,x - p\,xy) = a - p\,y, & f_y &= \frac{\partial}{\partial y}(a\,x - p\,xy) = 0 - p\,x, \\ g_x &= \frac{\partial}{\partial x}(q\,xy - b\,y) = q\,y - 0, & g_y &= \frac{\partial}{\partial y}(q\,xy - b\,y) = q\,x - b. \end{aligned}$$

The Jacobian matrix is given explicitly by

$$(3) \qquad J(x,y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} = \begin{pmatrix} a - p\,y & -p\,x \\ q\,y & q\,x - b \end{pmatrix}.$$

The matrix J is evaluated at equilibrium points $(0,0), (b/q, a/p)$ to obtain the 2×2 matrices for the linearized systems:

$$J(0,0) = \begin{pmatrix} a & 0 \\ 0 & -b \end{pmatrix}, \quad J(b/q, a/p) = \begin{pmatrix} 0 & -bp/q \\ aq/p & 0 \end{pmatrix}.$$

The linearized systems $\vec{v}'(t) = A\vec{v}(t)$ are:

Equilibrium $(0,0)$ $\qquad \frac{d}{dt}\vec{u}(t) = \begin{pmatrix} a & 0 \\ 0 & -b \end{pmatrix} \vec{u}(t)$

Equilibrium $(b/q, a/p)$ $\qquad \frac{d}{dt}\vec{u}(t) = \begin{pmatrix} 0 & -bp/q \\ aq/p & 0 \end{pmatrix} \vec{u}(t)$

Saddle $J(0,0)$. Matrix $\begin{pmatrix} a & 0 \\ 0 & -b \end{pmatrix}$ has unequal real eigenvalues $a, -b$ and associated Euler atoms e^{at}, e^{-bt}. No rotation implies a saddle or node, but limits at infinity imply a linear **saddle**. The **Paste Theorem** implies system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ has a saddle at equilibrium $(0,0)$.

Center $J(b/q, a/p)$. Matrix $\begin{pmatrix} 0 & -bp/q \\ aq/p & 0 \end{pmatrix}$ has eigenvalues $\lambda = \pm i\sqrt{ab}$ and associated Euler atoms $\cos(t\sqrt{ab}), \cos(t\sqrt{ab})$. Pure rotation (no exponential factor) implies a linear **center**. The **Paste Theorem** implies system $\frac{d}{dt}\vec{u}(t) = \vec{F}(\vec{u}(t))$ has either a center or a spiral at equilibrium $(b/q, a/p)$.

Shown below in Theorem 10.5 is that **the spiral case does not happen**. The proof of Lemma 10.2 is in the exercises.

Lemma 10.2 (Predator-Prey Implicit Solution)

Let $(x(t), y(t))$ be an orbit of the predator-prey system (1) with $x(0) > 0$ and $y(0) > 0$. Then for some constant C,

$$(4) \qquad a\,\ln|y(t)| + b\,\ln|x(t)| - q\,x(t) - p\,y(t) = C.$$

Theorem 10.5 (Spiral Case Eliminated)

Equilibrium $(b/q, a/p)$ of predator-prey system (1) cannot be a spiral.

Proof: Assume the equilibrium $(b/q, a/p)$ is a spiral point and some orbit touches the line $x = b/q$ in points $(b/q, u_1)$, $(b/q, u_2)$ with $u_1 \neq u_2$, $u_1 > a/p$, $u_2 > a/p$. Consider the energy function $E(u) = a \ln|u| - p\,u$. Due to relation (4), $E(u_1) = E(u_2) = E_0$, where $E_0 \equiv C + b - b \ln|b/q|$. By the Mean Value Theorem of calculus, $dE/du = 0$ at some u between u_1 and u_2. This is a contradiction, because $dE/du = (a - pu)/u$ is strictly negative for $a/p < u < \infty$. Therefore, equilibrium $(b/q, a/p)$ is **not a spiral**. ■

Rabbits and Foxes

An instance of predator-prey theory is a Volterra population model for x rabbits and y foxes given by the system of differential equations

$$\begin{aligned} x'(t) &= \frac{1}{250}\,x(t)(40 - y(t)), \\ y'(t) &= \frac{1}{50}\,y(t)(x(t) - 60). \end{aligned} \tag{5}$$

The equilibria of system (5) are $(0, 0)$ and $(60, 40)$. A phase portrait for system (5) appears in Figure 26.

The linearized system at $(60, 40)$ is

$$\begin{aligned} x'(t) &= -\frac{6}{25}y(t), \\ y'(t) &= \frac{4}{5}x(t). \end{aligned}$$

This system has eigenvalues $\pm i\sqrt{24/125}$. The Euler atoms are $\sin(t\sqrt{24/125})$ and $\cos(t\sqrt{24/125})$, which have period $2\pi/\sqrt{24/125} \approx 14.33934302$. The linear classification is a center.

The nonlinear classification at $(60, 40)$ is then a **center**, because of Theorem 10.5. Intuition dictates that the period of smaller and smaller nonlinear orbits enclosing the equilibrium $(60, 40)$ must approach a value that is approximately 14.3.

The fluctuations in population size $x(t)$ are measured graphically by the maximum and minimum values of x in the phase portrait, or more simply, by graphing t versus $x(t)$ in a planar graphic. To illustrate, the orbit for $x(0) = 60$, $y(0) = 100$ is graphed in Figure 27, from which it is determined that the rabbit population $x(t)$ fluctuates between 39 and 87. Similar remarks apply to foxes $y(t)$.

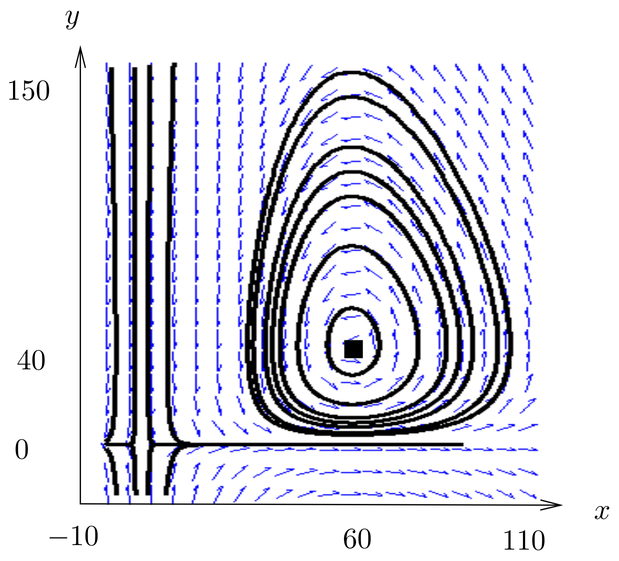


Figure 26. Rabbit and Fox System (5). Equilibria $(0,0)$ and $(60,40)$ are respectively a saddle and a center. The oscillation period is about 17 for the largest orbit and 14.5 for the smallest orbit.

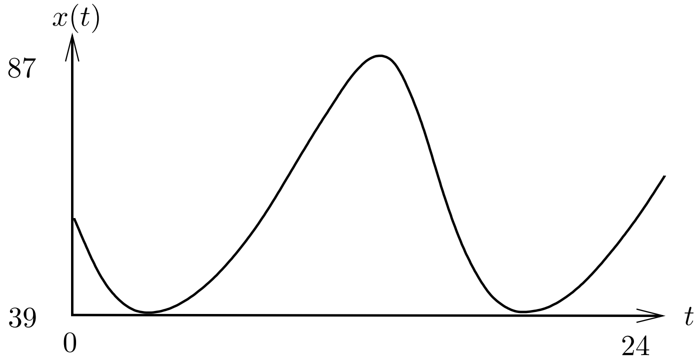


Figure 27. Scene Plot of $x(t)$ Rabbits.
An initial rabbit population of 60 and fox population of 100 causes the rabbit population $x(t)$ to fluctuate from 39 to 87. The plot uses nonlinear equations (5) with $x(0) = 60$, $y(0) = 100$.

Pesticides, Aphids and Ladybugs

The classical predator-prey equations apply for prey *Aphid* $x(t)$ and predator *Ladybug* $y(t)$, which for simplicity are assumed to be

$$\begin{aligned} x'(t) &= (1 - y(t))x(t), \\ y'(t) &= (x(t) - 1)y(t), \end{aligned} \tag{6}$$

with units in millions.

Consider deployment of an indiscriminate pesticide which kills a certain percentage of each insect. Typically available pesticide strengths are $s = 0.5$, $s = 0.75$. Strength $s = 0$ is no pesticide. We will assume hereafter that $0 \leq s < 1$. The predator-prey equations mutate by adding terms for pesticide-caused death rates, resulting in the **Pesticide Model**

$$(7)\qquad \begin{aligned} x'(t) &= (1-y(t))x(t) - s\,x(t), \\ y'(t) &= (x(t)-1)y(t) - s\,y(t). \end{aligned}$$

Explained below in Figures 28, 29 and 30 are the results in the following table.

Table 4. Effects of Pesticide on Aphids and Ladybugs

The aphids increase and the ladybugs decrease.
The insecticide had a counterproductive effect. Aphid damage to the garden plants increased by using a pesticide.

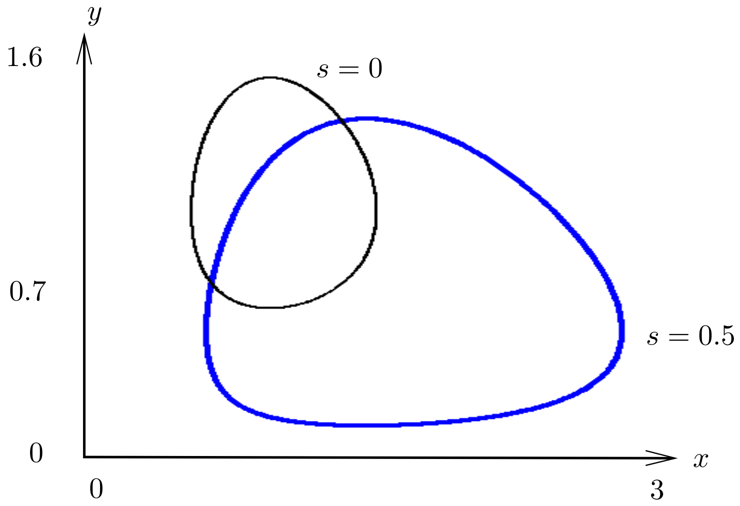


Figure 28. Aphid-Ladybug Portraits $s = 0$, $s = 0.5$.
Aphid population max and min are measured by the orbit width. Ladybug population max and min are measured by the orbit height. Both orbits use $x(0) = y(0) = 0.7$. Details appear in the x and y scene plots, *infra.*

Pesticide model (7) is equivalent to the classical predator-prey system (1) with replacements $a = 1 - s$, $b = 1 + s$. The nonlinear phase portrait for the pesticide model has according to predator-prey theory a saddle at $(0, 0)$ and a center at $(1 + s, 1 - s)$.

The scene plots in Figures 29 and 30 show that the aphids increase and the ladybugs decrease, for the two populations, $x(t)$ **aphids**, $y(t)$ **ladybugs** in pesticide system (7), with pesticide strengths $s = 0$ and $s = 0.5$ and initial populations $x(0) = 0.7$, $y(0) = 0.7$ (in millions).

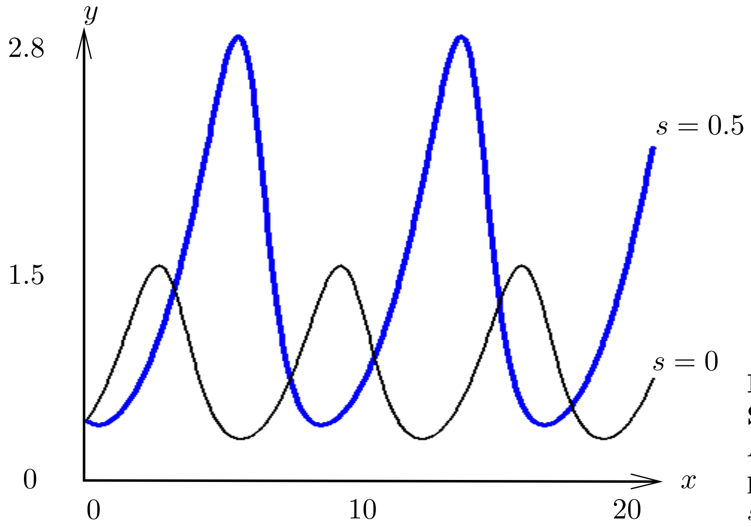


Figure 29. Aphid Scene $x(t)$.
Aphids increase when pesticide strength $s = 0.5$ is applied.

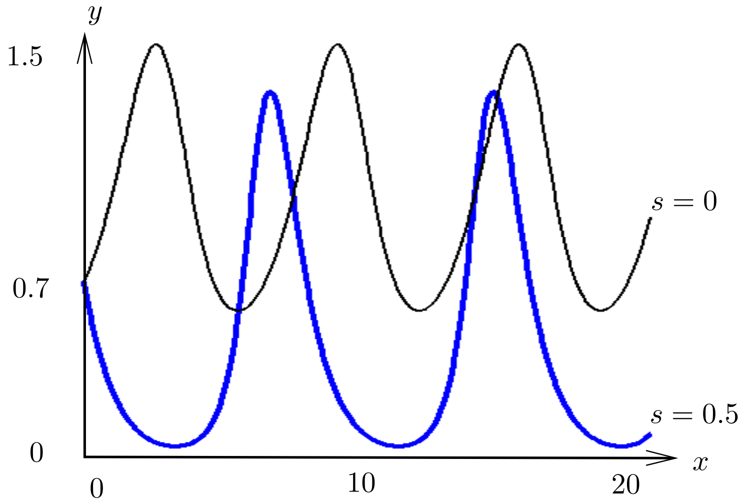


Figure 30. Ladybug Scene $y(t)$.
Ladybugs decrease when pesticide strength $s = 0.5$ is applied.

Competition Models

Two populations **1** and **2** feed on some constantly available food supply, e.g., two kinds of insects feed on fallen fruit. The following biological assumptions apply to model a two-population competition system.

Verhulst model **1**	Population **1** grows or decays according to the logistic equation $x'(t) = (a - bx(t))x(t)$, in the absence of population **2**.
Verhulst model **2**	Population **2** grows or decays according to the logistic equation $y'(t) = (c - dy(t))y(t)$, in the absence of population **1**.

Chance encounters	Population **1** decays at a rate $-pxy$, $p > 0$, due to chance encounters with population **2**. Population **2** decays at a rate $-qxy$, $q > 0$, due to chance encounters with population **1**.

Adding the Verhulst rates and the chance encounter rates gives the **Volterra competition system**

$$\begin{aligned} x'(t) &= (a - bx(t) - py(t))x(t), \\ y'(t) &= (c - dy(t) - qx(t))y(t). \end{aligned} \tag{8}$$

The equations show that each population satisfies a time-varying first order differential equation $u'(t) = r(t)u(t)$ in which the rate function $r(t)$ depends on time. For initial population sizes near zero, the two differential equations essentially reduce to the Malthusian growth models $x'(t) = ax(t)$ and $y'(t) = cy(t)$. As viewed from Malthus' law $u' = ru$, population **1** has growth rate $r = a - bx - py$ which decreases if population **2** grows, resulting in a reduction of population **1**. Likewise, population **2** has growth rate $r = c - dy - qx$, which reduces population **2** as population **1** grows. While a, c are Malthusian growth rates, constants b, d measure **inhibition** (due to lack of food or space) and constants p, q measure **competition**.

Equilibria

The equilibrium points $\vec{u}$ satisfy $\vec{F}(\vec{u}) = \vec{0}$ where $\vec{F}$ is defined by

$$\vec{F}(\vec{u}) = \begin{pmatrix} (a - bx - py)x \\ (c - dy - qx)y \end{pmatrix}, \quad \vec{u} = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{9}$$

To isolate the most important applications, the assumption will be made of exactly four roots in population quadrant I. This is equivalent to the condition $bd - qp \neq 0$ plus all equilibria have nonnegative coordinates.

Three of the four equilibria are found to be $(0, 0)$, $(a/b, 0)$, $(0, c/d)$. The last two represent the carrying capacities of the Verhulst models in the absence of the second population. The fourth equilibrium (x_0, y_0) is found as the *unique root* $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ of the linear system

$$\begin{pmatrix} b & p \\ q & d \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix},$$

which according to Cramer's rule is

$$x_0 = \frac{ad - pc}{bd - qp}, \quad y_0 = \frac{bc - qa}{bd - qp}.$$

Linearized Competition System

The Jacobian matrix $J(x, y)$ is computed from the partial derivatives of system variables f, g, which are found as follows.

$$\begin{array}{lclcr}
f(x,y) & = & (a - b\,x - p\,y)\,x, & = & a\,x - b\,x^2 - p\,xy \\
g(x,y) & = & (c - d\,y - q\,x)\,y & = & c\,y - d\,y^2 - q\,xy \\
f_x & = & \frac{\partial}{\partial x}(a\,x - b\,x^2 - p\,xy) & = & a - 2b\,x - p\,y \\
f_y & = & \frac{\partial}{\partial y}(a\,x - b\,x^2 - p\,xy) & = & -p\,x \\
g_x & = & \frac{\partial}{\partial x}(c\,y - d\,y^2 - q\,xy) & = & -q\,y \\
g_y & = & \frac{\partial}{\partial y}(c\,y - d\,y^2 - q\,xy) & = & c - 2d\,y - q\,x
\end{array}$$

The Jacobian matrix is given explicitly by

$$(10) \qquad J(x,y) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} = \begin{pmatrix} a - 2bx - py & -px \\ -qy & c - 2dy - qx \end{pmatrix}.$$

The matrix J is evaluated at an equilibrium point (a root of $\vec{F}(\vec{u}) = \vec{0}$) to obtain a 2×2 matrix A for the linearized system $\frac{d}{dt}\vec{v}(t) = A\,\vec{v}(t)$. The four linearized systems are:

Equilibrium $(0, 0)$
Nodal Repeller

$$\frac{d}{dt}\vec{u}(t) = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix} \vec{u}(t)$$

Equilibrium $(a/b, 0)$
Saddle or Nodal Attractor

$$\frac{d}{dt}\vec{u}(t) = \begin{pmatrix} -a & -ap/b \\ 0 & c - qa/b \end{pmatrix} \vec{u}(t)$$

Equilibrium $(0, c/d)$
Saddle or Nodal Attractor

$$\frac{d}{dt}\vec{u}(t) = \begin{pmatrix} a - cp/d & 0 \\ -qc/d & -c \end{pmatrix} \vec{u}(t)$$

Equilibrium (x_0, y_0)
Saddle or Nodal Attractor

$$\frac{d}{dt}\vec{u}(t) = \begin{pmatrix} -bx_0 & -px_0 \\ -qy_0 & -dy_0 \end{pmatrix} \vec{u}(t)$$

Equilibria $(a/b, 0)$ and $(0, c/d)$ are either both saddles or both nodal attractors, accordingly as $bd - qp > 0$ or $bd - qp < 0$, because of the requirement that a, b, c, d, p, q, x_0, y_0 be positive.

The analysis of equilibrium (x_0, y_0) is made by computing the eigenvalues λ of the linearized system, from characteristic equation $\lambda^2 + (bx_0 + dy_0)\lambda + (bd - pq)x_0y_0 = 0$, giving

$$\lambda = \frac{1}{2}\left(-(bx_0 + dy_0) \pm \sqrt{D}\right), \quad \text{where} \quad D = (bx_0 - dy_0)^2 + 4pqx_0y_0.$$

Because $D > 0$, the equilibrium is a saddle when the roots have opposite sign, and it is a nodal attractor when both roots are negative. The saddle case is $D > (bx_0 + dy_0)^2$ or equivalently $4x_0y_0(pq - bd) > 0$, which reduces to $bd - qp < 0$. In summary:

If $bd - qp > 0$, then equilibria $(a/b, 0)$, $(0, c/d)$, (x_0, y_0) are respectively a saddle, saddle, nodal attractor.

If $bd - qp < 0$, then equilibria $(a/b, 0)$, $(0, c/d)$, (x_0, y_0) are respectively a nodal attractor, nodal attractor, saddle.

Biological Meaning of $bd - qp$ Negative or Positive

The quantities bd and qp are measures of inhibition and competition.

Survival-Extinction — The inequality $bd - qp < 0$ means that competition qp is large compared with inhibition bd. The equilibrium point (x_0, y_0) is unstable in this case, which biologically means that the two species cannot coexist: **Survival** for one species and **Extinction** for the other species.

Co-existence — The inequality $bd - qp > 0$ means that competition qp is small compared with inhibition bd. The equilibrium point (x_0, y_0) is asymptotically stable in this case, which biologically means the two species **Co-exist**.

Survival of One Species

Consider populations $x(t)$ and $y(t)$ that satisfy the competition model

$$\begin{aligned} x'(t) &= x(t)(24 - x(t) - 2y(t)), \\ y'(t) &= y(t)(30 - y(t) - 2x(t)). \end{aligned} \tag{11}$$

We apply the general competition theory with $a = 24$, $b = 1$, $p = 2$, $c = 30$, $d = 1$, $q = 2$. The equilibrium points are $(0, 0)$, $(0, 30)$, $(24, 0)$, $(12, 6)$, shown in Figure 31 as solid circles and squares. Eigenvalues are computed from Jacobian matrix $J(x, y) = \begin{pmatrix} 24 - 2x - 2y & -2x \\ -2y & 30 - 2y - 2x \end{pmatrix}$ evaluated at the four equilibria. The answers:

Equilibrium $(0, 0)$**:** $\lambda = 24, 30$, nodal repeller.

Equilibrium $(0, 30)$**:** $\lambda = -36, -30$, nodal attractor.

Equilibrium $(24, 0)$**:** $\lambda = -24, -18$, nodal attractor.

Equilibrium $(12, 6)$**:** $\lambda = 8.23, -26.23$, saddle.

The **Paste Theorem** says that the linear portraits can be pasted atop the four equilibria in the nonlinear phase portrait. The tuned portrait appears in Figure 31, clipped to the population quadrant $x \geq 0, y \geq 0$.

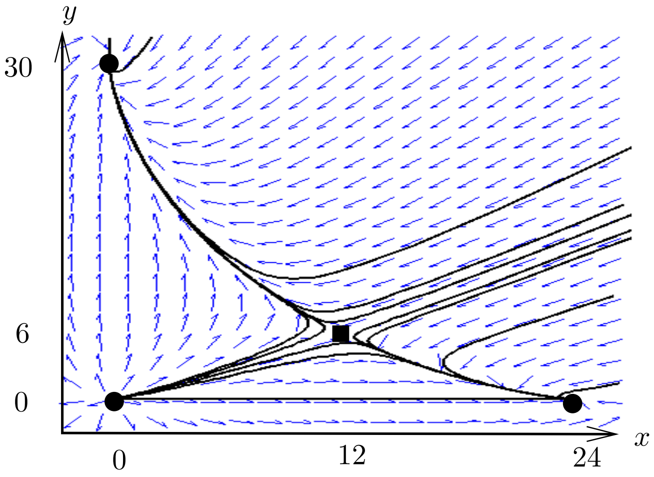


Figure 31. Survival of One Species. Portrait for system (11). Equilibria are $(0,0)$, $(0,30)$, $(24,0)$ and $(12,6)$, classified respectively as nodal repeller, nodal attractor, nodal attractor and saddle. The population with initial advantage survives, while the other dies out.

Co-existence

Consider populations $x(t)$ and $y(t)$ that satisfy the competition model

$$\begin{aligned} x'(t) &= x(t)(24-2x(t)-y(t)), \\ y'(t) &= y(t)(30-2y(t)-x(t)). \end{aligned} \tag{12}$$

We apply the general competition theory with $a=24$, $b=2$, $p=1$, $c=30$, $d=2$, $q=1$. The equilibrium points are $(0,0)$, $(0,15)$, $(12,0)$ and $(6,12)$, shown in Figure 32 as solid circles and squares. Eigenvalues are computed from Jacobian matrix $J(x,y) = \begin{pmatrix} 24-4x-y & -x \\ -y & 30-4y-x \end{pmatrix}$ evaluated at the four equilibria. The answers:

Equilibrium $(0,0)$**:** $\lambda = 24, 30$, nodal repeller.

Equilibrium $(0,30)$**:** $\lambda = 18, -24$, saddle.

Equilibrium $(24,0)$**:** $\lambda = 9, -30$, saddle.

Equilibrium $(12,6)$**:** $\lambda = -7.61, -28.39$, nodal attractor.

The linear portraits can be pasted atop the four equilibria in the nonlinear phase portrait, according to the **Paste Theorem**. Figure 32 is the tuned portrait.

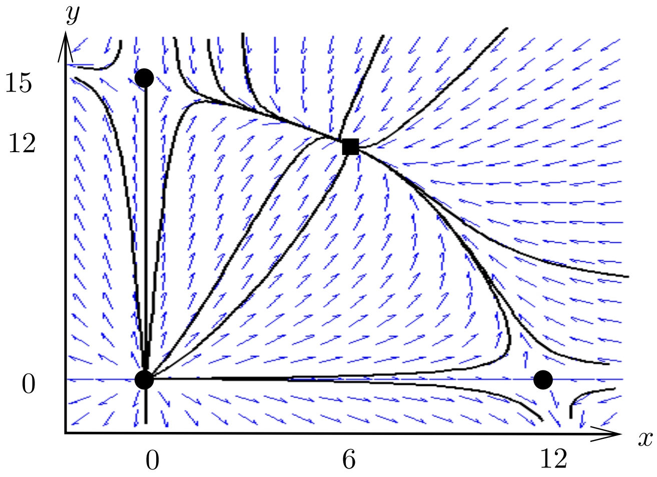


Figure 32. Coexistence.
Phase portrait of system (12). The equilibria are $(0,0)$, $(0,15)$, $(12,0)$ and $(6,12)$, classified respectively as nodal repeller, saddle, saddle, nodal attractor. A solution with $x(0) > 0$, $y(0) > 0$ limits at $t = \infty$ to the solid square $(6,12)$. **Coexistence states** are $x = 6, y = 12$.

Alligators, Explosion and Extinction

Let us assume a competition-type model (8) in which the Verhulst dynamics has explosion-extinction type. Accordingly, the signs of a, b, c, d in (8) are assumed to be negative, but p, q are still positive. The populations $x(t)$ and $y(t)$ are unsophisticated in the sense that each population in the absence of the other is subject to only the possibilities of population explosion or population extinction.

It can be verified for this general setting, although we shall not attempt to do so here, that the population quadrant $x(0) > 0$, $y(0) > 0$ is separated into two regions I and II, whose common boundary is a separatrix consisting of three equilibria and two orbits. An orbit starting in region I will have (a) $x(\infty) = 0$, $y(\infty) = \infty$, or (b) $x(\infty) = \infty$, $y(\infty) = 0$, or (c) $x(\infty) = \infty$, $y(\infty) = \infty$. Orbits starting in region II will satisfy (d) $x(\infty) = 0$, $y(\infty) = 0$. The biological conclusion is that either population explosion or extinction occurs for each population.

Consider the instance

$$\begin{aligned} x'(t) &= x(t)(x(t) - y(t) - 4), \\ y'(t) &= y(t)(x(t) + y(t) - 8). \end{aligned} \tag{13}$$

Let's apply the general competition theory with $a = 24$, $b = 2$, $p = 1$, $c = 30$, $d = 2$, $q = 1$. The equilibria are $(0,0)$, $(0,8)$, $(4,0)$ and $(6,2)$, shown in Figure 33 as solid circles and a square. Eigenvalues λ are computed from Jacobian matrix $J(x,y) = \begin{pmatrix} 2x - y - 4 & -x \\ -y & x + 2y - 8 \end{pmatrix}$ evaluated at the four equilibria. The answers below and the **Paste Theorem** predict the tuned portrait in Figure 33.

Equilibrium $(0,0)$**:** $\lambda = -4, -8$, nodal attractor.
Equilibrium $(0,30)$**:** $\lambda = 8, -12$, saddle.
Equilibrium $(24,0)$**:** $\lambda = 4, -4$, saddle.
Equilibrium $(12,6)$**:** $\lambda = 4 \pm 2.83\,i$, spiral repeller.

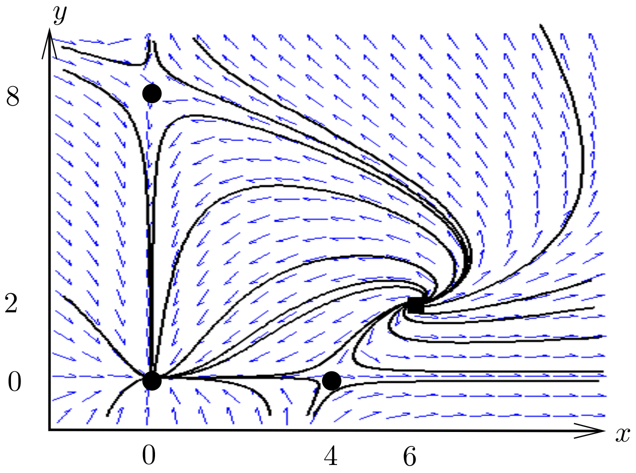


Figure 33. Population Explosion or Extinction. Phase portrait of system (13). The equilibria are $(0,0)$, $(0,8)$, $(4,0)$ and $(6,2)$, classified respectively as nodal attractor, saddle, saddle and spiral repeller. The node and two saddles are marked with a solid disk and the spiral repeller is marked with a solid square.

Exercises 10.4

Predator-Prey Models.

Consider the system

$$\begin{aligned} x'(t) &= \frac{1}{250}(1-2y(t))x(t), \\ y'(t) &= \frac{3}{500}(2x(t)-1)y(t). \end{aligned}$$

1. (System Variables) The system has vector-matrix form

$$\frac{d}{dt}\vec{u} = \vec{F}(\vec{u}(t)).$$

Display formulas for $\vec{u}$ and $\vec{F}$.

2. (System Parameters) Identify the values of a, b, c, d, p, q, as used in the textbook's predator-prey system.

3. (Identify Predator and Prey) Which of $x(t), y(t)$ is the predator?

4. (Switching Predator and Prey) Give an example of a predator-prey system in which $x(t)$ is the predator and $y(t)$ is the prey.

Implicit Solution Predator-Prey.

These exercises prove equation

$$a\,\ln|y| + b\,\ln|x| - q\,x - p\,y = C$$

for predator-prey system

$$\begin{aligned} x'(t) &= (a - p\,y(t))x(t), \\ y'(t) &= (q\,x(t) - b)y(t). \end{aligned}$$

5. (First Order Equation) Verify from the chain rule of calculus the first order equation

$$\frac{dy}{dx} = \frac{y'(t)}{x'(t)} = \frac{y}{x}\frac{qx-b}{a-py}.$$

6. (Separated Variables) Verify

$$\left(\frac{a}{y} - p\right) dy = \left(q - \frac{b}{x}\right) dx.$$

7. (Quadrature) Integrate the equation of Exercise 6 to obtain

$$a\,\ln|y| - p\,y = q\,x - b\,\ln|x| = C.$$

Then re-arrange to obtain the reported implicit solution.

8. (Energy Function) Define $E(t) = a\ln|u| - pu$. Show that $dE/du = (a - pu)/u$. Then show that $dE/du < 0$ for $a > 0$, $p > 0$ and $a/p < u < \infty$.

Linearized Predator-Prey System.

Consider

$$\begin{aligned} x'(t) &= (100 - 2y(t))x(t), \\ y'(t) &= (2x(t) - 160)y(t). \end{aligned}$$

9. (Find Equilibria) Verify equilibria $(0,0), (80,50)$.

10. (Jacobian Matrix) Compute $J(x,y)$ for each x, y. Then find $J(0,0)$ and $J(80,50)$.

11. **(Transit Time)** Find the transit time of an orbit for one loop about $(0,0)$ for system $\frac{d}{dt}\vec{v} = \begin{pmatrix} 0 & -160 \\ 100 & 0 \end{pmatrix} \vec{v}$, the linearization about $(80,50)$.

12. **(Paste Theorem)** Describe the local figures expected near equilibria in the nonlinear phase portrait.

Rabbits and Foxes. Consider

$$\begin{aligned} x'(t) &= \frac{1}{200}\, x(t)(50 - y(t)), \\ y'(t) &= \frac{1}{100}\, y(t)(x(t) - 40). \end{aligned}$$

13. **(Equilibria)** Verify equilibria $(0,0)$, $(40,50)$, showing all details.

14. **(Jacobian)** Compute Jacobian $J(x,y)$, then $J(0,0)$ and $J(40,50)$.

15. **(Rabbit Oscillation)** Find a graphical estimate for the period of oscillation of the rabbit population $x(t)$ for the nonlinear system, given $x(0) = 100$, $y(0) = 60$ and t is in weeks. Answer: about 23 weeks.

16. **(Rabbit-Gerbil Competing Species)** Consider system

$$\begin{aligned} x' &= \left(\tfrac{5}{4} - \tfrac{x}{160} - \tfrac{3y}{1000}\right) x, \\ y' &= \left(3 - \tfrac{3y}{500} - \tfrac{3x}{160}\right) y. \end{aligned}$$

Verify equilibria $(0,0)$, $(0,500)$, $(200,0)$, $(80,250)$. Show the first three are nodes and the last is a saddle.

Pesticides. Consider the system

$$\begin{aligned} x'(t) &= (10 - y(t))x(t) - s_1 x(t), \\ y'(t) &= (x(t) - 20)y(t) - s_2 y(t). \end{aligned}$$

17. **(Average Populations)** Explain: A field biologist should count, on the average, populations of about $20 + s_2$ prey and $10 - s_1$ predators.

18. **(Equilibria)** Show details for computing the pesticide system equilibria $(0,0)$, $(20+s_2, 10-s_1)$, where s_1, s_2 are the pesticide death rates.

Survival of One Species. Consider

$$\begin{aligned} x'(t) &= x(t)(24 - x(t) - 2y(t)), \\ y'(t) &= y(t)(30 - y(t) - 2x(t)). \end{aligned}$$

19. **(Equilibria)** Find all equilibria.

20. **(Interactions)** Show that doubling either x or y causes the interaction term $2xy$ to double.

21. **(Nonlinear Classification)** Classify each equilibrium point (x_0, y_0) as center, spiral, node, saddle, using the **Paste Theorem**. Determine stability for node and spiral. Make a computer phase portrait to confirm the classifications.

22. **(Extinction and Competing Species)** Equilibria for which either $x = 0$ or $y = 0$ signal extinction states. Discuss how the phase portrait of the nonlinear system shows extinction of one species but not both.

Co-existence

Find the equilibria, then classify them as node, saddle, spiral, center using the **Paste Theorem**. Determine stability for node and spiral. Make a computer phase portrait to confirm the classifications.

23. **(Node, Saddle, Saddle, Node)**

$$\begin{aligned} x' &= (144 - 2x - 3y)x, \\ y' &= (90 - 6y - x)y. \end{aligned}$$

24. **(Node, Saddle, Saddle, Node)**

$$\begin{aligned} x' &= (120 - 4x - 2y)x, \\ y' &= (60 - x - 2y)y. \end{aligned}$$

Explosion and Extinction

Find the equilibria, then classify them as node, saddle, spiral, center using the **Paste Theorem**. Determine stability for node and spiral. Make a computer phase portrait to confirm the classifications.

25. **(Node, Saddle, Saddle, Spiral)**

$$\begin{aligned} x' &= x(x - 2y - 4), \\ y' &= y(x + 2y - 8). \end{aligned}$$

26. **(Node, Saddle, Saddle, Spiral)**

$$\begin{aligned} x' &= x(x - y - 4), \\ y' &= y(x + y - 6). \end{aligned}$$

10.5 Mechanical Models

Nonlinear Spring-Mass System

The classical linear undamped spring-mass system is modeled by the equation $mx''(t) + kx(t) = 0$. This equation describes the excursion $x(t)$ from equilibrium $x = 0$ of a mass m attached to a spring of Hooke's constant k, with no damping and no external forces.

In the nonlinear theory, the Hooke's force term $-kx$ is replaced by a **Restoring Force** $F(x)$ which satisfies these four requirements:

Equilibrium 0**.** The equation $F(0) = 0$ is assumed, which gives $x = 0$ the status of a rest position.

Oddness. The equation $F(-x) = -F(x)$ is assumed, which says that the force F depends only upon the magnitude of the excursion from equilibrium, and not upon its direction. Then force F acts to **restore** the mass to its equilibrium position, like a Hooke's force $x \to kx$.

Zero damping. The damping effects always present in a real physical system are ignored. In linear approximations, it would be usual to assume a viscous damping effect $-cx'(t)$; from this viewpoint we assume $c = 0$.

Zero external force. There is no external force acting on the system. In short, only two forces act on the mass, (1) Newton's second law and (2) restoring force F.

The competition method applies to model the nonlinear spring-mass system via the two competing forces $mx''(t)$ and $F(x(t))$. The dynamical equation:

$$mx''(t) + F(x(t)) = 0. \tag{1}$$

Soft and Hard Springs

A restoring force F modeled upon Hooke's law is given by the equation $F(x) = kx$. With this force, the nonlinear spring-mass equation (1) becomes the undamped linear spring-mass system

$$mx''(t) + kx(t) = 0. \tag{2}$$

The linear equation can be thought to originate by replacing the actual spring force F by the first nonzero term of its Taylor series

$$F(x) = F(0) + F'(0)x + F''(0)\frac{x^2}{2!} + \cdots.$$

The assumptions $F(-x) = -F(x)$ and $F(0) = 0$ imply that $F(x)$ is a function of the form $F(x) = xG(x^2)$, hence all even terms in the Taylor series of F are zero.

Linear approximations to the force F drop the quadratic terms and higher from the Taylor series. More accurate nonlinear approximations are obtained by retaining extra Taylor series terms.

A restoring force F is called **hard** or **soft** provided it is given by a truncated Taylor series as follows.

Hard spring	$F(x) = kx + \beta x^3$, $\beta > 0$.
Soft spring	$F(x) = kx - \beta x^3$, $\beta > 0$.

For small excursions from equilibrium $x = 0$, a hard or soft spring force has magnitude approximately the same as the linear Hooke's force $F(x) = kx$.

Energy Conservation

Given nonlinear spring-mass equation $mx''(t) + F(x(t)) = 0$, each solution $x(t)$ satisfies on its domain of existence the **Conservation Law**

$$(3) \qquad \frac{m}{2}(x'(t))^2 + \int_{x(0)}^{x(t)} F(u)\,du = C, \quad C \equiv \frac{m}{2}(x'(0))^2.$$

To prove the law, multiply the nonlinear differential equation by $x'(t)$ to obtain $mx''(t)x'(t) + F(x(t))x'(t) = 0$, then apply quadrature to obtain (3).

Kinetic and Potential Energy

Using $v = x'(t)$, the term $mv^2/2$ in (3) is called the **Kinetic energy** (KE) and the term $\int_{x_0}^{x} F(u)du$ is called the **Potential energy** (PE). Equation (3) says that $KE + PE = C$ or that *energy is constant* along trajectories.

The conservation laws for the soft and hard nonlinear spring-mass systems, using position-velocity notation $x = x(t)$ and $y = x'(t)$, are therefore given by the equations

$$(4) \qquad my^2 + kx^2 + \frac{1}{2}\beta x^4 = C_1, \quad C_1 = \text{constant} > 0,$$

$$(5) \qquad my^2 + kx^2 - \frac{1}{2}\beta x^4 = C_2, \quad C_2 = \text{constant}.$$

Phase Plane and Scenes

Nonlinear behavior is commonly graphed in the **phase plane**, in which $x = x(t)$ and $y = x'(t)$ are the position and velocity of the mechanical system. The plots of t versus $x(t)$ or $x'(t)$ are called **Scenes**; these plots are invaluable for verifying periodic behavior and stability properties.

Hard spring

The only equilibrium for a hard spring $x' = y$, $my' = -kx - \beta x^3$ is the origin $x = y = 0$. Conservation law (4) describes a closed curve in the phase plane, which implies that trajectories are periodic orbits that encircle the equilibrium point $(0, 0)$. The classification of **center** applies. See Figures 34 and 35.

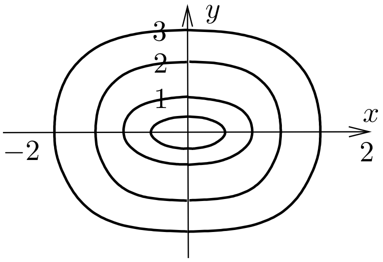


Figure 34. Hard spring $x''(t) + x(t) + 2x^3(t) = 0$.
Phase portrait for $x' = y$, $y' = -2x^3 - x$ on $|x| \le 2$, $|y| \le 3.5$. Initial data: $x(0) = 0$ and $y(0) = 1/2$, 1, 2, 3.

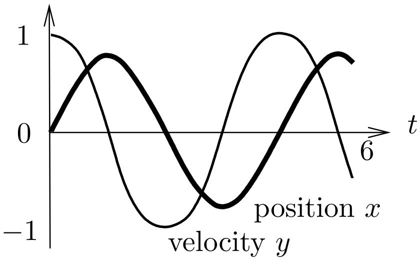


Figure 35. Hard spring $x''(t) + x(t) + 2x^3(t) = 0$.
Coordinate scenes for $x' = y$, $y' = -2x^3 - x$, $x(0) = 0$, $y(0) = 1$.

More intuition about the orbits can be obtained by finding the energy C_1 for each orbit. The value of C_1 decreases to zero as orbits close down upon the origin. Otherwise stated, the xyz-plot with $z = C_1$ has a minimum at the origin, which physically means that the equilibrium state $x = y = 0$ minimizes the energy. See Figure 36.

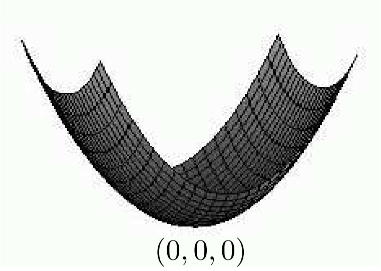


Figure 36. Hard spring energy minimization.
Plot for $x''(t) + x(t) + 2x^3(t) = 0$, using $z = y^2 + x^2 + x^4$ on $|x| \le 1/2$, $|y| \le 1$. The minimum is realized at $x = y = 0$.

Soft Spring

There are three equilibria for a soft spring

$$\begin{aligned} x' &= y, \\ my' &= -kx + \beta x^3. \end{aligned}$$

They are $(-\alpha, 0)$, $(0, 0)$, $(\alpha, 0)$, where $\alpha = \sqrt{k/\beta}$. If $(x(0), y(0))$ is given not at these points, then the mass undergoes motion. In short, the stationary mass positions are at the equilibria.

Linearization at the equilibria reveals part of the phase portrait. The linearized system at the origin is the system $x' = y$, $my' = -kx$, equivalent to the equation $mx'' + kx = 0$. It has a center at the origin. This implies the origin for the soft spring is either a center or a spiral. The other two equilibria have linearized systems equivalent to the equation $mx'' - 2kx = 0$; they are saddles.

The phase plot in Figure 37 shows separatrices, which are unions of solution curves and equilibrium points. Orbits in the phase plane, on either side of a separatrix, have physically different behavior. Shown is a center behavior interior to the union of the separatrices, while outside all orbits are unbounded.

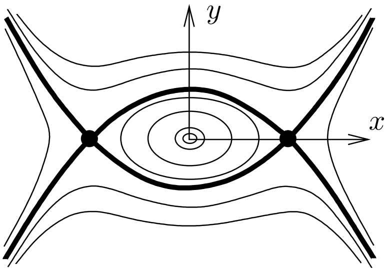


Figure 37. Soft spring $x''(t) + x(t) - 2x^3(t) = 0$.
A phase portrait for $x' = y$, $y' = 2x^3 - x$ on $|x| \le 1.2$, $|y| \le 1.2$. The 8 separatrices are the 6 bold curves plus the two equilibria $(\sqrt{0.5}, 0)$, $(-\sqrt{0.5}, 0)$.

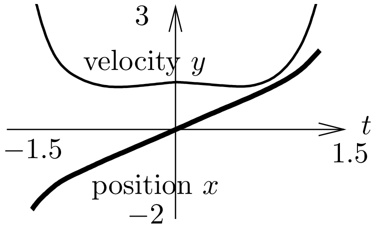


Figure 38. Soft spring $x''(t)+x(t)-2x^3(t) = 0$.
Coordinate scenes for $x' = y$, $y' = 2x^3 - x$, $x(0) = 0$, $y(0) = 4$.

Nonlinear Pendulum

Consider a nonlinear undamped pendulum of length L making angle $\theta(t)$ with the gravity vector. The **nonlinear pendulum equation** is given by

$$\frac{d^2\theta(t)}{dt^2} + \frac{g}{L}\sin(\theta(t)) = 0 \tag{6}$$

and its linearization at $\theta = 0$, called the **linearized pendulum equation**, is

$$\frac{d^2\theta(t)}{dt^2} + \frac{g}{L}\theta(t) = 0. \tag{7}$$

The linearized equation is valid only for small values of $\theta(t)$, because of the assumption $\sin\theta \approx \theta$ used to obtain (7) from (6).

Damped Pendulum

Physical pendulums are subject to friction forces, which we shall assume proportional to the velocity of the pendulum. The corresponding model which includes frictional forces is called the **damped pendulum equation**:

$$\frac{d^2\theta(t)}{dt^2} + c\frac{d\theta}{dt} + \frac{g}{L}\sin(\theta(t)) = 0. \tag{8}$$

It can be written as a first order system by setting $x(t) = \theta(t)$ and $y(t) = \theta'(t)$:

$$\begin{aligned} x'(t) &= y(t), \\ y'(t) &= -\frac{g}{L}\sin(x(t)) - cy(t). \end{aligned} \tag{9}$$

Undamped Pendulum

The position-velocity differential equations for the undamped pendulum are obtained by setting $x(t) = \theta(t)$ and $y(t) = \theta'(t)$:

$$\begin{aligned} x'(t) &= y(t), \\ y'(t) &= -\frac{g}{L}\sin(x(t)). \end{aligned} \tag{10}$$

Equilibrium points of nonlinear system (10) are at $y = 0$, $x = n\pi$, $n = 0$, ± 1, ± 2, ... with corresponding linearized system (see the exercises)

$$\begin{aligned} x'(t) &= y(t), \\ y'(t) &= -\frac{g}{L}\cos(n\pi)x(t). \end{aligned} \tag{11}$$

The characteristic equation of linear system (11) is $r^2 - \frac{g}{L}(-1)^n = 0$, because $\cos(n\pi) = (-1)^n$. The roots have different character depending on whether or not n is odd or even.

Even $n = 2m$. Then $r^2 + g/L = 0$ and the linearized system (11) is a **Center**. The orbits of (11) are concentric circles surrounding $x = n\pi$, $y = 0$.

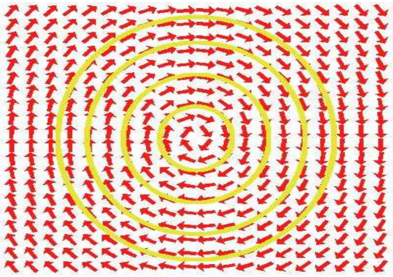


Figure 39. Linearized pendulum at equilibrium $x = 2m\pi$, $y = 0$.
Orbits are concentric circles.

Odd $n = 2m + 1$. Then $r^2 - g/L = 0$ and the linearized system (11) is classified as a **Saddle**. The orbits of (11) are hyperbolas with center $x = n\pi$, $y = 0$.

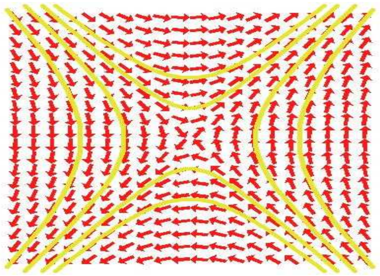


Figure 40. Linearized pendulum at $x = (2m+1)\pi$, $y = 0$.
Orbits are hyperbolas.

Drawing the Nonlinear Phase Diagram. The idea of the plot is to paste the linearized phase diagram onto the local region centered at the equilibrium point, when possible. The copying is guaranteed to be correct for the saddle case, but a center must be copied either as a spiral or a center. Extra analysis is needed to determine the figure to copy in the case of the center. The result appears in Figure 41.

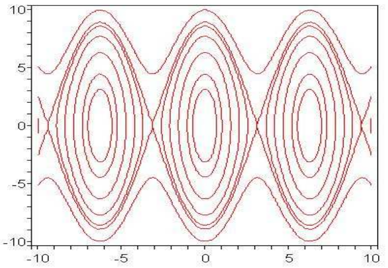


Figure 41. Nonlinear Pendulum.
Centers at $(-2\pi, 0)$, $(0, 0)$, $(2\pi, 0)$. Saddles at $(-3\pi, 0)$, $(-pi, 0)$, $(\pi, 0)$, $(3\pi, 0)$. Separatrices are unions of equilibria and conservation law curves $y^2 + \frac{4g}{L}\sin^2(x/2){=}2E$, with $E = 2\frac{g}{L}$ and $\frac{g}{L} = 10$.

We document the analysis used to produce Figure 41. The orbits trace an xy-curve given by integrating the separable equation

$$\frac{dy}{dx} = \frac{-g}{L}\frac{\sin x}{y}.$$

Then the conservation law for the mechanical system is

$$\frac{1}{2}y^2 + \frac{g}{L}(1 - \cos x) = E$$

where E is a constant of integration. This equation is arranged so that E is the sum of the kinetic energy $y^2/2$ and the potential energy $g(1 - \cos x)/L$, therefore E is the total mechanical energy. Using the double angle identity $\cos 2\phi = 1 - 2\sin^2\phi$ the conservation law can be written in the shorter form

$$y^2 + \frac{4g}{L}\sin^2(x/2) = 2E$$

When the energy E is small, $E < 2g/L$, then the pendulum never reaches the vertical position and it undergoes sustained periodic oscillation: the stable equilibria $(0, 2k\pi)$ have a local center structure.

When the energy E is large, $E > 2g/L$, then the pendulum reaches the vertical position and goes over the top repeatedly, represented by a saddle structure. The statement is verified from the two explicit solutions $y = \pm\sqrt{2E - 4g\sin^2(x/2)/L}$.

The energy equation $y^2 + \frac{4g}{L}\sin^2(x/2) = 4\frac{g}{L}$ (equivalent to $E = 2g/L$) produces the separatrix curves. **Separatrices** consist of equilibrium points plus solution curves which limit to the equilibria as $t \to \pm\infty$.

Exercises 10.5

Linear Mechanical Models

Consider the unforced linear model $mx'' + cx' + kx = 0$, where m, c, k are positive constants: m=mass, c=dashpot constant, k=Hooke's constant.

1. **(Dynamical System Form)** Write the scalar problem as $\vec{\mathbf{u}}\,' = A\vec{\mathbf{u}}$. Explicit definitions of $\vec{\mathbf{u}}(t)$ and A are expected.

2. **(Attractor to $\vec{\mathbf{u}} = \vec{\mathbf{0}}$)** Explain why $\lim_{t\to\infty} \vec{\mathbf{u}}(t) = \vec{\mathbf{0}}$, giving citations to theorems in this book.

3. **(Isolated Equilibrium)** Prove that $\vec{\mathbf{u}}\,' = A\vec{\mathbf{u}}$ has a unique equilibrium at $\vec{\mathbf{u}} = \vec{\mathbf{0}}$. Then explain why the equilibrium is isolated.

4. **(Phase Plots)** Classify the cases of **over-damped** and **under-damped** as a stable node or a stable spiral for $\vec{\mathbf{u}}\,' = A\vec{\mathbf{u}}$ at equilibrium $\vec{\mathbf{u}} = \vec{\mathbf{0}}$. Why are classifications *center* and *saddle* impossible?

Nonlinear Spring-Mass System

Consider the general model $x'' + F(x) = 0$ with the assumptions on page 804.

5. **(Harmonic Oscillator)** Let $F(x) = \omega^2 x$ with $\omega > 0$. Show F is odd and $F(0) = 0$. Then find the general solution $x(t)$ for $x'' + F(x) = 0$.

6. **(Taylor Series)** Show that an odd function $F(x)$ with Maclaurin series $\sum_{n=0}^{\infty} a_n x^n$ has all even order terms zero, that is, $a_n = 0$ for n even.

Soft and Hard Springs

Classify as a hard or soft spring. Then write the conservation law for the equation.

7. $x'' + x + x^3 = 0$

8. $x'' + x - x^3 = 0$

Hard spring

9. Prove that a hard spring has exactly one equilibrium $x = y = 0$.

10. Substitute $x = x(t), y = x'(t)$ into $z = y^2 + x^2 + x^4$ to obtain $z(t)$. Function $z(t)$ has a minimum when $\frac{dz}{dt} = 0$. Reduce this equation to $x'' + x + 2x^3 = 0$.

Soft Spring

Consider soft spring $x'' + kx - \beta x^3 = 0$, $k > 0$, $\beta > 0$.

11. **(Equilibria)** Verify the three equilibria $(0,0), (0, \sqrt{k\beta}), (0, -\sqrt{k\beta})$.

12. **(Saddles)** Verify by linearization and the **Paste Theorem** that nonlinear equilibria $(0, \sqrt{k\beta}), (0, -\sqrt{k\beta})$ are saddles.

13. **(Center or Spiral)** The **Paste Theorem** says that equilibrium $(0, 0)$ of the nonlinear system is a center or spiral. Verify by computer phase portrait $m = k = 1$ and $\beta = 2$ Figure 37, page 807.

14. **(Mass at Rest)** Verify that the only solutions with the mass at rest are the equilibria. **Mass at rest** means velocity zero: $\vec{\mathbf{u}}\,'(t_0) = \vec{\mathbf{0}}$ for some t_0, vector notation from Exercise 1.

15. **(Phase Portrait)** Solve for the equilibria of $x'' + 4x - x^3 = 0$. Draw a phase portrait similar to Figure 37, page 807.

16. **(Separatrix)** The energy equation for $x'' + 4x - x^3 = 0$ is $\frac{1}{2}y^2 + 2x^2 - \frac{1}{4}x^4 = E$. Substitute the saddle equilibria to find $E = 4$. Plot implicitly the energy equation curve. A separatrix is the union of the two saddle equilibria and this implicit curve.

Damped Nonlinear Pendulum

Consider $\frac{d^2\theta(t)}{dt^2} + c\frac{d\theta}{dt} + \frac{g}{L}\sin(\theta(t)) = 0$, which has vector-matrix form $\vec{\mathbf{u}}' = \vec{\mathbf{G}}(\vec{\mathbf{u}}(t))$.

17. Display both $\vec{\mathbf{u}}$ and $\vec{\mathbf{G}}$.

18. Find the Jacobian matrix of $\vec{\mathbf{G}}$ with respect to $\vec{\mathbf{u}}$.

Undamped Nonlinear Pendulum

Consider $\frac{d^2\theta(t)}{dt^2} + \frac{g}{L}\sin(\theta(t)) = 0$, having vector-matrix form $\vec{\mathbf{u}}' = \vec{\mathbf{F}}(\vec{\mathbf{u}}(t))$.

19. Find the Jacobian matrix of $\vec{\mathbf{F}}$ with respect to $\vec{\mathbf{u}}$.

20. Solve $\vec{\mathbf{F}}(\vec{\mathbf{u}}) = \vec{\mathbf{0}}$ for $\vec{\mathbf{u}}$, showing all details.

21. Evaluate the Jacobian matrix at the roots of $\vec{\mathbf{F}}(\vec{\mathbf{u}}) = \vec{\mathbf{0}}$.

22. Plot $y^2 + \frac{4g}{L}\sin^2(x/2) = 4\frac{g}{L}$ implicitly for $\frac{g}{L} = 10$. The separatrix is this curve plus equilibria.

Chapter 11

Systems of Differential Equations

Contents

11.1 Examples of Systems

Linear systems

A **linear system** is a system of differential equations of the form

$$\begin{array}{rcccccccc}
x_1' & = & a_{11}x_1 & + & \cdots & + & a_{1n}x_n & + & f_1, \\
x_2' & = & a_{21}x_1 & + & \cdots & + & a_{2n}x_n & + & f_2, \\
 & \vdots & & \vdots & \cdots & \vdots & & \vdots & \\
x_m' & = & a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & + & f_m,
\end{array} \tag{1}$$

where $' = d/dt$. Given are the functions $a_{ij}(t)$ and $f_j(t)$ on some interval $a < t < b$. The unknowns are the functions $x_1(t), \ldots, x_n(t)$.

The system is called **homogeneous** if all $f_j = 0$, otherwise it is called **non-homogeneous**.

Matrix Notation for Systems

A non-homogeneous system of linear equations (1) is written as the equivalent vector-matrix system

$$\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{f}}(t),$$

where

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \vec{\mathbf{f}} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ldots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

Brine Tank Cascade

Let brine tanks A, B, C be given of volumes 20, 40, 60, respectively, as in Figure 1.

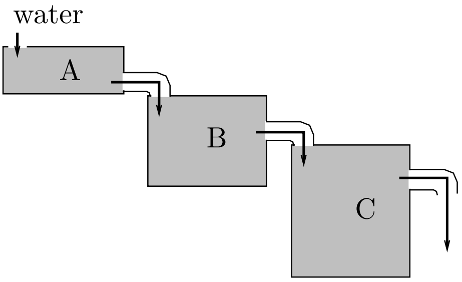


Figure 1. Three brine tanks in cascade.

It is supposed that fluid enters tank A at rate r, drains from A to B at rate r, drains from B to C at rate r, then drains from tank C at rate r. Hence the volumes of the tanks remain constant. Let $r = 10$, to illustrate the ideas.

Uniform stirring of each tank is assumed, which implies **uniform salt concentration** throughout each tank.

Let $x_1(t)$, $x_2(t)$, $x_3(t)$ denote the amount of salt at time t in each tank. We suppose **water containing no salt** is added to tank A . Therefore, the salt in all the tanks is eventually lost from the drains. The cascade is modeled by the **chemical balance law**

$$\text{rate of change} \quad = \quad \text{input rate} \quad - \quad \text{output rate}.$$

Application of the balance law, justified below in *compartment analysis*, results in the triangular differential system

$$\begin{aligned} x_1' &= -\frac{1}{2}x_1, \\ x_2' &= \frac{1}{2}x_1 - \frac{1}{4}x_2, \\ x_3' &= \frac{1}{4}x_2 - \frac{1}{6}x_3. \end{aligned}$$

The solution, to be justified later in this chapter, is given by the equations

$$\begin{aligned}
x_1(t) &= x_1(0)e^{-t/2},\\
x_2(t) &= -2x_1(0)e^{-t/2} + (x_2(0) + 2x_1(0))e^{-t/4},\\
x_3(t) &= \frac{3}{2}\,x_1(0)e^{-t/2} - 3(x_2(0) + 2x_1(0))e^{-t/4}\\
&\qquad + (x_3(0) - \frac{3}{2}x_1(0) + 3(x_2(0) + 2x_1(0)))e^{-t/6}.
\end{aligned}$$

Cascades and Compartment Analysis

A **Linear Cascade** is a diagram of **compartments** in which input and output rates have been assigned from one or more different compartments. The diagram is a succinct way to summarize and document the various rates.

The method of **compartment analysis** translates the diagram into a system of linear differential equations. The method has been used to derive applied models in diverse topics like ecology, chemistry, heating and cooling, kinetics, mechanics and electricity.

The method. Refer to Figure 2. A compartment diagram consists of the following components.

Variable Names — Each **compartment** is labelled with a variable X.

Arrows — Each arrow is labelled with a **Flow Rate** R.

Input Rate — An arrowhead pointing at compartment X documents **Input Rate** R.

Output Rate — An arrowhead pointing away from compartment X documents **Output Rate** R.

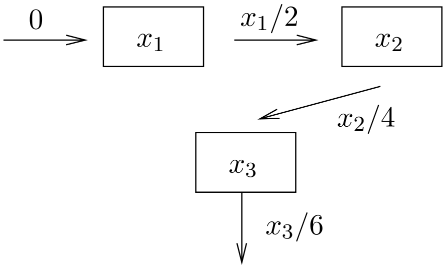


Figure 2. Compartment analysis diagram.
The diagram represents the classical brine tank problem of Figure 1.

Assembly of the single linear differential equation for a diagram compartment X is done by writing dX/dt for the left side of the differential equation and then algebraically adding the input and output rates to obtain the right side of the differential equation, according to the **balance law**

$$\frac{dX}{dt} = \text{sum of input rates} - \text{sum of output rates}$$

By convention, a compartment with no arriving arrowhead has input zero, and a compartment with no exiting arrowhead has output zero. Applying the balance law to Figure 2 gives one differential equation for each of the three compartments $\boxed{x_1}$, $\boxed{x_2}$, $\boxed{x_3}$.

$$\begin{aligned} x_1' &= 0 - \frac{1}{2}x_1, \\ x_2' &= \frac{1}{2}x_1 - \frac{1}{4}x_2, \\ x_3' &= \frac{1}{4}x_2 - \frac{1}{6}x_3. \end{aligned}$$

Recycled Brine Tank Cascade

Let brine tanks A, B, C be given of volumes 60, 30, 60, respectively, as in Figure 3.

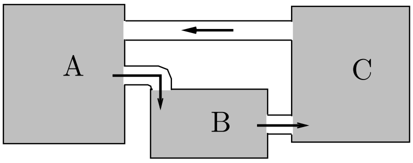


Figure 3. Three brine tanks in cascade with recycling.

Suppose that fluid drains from tank A to B at rate r, drains from tank B to C at rate r, then drains from tank C to A at rate r. The tank volumes remain constant due to constant recycling of fluid. For purposes of illustration, let $r = 10$.

Uniform stirring of each tank is assumed, which implies **uniform salt concentration** throughout each tank.

Let $x_1(t)$, $x_2(t)$, $x_3(t)$ denote the amount of salt at time t in each tank. No salt is lost from the system, due to recycling. Using compartment analysis, the recycled cascade is modeled by the non-triangular system

$$\begin{array}{rcrcrcr} x_1' &=& -\frac{1}{6}x_1 & & & + & \frac{1}{6}x_3, \\ x_2' &=& \frac{1}{6}x_1 & - & \frac{1}{3}x_2, & & \\ x_3' &=& & & \frac{1}{3}x_2 & - & \frac{1}{6}x_3. \end{array}$$

The solution is given by the equations

$$\begin{aligned} x_1(t) &= c_1 + (c_2 - 2c_3)e^{-t/3}\cos(t/6) + (2c_2 + c_3)e^{-t/3}\sin(t/6), \\ x_2(t) &= \frac{1}{2}c_1 + (-2c_2 - c_3)e^{-t/3}\cos(t/6) + (c_2 - 2c_3)e^{-t/3}\sin(t/6), \\ x_3(t) &= c_1 + (c_2 + 3c_3)e^{-t/3}\cos(t/6) + (-3c_2 + c_3)e^{-t/3}\sin(t/6). \end{aligned}$$

At infinity, $x_1 = x_3 = c_1$, $x_2 = c_1/2$. The meaning is that the total amount of salt is uniformly distributed in the tanks, in the ratio $2:1:2$.

Pond Pollution

Consider three ponds connected by streams, as in Figure 4. The first pond has a pollution source, which spreads via the connecting streams to the other ponds. The plan is to determine the amount of pollutant in each pond.

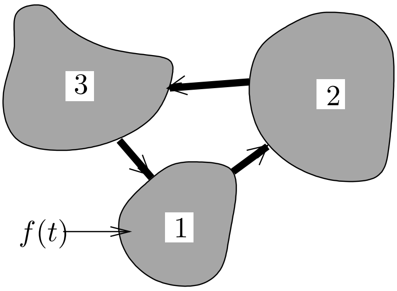


Figure 4. Three ponds 1, 2, 3 of volumes V_1, V_2, V_3 connected by streams. The pollution source $f(t)$ is in pond 1.

Assume the following.

- Symbol $f(t)$ is the pollutant flow rate into pond 1 (lb/min).
- Symbols f_1, f_2, f_3 denote the pollutant flow rates out of ponds 1, 2, 3, respectively (gal/min). It is assumed that the pollutant is well-mixed in each pond.
- The three ponds have volumes V_1, V_2, V_3 (gal), which remain constant.
- Symbols $x_1(t)$, $x_2(t)$, $x_3(t)$ denote the amount (lbs) of pollutant in ponds 1, 2, 3, respectively.

The pollutant flux is the flow rate times the pollutant concentration, e.g., pond 1 is emptied with flux f_1 times $x_1(t)/V_1$. A compartment analysis is summarized in the following diagram.

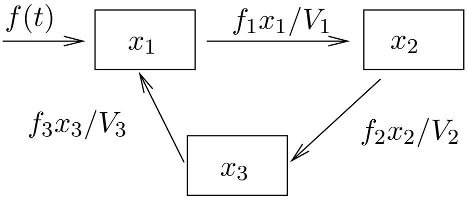


Figure 5. Pond diagram.
The compartment diagram represents the three-pond pollution problem of Figure 4.

The diagram plus compartment analysis gives the following differential equations.

$$\begin{aligned}
x_1'(t) &= \frac{f_3}{V_3}x_3(t) - \frac{f_1}{V_1}x_1(t) + f(t),\\
x_2'(t) &= \frac{f_1}{V_1}x_1(t) - \frac{f_2}{V_2}x_2(t),\\
x_3'(t) &= \frac{f_2}{V_2}x_2(t) - \frac{f_3}{V_3}x_3(t).
\end{aligned}$$

For a specific numerical example, take $f_i/V_i = 0.001$, $1 \le i \le 3$, and let $f(t) = 0.125$ lb/min for the first 48 hours (2880 minutes), thereafter $f(t) = 0$. We expect due to uniform mixing that after a long time there will be $(0.125)(2880) = 360$ pounds of pollutant uniformly deposited, which is 120 pounds per pond.

Initially, $x_1(0) = x_2(0) = x_3(0) = 0$, if the ponds were pristine. The specialized problem for the first 48 hours is

$$\begin{aligned} x_1'(t) &= 0.001\,x_3(t) - 0.001\,x_1(t) + 0.125,\\ x_2'(t) &= 0.001\,x_1(t) - 0.001\,x_2(t),\\ x_3'(t) &= 0.001\,x_2(t) - 0.001\,x_3(t),\\ x_1(0) &= x_2(0) = x_3(0) = 0. \end{aligned}$$

The solution to this system is

$$\begin{aligned} x_1(t) &= e^{-\frac{3t}{2000}}\left(\frac{125\sqrt{3}}{9}\sin\left(\frac{\sqrt{3}t}{2000}\right) - \frac{125}{3}\cos\left(\frac{\sqrt{3}t}{2000}\right)\right) + \frac{125}{3} + \frac{t}{24},\\ x_2(t) &= -\frac{250\sqrt{3}}{9}e^{-\frac{3t}{2000}}\sin\left(\frac{\sqrt{3}t}{2000}\right) + \frac{t}{24},\\ x_3(t) &= e^{-\frac{3t}{2000}}\left(\frac{125}{3}\cos\left(\frac{\sqrt{3}t}{2000}\right) + \frac{125\sqrt{3}}{9}\sin\left(\frac{\sqrt{3}t}{2000}\right)\right) + \frac{t}{24} - \frac{125}{3}. \end{aligned}$$

After 48 hours elapse, the approximate pollutant amounts in pounds are

$$x_1(2880) = 162.30, \quad x_2(2880) = 119.61, \quad x_3(2880) = 78.08.$$

It should be remarked that the system above and its solution will require a change in order to predict the state of the ponds after 48 hours/ The equations change by replacing constant 0.125 by zero. The corresponding homogeneous system has an equilibrium solution $x_1(t) = x_2(t) = x_3(t) = 120$. This constant solution, called the **steady-state**, is the limit at infinity of the solution to the homogeneous system using the initial values $x_1(0) \approx 162.30$, $x_2(0) \approx 119.61$, $x_3(0) \approx 78.08$, which are values from the forced system at $t = 48$ hours.

Home Heating

Consider a typical home with attic, basement and insulated main floor.

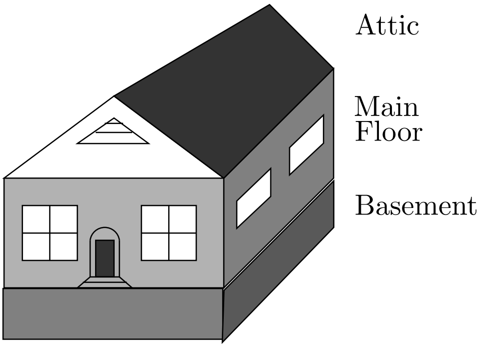


Figure 6. Typical home with attic and basement. The below-grade basement and the attic are uninsulated. Only the main living area is insulated.

It is usual to surround the main living area with insulation, but the attic area has walls and ceiling without insulation. The walls and floor in the basement are insulated by earth. The basement ceiling is insulated by air space in the joists, a layer of flooring on the main floor and a layer of drywall in the basement. We will analyze the changing temperatures in the three levels using Newton's cooling law and the variables

$$\begin{aligned}
z(t) &= \text{Temperature in the attic},\\
y(t) &= \text{Temperature in the main living area},\\
x(t) &= \text{Temperature in the basement},\\
t &= \text{Time in hours}.
\end{aligned}$$

Initial data. Assume it is winter time and the outside temperature in constantly 35°F during the day. Also assumed is a basement earth temperature of 45°F. Initially, the heat is off for several days. The initial values at noon ($t = 0$) are then $x(0) = 45$, $y(0) = z(0) = 35$.

Portable heater. A small electric heater is turned on at noon, with thermostat set for 100°F. When the heater is running, it provides a 20°F rise per hour, therefore it takes some time to reach 100°F (probably never!). Newton's cooling law

$$\text{Temperature rate} = \text{k(Temperature difference)}$$

will be applied to five boundary surfaces: (0) the basement walls and floor, (1) the basement ceiling, (2) the main floor walls, (3) the main floor ceiling, and (4) the attic walls and ceiling. Newton's cooling law gives positive cooling constants k_0, k_1, k_2, k_3, k_4 and the equations

$$\begin{aligned}
x' &= k_0(45 - x) + k_1(y - x),\\
y' &= k_1(x - y) + k_2(35 - y) + k_3(z - y) + 20,\\
z' &= k_3(y - z) + k_4(35 - z).
\end{aligned}$$

The insulation constants will be defined as $k_0 = 1/2$, $k_1 = 1/2$, $k_2 = 1/4$, $k_3 = 1/4$, $k_4 = 3/4$ to reflect insulation quality. The reciprocal $1/k$ is approximately the amount of time in hours required for 63% of the temperature difference to be exchanged. For instance, 4 hours elapse for the main floor. The model:

$$\begin{aligned}
x' &= \frac{1}{2}(45 - x) + \frac{1}{2}(y - x),\\
y' &= \frac{1}{2}(x - y) + \frac{1}{4}(35 - y) + \frac{1}{4}(z - y) + 20,\\
z' &= \frac{1}{4}(y - z) + \frac{3}{4}(35 - z).
\end{aligned}$$

The homogeneous solution in vector form is given in terms of constants $a = 1 + \sqrt{5}/4$, $b = 1 - \sqrt{5}/4$, and arbitrary constants c_1, c_2, c_3 by the formula

$$\begin{pmatrix} x_h(t) \\ y_h(t) \\ z_h(t) \end{pmatrix} = c_1 e^{-t} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} + c_2 e^{-at} \begin{pmatrix} 2 \\ \sqrt{5} \\ 1 \end{pmatrix} + c_3 e^{-bt} \begin{pmatrix} 2 \\ -\sqrt{5} \\ 1 \end{pmatrix}.$$

A particular solution is an equilibrium solution

$$\begin{pmatrix} x_p(t) \\ y_p(t) \\ z_p(t) \end{pmatrix} = \begin{pmatrix} \frac{620}{11} \\ \frac{745}{11} \\ \frac{475}{11} \end{pmatrix}.$$

The homogeneous solution has limit zero at infinity, hence the temperatures of the three spaces hover around $x = 56.4$, $y = 67.7$, $z = 43.2$ degrees Fahrenheit. Specific information can be gathered by solving for c_1, c_2, c_3 according to the initial data $x(0) = 45$, $y(0) = z(0) = 35$. The answers are

$$c_1 = 5, \quad c_2 = \frac{25}{2} + \frac{7}{2}\sqrt{5}, \quad c_3 = \frac{25}{2} - \frac{7}{2}\sqrt{5}.$$

Underpowered heater. To the main floor each hour is added 20°F, but the heat escapes at a substantial rate, so that after one hour $y \approx 68$°F. After five hours, $y \approx 68$°F. The heater in this example is so inadequate that even after many hours, the main living area is still under 69°F.

Forced air furnace. Replacing the space heater by a normal furnace adds the difficulty of **switches** in the input, namely, the thermostat turns off the furnace when the main floor temperature reaches 70°F, and it turns it on again after a 4°F temperature drop. We will suppose that the furnace has four times the BTU rating of the space heater, which translates to an 80°F temperature rise per hour. The study of the forced air furnace requires two differential equations, one with 20 replaced by 80 (DE 1, furnace on) and the other with 20 replaced by 0 (DE 2, furnace off). The plan is to use the first differential equation on time interval $0 \le t \le t_1$, then switch to the second differential equation for time interval $t_1 \le t \le t_2$. The time intervals are selected so that $y(t_1) = 70$ (the thermostat setting) and $y(t_2) = 66$ (thermostat setting less 4 degrees). Numerical work gives the following results.

Time in minutes	Main floor temperature	Model	Furnace
31.6	70	DE 1	on
40.9	66	DE 2	off
45.3	70	DE 1	on
54.6	66	DE 2	off

The reason for the non-uniform times between furnace cycles can be seen from the model. Each time the furnace cycles, heat enters the main floor, then escapes through the other two levels. Consequently, the initial conditions on each floor applied to models 1 and 2 are changing, resulting in different solutions to the models on each switch.

Chemostats and Microorganism Culturing

A vessel into which nutrients are pumped, to feed a microorganism, is called a **Chemostat**[1]. Uniform distributions of microorganisms and nutrients are assumed, for example, due to stirring effects. The pumping is matched by draining to keep the volume constant.

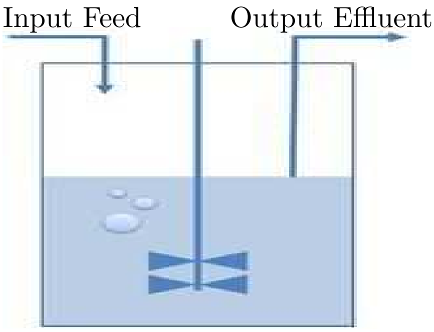


Figure 7. A Basic Chemostat. A stirred bio-reactor operated as a chemostat, with continuous inflow and outflow. The flow rates are controlled to maintain a constant culture volume.

In a typical chemostat, one nutrient is kept in short supply while all others are abundant. We consider here the question of **survival** of the organism subject to the limited resource. The problem is quantified as follows:

$$\begin{aligned} x(t) &= \text{the concentration of the limited nutrient in the vessel,}\\ y(t) &= \text{the concentration of organisms in the vessel.}\end{aligned}$$

A special case of the derivation in J.M. Cushing's text [Cushing] for the organism *E. Coli*[2] is the set of **nonlinear** differential equations[3]

$$\begin{aligned} x' &= -0.075x + (0.075)(0.005) - \frac{1}{63}g(x)y,\\ y' &= -0.075y + g(x)y, \end{aligned} \tag{2}$$

where $g(x) = 0.68x(0.0016+x)^{-1}$. Of special interest to the study of this equation are two linearized equations at equilibria, given by

$$\begin{aligned} u_1' &= -0.075\,u_1 - 0.008177008175\,u_2,\\ u_2' &= 0.4401515152\,u_2, \end{aligned} \tag{3}$$

[1] The October 14, 2004 issue of the journal *Nature* featured a study of the co-evolution of a common type of bacteria, Escherichia coli, and a virus that infects it, called bacteriophage T7. Postdoctoral researcher Samantha Forde set up ”microbial communities of bacteria and viruses with different nutrient levels in a series of chemostats – glass culture tubes that provide nutrients and oxygen and siphon off wastes.”

[2] In a biology Master's thesis, two strains of Escherichia coli were grown in a glucose-limited chemostat coupled to a modified Robbins device containing plugs of silicone rubber urinary catheter material. Reference: Jennifer L. Adams and Robert J. C. McLean, Applied and Environmental Microbiology, September 1999, p. 4285-4287, Vol. 65, No. 9.

[3] More details can be found in *The Theory of the Chemostat Dynamics of Microbial Competition*, ISBN-13: 9780521067348, by Hal Smith and Paul Waltman, June 2008.

$$\begin{aligned} v_1' &= -1.690372243\, v_1 - 0.001190476190\, v_2, \\ v_2' &= 101.7684513\, v_1. \end{aligned} \tag{4}$$

Although we cannot solve the nonlinear system explicitly, nevertheless there are explicit formulas for u_1, u_2, v_1, v_2 that complete the picture of how solutions $x(t)$, $y(t)$ behave at $t = \infty$. The result of the analysis is that *E. Coli* survives indefinitely in this vessel at concentration $y \approx 0.3$.

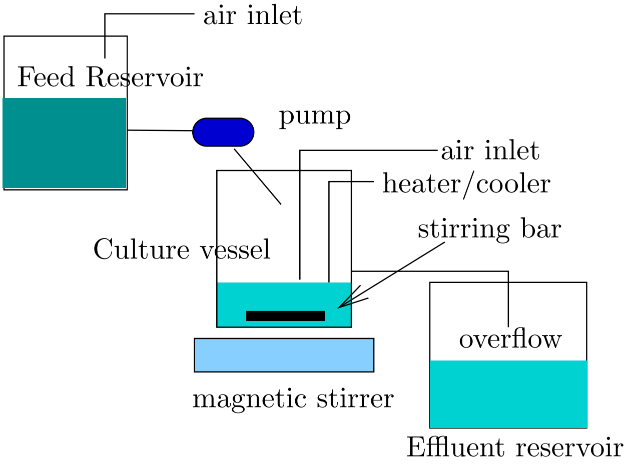


Figure 8. Laboratory Chemostat. The components are the **Feed reservoir**, which contains the nutrients, a stirred chemical reactor labeled the **Culture vessel**, and the **Effluent reservoir**, which holds the effluent overflow from the reactor.

Irregular Heartbeats and Lidocaine

The human malady of **Ventricular Arrhythmia** or irregular heartbeat is treated clinically using the drug **lidocaine**.

Figure 9. Xylocaine label, a brand name for the drug lidocaine.

To be effective, the drug has to be maintained at a bloodstream concentration of 1.5 milligrams per liter, but concentrations above 6 milligrams per liter are considered lethal in some patients. The actual dosage depends upon body weight. The adult dosage maximum for ventricular tachycardia is reported at 3 mg/kg.[4] The drug is supplied in 0.5%, 1% and 2% solutions, which are stored at room temperature.

A differential equation model for the dynamics of the drug therapy uses

[4] Source: **Family Practice Notebook**, `http://www.fpnotebook.com/`. The author is Scott Moses, MD, who practises in Lino Lakes, Minnesota.

$$\begin{aligned} x(t) &= \text{amount of } \textit{lidocaine} \text{ in the bloodstream}, \\ y(t) &= \text{amount of } \textit{lidocaine} \text{ in body tissue}. \end{aligned}$$

A typical set of equations, valid for a special body weight only, appears below; for more detail see J.M. Cushing's text [Cushing].

$$\begin{aligned} x'(t) &= -0.09x(t) + 0.038y(t), \\ y'(t) &= 0.066x(t) - 0.038y(t). \end{aligned} \tag{5}$$

The physically significant initial data is zero drug in the bloodstream $x(0) = 0$ and injection dosage $y(0) = y_0$. The answers:

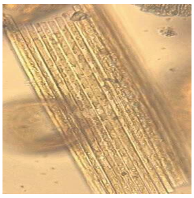
$$\begin{aligned} x(t) &= -0.3367y_0e^{-0.1204t} + 0.3367y_0e^{-0.0076t}, \\ y(t) &= 0.2696y_0e^{-0.1204t} + 0.7304y_0e^{-0.0076t}. \end{aligned}$$

The answers can be used to estimate the maximum possible safe dosage y_0 and the duration of time that the drug *lidocaine* is effective.

Nutrient Flow in an Aquarium

Consider a vessel of water containing a radioactive isotope, to be used as a tracer for the food chain, which consists of aquatic plankton varieties A and B.

Plankton are aquatic organisms that drift with the currents, typically in an environment like Chesapeake Bay. Plankton can be divided into two groups, phytoplankton and zooplankton. The phytoplankton are *plant-like* drifters: diatoms and other alga. Zooplankton are *animal-like* drifters: copepods, larvae, and small crustaceans.



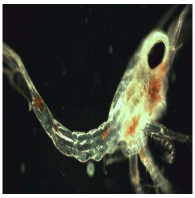


Figure 10. Left: Bacillaria paxillifera, phytoplankton. Right: Anomura Galathea zoea, zooplankton.

Let

$$\begin{aligned} x(t) &= \text{isotope concentration in the water}, \\ y(t) &= \text{isotope concentration in } A, \\ z(t) &= \text{isotope concentration in } B. \end{aligned}$$

Typical differential equations are

$$\begin{aligned} x'(t) &= -3x(t) + 6y(t) + 5z(t), \\ y'(t) &= 2x(t) - 12y(t), \\ z'(t) &= x(t) + 6y(t) - 5z(t). \end{aligned}$$

The answers are

$$\begin{aligned}
x(t) &= 6c_1 + (1+\sqrt{6})c_2 e^{(-10+\sqrt{6})t} + (1-\sqrt{6})c_3 e^{(-10-\sqrt{6})t},\\
y(t) &= c_1 + c_2 e^{(-10+\sqrt{6})t} + c_3 e^{(-10-\sqrt{6})t},\\
z(t) &= \frac{12}{5}c_1 - \left(2+\sqrt{1.5}\right) c_2 e^{(-10+\sqrt{6})t} + \left(-2+\sqrt{1.5}\right) c_3 e^{(-10-\sqrt{6})t}.
\end{aligned}$$

The constants c_1, c_2, c_3 are related to the initial radioactive isotope concentrations $x(0) = x_0$, $y(0) = 0$, $z(0) = 0$, by the 3×3 system of linear algebraic equations

$$\begin{array}{rcrcrcr}
6c_1 & + & (1+\sqrt{6})c_2 & + & (1-\sqrt{6})c_3 & = & x_0,\\
c_1 & + & c_2 & + & c_3 & = & 0,\\
\frac{12}{5}c_1 & - & \left(2+\sqrt{1.5}\right) c_2 & + & \left(-2+\sqrt{1.5}\right) c_3 & = & 0.
\end{array}$$

Biomass Transfer

Consider a European forest having one or two varieties of trees. We select some of the oldest trees, those expected to die off in the next few years, then follow the cycle of living trees into dead trees. The dead trees eventually decay and fall from seasonal and biological events. Finally, the fallen trees become humus.

Figure 11. Forest Biomass. Total biomass is a parameter used to assess atmospheric carbon that is harvested by trees. Forest management uses biomass subclasses to classify fire risk.

Let variables x, y, z, t be defined by

$$\begin{aligned}
x(t) &= \text{biomass decayed into humus},\\
y(t) &= \text{biomass of dead trees},\\
z(t) &= \text{biomass of living trees},\\
t &= \text{time in decades } (\textit{decade} = 10 \text{ years}).
\end{aligned}$$

A typical biological model is

$$\begin{aligned} x'(t) &= -x(t) + 3y(t), \\ y'(t) &= -3y(t) + 5z(t), \\ z'(t) &= -5z(t). \end{aligned}$$

Suppose there are no dead trees and no humus at $t = 0$, with initially z_0 units of living tree biomass. These assumptions imply initial conditions $x(0) = y(0) = 0$, $z(0) = z_0$. The solution is

$$\begin{aligned} x(t) &= \frac{15}{8} z_0 \left(e^{-5t} - 2e^{-3t} + e^{-t}\right), \\ y(t) &= \frac{5}{2} z_0 \left(-e^{-5t} + e^{-3t}\right), \\ z(t) &= z_0 e^{-5t}. \end{aligned}$$

The live tree biomass $z(t) = z_0 e^{-5t}$ decreases according to a Malthusian decay law from its initial size z_0. It decays to 60% of its original biomass in one year. Interesting calculations that can be made from the other formulas include the future dates when the dead tree biomass and the humus biomass are maximum. The predicted dates are approximately 2.5 and 8 years hence, respectively.

The predictions made by this model are trends extrapolated from rate observations in the forest. Like weather prediction, it is a calculated guess that disappoints on a given day and from the outset has no predictable answer.

Total biomass is considered an important parameter to assess atmospheric carbon that is harvested by trees. Biomass estimates for forests since 1980 have been made by satellite remote sensing data with instances of 90% accuracy (*Science* 87(5), September 2004).

Pesticides in Soil and Trees

A Washington cherry orchard was sprayed with pesticides.

Figure 12. June Cherries.

Assume that a negligible amount of pesticide was sprayed on the soil. Pesticide applied to the trees has a certain outflow rate to the soil, and conversely, pesticide in the soil has a certain uptake rate into the trees. Repeated applications of the pesticide are required to control the insects, which implies the pesticide levels in the trees varies with time. Quantize the pesticide spraying as follows.

$$
\begin{aligned}
x(t) &= \text{amount of pesticide in the trees,}\\
y(t) &= \text{amount of pesticide in the soil,}\\
r(t) &= \text{amount of pesticide applied to the trees,}\\
t &= \text{time in years.}
\end{aligned}
$$

A typical model is obtained from input-output analysis, similar to the brine tank models:

$$
\begin{aligned}
x'(t) &= 2x(t) - y(t) + r(t),\\
y'(t) &= 2x(t) - 3y(t).
\end{aligned}
$$

In a pristine orchard, the initial data is $x(0) = 0$, $y(0) = 0$, because the trees and the soil initially harbor no pesticide. The solution of the model obviously depends on $r(t)$. The nonhomogeneous dependence is treated by the method of variation of parameters *infra*. Approximate formulas are

$$
\begin{aligned}
x(t) &\approx \int_0^t \left(1.10e^{1.6(t-u)} - 0.12e^{-2.6(t-u)}\right) r(u)du,\\
y(t) &\approx \int_0^t \left(0.49e^{1.6(t-u)} - 0.49e^{-2.6(t-u)}\right) r(u)du.
\end{aligned}
$$

The exponential rates 1.6 and -2.6 represent respectively the accumulation of the pesticide into the soil and the decay of the pesticide from the trees. The application rate $r(t)$ is typically a step function equal to a positive constant over a small interval of time and zero elsewhere, or a sum of such functions, representing periodic applications of pesticide.

Forecasting Prices

A manufacturer has a marketing policy based upon the price $x(t)$ of its product.

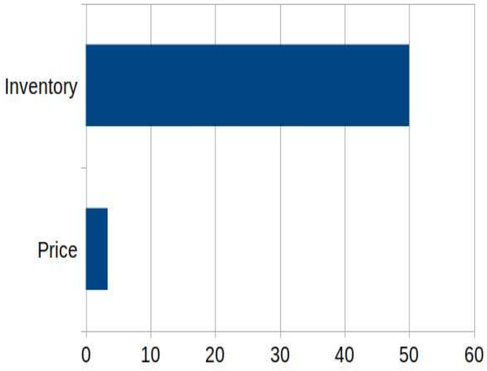


Figure 13. Pricing and Inventory.
Dynamic pricing reflects demand for the product, predicted by sales data.

The **Production** $P(t)$ and the **Sales** $S(t)$ are given in terms of the **Price** $x(t)$ and the **Change in Price** $x'(t)$ by the equations

$$\begin{aligned} P(t) &= 4 - \frac{3}{4}x(t) - 8x'(t) \quad \text{(Production)}, \\ S(t) &= 15 - 4x(t) - 2x'(t) \quad \text{(Sales)}. \end{aligned}$$

The differential equations for the price $x(t)$ and inventory level $I(t)$ are

$$\begin{aligned} x'(t) &= k(I(t) - I_0), \\ I'(t) &= P(t) - S(t). \end{aligned}$$

The inventory level $I_0 = 50$ represents the desired level. The equations can be written in terms of $x(t)$, $I(t)$ as follows.

$$\begin{array}{rclcrcl} x'(t) & = & & & kI(t) & - & kI_0, \\ I'(t) & = & \frac{13}{4}x(t) & - & 6kI(t) & + & 6kI_0 - 11. \end{array}$$

If $k = 1$, $x(0) = 10$ and $I(0) = 7$, then the solution is given by

$$\begin{aligned} x(t) &= \frac{44}{13} + \frac{86}{13}e^{-13t/2}, \\ I(t) &= 50 - 43e^{-13t/2}. \end{aligned}$$

The **Forecast** of price $x(t) \approx 3.38$ dollars at inventory level $I(t) \approx 50$ is based upon the two limits

$$\lim_{t\to\infty} x(t) = \frac{44}{13}, \quad \lim_{t\to\infty} I(t) = 50.$$

Coupled Spring-Mass Systems

Three masses are attached to each other by four springs as in Figure 14.

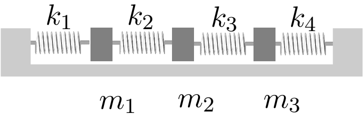


Figure 14. Three masses connected by springs. The masses slide along a frictionless horizontal surface.

The analysis uses the following constants, variables and assumptions.

Mass Constants The masses m_1, m_2, m_3 are assumed to be point masses concentrated at their center of gravity.

Spring Constants The mass of each spring is negligible. The springs operate according to Hooke's law: Force = k(elongation). Constants k_1, k_2, k_3, k_4 denote the Hooke's constants. The springs restore after compression and extension.

Position Variables The symbols $x_1(t)$, $x_2(t)$, $x_3(t)$ denote the mass positions along the horizontal surface, measured from their equilibrium positions, plus right and minus left.

Fixed Ends The first and last spring are attached to fixed walls.

The **competition method** is used to derive the equations of motion. In this case, the law is

$$\text{Newton's Second Law Force} = \text{Sum of the Hooke's Forces.}$$

The model equations are

$$\begin{array}{lrcl} & m_1x_1''(t) & = & -k_1x_1(t) + k_2[x_2(t) - x_1(t)], \\ (6) & m_2x_2''(t) & = & -k_2[x_2(t) - x_1(t)] + k_3[x_3(t) - x_2(t)], \\ & m_3x_3''(t) & = & -k_3[x_3(t) - x_2(t)] - k_4x_3(t). \end{array}$$

The equations are justified in the case of all positive variables by observing that the first three springs are elongated by x_1, $x_2 - x_1$, $x_3 - x_2$, respectively. The last spring is compressed by x_3, which accounts for the minus sign.

Another way to justify the equations is through mirror-image symmetry: interchange $k_1 \longleftrightarrow k_4$, $k_2 \longleftrightarrow k_3$, $x_1 \longleftrightarrow x_3$, then equation 2 should be unchanged and equation 3 should become equation 1.

Matrix Formulation. System (6) can be written as a second order vector-matrix system

$$\begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix} \begin{pmatrix} x_1'' \\ x_2'' \\ x_3'' \end{pmatrix} = \begin{pmatrix} -k_1 - k_2 & k_2 & 0 \\ k_2 & -k_2 - k_3 & k_3 \\ 0 & k_3 & -k_3 - k_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

More succinctly, the system is written as

$$M\vec{\mathbf{x}}''(t) = K\vec{\mathbf{x}}(t)$$

where the **displacement** $\vec{\mathbf{x}}$, **mass matrix** M and **stiffness matrix** K are defined by the formulas

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad M = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}, \quad K = \begin{pmatrix} -k_1 - k_2 & k_2 & 0 \\ k_2 & -k_2 - k_3 & k_3 \\ 0 & k_3 & -k_3 - k_4 \end{pmatrix}.$$

Numerical example. Let $m_1 = 1$, $m_2 = 1$, $m_3 = 1$, $k_1 = 2$, $k_2 = 1$, $k_3 = 1$, $k_4 = 2$. Then the system is given by

$$\begin{pmatrix} x_1'' \\ x_2'' \\ x_3'' \end{pmatrix} = \begin{pmatrix} -3 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The vector solution is given by the formula

$$\begin{aligned}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} &= (a_1\cos t + b_1\sin t)\begin{pmatrix}1\\2\\1\end{pmatrix}\\ &\quad + \left(a_2\cos\sqrt{3}t + b_2\sin\sqrt{3}t\right)\begin{pmatrix}1\\0\\-1\end{pmatrix}\\ &\quad + (a_3\cos 2t + b_3\sin 2t)\begin{pmatrix}1\\-1\\1\end{pmatrix},\end{aligned}$$

where a_1, a_2, a_3, b_1, b_2, b_3 are arbitrary constants.

Railway Cars

A special case of the coupled spring-mass system is three flatbed rail cars on a level frictionless track connected by springs, as in Figure 15.

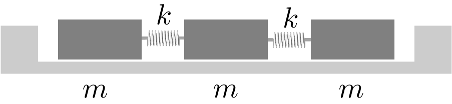


Figure 15. Three identical rail cars connected by identical springs.

Except for the springs on fixed ends, this problem is the same as the one in equation (6), therefore taking $k_1 = k_4 = 0$, $k_2 = k_3 = k$, $m_1 = m_2 = m_3 = m$ gives the system

$$\begin{pmatrix}m&0&0\\0&m&0\\0&0&m\end{pmatrix}\begin{pmatrix}x_1''\\x_2''\\x_3''\end{pmatrix} = \begin{pmatrix}-k&k&0\\k&-2k&k\\0&k&-k\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}.$$

Take $k/m = 1$ to obtain the illustration

$$\vec{\mathbf{x}}'' = \begin{pmatrix}-1&1&0\\1&-2&1\\0&1&-1\end{pmatrix}\vec{\mathbf{x}},$$

which has vector solution

$$\begin{aligned}\vec{\mathbf{x}} &= (a_1 + b_1 t)\begin{pmatrix}1\\1\\1\end{pmatrix} + (a_2\cos t + b_2\sin t)\begin{pmatrix}1\\0\\-1\end{pmatrix}\\ &\quad + \left(a_3\cos\sqrt{3}t + b_3\sin\sqrt{3}t\right)\begin{pmatrix}1\\-2\\1\end{pmatrix},\end{aligned}$$

where a_1, a_2, a_3, b_1, b_2, b_3 are arbitrary constants.

The solution expression can be used to discover what happens to the rail cars when the springs act normally upon compression but disengage upon expansion. An interesting physical situation is when one car moves along the track, contacts two stationary cars, then transfers its momentum to the other cars, followed by disengagement.

Monatomic Crystals

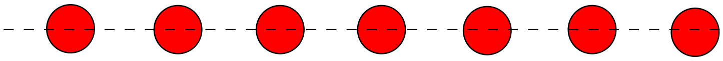


Figure 16. A Crystal Model.
The n crystals are identical masses m assumed connected by equal springs of Hooke's constant k. The last mass is connected to the first mass.

The scalar differential equations for Figure 16 are written for mass positions $x_1, \ldots, x_n$, with $x_0 = x_n$, $x_{n+1} = x_1$ to account for the ring of identical masses (periodic boundary condition). Then for $k = 1, \ldots, n$

$$m\frac{d^2x_k}{dt^2} = k(x_{k+1} - x_k) + k(x_{k-1} - x_k) = k(x_{k-1} - 2x_k + x_{k+1}).$$

These equations represent a system $x'' = Ax$, where the symmetric matrix of coefficients $A = M^{-1}K$ is given for $n = 5$ and $k/m = 1$ by

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & 1 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 1 & 0 & 0 & 1 & -2 \end{pmatrix}.$$

If $n = 3$ and $k/m = 1$, then $A = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$ and the solutions x_1, x_2, x_3 are linear combinations of the functions 1, t, $\cos\sqrt{3}t$, $\sin\sqrt{3}t$.

Electrical LR–Network no EMF

Consider the LR-network of Figure 17.

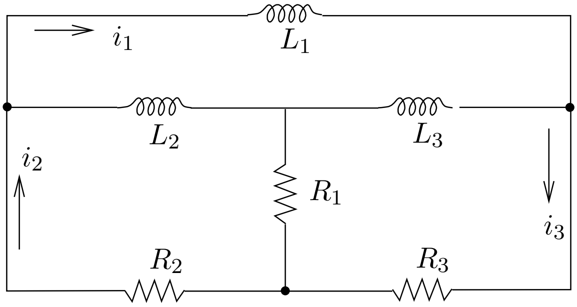


Figure 17. An electrical network.
There are three resistors R_1, R_2, R_3 and three inductors L_1, L_2, L_3. The currents i_1, i_2, i_3 are defined between nodes (black dots).

The derivation of the differential equations for the loop currents i_1, i_2, i_3 uses Kirchhoff's laws and the voltage drop formulas for resistors and inductors. The black dots in the diagram are the **nodes** that determine the beginning and end of each of the currents i_1, i_2, i_3. Currents are defined only on the outer boundary of the network. Kirchhoff's node law determines the currents across L_2, L_3

(arrowhead right) as $i_2 - i_1$ and $i_3 - i_1$, respectively. Similarly, $i_2 - i_3$ is the current across R_1 (arrowhead down). Using Ohm's law $V_R = RI$ and Faraday's law $V_L = LI'$ plus Kirchhoff's loop law *algebraic sum of the voltage drops is zero* around a closed loop (see the `maple` code below), we arrive at the model

$$\begin{aligned}
i_1' &= - \left(\frac{R_2}{L_1}\right) i_2 - \left(\frac{R_3}{L_1}\right) i_3, \\
i_2' &= - \left(\frac{R_2}{L_2} + \frac{R_2}{L_1}\right) i_2 + \left(\frac{R_1}{L_2} - \frac{R_3}{L_1}\right) i_3, \\
i_3' &= \left(\frac{R_1}{L_3} - \frac{R_2}{L_1}\right) i_2 - \left(\frac{R_1}{L_3} + \frac{R_3}{L_1} + \frac{R_3}{L_3}\right) i_3
\end{aligned}$$

A computer algebra system is helpful to obtain the differential equations from the closed loop formulas. Part of the theory is that the number of equations equals the number of *holes* in the network, called the **connectivity**. Here's some `maple` code for determining the equations in scalar and also in vector-matrix form.

```
loop1:=L1*D(i1)+R3*i3+R2*i2=0;
loop2:=L2*D(i2)-L2*D(i1)+R1*(i2-i3)+R2*i2=0;
loop3:=L3*D(i3)-L3*D(i1)+R3*i3+R1*(i3-i2)=0;
f1:=solve(loop1,D(i1));
f2:=solve(subs(D(i1)=f1,loop2),D(i2));
f3:=solve(subs(D(i1)=f1,loop3),D(i3));
with(linalg):
jacobian([f1,f2,f3],[i1,i2,i3]);
```

Electrical LR–Network with EMF

Consider the LR-network of Figure 18. This network produces only two differential equations, even though there are three *holes* (connectivity 3). The derivation of the differential equations parallels the previous network, so nothing will be repeated here.

A computer algebra system is used to obtain the differential equations from the closed loop formulas. Below is `maple` code to generate the equations $i_1' = f_1$, $i_2' = f_2$, $i_3 = f_3$.

```
loop1:=L1*D(i1)+R2*(i1-i2)+R1*(i1-i3)=0;
loop2:=L2*D(i2)+R3*(i2-i3)+R2*(i2-i1)=0;
loop3:=R3*(i3-i2)+R1*(i3-i1)=E;
f3:=solve(loop3,i3);
f1:=solve(subs(i3=f3,loop1),D(i1));
f2:=solve(subs(i3=f3,loop2),D(i2));
```

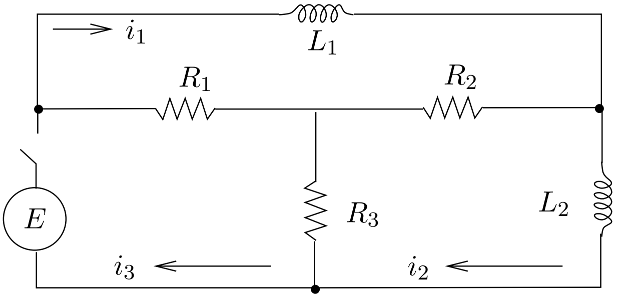


Figure 18. An electrical network.
There are three resistors R_1, R_2, R_3, two inductors L_1, L_2 and a battery E. The currents i_1, i_2, i_3 are defined between nodes (black dots).

The model, in the special case $L_1 = L_2 = 1$ and $R_1 = R_2 = R_3 = R$:

$$\begin{array}{rcrcrcr} i_1' & = & -\dfrac{3R}{2}i_1 & + & \dfrac{3R}{2}i_2 & + & \dfrac{E}{2}, \\ i_2' & = & \dfrac{3R}{2}i_1 & - & \dfrac{3R}{2}i_2 & + & \dfrac{E}{2}, \\ i_3 & = & \dfrac{1}{2}i_1 & + & \dfrac{1}{2}i_2 & + & \dfrac{E}{2R}. \end{array}$$

It is easily justified that the solution of the differential equations for initial conditions $i_1(0) = i_2(0) = 0$ is given by

$$i_1(t) = \frac{E}{2}t, \quad i_2(t) = \frac{E}{2}t.$$

Logging Timber by Helicopter

Certain sections of National Forest in the USA do not have logging access roads. In order to log the timber in these areas, helicopters are employed to move the felled trees to a nearby loading area, where they are transported by truck to the mill. The felled trees are slung beneath the helicopter on cables.

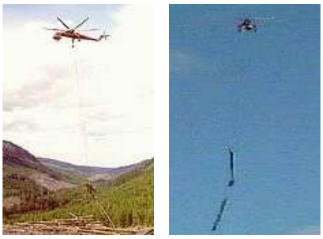


Figure 19. Helicopter logging.
Left: An Erickson helicopter lifts felled trees.
Right: Two trees are attached to the cable to lower transportation costs.

The payload for two trees approximates a double pendulum, which oscillates during flight. The angles of oscillation θ_1, θ_2 of the two connecting cables, measured from the gravity vector direction, satisfy the following differential equations, in which g is the gravitation constant, m_1, m_2 denote the masses of the two trees and L_1, L_2 are the cable lengths.

$$\begin{array}{rcrcr} (m_1+m_2)L_1^2\theta_1'' & + & m_2L_1L_2\theta_2'' & + & (m1+m_2)L_1g\theta_1 = 0, \\ m_2L_1L_2\theta_1'' & + & m_2L_2^2\theta_2'' & + & m_2L_2g\theta_2 = 0. \end{array}$$

This model is derived assuming small displacements θ_1, θ_2, that is, $\sin\theta \approx \theta$ for both angles, using the following diagram.

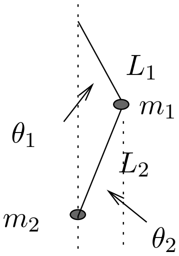


Figure 20. Logging Timber by Helicopter.
The cables have lengths L_1, L_2. The angles θ_1, θ_2 are measured from vertical.

The lengths L_1, L_2 are adjusted on each trip for the length of the trees, so that the trees do not collide in flight with each other nor with the helicopter. Sometimes, three or more smaller trees are bundled together in a package, which is treated here as identical to a single, very thick tree hanging on the cable.

Vector-matrix model. The angles θ_1, θ_2 satisfy the second-order vector-matrix equation

$$\begin{pmatrix} (m_1+m_2)L_1 & m_2L_2 \\ L_1 & L_2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}'' = -\begin{pmatrix} m_1g+m_2g & 0 \\ 0 & g \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.$$

This system is equivalent to the second-order system

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}'' = \begin{pmatrix} -\dfrac{m_1g+m_2g}{L_1m_1} & \dfrac{m_2g}{L_1m_1} \\ \dfrac{m_1g+m_2\,g}{L_2m_1} & -\dfrac{(m_1+m_2)\,g}{L_2m_1} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.$$

Earthquake Effects on Buildings

A horizontal earthquake oscillation $F(t) = F_0 \cos \omega t$ affects each floor of a 5-floor building; see Figure 21. The effect of the earthquake depends upon the natural frequencies of oscillation of the floors.

In the case of a single-floor building, the center-of-mass position $x(t)$ of the building satisfies $mx''+kx = E$ and the natural frequency of oscillation is $\sqrt{k/m}$. The earthquake force E is given by Newton's second law: $E(t) = -mF''(t)$. If $\omega \approx \sqrt{k/m}$, then the amplitude of $x(t)$ is large compared to the amplitude of the force E. The amplitude increase in $x(t)$ means that a small-amplitude earthquake wave can resonant with the building and possibly demolish the structure.

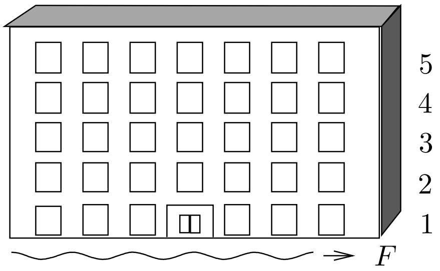


Figure 21. A 5-Floor Building.
A horizontal earthquake wave F affects every floor. The actual wave has wavelength many times larger than the illustration.

The following assumptions and symbols are used to quantize the oscillation of the 5-floor building.

- Each floor is considered a point mass located at its center-of-mass. The floors have masses $m_1, \ldots, m_5$.

- Each floor is restored to its equilibrium position by a linear restoring force or Hooke's force $-k$(elongation). The Hooke's constants are $k_1, \ldots, k_5$.

- The locations of masses representing the 5 floors are $x_1, \ldots, x_5$. The equilibrium position is $x_1 = \cdots = x_5 = 0$.

- Damping effects of the floors are ignored. This is a *frictionless* system.

The differential equations for the model are obtained by **competition**: the Newton's second law force is set equal to the sum of the Hooke's forces and the external force due to the earthquake wave. This results in the following system, where $k_6 = 0$, $E_j = m_j F''$ for $j = 1, 2, 3, 4, 5$ and $F = F_0 \cos \omega t$.

$$\begin{aligned}
m_1 x_1'' &= -(k_1 + k_2)x_1 + k_2 x_2 + E_1, \\
m_2 x_2'' &= k_2 x_1 - (k_2 + k_3)x_2 + k_3 x_3 + E_2, \\
m_3 x_3'' &= k_3 x_2 - (k_3 + k_4)x_3 + k_4 x_4 + E_3, \\
m_4 x_4'' &= k_4 x_3 - (k_4 + k_5)x_4 + k_5 x_5 + E_4, \\
m_5 x_5'' &= k_5 x_4 - (k_5 + k_6)x_5 + E_5.
\end{aligned}$$

In particular, the equations for a floor depend only upon the neighboring floors. The bottom floor and the top floor are exceptions: they have just one neighboring floor.

Vector-matrix second order system. Define

$$M = \begin{pmatrix} m_1 & 0 & 0 & 0 & 0 \\ 0 & m_2 & 0 & 0 & 0 \\ 0 & 0 & m_3 & 0 & 0 \\ 0 & 0 & 0 & m_4 & 0 \\ 0 & 0 & 0 & 0 & m_5 \end{pmatrix}, \quad \vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \quad \vec{\mathbf{H}} = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{pmatrix},$$

$$K = \begin{pmatrix} -k_1 - k_2 & k_2 & 0 & 0 & 0 \\ k_2 & -k_2 - k_3 & k_3 & 0 & 0 \\ 0 & k_3 & -k_3 - k_4 & k_4 & 0 \\ 0 & 0 & k_4 & -k_4 - k_5 & k_5 \\ 0 & 0 & 0 & k_5 & -k_5 - k_6 \end{pmatrix}.$$

In the last row, $k_6 = 0$, to reflect the absence of a floor above the fifth. The second order system is

$$M\vec{\mathbf{x}}''(t) = K\vec{\mathbf{x}}(t) + \vec{\mathbf{H}}(t).$$

The matrix M is called the **mass matrix** and the matrix K is called the **Hooke's matrix**. The **external force** $\vec{\mathbf{H}}(t)$ can be written as a scalar function $E(t) =$

$-F''(t)$ times a constant vector:

$$\vec{\mathbf{H}}(t) = -\omega^2 F_0 \cos\omega t \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{pmatrix}.$$

Identical floors. Let us assume that all floors have the same mass m and the same Hooke's constant k. Then $M = mI$ and the equation becomes

$$\vec{\mathbf{x}}'' = m^{-1} \begin{pmatrix} -2k & k & 0 & 0 & 0 \\ k & -2k & k & 0 & 0 \\ 0 & k & -2k & k & 0 \\ 0 & 0 & k & -2k & k \\ 0 & 0 & 0 & k & -k \end{pmatrix} \vec{\mathbf{x}} - F_0\omega^2 \cos(\omega t) \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

The Hooke's matrix K is symmetric ($K^T = K$) with negative entries only on the diagonal. The last diagonal entry is $-k$ (a common error is to write $-2k$).

Particular solution. The method of undetermined coefficients predicts a trial solution $\vec{\mathbf{x}}_p(t) = \vec{\mathbf{c}} \cos\omega t$, because each differential equation has nonhomogeneous term $-F_0\omega^2 \cos\omega t$. The constant vector $\vec{\mathbf{c}}$ is found by trial solution substitution. Cancel the common factor $\cos\omega t$ in the substituted equation to obtain the equation $\left(m^{-1}K + \omega^2\, I\right)\vec{\mathbf{c}} = F_0\omega^2\vec{\mathbf{b}}$, where $\vec{\mathbf{b}}$ is column vector of ones in the preceding display. Let $B(\omega) = m^{-1}K + \omega^2\, I$. Then the formula $B^{-1} = \dfrac{\textsf{adj}(B)}{\det(B)}$ gives

$$\vec{\mathbf{c}} = F_0\omega^2 \frac{\textsf{adj}(B(\omega))}{\det(B(\omega))}\vec{\mathbf{b}}.$$

The constant vector $\vec{\mathbf{c}}$ can have a large magnitude when $\det(B(\omega)) \approx 0$. This occurs when $-\omega^2$ is nearly an eigenvalue of $m^{-1}K$.

Homogeneous solution. The theory of this chapter gives the homogeneous solution $\vec{\mathbf{x}}_h(t)$ as the sum

$$\vec{\mathbf{x}}_h(t) = \sum_{j=1}^{5}(a_j \cos\omega_j t + b_j \sin\omega_j t)\vec{\mathbf{v}}_j$$

where $r = \omega_j$ and $\vec{\mathbf{v}} = \vec{\mathbf{v}}_j \neq \vec{\mathbf{0}}$ satisfy

$$\left(\frac{1}{m}K + r^2\, I\right)\vec{\mathbf{v}} = \vec{\mathbf{0}}.$$

Special case $k/m = 10$. Then

$$\frac{1}{m}K = \begin{pmatrix} -20 & 10 & 0 & 0 & 0 \\ 10 & -20 & 10 & 0 & 0 \\ 0 & 10 & -20 & 10 & 0 \\ 0 & 0 & 10 & -20 & 10 \\ 0 & 0 & 0 & 10 & -10 \end{pmatrix}$$

and the values $\omega_1, \ldots, \omega_5$ are found by solving equation $\det((1/m)K + \omega^2 I) = 0$, to obtain the values in Table 1.

Table 1. Natural Frequencies for the Special Case $k/m = 10$.

Frequency	Value
ω_1	0.900078068
ω_2	2.627315231
ω_3	4.141702938
ω_4	5.320554507
ω_5	6.068366391

General solution. Superposition implies $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_h(t) + \vec{\mathbf{x}}_p(t)$. Both terms of the general solution represent bounded oscillations.

Resonance effects. The special solution $\vec{\mathbf{x}}_p(t)$ can be used to obtain some insight into practical resonance effects between the incoming earthquake wave and the building floors. When ω is close to one of the frequencies $\omega_1, \ldots, \omega_5$, then the amplitude of a component of $\vec{\mathbf{x}}_p$ can be very large, causing the floor to take an excursion that is too large to maintain the structural integrity of the floor.

The **physical interpretation** is that an earthquake wave of the proper frequency, having time duration sufficiently long, can demolish a floor and hence demolish the entire building. The amplitude of the earthquake wave does not have to be large: a fraction of a centimeter might be enough to start the oscillation of the floors.

Earthquakes and Tsunamis

Seismic wave shape was studied for first order equations in Chapter 2 Section 8. Recorded here are some historical notes about seismic waves and earthquake events.

The original **Richter scale**, with deprecated use in seismology, was invented by seismologist C. Richter to rank earthquake power.

The moment magnitude scale (M_W) has largely replaced the original Richter scale and its modified versions. The highest reported magnitude is 9.5 M_W by the United States Geological Survey for the Concepción, Chile earthquake of May 22, 1960. News reports and the general public still refer to earthquake magnitude using the term *Richter Scale*.

The Sumatra earthquake of December 26, 2004 occurred close to a deep-sea trench, a subduction zone where one tectonic plate slips beneath another. Most of the earthquake energy is released in these areas as the two plates grind towards each other. Estimates of magnitude 8.8 M_W to 9.3 M_W followed the event. The US Geological Survey estimated 9.2 M_W.

The largest earthquake ever recorded was the 1960 Chile earthquake. There were three earthquakes May 21-22, 1960, estimated magnitudes 9.4 to 9.6. The tsunami caused by the Chile earthquake has been well-documented by Dr. Gerard Fryer of the Hawaii Institute of Geophysics and Planetology in Honolulu.

> What happened in the earthquake was that a piece of the Pacific seafloor (or strictly speaking, the Nazca Plate) about the size of California slid fifty feet beneath the continent of South America. Like a spring, the lower slopes of the South American continent offshore snapped upwards as much as twenty feet while land along the Chile coast dropped about ten feet. This change in the shape of the ocean bottom changed the shape of the sea surface. Since the sea surface likes to be flat, the pile of excess water at the surface collapsed to create a series of waves — the tsunami.
>
> The tsunami, together with the coastal subsidence and flooding, caused tremendous damage along the Chile coast, where about 2,000 people died. The waves spread outwards across the Pacific. About 15 hours later the waves flooded Hilo, on the island of Hawaii, where they built up to 30 feet and caused 61 deaths along the waterfront. Seven hours after that, 22 hours after the earthquake, the waves flooded the coastline of Japan where 10-foot waves caused 200 deaths. The waves also caused damage in the Marquesas, in Samoa, and in New Zealand. Tidal gauges throughout the Pacific measured anomalous oscillations for about three days as the waves bounced from one side of the ocean to the other.

Valdivia after the 22 May earthquake

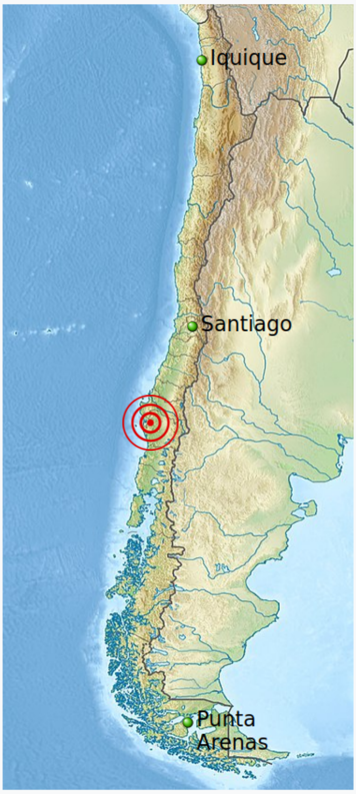


Image Source: Wikipedia 1960 Valdivia Chile Earthquakes

Exercises 11.1

There are no exercises for this section of examples. Later sections use this section for definitions, equations and key examples.

11.2 Fundamental System Methods

Solving 2×2 Systems

It is shown here that *any* constant linear system

$$\vec{\mathbf{x}}' = A\vec{\mathbf{x}}, \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

can be solved by one of the following elementary methods.

(a) The integrating factor method for $y' = p(x)y + q(x)$.

(b) The second order constant coefficient formulas in Chapter 6, Theorem 6.1.

Triangular 2×2 Matrix A

Let's assume $b = 0$ in matrix $A = \begin{pmatrix} a\, b \\ c\, d \end{pmatrix}$ making A lower triangular. The upper triangular case is handled similarly. Then $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ has the scalar form

$$\begin{aligned} x_1' &= ax_1, \\ x_2' &= cx_1 + dx_2. \end{aligned}$$

The first differential equation is solved by the growth/decay formula:

$$x_1(t) = x_0 e^{at}.$$

Then substitute the answer just found into the second differential equation to give

$$x_2' = dx_2 + cx_0 e^{at}.$$

This is a linear first order equation of the form $y' = p(x)y + q(x)$, to be solved by the integrating factor method. Therefore, a triangular system can always be solved by the first order integrating factor method.

An illustration. Let us solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ for the triangular matrix

$$A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \quad \text{representing} \quad \begin{cases} x_1' &= x_1, \\ x_2' &= 2x_1 + x_2. \end{cases}$$

The first equation $x_1' = x_1$ has solution $x_1 = c_1 e^t$. The second equation $x_2' = 2x_1 + x_2$ becomes upon substitution of $x_1 = c_1 e^t$ the new equation

$$x_2' = 2c_1 e^t + x_2,$$

which is a first order linear differential equation with linear integrating factor method solution $x_2 = (2c_1 t + c_2)e^t$. The general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ in scalar form is

$$x_1 = c_1 e^t, \quad x_2 = 2c_1 t e^t + c_2 e^t.$$

The **General Solution vector form** for $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is

$$\vec{\mathbf{x}}(t) = c_1 \begin{pmatrix} e^t \\ 2te^t \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ e^t \end{pmatrix}.$$

A **vector basis** $\mathcal{B}$ for the solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is

$$\mathcal{B} = \left\{ \begin{pmatrix} e^t \\ 2te^t \end{pmatrix}, \begin{pmatrix} 0 \\ e^t \end{pmatrix} \right\}.$$

Non-Triangular 2×2 Matrix A

In order that A be non-triangular, both $b \neq 0$ and $c \neq 0$ must be satisfied. The scalar form of the system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is

$$\left\{ \begin{array}{lcl} x_1' &=& ax_1 + bx_2, \\ x_2' &=& cx_1 + dx_2, \end{array} \right. \quad \vec{\mathbf{x}}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Theorem 11.1 (Solving 2×2 Non-Triangular $\vec{\mathrm{x}}\,' = A\,\vec{\mathrm{x}}$)
Solutions x_1, x_2 of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ are linear combinations of the list of Euler solution atoms obtained from roots r of $\det(A - rI) = 0$, which is the characteristic equation of A.

This result is called **Cayley-Hamilton-Ziebur** (abbreviated **CHZ**).

Proof: The method: differentiate the first equation, then use the equations to eliminate x_2, x_2'. The result is a second order differential equation for x_1. The same differential equation is satisfied also for x_2. The details:

$$\begin{aligned} x_1'' &= ax_1' + bx_2' && \text{Differentiate the first equation.} \\ &= ax_1' + bcx_1 + bdx_2 && \text{Use equation } x_2' = cx_1 + dx_2. \\ &= ax_1' + bcx_1 + d(x_1' - ax_1) && \text{Use equation } x_1' = ax_1 + bx_2. \\ &= (a+d)x_1' + (bc - ad)x_1 && \text{Second order equation for } x_1 \text{ found} \end{aligned}$$

The characteristic equation of $x_1'' - (a+d)x_1' + (ad - bc)x_1 = 0$ is

$$r^2 - (a+d)r + (bc - ad) = 0.$$

Finally, we show the expansion of $\det(A - rI)$ is the same characteristic polynomial:

$$\begin{array}{rcl} \det(A - rI) &=& \begin{vmatrix} a - r & b \\ c & d - r \end{vmatrix} \\ &=& (a-r)(d-r) - bc \\ &=& r^2 - (a+d)r + ad - bc. \end{array}$$

■

Proposition 11.1 (Differential Equation for x_1 and x_2)
Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then for $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$:

$$\begin{aligned}&\det(A - rI) = r^2 - \textbf{trace}(A)r + \det(A)\\&u'' - \textbf{trace}(A)u' + \det(A)u = 0 \text{ for } u = x_1, x_2\end{aligned}$$

Proof: The trace of A is $a+d$ and $\det(A) = ad-bc$. Apply proof details from Theorem 11.1. ■

Assume below that A is non-triangular, meaning $b \neq 0$ and $c \neq 0$.

How to Find x_1. Apply Chapter 6 Theorem 6.1 for equation $Ay''+By'+Cy = 0$ to solve for x_1. This involves writing a list of Euler solution atoms corresponding to the two roots of the characteristic equation $r^2-(a+d)r+ad-bc = 0$, followed by expressing x_1 as a linear combination of the two Euler atoms.

How to Find x_2. Isolate x_2 in the first differential equation by division:

$$x_2 = \frac{1}{b}(x_1' - ax_1).$$

The two formulas for x_1, x_2 represent the general solution of the system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, when A is 2×2.

An illustration. Let's solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ when

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \text{representing} \quad \begin{cases} x_1' &= x_1 + 2x_2, \\ x_2' &= 2x_1 + x_2. \end{cases}$$

The equation $\det(A-rI) = 0$ is $(1-r)^2-4 = 0$ with roots $r = -1$ and $r = 3$. The Euler solution atoms are e^{-t}, e^{3t}. Then $x_1 = c_1e^{-t} + c_2e^{3t}$, a linear combination of Euler solution atoms. The first equation $x_1' = x_1+2x_2$ implies $x_2 = \frac{1}{2}(x_1'-x_1)$ (we solve the first equation for symbol x_2). Insert $x_1 = c_1e^{-t}+c_2e^{3t}$ and simplify to find x_2 explicitly. The scalar general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is then

$$x_1 = c_1e^{-t} + c_2e^{3t}, \quad x_2 = -c_1e^{-t} + c_2e^{3t}.$$

In vector form, the general solution is

$$\vec{\mathbf{x}} = c_1\begin{pmatrix} e^{-t} \\ -e^{-t} \end{pmatrix} + c_2\begin{pmatrix} e^{3t} \\ e^{3t} \end{pmatrix}.$$

History. The fundamental idea in the illustration was developed by Ziebur using the classical **Cayley-Hamilton theorem**, which says that a square matrix satisfies its own characteristic equation. History suggests the name **Cayley-Hamilton-Ziebur** (abbreviated **CHZ**).

The Cayley-Hamilton theorem is the foundation for spectral methods developed in this chapter. Computer algebra systems provide algorithms for solving any system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$, possible because of the foundation provided by Cayley-Hamilton.

Method for $n \times n$ Diagonal A

If an $n \times n$ matrix A is diagonal, $A = \textbf{diag}(a_1, \ldots, a_n)$, then the system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is a set of uncoupled scalar growth/decay equations:

$$\begin{array}{rcl} x_1'(t) & = & a_1 x_1(t), \\ x_2'(t) & = & a_2 x_2(t), \\ & \vdots & \\ x_n'(t) & = & a_n x_n(t). \end{array}$$

The solution to the system is given by the formulas

$$\begin{array}{rcl} x_1(t) & = & c_1 e^{a_1 t}, \\ x_2(t) & = & c_2 e^{a_2 t}, \\ & \vdots & \\ x_n(t) & = & c_n e^{a_n t}. \end{array}$$

The numbers $c_1, \ldots, c_n$ are arbitrary constants.

Method for $n \times n$ Lower Triangular A

Assume a linear system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ has a square lower triangular matrix A. The system can be solved by first order scalar methods. To illustrate the ideas, consider the 3×3 lower triangular linear system

$$\vec{\mathbf{x}}' = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 3 & 0 \\ 4 & 4 & 4 \end{pmatrix} \vec{\mathbf{x}}.$$

In scalar form, the system is given by the equations

$$\begin{array}{rcl} x_1'(t) & = & 2x_1(t), \\ x_2'(t) & = & 3x_1(t) + 3x_2(t), \\ x_3'(t) & = & 4x_1(t) + 4x_2(t) + 4x_3(t). \end{array}$$

A recursive method. The system is solved recursively by first order scalar methods only, starting with the first equation $x_1'(t) = 2x_1(t)$. This growth equation has general solution $x_1(t) = c_1 e^{2t}$. The second equation then becomes the first order linear equation

$$\begin{array}{rcl} x_2'(t) & = & 3x_1(t) + 3x_2(t) \\ & = & 3x_2(t) + 3c_1 e^{2t}. \end{array}$$

The integrating factor method applies: $x_2(t) = -3c_1 e^{2t} + c_2 e^{3t}$ is the general solution. The third and last equation becomes the first order linear equation

$$\begin{array}{rcl} x_3'(t) & = & 4x_1(t) + 4x_2(t) + 4x_3(t) \\ & = & 4x_3(t) + 4c_1 e^{2t} + 4(-3c_1 e^{2t} + c_2 e^{3t}). \end{array}$$

The integrating factor method is repeated to find the general solution $x_3(t) = 4c_1e^{2t} - 4c_2e^{3t} + c_3e^{4t}$.

In summary, the scalar general solution to the system is given by the formulas

$$\begin{aligned} x_1(t) &= c_1e^{2t}, \\ x_2(t) &= -3c_1e^{2t} + c_2e^{3t}, \\ x_3(t) &= 4c_1e^{2t} - 4c_2e^{3t} + c_3e^{4t}. \end{aligned}$$

Structure of solutions. A system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ for $n \times n$ triangular A has component solutions $x_1(t)$, …, $x_n(t)$ given as polynomials times exponentials. The exponential factors $e^{a_{11}t}$, …, $e^{a_{nn}t}$ are expressed in terms of the diagonal elements a_{11}, …, a_{nn} of the matrix A. Fewer than n distinct exponential factors may appear, due to duplicate diagonal elements. These duplications cause the polynomial factors to appear. The reader is invited to work out the solution to the system below, which has duplicate diagonal entries $a_{11} = a_{22} = a_{33} = 2$.

$$\begin{aligned} x_1'(t) &= 2x_1(t), \\ x_2'(t) &= 3x_1(t) + 2x_2(t), \\ x_3'(t) &= 4x_1(t) + 4x_2(t) + 2x_3(t). \end{aligned}$$

The solution, given below, has polynomial factors t and t^2, appearing because of the duplicate diagonal entries $2, 2, 2$, and only one exponential factor e^{2t}.

$$\begin{aligned} x_1(t) &= c_1e^{2t}, \\ x_2(t) &= 3c_1te^{2t} + c_2e^{2t}, \\ x_3(t) &= 4c_1te^{2t} + 6c_1t^2e^{2t} + 4c_2te^{2t} + c_3e^{2t}. \end{aligned}$$

Method for $n \times n$ Upper Triangular A

A matrix differential system $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$ with T upper triangular splits into scalar equations which can be solved by elementary methods for first order scalar differential equations. To illustrate, consider the system

$$\begin{aligned} y_1' &= 3y_1 + y_2 + y_3, \\ y_2' &= 3y_2 + y_3, \\ y_3' &= 2y_3. \end{aligned}$$

The techniques that apply are the growth-decay formula for $u' = ku$ and the integrating factor method for $u' = ku + p(t)$. Working backwards from the last equation with back-substitution gives

$$\begin{aligned} y_3 &= c_3e^{2t}, \\ y_2 &= c_2e^{3t} - c_3e^{2t}, \\ y_1 &= (c_1 + c_2t)e^{3t}. \end{aligned}$$

Jordan's $n \times n$ Variable Change for $\vec{x}' = A\vec{x}$

What has been said above applies to any triangular system $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$, in order to write an exact formula for the solution $\vec{\mathbf{y}}(t)$.

If A is an $n \times n$ matrix, then Jordan's theorem gives $A = PTP^{-1}$ with T upper triangular and P invertible. The change of variable $\vec{\mathbf{x}}(t) = P\vec{\mathbf{y}}(t)$ changes $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$ into the triangular system $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$.

There is no special condition on A to effect the change of variable $\vec{\mathbf{x}}(t) = P\vec{\mathbf{y}}(t)$. The solution $\vec{\mathbf{x}}(t)$ of $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$ is a product of the invertible matrix P and a column vector $\vec{\mathbf{y}}(t)$; the latter is the solution of the triangular system $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$, obtained by growth-decay and integrating factor methods.

The *importance of this idea* is to provide a reliable method for solving any system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$. Later in this chapter, we outline how to find the matrix P and the matrix T in Jordan's theorem $A = PTP^{-1}$. The additional theory provides both desktop paper-and-pencil and computer matrix methods for solving any system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$.

Differential Equation Conversion to $\vec{x}' = A\vec{x}$

Considered here are source equations in scalar form or in vector form. The object is to define a new vector variable $\vec{\mathbf{x}}(t)$ and a matrix A which converts the source equations into the system form $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$. The ideas apply as well to systems of nonlinear and/or non-homogeneous equations with higher derivatives, the converted system having the nonlinear form $\vec{\mathbf{x}}' = \vec{\mathbf{f}}(t, \vec{\mathbf{x}})$, a form precursor to applying computer numerical methods. The list of source equations to be considered:

Scalar linear 2nd order $au'' + bu' + cu = f$

Scalar linear 2nd order system $\left\{\begin{array}{rcl} a_1x_1'' + b_1x_1' + c_1x_1 &=& f_1, \\ a_2x_2'' + b_2x_2' + c_2x_2 &=& f_2. \end{array}\right.$

Coupled spring-mass $\begin{cases} m_1x_1''(t) = -k_1x_1(t) + k_2(x_2(t) - x_1(t)), \\ m_2x_2''(t) = -k_2(x_2(t) - x_1(t)) + k_3(u_3(t) - x_2(t)), \\ m_3u_3''(t) = -k_3(u_3(t) - x_2(t)) - k_4u_3(t). \end{cases}$

Scalar linear nth order $y^{(n)} = p_0y + \cdots + p_{n-1}y^{(n-1)}$

Scalar continuous coefficients $y^{iv} = a(x)y + b(x)y' + c(x)y'' + d(x)y'''$

Forced higher order $y^{iv} = 2y + \sin(x)y' + \cos(x)y'' + x^2y''' + f(x)$.

Second order system $M\vec{\mathbf{x}}'' = K\vec{\mathbf{x}}$

Forced second order system $M\vec{\mathbf{x}}'' = K\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$.

Damped Forced system $M\vec{\mathbf{x}}'' = B\vec{\mathbf{x}}' + K\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

Convert Scalar Linear 2nd Order to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

Consider an equation $au'' + bu' + cu = f$ where $a \neq 0$, b, c, f are allowed to depend on t, $' = d/dt$. Define the **Position-Velocity substitution**

$$x_1(t) = u(t), \quad x_2(t) = u'(t).$$

Then $x_1' = u' = x_2$ and $x_2' = u'' = (-bu' - cu + f)/a = -(b/a)x_2 - (c/a)x_1 + f/a$. The resulting system is equivalent to the second order equation, in the sense that the position-velocity substitution transforms solutions of one system to the other:

$$\begin{cases} x_1'(t) &= &(0)\,x_1(t) &+ &(1)\,x_2(t), \\ x_2'(t) &= &-\left(\frac{c(t)}{a(t)}\right)x_1(t) &- &\left(\frac{b(t)}{a(t)}\right)x_2(t) &+ &\frac{f(t)}{a(t)}. \end{cases}$$

The case of constant coefficients and f a function of t arises often enough to isolate the result for further reference.

Theorem 11.2 (Constant-Coefficient 2nd Order Conversion)
Let $a \neq 0$, b, c be constants and $f(t)$ continuous. Then $au'' + bu' + cu = f(t)$ is equivalent to the first order system

$$a\vec{\mathbf{x}}'(t) = \begin{pmatrix} 0 & a \\ -c & -b \end{pmatrix} \vec{\mathbf{w}}(t) + \begin{pmatrix} 0 \\ f(t) \end{pmatrix}, \quad \vec{\mathbf{x}}(t) = \begin{pmatrix} u(t) \\ u'(t) \end{pmatrix}.$$

Convert Second Order Scalar Systems to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

A position-velocity substitution can be carried out on a system of two second order linear differential equations. Assume

$$\begin{cases} a_1x_1'' + b_1x_1' + c_1x_1 &= f_1, \\ a_2x_2'' + b_2x_2' + c_2x_2 &= f_2. \end{cases}$$

Then the preceding methods for the scalar case give the equivalence

$$\begin{pmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_1 & 0 & 0 \\ 0 & 0 & a_2 & 0 \\ 0 & 0 & 0 & a_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_1' \\ x_2 \\ x_2' \end{pmatrix}' = \begin{pmatrix} 0 & a_1 & 0 & 0 \\ -c_1 & -b_1 & 0 & 0 \\ 0 & 0 & 0 & a_2 \\ 0 & 0 & -c_2 & -b_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_1' \\ x_2 \\ x_2' \end{pmatrix} + \begin{pmatrix} 0 \\ f_1 \\ 0 \\ f_2 \end{pmatrix}.$$

Convert Coupled Spring-Mass Systems to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

Springs connecting undamped coupled masses were considered at the beginning of this chapter, page 827. Typical equations are

$$(1) \qquad \begin{cases} m_1x_1''(t) &= -k_1x_1(t) + k_2(x_2(t) - x_1(t)), \\ m_2x_2''(t) &= -k_2(x_2(t) - x_1(t)) + k_3(u_3(t) - x_2(t)), \\ m_3u_3''(t) &= -k_3(u_3(t) - x_2(t)) - k_4u_3(t). \end{cases}$$

The equations can be represented by a second order linear system of dimension 3 of the form $M\vec{\mathbf{x}}'' = K\vec{\mathbf{x}}$, where the **Vector Position** $\vec{\mathbf{x}}$, the **mass matrix** M and the **Hooke's matrix** K are given by the equalities

$$\vec{\mathbf{x}}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ u_3(t) \end{pmatrix}, \quad M = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix},$$

$$K = \begin{pmatrix} -(k_1+k_2) & k_2 & 0 \\ k_2 & -(k_2+k_3) & k_3 \\ 0 & -k_3 & -(k_3+k_4) \end{pmatrix}.$$

Conversion to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ uses a position-velocity substitution to obtain the block matrix multiply equation (I = identity matrix, 0 = zero matrix)

$$\vec{\mathbf{x}}(t) = \begin{pmatrix} \vec{\mathbf{x}}(t) \\ \vec{\mathbf{x}}'(t) \end{pmatrix}, \quad \vec{\mathbf{x}}'(t) = \left(\begin{array}{c|c} 0 & I \\ \hline M^{-1}K & 0 \end{array}\right) \vec{\mathbf{x}}(t).$$

Convert Higher Order Linear Equations to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

Every homogeneous nth order linear differential equation

$$y^{(n)} = p_0 y + \cdots + p_{n-1} y^{(n-1)}$$

with constant coefficients can be converted to a linear homogeneous vector-matrix system

$$\frac{d}{dx}\begin{pmatrix} y \\ y' \\ y'' \\ \vdots \\ y^{(n-1)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \\ p_0 & p_1 & p_2 & \cdots & p_{n-1} \end{pmatrix} \begin{pmatrix} y \\ y' \\ y'' \\ \vdots \\ y^{(n-1)} \end{pmatrix}.$$

This is a linear system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ where $\vec{\mathbf{x}}(t)$ is the $n \times 1$ column vector consisting of $y(t)$ and its successive derivatives, while the $n \times n$ matrix A is the classical **Companion Matrix**[5] of the characteristic polynomial

$$r^n = p_0 + p_1 r + p_2 r^2 + \cdots + p_{n-1} r^{n-1}.$$

To illustrate, the companion matrix (page 846) for $r^4 = a + br + cr^2 + dr^3$ is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a & b & c & d \end{pmatrix}.$$

[5] The transpose of the companion matrix defined in Wikipedia. The companion matrix or its transpose appears in advanced topics in linear algebra, e.g. the Frobenius Rational Form.

The preceding companion matrix has the following block matrix form, which is representative of all companion matrices.

$$A = \left(\begin{array}{c|ccc} \vec{\mathbf{0}} & & I & \\ \hline a & b & c & d \end{array}\right).$$

Convert Scalar Continuous-Coefficient Equations to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

. Methods above apply equally to higher order linear differential equations with continuous coefficients. To illustrate, the fourth order linear equation $y^{iv} = a(x)y + b(x)y' + c(x)y'' + d(x)y'''$ has first order system form $\vec{\mathbf{x}}' = A(x)\vec{\mathbf{x}}$ where $A(x)$ is the companion matrix (page 846) for the polynomial $r^4 = a(x) + b(x)r + c(x)r^2 + d(x)r^3$, x held fixed:

$$A(x) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a(x) & b(x) & c(x) & d(x) \end{pmatrix}.$$

Convert Forced Higher Order Equations to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

All that has been said above applies equally to a forced linear equation like

$$y^{iv} = 2y + \sin(x)y' + \cos(x)y'' + x^2y''' + f(x).$$

It has a conversion to a first order nonhomogeneous linear system

$$\vec{\mathbf{x}}' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 2 & \sin x & \cos x & x^2 \end{pmatrix}\vec{\mathbf{x}} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ f(x) \end{pmatrix}, \quad \vec{\mathbf{x}} = \begin{pmatrix} y \\ y' \\ y'' \\ y''' \end{pmatrix}.$$

Convert 2nd Order System to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

A second order system $M\vec{\mathbf{x}}'' = K\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ is called a **forced system** and $\vec{\mathbf{F}}$ is called the **external vector force**. Such a system can always be converted to a second order system where the mass matrix is the identity, by multiplying by M^{-1}:

$$\vec{\mathbf{x}}'' = M^{-1}K\vec{\mathbf{x}} + M^{-1}\vec{\mathbf{F}}(t).$$

The benign form $\vec{\mathbf{x}}'' = B\vec{\mathbf{x}} + \vec{\mathbf{G}}(t)$, where $B = M^{-1}K$ and $\vec{\mathbf{G}} = M^{-1}\vec{\mathbf{F}}$, admits a block matrix conversion into a forced first order system of the form $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{f}}(t)$:

$$\vec{\mathbf{x}}(t) = \begin{pmatrix} \vec{\mathbf{x}}(t) \\ \vec{\mathbf{x}}'(t) \end{pmatrix}, \quad \frac{d}{dt}\vec{\mathbf{x}}(t) = \left(\begin{array}{c|c} 0 & I \\ \hline M^{-1}K & 0 \end{array}\right)\vec{\mathbf{x}}(t) + \begin{pmatrix} \vec{\mathbf{0}} \\ M^{-1}\vec{\mathbf{F}}(t) \end{pmatrix}.$$

Convert Damped 2nd Order System to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

The addition of a dashpot to each of the masses gives a **damped** second order system with forcing term

$$M\vec{\mathbf{x}}'' = B\vec{\mathbf{x}}' + K\vec{\mathbf{x}} + \vec{\mathbf{F}}(t).$$

In the case of one scalar equation, the matrices M, B, K are constants m, $-c$, $-k$ and the external force is a scalar function $f(t)$, hence the system becomes the classical damped spring-mass equation

$$mu'' + cu' + ku = f(t).$$

A standard way to write the first order system $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{G}}(t)$ is to introduce variable $\vec{\mathbf{u}} = M\begin{pmatrix}\vec{\mathbf{x}}\\ \vec{\mathbf{x}}'\end{pmatrix}$, in order to obtain

$$\vec{\mathbf{u}}' = M\frac{d}{dt}\begin{pmatrix}M\vec{\mathbf{x}}\\ \vec{\mathbf{x}}'\end{pmatrix} = M\begin{pmatrix}\vec{\mathbf{x}}'\\ \vec{\mathbf{x}}''\end{pmatrix} = M\begin{pmatrix}\vec{\mathbf{x}}'\\ B\vec{\mathbf{x}}' + K\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)\end{pmatrix}$$

Then a first order system in block matrix form is given by

$$\left(\begin{array}{c|c} M & 0 \\ \hline 0 & M \end{array}\right)\frac{d}{dt}\begin{pmatrix}\vec{\mathbf{x}}(t)\\ \vec{\mathbf{x}}'(t)\end{pmatrix} = \left(\begin{array}{c|c} 0 & M \\ \hline K & B \end{array}\right)\begin{pmatrix}\vec{\mathbf{x}}(t)\\ \vec{\mathbf{x}}'(t)\end{pmatrix} + \begin{pmatrix}\vec{\mathbf{0}}\\ \vec{\mathbf{F}}(t)\end{pmatrix}.$$

The benign form $\vec{\mathbf{x}}'' = M^{-1}B\vec{\mathbf{x}}' + M^{-1}K\vec{\mathbf{x}} + M^{-1}\vec{\mathbf{F}}(t)$, which is obtained from left-multiplication by M^{-1}, can be similarly written as a first order system in block matrix form.

$$\frac{d}{dt}\begin{pmatrix}\vec{\mathbf{x}}(t)\\ \vec{\mathbf{x}}'(t)\end{pmatrix} = \left(\begin{array}{c|c} 0 & I \\ \hline M^{-1}K & M^{-1}B \end{array}\right)\begin{pmatrix}\vec{\mathbf{x}}(t)\\ \vec{\mathbf{x}}'(t)\end{pmatrix} + \begin{pmatrix}\vec{\mathbf{0}}\\ M^{-1}\vec{\mathbf{F}}(t)\end{pmatrix}$$

Exercises 11.2

Solving 2 × 2 Systems

1. Solve $x_1' = 2x_1 + x_2$, $x_2' = x_2$. Ans: $x_1 = c_1\,\mathrm{e}^{2t} - c_2\,\mathrm{e}^{t}$, $x_2 = c_2\,\mathrm{e}^{t}$

2. Discuss how to solve $\vec{\mathbf{x}}' = \begin{pmatrix}a & b\\ 0 & d\end{pmatrix}\vec{\mathbf{x}}$.

Triangular 2 × 2 Matrix A

3. Solve $\vec{\mathbf{x}}' = \begin{pmatrix}2 & 1\\ 0 & 3\end{pmatrix}\vec{\mathbf{x}}$.

4. Solve $\vec{\mathbf{x}}' = \begin{pmatrix}2 & 0\\ 2 & 3\end{pmatrix}\vec{\mathbf{x}}$.

Non-Triangular 2 × 2 Matrix A

5. Solve $\vec{\mathbf{x}}' = \begin{pmatrix}1 & 3\\ 3 & 1\end{pmatrix}\vec{\mathbf{x}}$.

6. Solve $\vec{\mathbf{x}}' = \begin{pmatrix}1 & 3\\ -3 & 1\end{pmatrix}\vec{\mathbf{x}}$.

Method for $n \times n$ Diagonal A

7. Solve $\vec{\mathbf{x}}' = \begin{pmatrix}1 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 2\end{pmatrix}\vec{\mathbf{x}}$.

8. Solve $\vec{\mathbf{x}}' = \begin{pmatrix} 1&0&0&0\\ 0&3&0&0\\ 0&0&4&0\\ 0&0&0&2 \end{pmatrix} \vec{\mathbf{x}}$.

Method for $n \times n$ Lower Triangular

9. Solve $\vec{\mathbf{x}}' = \begin{pmatrix} 1&0&0\\ 1&3&0\\ 1&0&2 \end{pmatrix} \vec{\mathbf{x}}$.

10. Solve $\vec{\mathbf{x}}' = \begin{pmatrix} 1&0&0\\ 0&3&0\\ 1&0&2 \end{pmatrix} \vec{\mathbf{x}}$.

Method for $n \times n$ Upper Triangular

11. Solve $\vec{\mathbf{x}}' = \begin{pmatrix} 1&0&1\\ 0&3&1\\ 0&0&2 \end{pmatrix} \vec{\mathbf{x}}$.

12. Solve $\vec{\mathbf{x}}' = \begin{pmatrix} 1&1&0\\ 0&3&1\\ 0&0&2 \end{pmatrix} \vec{\mathbf{x}}$.

Jordan's $n \times n$ Variable Change

Let $A = PTP^{-1}$ with T upper triangular and P invertible. Define change of variable $\vec{\mathbf{x}}(t) = P\vec{\mathbf{y}}(t)$. Prove these results:

13. If $\vec{\mathbf{x}}(t)$ solves $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$, then $\vec{\mathbf{y}}(t) = P^{-1}\vec{\mathbf{x}}(t)$ solves $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$.

14. If $\vec{\mathbf{y}}'(t) = T\vec{\mathbf{y}}(t)$, then $\vec{\mathbf{x}}(t) = P\vec{\mathbf{y}}(t)$ solves $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$.

Convert Scalar Linear 2nd Order to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$

15. $x'' + 2x' + x = \sin t$

16. $2x'' + 3x' + 8x = 4\cos t$

Convert Second Order Scalar System to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$

17. $x'' = x + y$, $y'' = x - y$

18. $x'' = x + y + \sin t$, $y'' + y = x + \cos t$

Convert Coupled Spring-Mass System to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}$

19. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1\\ 1&-1 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 0\\ \sin t \end{pmatrix}$

20. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1&0\\ 1&-2&1\\ 0&-1&-2 \end{pmatrix} \vec{\mathbf{x}}$

Convert Higher Order Linear Equations to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$

21. $x''' = x$

22. $\dfrac{d^4y}{dx^4} + 16y = 0$

Convert Scalar Continuous-Coefficient Equation to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$

23. $x^2y'' + 3xy' + 2y = 0$

24. $y''' + xy'' + x^2y + y = 0$

Convert Forced Higher Order Equation to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$

25. $\dfrac{d^4y}{dx^4} = y''' + y + \sin x$

26. $\dfrac{d^6y}{dx^6} = \dfrac{d^4y}{dx^4} + y + \cos t$

Convert 2nd Order System to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{G}}(t)$

27. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1\\ 1&-1 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 1\\ -1 \end{pmatrix}$

28. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1&0\\ 1&-2&1\\ 0&-1&-2 \end{pmatrix} \vec{\mathbf{x}} + e^t \begin{pmatrix} 1\\ 1\\ 1 \end{pmatrix}$

Convert Damped 2nd Order System to $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{G}}(t)$

29. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1\\ 1&-1 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 0&1\\ 1&0 \end{pmatrix} \vec{\mathbf{x}}' + \begin{pmatrix} 1\\ -1 \end{pmatrix}$

30. $\vec{\mathbf{x}}'' = \begin{pmatrix} -2&1&0\\ 1&-2&1\\ 0&-1&-2 \end{pmatrix} \vec{\mathbf{x}} + \vec{\mathbf{x}}' + e^t \begin{pmatrix} 1\\ 1\\ 1 \end{pmatrix}$

11.3 Structure of Linear Systems

Notation for Linear Systems

A **linear system** is a system of differential equations of the form

$$
\begin{array}{ccccccccc}
x_1' & = & a_{11}x_1 & + & \cdots & + & a_{1n}x_n & + & f_1, \\
x_2' & = & a_{21}x_1 & + & \cdots & + & a_{2n}x_n & + & f_2, \\
\vdots & & & \vdots & \cdots & \vdots & & \vdots & \\
x_m' & = & a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & + & f_m,
\end{array}
\tag{1}
$$

where $' = d/dt$. Given are the functions $a_{ij}(t)$ and $f_j(t)$ on some interval $a < t < b$. The unknowns are the functions $x_1(t), \ldots, x_n(t)$.

The system is called **homogeneous** if all $f_j = 0$, otherwise it is called **non-homogeneous**.

Matrix Notation. A non-homogeneous system of linear equations (1) is written as the equivalent vector-matrix system

$$\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$$

where

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \vec{\mathbf{F}} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

Existence-Uniqueness

Special results are isolated to illustrate how Picard-Lindelöf theory is applied to linear systems. Proofs start on page 859.

Theorem 11.3 (Gronwall's Lemma)
Let $u(t), v(t)$ be continuous functions with $v(t) \ge 0$ on interval $t_0 \le t \le t_0 + H$. Assume $u(t) \le c + \int_{t_0}^t u(r)v(r)dr$ for t for $t_0 \le t \le t_0 + H$. Then:

$$u(t) \le c\,e^{-\int_{t_0}^t v(r)dr}, \quad t_0 \le t \le t_0 + H.$$

Theorem 11.4 (Unique Zero Solution)
Let $A(t)$ be an $m \times n$ matrix with entries continuous on $t_0 \le t \le t_0 + H$. Then the initial value problem

$$\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}, \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{0}}$$

has unique solution $\vec{\mathbf{x}}(t) = \vec{\mathbf{0}}$ on $t_0 \le t \le t_0 + H$.

Theorem 11.5 (Picard-Lindelöf)
Let n-vector $\vec{\mathbf{F}}(t)$ and $n \times n$ matrix $A(t)$ be continuous on interval J: $a < t < b$. Let t_0 be in J. Let $\vec{\mathbf{x}}_0$ be in $\mathcal{R}^n$. Then the initial value problem

$$\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t), \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$$

has a unique solution $\vec{\mathbf{x}}(t)$ defined on all of J.

Theorem 11.6 (Existence-Uniqueness for Constant Linear Systems)
Let $A(t) = A$ be an $m \times n$ matrix with constant entries and let t_0 be any real number and let $\vec{\mathbf{x}}_0$ be any n-vector. Then the initial value problem

$$\vec{\mathbf{x}}' = A\vec{\mathbf{x}}, \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$$

has a unique solution $\vec{\mathbf{x}}(t)$ defined for all values of t.

Theorem 11.7 (Uniqueness and Solution Crossings)
Let $A(t)$ be an $m \times n$ matrix with entries continuous on $a < t < b$ and assume $\vec{\mathbf{F}}(t)$ is also continuous on $a < t < b$. If $\vec{\mathbf{x}}_1(t)$ and $\vec{\mathbf{x}}_2(t)$ are solutions of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ on $a < t < b$ and $\vec{\mathbf{x}}_1(t_0) = \vec{\mathbf{x}}_2(t_0)$ for some t_0, $a < t_0 < b$, then $\vec{\mathbf{x}}_1(t) = \vec{\mathbf{x}}_2(t)$ for $a < t < b$.

Linearity and Superposition

Linear homogeneous systems have **linear structure** and nonhomogeneous systems obey a **Principle of Superposition**.

Theorem 11.8 (Linear Structure)
Let $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ have two solutions $\vec{\mathbf{x}}_1(t)$, $\vec{\mathbf{x}}_2(t)$. If k_1, k_2 are constants, then $\vec{\mathbf{x}}(t) = k_1\vec{\mathbf{x}}_1(t) + k_2\vec{\mathbf{x}}_2(t)$ is also a solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$.

Theorem 11.9 (Basis)
The solution set of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ is an n-dimensional subspace of the vector space of all vector-valued functions $\vec{\mathbf{x}}(t)$ on $a < t < b$.

Let $a < t_0 < b$. A **standard basis** $\vec{\mathbf{w}}_1(t), \ldots, \vec{\mathbf{w}}_n(t)$ is defined by $\vec{\mathbf{w}}_j'(t) = A(t)\vec{\mathbf{w}}_j(t)$, $\vec{\mathbf{w}}_j(t_0) = \vec{\mathbf{e}}_j =$ column j of the identity matrix I, $1 \le j \le n$.

Every solution $\vec{\mathbf{x}}(t)$ of $\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t)$ has a unique basis expansion:

$$\vec{\mathbf{x}}(t) = c_1\vec{\mathbf{w}}_1(t) + c_2\vec{\mathbf{w}}_2(t) + \cdots + c_n\vec{\mathbf{w}}_n(t)$$

Theorem 11.10 (Superposition Principle)
Let $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ have a particular solution $\vec{\mathbf{x}}_p(t)$. If $\vec{\mathbf{x}}(t)$ is any solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$, then $\vec{\mathbf{x}}(t)$ can be decomposed as **homogeneous plus particular**:

$$\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_h(t) + \vec{\mathbf{x}}_p(t).$$

Term $\vec{\mathbf{x}}_h(t)$ is a certain solution of the homogeneous differential equation $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$, which means arbitrary constants c_1, c_2, ... have been assigned specific values. The shortest particular solution $\vec{\mathbf{x}}_p(t)$ excludes any term $\vec{\mathbf{y}}(t)$ satisfying $\vec{\mathbf{y}}'t() = A(t)\vec{\mathbf{y}}(t)$, such terms being absorbed into $\vec{\mathbf{x}}_h(t)$.

Theorem 11.11 (Difference of Solutions)
Let $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ have two solutions $\vec{\mathbf{x}} = \vec{\mathbf{u}}(t)$ and $\vec{\mathbf{x}} = \vec{\mathbf{v}}(t)$. Define $\vec{\mathbf{y}}(t) = \vec{\mathbf{u}}(t) - \vec{\mathbf{v}}(t)$. Then $\vec{\mathbf{y}}(t)$ satisfies the homogeneous equation

$$\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}.$$

General Solution

The general solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ is an expression involving arbitrary constants c_1, c_2, ... and certain functions. The expression may be given in vector notation, although scalar expressions are commonplace and perfectly acceptable. Required is that the expression represents all solutions of the differential equation, in the following sense:

Definition 11.1 (General Solution of $\vec{\mathbf{x}}\,' = A(t)\,\vec{\mathbf{x}} + \vec{\mathbf{F}}\,(t)$)
An expression is called a **general solution** of system $\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t)$ provided:

(a) Every **assignment of constants** in the expression produces a solution of the differential equation.

(b) Every possible solution is **uniquely** obtained from the expression by **specializing the constants**.

Superposition Theorem 11.10 implies that the constants in the general solution are identified as multipliers against solutions of the homogeneous differential equation. The general solution has recognizable structure:

Theorem 11.12 (General Solution)
Let $A(t)$ be an $n \times n$ matrix. Let $\vec{\mathbf{F}}(t)$ be an $n \times 1$ vector. Assume $A(t)$ and $\vec{\mathbf{F}}(t)$ are continuous on an interval $a < t < b$. Then linear nonhomogeneous system $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ has general solution $\vec{\mathbf{x}}$ given by the expression

$$\vec{\mathbf{x}} = \vec{\mathbf{x}}_h(t) + \vec{\mathbf{x}}_p(t).$$

1. Term $\vec{\mathbf{y}} = \vec{\mathbf{x}}_h(t)$ is a general solution of the homogeneous equation $\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}$ which contains n arbitrary constants c_1, ..., c_n.

2. Term $\vec{\mathbf{x}} = \vec{\mathbf{x}}_p(t)$ is a particular solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$.

Recognition of Homogeneous Solution Terms

Assume given an expression $\vec{\mathbf{x}}$ for the general solution of vector-matrix equation $\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t)$, either in scalar component form or in vector form. Expression $\vec{\mathbf{x}}$ contains arbitrary constants $c_1, \ldots, c_n$. It is possible to isolate both terms $\vec{\mathbf{x}}_h$ and $\vec{\mathbf{x}}_p$ by a simple procedure.

Finding $\vec{\mathbf{x}}_p$. The first step: **set to zero** all arbitrary constants $c_1, c_2, \ldots, c_n$. The resulting expression is free of unresolved constants. The answer sought for $\vec{\mathbf{x}}_p(t)$ has no term $\vec{\mathbf{y}}(t)$ with $A(t)\vec{\mathbf{y}}(t) = \vec{\mathbf{0}}$. If the expression contains such a term $\vec{\mathbf{y}}$, then remove it. Repeat inspection and removal until no such term $\vec{\mathbf{y}}$ appears. If the expression $\vec{\mathbf{x}}$ consists of equations in scalar component form, then assemble the modified equations into vector $\vec{\mathbf{x}}_p$. Otherwise, the modified $\vec{\mathbf{x}}$ is vector $\vec{\mathbf{x}}_p$.

Finding $\vec{\mathbf{x}}_h$. The first step: take partial derivatives on the general solution expression $\vec{\mathbf{x}}$ with respect to the symbols $c_1, \ldots, c_n$. The formula:

$$\vec{\mathbf{u}}_k(t) = \frac{\partial}{\partial c_k}\vec{\mathbf{x}}, \quad 1 \le k \le n.$$

A vector solution basis for $\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}$ is $\{\vec{\mathbf{u}}_k\}_{k=1}^n$. The technique isolates the vector components of the homogeneous solution from any form of the general solution, including scalar formulas for the components of $\vec{\mathbf{x}}$. Then:

$$\vec{\mathbf{x}}_h(t) = c_1\vec{\mathbf{u}}_1(t) + c_2\vec{\mathbf{u}}_2(t) + \cdots + c_n\vec{\mathbf{u}}_n(t).$$

Vector General Solution. A general solution $\vec{\mathbf{x}}$ of the nonhomogeneous linear system $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ is given by the expression

$$\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1(t) + c_2\vec{\mathbf{u}}_2(t) + \cdots + c_n\vec{\mathbf{u}}_n(t) + \vec{\mathbf{x}}_p(t).$$

In this expression, each *assignment* of the constants $c_1, \ldots, c_n$ produces a solution of the nonhomogeneous system, and conversely, each possible solution of the nonhomogeneous system is obtained by a unique *specialization* of the constants $c_1, \ldots, c_n$.

Independence

Constants $c_1, \ldots, c_n$ in the general solution $\vec{\mathbf{x}} = \vec{\mathbf{x}}_h + \vec{\mathbf{x}}_p$ appear exactly in the expression $\vec{\mathbf{x}}_h$, which has the form

$$\vec{\mathbf{x}}_h = c_1\vec{\mathbf{u}}_1 + c_2\vec{\mathbf{u}}_2 + \cdots + c_n\vec{\mathbf{u}}_n.$$

A solution $\vec{\mathbf{x}}$ of $\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t)$ uniquely determines the constants. In particular, the zero solution of the homogeneous equation is uniquely represented, which can be stated this way:

$$c_1\vec{\mathbf{u}}_1 + c_2\vec{\mathbf{u}}_2 + \cdots + c_n\vec{\mathbf{u}}_n = \vec{\mathbf{0}} \quad \text{implies} \quad c_1 = c_2 = \cdots = c_n = 0.$$

This statement equivalently says that the list of n vector-valued functions $\vec{\mathbf{u}}_1(t)$, …, $\vec{\mathbf{u}}_n(t)$ is **Linearly Independent**, as defined in linear algebra.

Hand calculations might write down a candidate general solution to some 3×3 linear system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, the resulting equations looking like

$$\begin{aligned} x_1 &= c_1 e^t + c_2 e^t + c_3 e^{2t}, \\ x_2 &= c_1 e^t + c_2 e^t + 2c_3 e^{2t}, \\ x_3 &= c_1 e^t + c_2 e^t + 4c_3 e^{2t}. \end{aligned}$$

The example illustrates a classic mistake made in calculations: it is not a general solution, even though it satisfies $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$!

How can we detect the mistake, given only that this expression is supposed to represent the general solution? A required step is to test that $\vec{\mathbf{u}}_1 = \partial\vec{\mathbf{x}}/\partial c_1$, $\vec{\mathbf{u}}_2 = \partial\vec{\mathbf{x}}/\partial c_2$, $\vec{\mathbf{u}}_3 = \partial\vec{\mathbf{x}}/\partial c_3$ are indeed solutions. To insure the **unique representation requirement** of a general solution (**(b)** page 852), the vector functions $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$, $\vec{\mathbf{u}}_3$ must be linearly independent. Compute partial derivatives on symbols c_1, c_2, c_3:

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} e^t \\ e^t \\ e^t \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} e^t \\ e^t \\ e^t \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} e^{2t} \\ 2e^{2t} \\ 4e^{2t} \end{pmatrix}.$$

Then $\vec{\mathbf{u}}_1 = \vec{\mathbf{u}}_2$, which implies that the functions $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$, $\vec{\mathbf{u}}_3$ fail to be independent. While it is possible to test independence by a rudimentary test based upon the definition of independence, the preferred method uses following tests due to Norwegian mathematician N. H. Abel (1802-1829).

Definition 11.2 (Wronskian Determinant of Vector Functions)

Let $\vec{\mathbf{u}}_j(t) : a < t < b \to \mathcal{R}^n$ be given, $1 \le j \le n$. The **Wronskian determinant** is $W(t) = \det(U)$, where U is the augmented matrix of $\vec{\mathbf{u}}_1(t), \ldots, \vec{\mathbf{u}}_n(t)$. In terms of components u_{ij} of vector $\vec{\mathbf{u}}_j$, $1 \le i, j \le n$:

$$W(t) = \begin{vmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{vmatrix}$$

Theorem 11.13 (Abel-Liouville Formula)

Let vector functions $\vec{\mathbf{u}}_1(t)$, …, $\vec{\mathbf{u}}_n(t)$ be solutions of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$, $a < t < b$. Let $W(t)$ be the Wronskian determinant of these solutions. Assume $a < t_0 < b$. Then the **Abel-Liouville formula** holds:

$$W(t) = \mathrm{e}^{\int_{t_0}^t \mathbf{trace}(A(s))ds} W(t_0).^{6}$$

In particular, the Wronskian determinant $W(t)$ is either everywhere nonzero or everywhere zero, accordingly as $W(t_0) \ne 0$ or $W(t_0) = 0$.

[6] The **trace** of a square matrix is the sum of its diagonal elements.

Theorem 11.14 (Abel's Wronskian Independence Test)
Vector solutions $\vec{\mathbf{x}} = \vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n$ of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ are linearly independent on $a < t < b$ if and only if the Wronskian determinant $W(t_0)$ is nonzero for some $a < t_0 < b$.

Initial Value Problems and Reduced Echelon Form

An **initial value problem** is the problem of solving for $\vec{\mathbf{x}}$, given

$$\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t), \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0.$$

Assume general solution

$$\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1(t) + \cdots + c_n\vec{\mathbf{u}}_n(t) + \vec{\mathbf{x}}_p(t),$$

then the problem of finding $\vec{\mathbf{x}}$ reduces to finding $c_1, \ldots, c_n$ in the relation

$$c_1\vec{\mathbf{u}}_1(t_0) + \cdots + c_n\vec{\mathbf{u}}_n(t_0) + \vec{\mathbf{x}}_p(t_0) = \vec{\mathbf{x}}_0.$$

This is a matrix equation for the unknown constants $c_1, \ldots, c_n$ of the form $B\vec{\mathbf{c}} = \vec{\mathbf{d}}$, where B is the augmented matrix of $\vec{\mathbf{u}}_1(t_0), \ldots, \vec{\mathbf{u}}_n(t_0)$:

$$B = \Big\langle \vec{\mathbf{u}}_1(t_0)|\cdots|\vec{\mathbf{u}}_n(t_0)\Big\rangle, \quad \vec{\mathbf{c}} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \quad \vec{\mathbf{d}} = \vec{\mathbf{x}}_0 - \vec{\mathbf{x}}_p(t_0).$$

The reduced row echelon form or **rref** provides a method to find $\vec{\mathbf{c}}$. The method: perform swap, combination and multiply operations to the augmented matrix $C = \Big\langle B|\vec{\mathbf{d}}\Big\rangle$ until $\mathbf{rref}(C) = \Big\langle I|\vec{\mathbf{c}}\Big\rangle$.

Equilibria of $\vec{x}' = A(t)\vec{x}$

An equilibrium point $\vec{\mathbf{x}}_0$ of a linear system $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ is a constant solution, $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_0$ for all t. Equilibria make sense when $A(t)$ is constant, although the definition applies to continuous systems. For a solution $\vec{\mathbf{x}}$ to be constant means $\vec{\mathbf{x}}' = \vec{\mathbf{0}}$, hence all equilibria are determined from the equation

$$A(t)\vec{\mathbf{x}}_0 = \vec{\mathbf{0}} \quad \text{for all } t.$$

This homogeneous system of linear algebraic equations is to be solved for $\vec{\mathbf{x}}_0$. It is not allowed for the answer $\vec{\mathbf{x}}_0$ to depend on t: if it does, then it is **not** an equilibrium.

The theory for a constant matrix $A(t) \equiv A$ says that either $\vec{\mathbf{x}}_0 = \vec{\mathbf{0}}$ is the unique solution or else there are infinitely many nonzero answers for $\vec{\mathbf{x}}_0$. Expectations for any matrix $A(t)$ are similar but an algorithm is lacking for finding nonzero $\vec{\mathbf{x}}_0$.

Examples and Methods

Example 11.1 (Vector Form of the General Solution)

Consider a 3×3 linear system $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ with general solution $\vec{\mathbf{x}}$ (components x_1, x_2, x_3) given in scalar form by the expressions

$$
\begin{array}{lrcl}
(2) \qquad & x_1 & = & c_1e^t + c_2e^{-t} + t, \\
& x_2 & = & (c_1 + c_2)e^t + c_3e^{2t}, \\
& x_3 & = & (2c_2 - c_1)e^{-t} + (4c_1 - 2c_3)e^{2t} + 2t.
\end{array}
$$

Find the vector form of the general solution.

Solution to Example 11.1

Find $\vec{\mathbf{x}}_p(t)$. Set $c_1 = c_2 = c_3 = 0$ in scalar equations (2):

$$\vec{\mathbf{x}}_p(t) = \begin{pmatrix} t \\ 0 \\ 2t \end{pmatrix}.$$

Find $\vec{\mathbf{x}}_h$. Take partial derivatives in scalar equations (2) with respect to the variable names c_1, c_2, c_3 to determine $\vec{\mathbf{u}}_k = \dfrac{\partial \vec{\mathbf{x}}}{\partial c_k}$:

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} e^t \\ e^t \\ -e^{-t} + 4e^{2t} \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} e^{-t} \\ e^t \\ 2e^{-t} \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} 0 \\ e^{2t} \\ -2e^{2t} \end{pmatrix}.$$

The homogeneous system **vector** solution:

$$\vec{\mathbf{x}}_h(t) = c_1\vec{\mathbf{u}}_1(t) + c_2\vec{\mathbf{u}}_2(t) + c_3\vec{\mathbf{u}}_3(t)$$

The nonhomogeneous system **vector** general solution:

$$
\begin{aligned}
\vec{\mathbf{x}}(t) &= c_1\vec{\mathbf{u}}_1(t) + c_2\vec{\mathbf{u}}_2(t) + c_3\vec{\mathbf{u}}_3(t) + \vec{\mathbf{x}}_p(t) \\
&= c_1 \begin{pmatrix} e^t \\ e^t \\ -e^{-t} + 4e^{2t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-t} \\ e^t \\ 2e^{-t} \end{pmatrix} + c_3 \begin{pmatrix} 0 \\ e^{2t} \\ -2e^{2t} \end{pmatrix} + \begin{pmatrix} t \\ 0 \\ 2t \end{pmatrix}.
\end{aligned}
$$

To be a general solution, expression $\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1(t) + c_2\vec{\mathbf{u}}_2(t) + c_3\vec{\mathbf{u}}_3(t) + \vec{\mathbf{x}}_p(t)$ must satisfy required elements **(a)** and **(b)** in the definition of general solution (page 852). Already **(a)** is satisfied. Issue **(b)** is not settled: vectors $\vec{\mathbf{u}}_1, \vec{\mathbf{u}}_2, \vec{\mathbf{u}}_3$ must be independent, to be settled by Abel's formula and the Wronskian test *infra*, details delayed to a further example.

Example 11.2 (Dependence by Abel's Wronskian Test)

Assume a 3×3 system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ was solved by hand for general solution $\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1 + c_2\vec{\mathbf{u}}_2 + c_3\vec{\mathbf{u}}_3$ where

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} e^t \\ e^t \\ e^t \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} e^t \\ e^t \\ e^t \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} e^{2t} \\ 2e^{2t} \\ 4e^{2t} \end{pmatrix}.$$

Choose t_0 in Abel's Wronskian Test to establish **dependence**. The reported expression $\vec{\mathbf{x}}$ is **not** a general solution.

Details for Example 11.2

Wronskian determinant $W(t)$ is quite complicated, but $W(0)$ is zero because it has two duplicate columns. Choice $t_0 = 0$ in Abel's Wronskian test detects **dependence** of solutions $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$, $\vec{\mathbf{u}}_3$.

Example 11.3 (Abel's Wronskian Test Detects Independence)

Assume a 3×3 system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ was solved by hand for general solution $\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1 + c_2\vec{\mathbf{u}}_2 + c_3\vec{\mathbf{u}}_3$ where

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} 2e^{-t} \\ -e^{2t} + 2e^t \\ 4e^{-t} + 2e^{2t} \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} e^{-t} \\ e^{-t} - e^{2t} \\ 2e^{2t} + 2e^{-t} \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} e^t \\ e^t \\ 3e^t \end{pmatrix}.$$

Choose t_0 in Abel's Wronskian Test to establish independence. The expression $\vec{\mathbf{x}}$ is the general solution.

Details for Example 11.3

At $t = 0$ the solutions become the column vectors

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} 2 \\ 1 \\ 6 \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}.$$

Then $W(0) = \det\left(\left\langle \vec{\mathbf{u}}_1(0)|\vec{\mathbf{u}}_2(0)|\vec{\mathbf{u}}_3(0)\right\rangle\right) = -1$ is nonzero. Vectors $\vec{\mathbf{u}}_1$, $\vec{\mathbf{u}}_2$, $\vec{\mathbf{u}}_3$ are **independent** and $\vec{\mathbf{x}}$ is the general solution.

Example 11.4 (Find A and $\vec{\mathbf{F}}$ from a General Solution)

Assume a 3×3 system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t)$ has general solution $\vec{\mathbf{x}} = c_1\vec{\mathbf{u}}_1 + c_2\vec{\mathbf{u}}_2 + c_3\vec{\mathbf{u}}_3 + \vec{\mathbf{x}}_p$ where

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} 2e^{-t} \\ -e^{2t} + 2e^t \\ 4e^{-t} + 2e^{2t} \end{pmatrix}, \vec{\mathbf{u}}_2 = \begin{pmatrix} e^{-t} \\ e^{-t} - e^{2t} \\ 2e^{2t} + 2e^{-t} \end{pmatrix}, \vec{\mathbf{u}}_3 = \begin{pmatrix} e^t \\ e^t \\ 3e^t \end{pmatrix}, \vec{\mathbf{x}}_p = \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix}.$$

Find matrix A and vector function $\vec{\mathbf{F}}(t)$.

Solution to Example 11.4

Superposition implies $\vec{\mathbf{u}}_k'(t) = A\vec{\mathbf{u}}_k(t)$, $1 \le k \le 3$. Let $t = 0$ in these equations and then re-assemble the equations into a single matrix equation:

$$\left\langle \vec{\mathbf{u}}_1'(0)|\vec{\mathbf{u}}_2'(0)|\vec{\mathbf{u}}_3'(0)\right\rangle = A\left\langle \vec{\mathbf{u}}_1(0)|\vec{\mathbf{u}}_2(0)|\vec{\mathbf{u}}_3(0)\right\rangle$$

$$\begin{pmatrix} -2 & -1 & 1 \\ -4 & -3 & 1 \\ 0 & 2 & 3 \end{pmatrix} = A \begin{pmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 6 & 4 & 3 \end{pmatrix}$$

Solve the matrix equation by inversion:

$$A = \begin{pmatrix} -9 & 4 & 2 \\ -11 & 6 & 2 \\ -18 & 6 & 5 \end{pmatrix}$$

Vector $\vec{\mathbf{F}}(t)$ can be found from $\vec{\mathbf{x}}_p'(t) = A\vec{\mathbf{x}}_p(t) + \vec{\mathbf{F}}(t)$ by solving for $\vec{\mathbf{F}}$:

$$\vec{\mathbf{F}}(t) = \begin{pmatrix} -2t^2 - 4t + 9 \\ -2t^2 - 6t + 12 \\ -5t^2 - 4t + 18 \end{pmatrix}$$

Example 11.5 (Solve $\overset{\rightarrow}{\mathbf{x}}\,'(t) = A\overset{\rightarrow}{\mathbf{x}}(t) + \overset{\rightarrow}{\mathbf{F}}(t)$ with Initial Conditions)

Assume:

$$\vec{\mathbf{x}}'(t) = \begin{pmatrix} -3 & 4 & 2 \\ -2 & 6 & 2 \\ -12 & 6 & 7 \end{pmatrix} + \begin{pmatrix} t \\ 0 \\ 2t \end{pmatrix}$$

$$x_1(0) = 1,\ x_2(0) = 0,\ x_3(0) = -1$$

$$\begin{aligned} x_1 &= c_1e^t + c_2e^{-t} + t \\ x_2 &= (c_1 + c_2)e^t + c_3e^{2t} \\ x_3 &= (2c_2 - c_1)e^{-t} + (4c_1 - 2c_3)e^{2t} + 2t \end{aligned}$$

Solve for c_1, c_2, c_3.

Solution to Example 11.5

The equations for x_1, x_2, x_3 evaluated at $t = 0$ give the system of linear algebraic equations

$$\begin{aligned} 1 &= c_1e^0 + c_2e^0 + 0, \\ 0 &= (c_1 + c_2)e^0 + c_3e^0, \\ -1 &= (2c_2 - c_1)e^0 + (4c_1 - 2c_3)e^0 + 0. \end{aligned}$$

In standard form it is the 3×3 linear system

$$\begin{array}{rcrcrcr} c_1 & + & c_2 & & & = & 1, \\ c_1 & + & c_2 & + & c_3 & = & 0, \\ 3c_1 & + & 2c_2 & - & 2c_3 & = & -1. \end{array}$$

The augmented matrix C:

$$C = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 3 & 2 & -2 & -1 \end{pmatrix}, \quad \mathbf{rref}(C) = \begin{pmatrix} 1 & 0 & 0 & -5 \\ 0 & 1 & 0 & 6 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Then $c_1 = -5$, $c_2 = 6$, $c_3 = -1$.

The final answer:

$$\begin{aligned} x_1 &= -5e^t + 6e^{-t} + t, \\ x_2 &= e^t - e^{2t}, \\ x_3 &= 17e^{-t} - 18e^{2t} + 2t. \end{aligned}$$

Example 11.6 (Equilibria for $\overrightarrow{\mathbf{x}}\,'(t) = A(t)\,\overrightarrow{\mathbf{x}}\,(t)$)

Find all equilibria for system

$$\begin{array}{rcrcr} x_1' &=& x_1 &+& x_2 \\ x_2' &=& \sin(t)\,x_1 &+& \sin(t)\,x_2 \end{array}$$

Solution to Example 11.6

Let $A(t) = \begin{pmatrix} 1 & 1 \\ \sin(t) & \sin(t) \end{pmatrix}$. Let vector $\vec{\mathbf{x}}_0$ have components x_1, x_2. Then $A(t)\vec{\mathbf{x}}_0 = \vec{\mathbf{0}}$ has scalar form:

$$\left\{ \begin{array}{rcrcl} x_1 &+& x_2 &=& 0 \\ \sin(t)\,x_1 &+& \sin(t)\,x_2 &=& 0 \end{array} \right.$$

The equations must hold for all values of t. Because $\sin(t) \neq 0$ except for $t = n\pi$, an equivalent system for x_1, x_2 is

$$\left\{ \begin{array}{rcl} x_1 + x_2 &=& 0 \\ x_1 + x_2 &=& 0 \end{array} \right.$$

Solve the linear system. Then all constant solutions of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ are:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = t_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad -\infty < t_1 < \infty$$

It is an error to report $t = \pi$, $x_1 = 1$, $x_2 = -1$ as an equilibrium solution. Reports of equilibria are constants for x_1, x_2 which produce a solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ **for all values of** t.

Proofs for Theorems 11.3 to 11.14

Proof of Theorem 11.3: Gronwall's Lemma

Let $w(t) = c + \int_{t_0}^{t} u(r)v(r)dr$ and $F(t) = e^{\int_{t_0}^{t} v(r)dr}$. Then:

$w'(t) = u(t)v(t)$	Fundamental Theorem of calculus.
$w'(t) \leq v(t)w(t)$	Hypothesis $u(t) \leq w(t)$.
$\dfrac{(F(t)w(t))'}{F(t)} \leq 0$	Integrating factor identity, for $t > t_0$.
$(F(t)w(t))' \leq 0$	For t in J.
$F(t)w(t) \leq F(t_0)w(t_0)$	Integrate across the inequality on J.
$F(t)w(t) \leq c$	Because $F(t_0) = 1$ and $w(t_0) = c$.
$w(t) \leq c\,e^{-\int_{t_0}^{t} v(r)dr}$	Divide by $F(t)$.
$u(t) \leq c\,e^{-\int_{t_0}^{t} v(r)dr}$	Hypothesis $u(t) \leq w(t)$.

Proof of Theorem 11.4: Unique Zero Solution

Zero is a solution because it satisfies both the differential equation and the initial condition. It remains to prove that zero is the unique *global* solution.

Assume $\vec{\mathbf{x}}(t)$ is another solution to the initial value problem. Let $\|B\|$ denote the Euclidean matrix norm. Then $|\vec{\mathbf{x}}(t)| \le \int_0^t \|A(r)\||\vec{\mathbf{x}}(r)|dr$ for $t \ge t_0$. Define $u(t) = |\vec{\mathbf{x}}(t)|$ and $v(t) = \|A(t)\|$. Then $u(t) \le c + \int_{t_0}^t u(r)v(r)dr$ for $c = 0$. Apply Gronwall's Lemma 11.3. Then $u(t) \le 0$, which implies $\vec{\mathbf{x}}(t) = 0$ for $t_0 \le t \le t_0 + H$.

Proof of Theorem 11.5: Picard-Lindelöf

Uniqueness is proved by subtracting two possible solutions: $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_1(t) - \vec{\mathbf{x}}_2(t)$. Then $\vec{\mathbf{x}}$ satisfies the hypotheses of Theorem 11.4, implying $\vec{\mathbf{x}}(t) = 0$ and then $\vec{\mathbf{x}}_1(t) = \vec{\mathbf{x}}_2(t)$ for all t in J.

Existence is proved by modification of the classical Picard-Lindelöf proof. The Picard iterates are constructed for the associated integral equation:

$$\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}(t_0) + \int_{t_0}^t A(r)\vec{\mathbf{F}}(r)\,dr$$

The essential step proves that the iterates converge uniformly to a solution $\vec{\mathbf{x}}(t)$ on the **entire interval** J. Details are in the exercises (Advanced Calculus required).

Proof of Theorem 11.6: Existence-Uniqueness for Constant Linear Systems

Picard-Lindelöf Theorem 11.5 applies to any interval $a < t < b$. Therefore, the unique solution is defined for all values of t.

Proof of Theorem 11.7: Uniqueness and Solution Crossings

The crossing theorem restates uniqueness in Picard-Lindelöf Theorem 11.5.

Proof of Theorem 11.8: Linear Structure

Let $\vec{\mathbf{x}}(t) = k_1\vec{\mathbf{x}}_1(t) + k_2\vec{\mathbf{x}}_2(t)$. Then:

$$\begin{aligned} A(t)\vec{\mathbf{x}}(t) &= k_1\,A(t)\vec{\mathbf{x}}_1(t) + k_2\,A(t)\vec{\mathbf{x}}_2(t) && \text{Matrix multiply} \\ &= k_1\,\vec{\mathbf{x}}_1'(t) + k_2\,\vec{\mathbf{x}}_2'(t) && \text{Because } \vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2 \text{ are solutions.} \\ &= \vec{\mathbf{x}}'(t) && \text{Differential equation verified.} \end{aligned}$$

Proof of Theorem 11.9: Basis

Let V be the vector space of all real-valued vector functions $\vec{\mathbf{x}}(t)$ defined on $a < t < b$.

Let S be the set of all solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, a subset of V.

Construct a **standard basis** $\{\vec{\mathbf{w}}_k\}_{k=1}^n$ for S by applying the Picard-Lindelöf theorem to initial value problem $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$, with $\vec{\mathbf{x}}_0$ successively set equal to the columns of the $n \times n$ identity matrix. This produces n solutions $\vec{\mathbf{w}}_1, \ldots, \vec{\mathbf{w}}_n$ to the equation $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$, all of which exist on the same interval $a < t < b$.

It will be shown that the span in V of $W = \{\vec{\mathbf{w}} + 1, \ldots, \vec{\mathbf{w}}_n\}$ equals S. Then S has a basis of n elements, which proves the theorem.

span$(W) \subset S$: Let linear combination

$$\vec{\mathbf{x}}(t) = c_1\vec{\mathbf{w}}_1(t) + c_2\vec{\mathbf{w}}_2(t) + \cdots + c_n\vec{\mathbf{w}}_n(t) \tag{3}$$

belong to **span**(W). Theorem 11.8 implies that the linear combination $\vec{\mathbf{x}}(t)$ is a solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$. Then $\vec{\mathbf{x}}(t)$ is in S.

$S \subset$ **span**(W): if $\vec{\mathbf{x}}(t)$ is in S, then $\vec{\mathbf{x}}(t_0)$ has components $c_1, \ldots, c_n$. Function $\vec{\mathbf{y}}(t) = c_1\vec{\mathbf{w}}_1(t) + c_2\vec{\mathbf{w}}_2(t) + \cdots + c_n\vec{\mathbf{w}}_n(t)$ is in **span**(W), hence in S, and it has the same initial condition: $\vec{\mathbf{y}}(t_0)$ equals $\vec{\mathbf{x}}(t_0)$. The Picard theorem says $\vec{\mathbf{x}}(t) = \vec{\mathbf{y}}(t)$, therefore $\vec{\mathbf{x}}(t)$ is in $\mathbf{span}(W)$.

Proof of Theorem 11.10: Superposition

Assume $\vec{\mathbf{x}}'_h(t) = A(t)\vec{\mathbf{x}}_h(t)$ and $\vec{\mathbf{x}}'_p(t) = A(t)\vec{\mathbf{x}}p(t) + \vec{\mathbf{F}}(t)$.

Let $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_h(t)+\vec{\mathbf{x}}_p(t)$. Let's prove $\vec{\mathbf{x}}(t)$ is a solution of the nonhomogeneous equation.

$$\begin{aligned}
\vec{\mathbf{x}}'(t) &= \vec{\mathbf{x}}'_h(t) + \vec{\mathbf{x}}'_p(t) && \text{Differential calculus.}\\
&= A(t)\vec{\mathbf{x}}_h(t) + A(t)\vec{\mathbf{x}}_p(t) + \vec{\mathbf{F}}(t) && \text{Use the two differential equations.}\\
&= A(t)\left(\vec{\mathbf{x}}_h(t) + \vec{\mathbf{x}}_p(t)\right) + \vec{\mathbf{F}}(t) && \text{Matrix algebra.}\\
&= A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t) && \text{Definition of } \vec{\mathbf{x}}(t).
\end{aligned}$$

Let $\vec{\mathbf{x}}(t)$ denote any solution of $\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t)$. To prove: $\vec{\mathbf{y}}(t) = \vec{\mathbf{x}}(t) - \vec{\mathbf{x}}_p(t)$ is a solution of the homogeneous equation $\vec{\mathbf{y}}'(t) = A(t)\vec{\mathbf{y}}(t)$. Then for some assignment of constants $\vec{\mathbf{y}}(t)$ equals $\vec{\mathbf{x}}_h(t)$ and $\vec{\mathbf{x}} = \vec{\mathbf{y}} + \vec{\mathbf{x}}_p = \vec{\mathbf{x}}_h + \vec{\mathbf{x}}_p$.

$$\begin{aligned}
\vec{\mathbf{y}}'(t) &= \vec{\mathbf{x}}'(t) - \vec{\mathbf{x}}'_p(t) && \text{Differential calculus.}\\
&= A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t) - \vec{\mathbf{x}}'_p(t) && \text{Differential equation for } \vec{\mathbf{x}}(t).\\
&= A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t) - A(t)\vec{\mathbf{x}}_p(t) - \vec{\mathbf{F}}(t) && \text{Differential equation for } \vec{\mathbf{x}}_p(t).\\
&= A(t)\left(\vec{\mathbf{x}}(t) - \vec{\mathbf{x}}_p(t)\right) && \text{Matrix algebra.}\\
&= A(t)\vec{\mathbf{y}}(t) && \text{Definition of } \vec{\mathbf{y}}(t).
\end{aligned}$$

Proof of Theorem 11.11: Difference of Solutions

$$\begin{aligned}
\vec{\mathbf{y}}'(t) &= \vec{\mathbf{u}}'(t) - \vec{\mathbf{v}}'(t) && \text{Differential calculus.}\\
&= A(t)\vec{\mathbf{u}}(t) + \vec{\mathbf{F}}(t) - \vec{\mathbf{v}}'(t) && \text{Differential equation for } \vec{\mathbf{u}}(t).\\
&= A(t)\vec{\mathbf{u}}(t) + \vec{\mathbf{F}}(t) - A(t)\vec{\mathbf{v}}(t) - \vec{\mathbf{F}}(t) && \text{Differential equation for } \vec{\mathbf{v}}(t).\\
&= A(t)\left(\vec{\mathbf{u}}(t) - \vec{\mathbf{v}}(t)\right) && \text{Matrix algebra.}\\
&= A(t)\vec{\mathbf{y}}(t) && \text{Definition of } \vec{\mathbf{y}}(t).
\end{aligned}$$

Proof of Theorem 11.12: General Solution

Claim **1**. Term $\vec{\mathbf{y}} = \vec{\mathbf{x}}_h(t)$ is a general solution of the homogeneous equation $\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}$ which contains n arbitrary constants $c_1, \ldots, c_n$.

Each solution $\vec{\mathbf{y}} = \vec{\mathbf{x}}_h(t)$ of $\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}$ can be expanded uniquely as a linear combination of basis elements $\vec{\mathbf{w}}_1(t), \vec{\mathbf{w}}_2(t), \ldots, \vec{\mathbf{w}}_n(t)$ because of the Picard-Lindelöf Theorem 11.5 and Theorem 11.9. Then $\vec{\mathbf{y}}(t) = c_1\vec{\mathbf{w}}_1(t)+c_2\vec{\mathbf{w}}_2(t)+\cdots+c_n\vec{\mathbf{w}}_n(t)$ for weights $c_1,\ldots,c_n$ is a general solution of $\vec{\mathbf{y}}' = A(t)\vec{\mathbf{y}}$. The weights $c_1,\ldots,c_n$ are the n arbitrary constants required in the general solution.

Claim **2**. Term $\vec{\mathbf{x}} = \vec{\mathbf{x}}_p(t)$ is a particular solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$.

Let $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_h(t)+\vec{\mathbf{x}}_p(t)$ be a general solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$. Then $\vec{\mathbf{x}}'_h = A(t)\vec{\mathbf{x}}_h$ implies $\vec{\mathbf{x}}'_p = \vec{\mathbf{x}}' - \vec{\mathbf{x}}'_h = A(t)(\vec{\mathbf{x}}_h + \vec{\mathbf{x}}_p) + \vec{\mathbf{F}}(t) - A(t)\vec{\mathbf{x}}_h = A(t)\vec{\mathbf{x}}_p + \vec{\mathbf{F}}(t)$. Then $\vec{\mathbf{x}}_p$ is a particular solution.

Proof of Theorem 11.13: Abel's Formula

Let determinant $D_j(t)$ equal $W(t)$ with row j replaced by its derivative, $1 \le j \le n$. The derivative of determinant $W(t)$ is the sum of these determinants:

$$W'(t) = D_1(t) + \cdots + D_n(t)$$

Determinant $D_1(t) = a_{11}W(t)$, discovered as follows. Element d_{1j} of $D_1(t)$ is expressed as a summation $\sum_{i=1}^n a_{1i}(t)u_{ij}(t)$. The details: $\vec{\mathbf{u}}_j'(t) = A(t)\vec{\mathbf{u}}_j(t)$ and $\vec{\mathbf{u}}_j(t)$ has components $u_{1j}, \ldots, u_{nj}$. Determinant $D_1(t)$ has value unchanged by adding to row 1 a linear combination of rows 2 to n. The selected combination adds $-\sum_{i=2}^n a_{1i}(t)u_{ij}(t)$ to d_{1j}, effectively replacing d_{1j} by $a_{11}u_{1j}$. Then $a_{11}(t)$ is a common factor in row 1 of the modified determinant $D_1(t)$. Factor out $a_{11}(t)$ from row 1, leaving determinant $W(t)$. Then $D_1(t) = a_{11}(t)W(t)$.

Proceeding similarly: $D_j(t) = a_{jj}(t)W(t)$ for $2 \le j \le n$. Then:

$$\begin{aligned} W'(t) &= D_1(t) + \cdots + D_n(t) \\ &= (a_{11}(t) + \cdots + a_{nn}(t))\, W(t) \\ &= \textbf{trace}(A(t))\, W(t) \end{aligned}$$

The claimed expression for $W(t)$ is the solution of the first order linear differential equation $W' = \textbf{trace}(A(t))W$, by the linear integrating factor method.

If $W(t_0) = 0$, then the formula implies $W(t) = 0$ for all t. Conversely, if $W(t_0) \ne 0$ for some t_0, then the formula implies $W(t)$ is never zero, because exponentials are never zero.

Proof of Theorem 11.14: Abel's Wronskian Test Linear combination $\sum_{i=1}^n c_i\vec{\mathbf{u}}_i(t)$ is the zero function if and only if the matrix equation $U(t)\vec{\mathbf{c}} = \vec{\mathbf{0}}$ has only the zero solution $\vec{\mathbf{c}} = \vec{\mathbf{0}}$, where $U(t)$ is the augmented matrix of $\vec{\mathbf{u}}_1(t), \ldots, \vec{\mathbf{u}}_n((t)$ and vector $\vec{\mathbf{c}}$ has components $c_1, \ldots, c_n$. The matrix equation has only the zero solution $\vec{\mathbf{c}} = \vec{\mathbf{0}}$ if and only if $\det(U(t)) \ne 0$. The Abel-Liouville formula completes the proof, because $\det(U(t)) = W(t)$, the Wronskian of the n solutions.

Exercises 11.3

Linear Systems

Convert to matrix notation $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$.

1. $x_1' = 2x_1 + x_2 + e^t$,
$x_2' + x_1 - 2x_2 = \sinh(t)$

2. $x_1' = x_1 + x_2 + x_3$,
$x_2' + x_1 - 2x_2 + x_3 = \ln|1 + t^2|$,
$x_3' = x_2 + x_3 + \cosh(t)$

Existence-Uniqueness

3. Apply Gronwall's inequality to
$|y(t)| \le 4 + \int_0^t (1 + r^2)|y(r|\, dr$, $t \ge 0$.

4. Solve with $x_1(0) = x_2(0) = 0$:
$x_1' = e^t x + e^{-t}x_2$,
$x_2' = \ln|1 + \sinh^2(t)|\, x_1 + x_2$

5. Find the interval on which the solution is defined:
$x_1' = tx_1 + x_2$, $x_2' = x_1 + \tan(t)\, x_2$

6. Let matrix A be 2×2 constant. Find A, given $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ has general solution $x_1 = c_1e^t + c_2e^{2t}$, $x_2 = 5c_12e^t + 4c_2e^{2t}$.

7. Let $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ have two solutions : $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\begin{pmatrix} e^t \\ e^t \end{pmatrix}$. Solve $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$.

8. Let $A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

9. Let constant matrix A be 10×10. Two solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ have equal value at $t = 100$. Are they the same solution?

10. Solutions y_1, y_2 of $y' + p(x)y = q(x)$ are zero at $x = -2$. What assumptions on p, q imply $y_1 \equiv y_2$?

Superposition

11. Explain: e^t is a solution of $y'' - y = 0$ because $\cosh(t)$, $\sinh(t)$ are a solution basis.

12. Explain: $e^t + 10$ is a solution of $y'' - y = -10$, therefore 10 is a particular solution.

13. The shortest solution of $y' + y = 100$ is $y = 100$. Explain why.

14. Let $x_1' = 2x_1$, $x_2' = -x_2$. Report the matrix form $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ and the vector general solution.

15. Let 2-dimensional $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ have general solution $x_1 = c_1 e^t + c_2 e^{3t}$, $x_2 = (c_1 + c_2)e^t + 2c_2 e^{3t} + \cos(t)$. Find formulas for vectors $\vec{\mathbf{x}}_h$ and $\vec{\mathbf{x}}_p$.

16. Let $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ have two solutions $x_1 = e^t + e^{3t}$, $x_2 = 2e^t + \sin(t)$ and $x_1 = e^{3t}$, $x_2 = e^{3t} + \sin(t)$. Find a solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

Superposition $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$

17. Let $\vec{\mathbf{u}}_1(t), \ldots, \vec{\mathbf{u}}_k(t)$ be solutions of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$. Let $c_1, \ldots, c_k$ be constants. Prove: $\vec{\mathbf{u}}(t) = \sum_{i=1}^k c_i \vec{\mathbf{u}}_i(t)$ is a solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$.

18. Find the **standard basis** $\vec{\mathbf{w}}_1(t), \vec{\mathbf{w}}_2(t)$:
$$\vec{\mathbf{x}}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \vec{\mathbf{x}}$$

19. Let matrix A be 2×2. For $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$, find $\vec{\mathbf{x}}_h(t)$, $\vec{\mathbf{x}}_p(t)$:
$x_1 = c_1 + c_2 t + e^t$, $x_2 = (c_1 - c_2)t + e^{2t}$

20. Let matrix $A(t)$ be 2×2. Let $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ have two solutions $\begin{pmatrix} 1 + e^t \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 + e^{-t} \\ -1 \end{pmatrix}$. Find a solution of $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$.

General Solution

21. Assume A is 2×2 and $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ has solutions $e^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $e^{-t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Find the general solution and explain.

22. Assume $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Prove that zero is not a solution.

23. Assume $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_0 =$ constant. Find an equation for $\vec{\mathbf{x}}_0$.

24. Find the vector general solution:
$$\vec{\mathbf{x}}' = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

25. Given 3 $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ with scalar general solution $x_1 = c_1 + c_2 t + c_3 t^2$, $x_2 = c_2 + c_3 t$, $x_3 = c_3$, find the vector general solution.

26. Given 3 $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}}$ with scalar general solution $x_1 = c_1 + c_2 t + c_3 t^2$, $x_2 = c_2 + c_3 t$, $x_3 = c_3$, find $A(t)$.

27. Find the vector general solution:
$$\vec{\mathbf{x}}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

28. Find the vector general solution:
$$\vec{\mathbf{x}}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Independence

29. Assume A is 2×2 and $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ has solutions $e^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $e^{-t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Prove they are independent directly from the definition.

30. Compute the Wronskian:
$$e^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; e^{-t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Abel-Liouville Formula

31. Apply Abel's Independence Test:
$$e^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; e^{-t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

32. Let $\Phi(t)$ an invertible matrix satisfying $\Phi'(t) = A\Phi(t)$. Prove that the columns of $\Phi(t)$ are independent solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

33. Let $\Phi(t)$ an invertible matrix satisfying $\Phi'(t) = A\Phi(t)$. Prove that the columns of $\Phi(t)$ are independent solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

34. Let $\Phi(t)$ any matrix satisfying $\Phi'(t) = A\Phi(t)$. Assume the determinant of $\Phi(t_0)$ is nonzero. Prove that the columns of $\Phi(t)$ are independent solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

35. Let $\Phi(t)$ any matrix satisfying $\Phi'(t) = A\Phi(t)$. Let C be a constant matrix. Prove that the columns of $\Phi(t)C$ are solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

36. Assume continuous coefficients:
$y^{(n)}+p_{n-1}y^{(n-1)}+\cdots+p_0y=0$
Prove from the Abel-Liouville formula for the companion system
that the Wronskian $W(t)$ of
solutions $y_1, \ldots, y_n$ satisfies
$W' + p_{n-1}(t)W = 0$.

Initial Value Problem

37. Let matrix A be 3×3. Assume $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ has scalar general solution $x_1 = c_1e^t + c_2e^{-t} + t$, $x_2 = (c_1 + c_2)e^t + c_3e^{2t}$, $x_3 = (c_1 + c_2)e^t - 2c_2e^{-t} + c_3e^{2t} + t$. Given initial conditions $x_1(0) = x_2(0) = 0$, $x_3(0) = 1$, solve for c_1, c_2, c_3.

38. Let matrix A be 3×3. Assume $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ has scalar general solution $x_1 = c_1 + c_2t + c_3t^2 + e^t$, $x_2 = c_2 + c_3t + e^{2t}$, $x_3 = c_3$. Find the **vector** particular solution $\vec{\mathbf{x}}$ for initial conditions $x_1(0) = x_2(0) = 0$, $x_3(0) = 1$.

Equilibria

39. Find all equilibria:
$$\vec{\mathbf{x}}' = \begin{pmatrix} \cos(t) & \cos(t) \\ 2 & 2 \end{pmatrix} \vec{\mathbf{x}}$$

40. Find all equilibria:
$$\vec{\mathbf{x}}' = \begin{pmatrix} \sin(t) & \sin^2(t) \\ 2 & 2 \end{pmatrix} \vec{\mathbf{x}}$$

11.4 Matrix Exponential

The problem

$$\frac{d}{dt}\vec{\mathbf{x}}(t) = A\vec{\mathbf{x}}(t), \quad \vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$$

has a unique solution, according to the Picard-Lindelöf theorem. Solve the problem n times, when $\vec{\mathbf{x}}_0$ equals a column of the identity matrix, and write $\vec{\mathbf{w}}_1(t)$, $\ldots$, $\vec{\mathbf{w}}_n(t)$ for the n solutions so obtained. The solutions form the **standard basis**. Define the **matrix exponential** e^{At} by packaging these n solutions into the columns of a matrix:

$$e^{At} \equiv \left\langle \vec{\mathbf{w}}_1(t) | \ldots | \vec{\mathbf{w}}_n(t) \right\rangle.$$

By construction, any possible solution of $\frac{d}{dt}\vec{\mathbf{x}} = A\vec{\mathbf{x}}$ can be uniquely expressed in terms of the matrix exponential e^{At} by the formula

$$\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}(0).$$

Matrix Exponential Identities

Announced here are formulas and identities for e^{At}, the matrix exponential. Most details are delayed to page 869.

$e^{At} = e^{\lambda_1 t} I + \dfrac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2}(A - \lambda_1 I)$	A is 2×2, $\lambda_1 \neq \lambda_2$ real, Theorem page 866.
$e^{At} = e^{\lambda_1 t} I + t e^{\lambda_1 t}(A - \lambda_1 I)$	A is 2×2, $\lambda_1 = \lambda_2$ real.
$e^{At} = e^{at}\cos bt \, I + \dfrac{e^{at}\sin bt}{b}(A - aI)$	A is 2×2, $\lambda_1 = \overline{\lambda_2} = a + ib$, $b > 0$.
$e^{At} = r_1(t)P_1 + \cdots + r_n(t)P_n$	Putzer's $n \times n$ spectral formula, Theorem page 868.
$\dfrac{d}{dt}\left(e^{At}\right) = Ae^{At}$	Columns of e^{At} satisfy $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$. Page 869.
$e^{\mathbf{0}} = I$	Where $\mathbf{0}$ is the zero matrix.
$Be^{At} = e^{At}B$	If $AB = BA$.
$e^{At}e^{Bt} = e^{(A+B)t}$	If $AB = BA$.
$e^{At}e^{As} = e^{A(t+s)}$	Since At and As commute.
$\left(e^{At}\right)^{-1} = e^{-At}$	Equivalently, $e^{At}e^{-At} = I$.
$e^{At} = P^{-1}e^{Jt}P$	Jordan form $J = PAP^{-1}$
$e^{At} = \displaystyle\sum_{n=0}^{\infty} A^n \frac{t^n}{n!}$	Picard series identity, proof on page 870

Putzer's Spectral Formula

The spectral formula of Putzer applies to a system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ to find its general solution. The method uses matrices $P_1, \ldots, P_n$ constructed from A, the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, matrix multiplication and the solution $\vec{\mathbf{r}}(t)$ of the first order $n \times n$ initial value problem

$$\vec{\mathbf{r}}'(t) = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda_2 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda_3 & \cdots & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 & \lambda_n \end{pmatrix} \vec{\mathbf{r}}(t), \quad \vec{\mathbf{r}}(0) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The system is solved by first order scalar methods and back-substitution. The formula will be derived separately for the 2×2 case (the one used most often) and the $n \times n$ case.

Theorem 11.15 (Putzer's 2×2 Spectral Formula)
Let A be a 2×2 matrix. Let $r = \lambda_1, \lambda_2$ be the two real or complex roots of the characteristic equation $\det(A - rI) = 0$. Let $P_1 = I$, $P_2 = A - \lambda_1 I$. Let functions $r_1(t)$, $r_2(t)$ be defined by the scalar system

$$\begin{cases} r_1' &= \lambda_1 r_1, & r_1(0) = 1, \\ r_2' &= \lambda_2 r_2 + r_1, & r_2(0) = 0. \end{cases}$$

Then the 2×2 system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$ has solution

$$\vec{\mathbf{x}}(t) = (r_1(t)P_1 + r_2(t)P_2)\,\vec{\mathbf{x}}_0$$

Proof: The Cayley-Hamilton formula $(A - \lambda_1 I)(A - \lambda_2 I) = \vec{\mathbf{0}}$ is valid for any 2×2 matrix A, if $r = \lambda_1, \lambda_2$ are the two roots of the determinant equation $\det(A - rI) = 0$. See page 721. The Cayley-Hamilton formula is the same as $(A - \lambda_2 I)P_2 = \vec{\mathbf{0}}$, which implies the identity $AP_2 = \lambda_2 P_2$. Compute as follows.

$$\begin{aligned} \vec{\mathbf{x}}'(t) &= (r_1'(t)P_1 + r_2'(t)P_2)\,\vec{\mathbf{x}}_0 \\ &= (\lambda_1 r_1(t)P_1 + r_1(t)P_2 + \lambda_2 r_2(t)P_2)\,\vec{\mathbf{x}}_0 \\ &= (r_1(t)A + \lambda_2 r_2(t)P_2)\,\vec{\mathbf{x}}_0 \\ &= (r_1(t)A + r_2(t)AP_2)\,\vec{\mathbf{x}}_0 \\ &= A\,(r_1(t)I + r_2(t)P_2)\,\vec{\mathbf{x}}_0 \\ &= A\vec{\mathbf{x}}(t). \end{aligned}$$

This proves that $\vec{\mathbf{x}}(t)$ is a solution. Because $\Phi(t) \equiv r_1(t)P_1 + r_2(t)P_2$ satisfies $\Phi(0) = I$, then $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$ is satisfied. ■

Real Distinct Eigenvalues

Suppose A is 2×2 having real distinct eigenvalues λ_1, λ_2 and $\vec{\mathbf{x}}(0)$ is real. Then

$$r_1 = e^{\lambda_1 t}, \quad r_2 = \frac{e^{\lambda_1 t} - e^{\lambda_2 T}}{\lambda_1 - \lambda_2}$$

and

$$\vec{\mathbf{x}}(t) = \left(e^{\lambda_1 t} I + \frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2}(A - \lambda_1 I) \right) \vec{\mathbf{x}}(0).$$

The matrix exponential formula for real distinct eigenvalues:

$$e^{At} = e^{\lambda_1 t} I + \frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2}(A - \lambda_1 I).$$

Real Equal Eigenvalues

Suppose A is 2×2 having real equal eigenvalues $\lambda_1 = \lambda_2$ and $\vec{\mathbf{x}}(0)$ is real. Then $r_1 = e^{\lambda_1 t}$, $r_2 = te^{\lambda_1 t}$ and

$$\vec{\mathbf{x}}(t) = \left(e^{\lambda_1 t} I + te^{\lambda_1 t}(A - \lambda_1 I) \right) \vec{\mathbf{x}}(0).$$

The matrix exponential formula for real equal eigenvalues:

$$e^{At} = e^{\lambda_1 t} I + te^{\lambda_1 t}(A - \lambda_1 I).$$

Complex Eigenvalues

Suppose A is 2×2 having complex eigenvalues $\lambda_1 = a + bi$ with $b > 0$ and $\lambda_2 = a - bi$. If $\vec{\mathbf{x}}(0)$ is real, then a real solution is obtained by taking the real part of the spectral formula. This formula is formally identical to the case of real distinct eigenvalues. Then

$$\begin{aligned}
\mathcal{R}\mathrm{e}(\vec{\mathbf{x}}(t)) &= (\mathcal{R}\mathrm{e}(r_1(t)) I + \mathcal{R}\mathrm{e}(r_2(t)(A - \lambda_1 I))) \vec{\mathbf{x}}(0) \\
&= \left(\mathcal{R}\mathrm{e}(e^{(a+ib)t}) I + \mathcal{R}\mathrm{e}(e^{at} \frac{\sin bt}{b}(A - (a + ib) I)) \right) \vec{\mathbf{x}}(0) \\
&= \left(e^{at} \cos bt \, I + e^{at} \frac{\sin bt}{b}(A - aI) \right) \vec{\mathbf{x}}(0)
\end{aligned}$$

The matrix exponential formula for complex conjugate eigenvalues:

$$e^{At} = e^{at} \left(\cos bt \, I + \frac{\sin bt}{b}(A - aI) \right).$$

How to Remember Putzer's 2×2 Formula

The expressions

$$\text{(1)} \qquad \begin{aligned} e^{At} &= r_1(t)I + r_2(t)(A - \lambda_1 I), \\ r_1(t) &= e^{\lambda_1 t}, \quad r_2(t) = \frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2} \end{aligned}$$

are enough to generate all three formulas. Fraction r_2 is the $d/d\lambda$-Newton quotient for r_1. It has limit $te^{\lambda_1 t}$ as $\lambda_2 \to \lambda_1$, therefore the formula includes the case $\lambda_1 = \lambda_2$ by limiting. If $\lambda_1 = \overline{\lambda_2} = a + ib$ with $b > 0$, then the fraction r_2 is already real, because it has for $z = e^{\lambda_1 t}$ and $w = \lambda_1$ the form

$$r_2(t) = \frac{z - \overline{z}}{w - \overline{w}} = \frac{\sin bt}{b}.$$

Taking real parts of expression (1) gives the complex case formula.

Theorem 11.16 (Putzer's $n \times n$ Spectral Formula)
Let A be an $n \times n$ matrix. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, the real or complex roots r of $\det(A - rI) = 0$. Let

$$P_1 = I, \quad P_k = P_{k-1}(A - \lambda_{k-1}I) = \Pi_{j=1}^{k-1}(A - \lambda_j I), \quad k = 2, \ldots, n.$$

Let functions $r_1(t)$, …, $r_n(t)$ be defined by the differential system

$$\begin{array}{rcll} r_1' &=& \lambda_1 r_1, & r_1(0) = 1, \\ r_2' &=& \lambda_2 r_2 + r_1, & r_2(0) = 0, \\ &\vdots&& \\ r_n' &=& \lambda_n r_n + r_{n-1}, & r_n(0) = 0. \end{array}$$

Then system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$ has solution

$$\vec{\mathbf{x}}(t) = (r_1(t)P_1 + r_2(t)P_2 + \cdots + r_n(t)P_n)\,\vec{\mathbf{x}}_0,$$

Proof on page 871

Theorem 11.17 (Compute e^{Jt} for J Triangular)
If J is an upper triangular matrix, then a column $\vec{\mathbf{u}}(t)$ of e^{Jt} can be computed by solving the system $\vec{\mathbf{u}}'(t) = J\vec{\mathbf{u}}(t)$, $\vec{\mathbf{u}}(0) = \vec{\mathbf{v}}$, where $\vec{\mathbf{v}}$ is the corresponding column of the identity matrix. This problem can always be solved by first-order scalar methods of growth-decay theory and the integrating factor method. Proof on page 872.

Theorem 11.18 (Exponential of a Diagonal Matrix)
For real or complex constants λ_1, …, λ_n,

$$e^{\mathbf{diag}(\lambda_1,\ldots,\lambda_n)t} = \mathbf{diag}\left(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\right).$$

Proof on page 872.

Theorem 11.19 (Block Diagonal Matrix)
If $A = \textbf{diag}(B_1, \ldots, B_k)$ and each of $B_1, \ldots, B_k$ is a square matrix, then

$$e^{At} = \textbf{diag}\left(e^{B_1 t}, \ldots, e^{B_k t}\right).$$

Proof on page 872.

Theorem 11.20 (Complex Exponential)

Given real a, b, then $e^{\begin{pmatrix} a & b \\ -b & a \end{pmatrix} t} = e^{at}\begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}.$

Proof on page 872

Proofs of Matrix Exponential Identities

The 2×2 Putzer identities have proofs in the text page 866. Proofs of theorems are on page 871. The remaining proofs are here.

Verify $\left(e^{At}\right)' = Ae^{At}$.

Let $\vec{\mathbf{x}}_0$ denote a column of the identity matrix. Define $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}_0$. Then

$$\begin{aligned} \left(e^{At}\right)'\vec{\mathbf{x}}_0 &= \vec{\mathbf{x}}'(t) \\ &= A\vec{\mathbf{x}}(t) \\ &= Ae^{At}\vec{\mathbf{x}}_0. \end{aligned}$$

Because this identity holds for all columns of the identity matrix, then $(e^{At})'$ and Ae^{At} have identical columns. Identity $\left(e^{At}\right)' = Ae^{At}$ is proved. ■

Verify $e^{\mathbf{0}} = I$.

$e^{\mathbf{0}} = \left\langle \vec{\mathbf{w}}_1(0)|\ldots|\vec{\mathbf{w}}_n(0)\right\rangle = I.$ ■

Verify $Be^{At} = e^{At}B$ if $AB = BA$.

Define $\vec{\mathbf{w}}_1(t) = e^{At}B\vec{\mathbf{w}}_0$ and $\vec{\mathbf{w}}_2(t) = Be^{At}\vec{\mathbf{w}}_0$. Calculate $\vec{\mathbf{w}}_1'(t) = A\vec{\mathbf{w}}_1(t)$ and $\vec{\mathbf{w}}_2'(t) = BAe^{At}\vec{\mathbf{w}}_0 = ABe^{At}\vec{\mathbf{w}}_0 = A\vec{\mathbf{w}}_2(t)$, due to $BA = AB$. Because $\vec{\mathbf{w}}_1(0) = \vec{\mathbf{w}}_2(0) = \vec{\mathbf{w}}_0$, then the uniqueness assertion of the Picard-Lindelöf theorem implies that $\vec{\mathbf{w}}_1(t) = \vec{\mathbf{w}}_2(t)$. Because $\vec{\mathbf{w}}_0$ is any vector, then $e^{At}B = Be^{At}$. ■

Verify $e^{At}e^{Bt} = e^{(A+B)t}$.

Let $\vec{\mathbf{x}}_0$ be a column of the identity matrix. Define $\vec{\mathbf{x}}(t) = e^{At}e^{Bt}\vec{\mathbf{x}}_0$ and $\vec{\mathbf{y}}(t) = e^{(A+B)t}\vec{\mathbf{x}}_0$. We must show that $\vec{\mathbf{x}}(t) = \vec{\mathbf{y}}(t)$ for all t. Define $\vec{\mathbf{u}}(t) = e^{Bt}\vec{\mathbf{x}}_0$. We will apply the result $e^{At}B = Be^{At}$, valid for $BA = AB$. The details:

$$\begin{aligned} \vec{\mathbf{x}}'(t) &= \left(e^{At}\vec{\mathbf{u}}(t)\right)' \\ &= Ae^{At}\vec{\mathbf{u}}(t) + e^{At}\vec{\mathbf{u}}'(t) \\ &= A\vec{\mathbf{x}}(t) + e^{At}B\vec{\mathbf{u}}(t) \\ &= A\vec{\mathbf{x}}(t) + Be^{At}\vec{\mathbf{u}}(t) \\ &= (A+B)\vec{\mathbf{x}}(t). \end{aligned}$$

Also known is that $\vec{\mathbf{y}}'(t) = (A+B)\vec{\mathbf{y}}(t)$ and since $\vec{\mathbf{x}}(0) = \vec{\mathbf{y}}(0) = \vec{\mathbf{x}}_0$, then the Picard-Lindelöf theorem implies that $\vec{\mathbf{x}}(t) = \vec{\mathbf{y}}(t)$ for all t. ■

Verify $e^{At}e^{As} = e^{A(t+s)}$.

Let t be a variable and consider s fixed. Define $\vec{\mathbf{x}}(t) = e^{At}e^{As}\vec{\mathbf{x}}_0$ and $\vec{\mathbf{y}}(t) = e^{A(t+s)}\vec{\mathbf{x}}_0$. Then $\vec{\mathbf{x}}(0) = \vec{\mathbf{y}}(0)$ and both satisfy the differential equation $\vec{\mathbf{u}}'(t) = A\vec{\mathbf{u}}(t)$. By the uniqueness in the Picard-Lindelöf theorem, $\vec{\mathbf{x}}(t) = \vec{\mathbf{y}}(t)$, which implies $e^{At}e^{As} = e^{A(t+s)}$. ■

Verify $\left(e^{At}\right)^{-1} = e^{-At}$.

Let $s = -t$ in the preceding identity $e^{At}e^{As} = e^{A(t+s)}$. The right side is $e^{\mathbf{0}} = I$. The inverse test Chapter 5 Section 2, Theorem 5.9, implies that the two matrices e^{At} and e^{-At} are inverses of one another. ■

Verify $e^{At} = P^{-1}e^{Jt}P$ if $J = PAP^{-1}$.

The proof uses the Picard series identity $e^{At} = \sum_{n=0}^{\infty} A^n \frac{t^n}{n!}$, which is proved below. The issue is the simplification of A^n using $A = P^{-1}JP$. Induction is used to derive the following identities, in which $Q = P^{-1}$ (then $QP = PQ = I$):

$$\begin{array}{lclcl} A & = & P^{-1}JP & = & QJP \\ A^2 & = & QJPQJP & = & QJ^2P \\ & \vdots & & & \\ A^n & = & (QJP)\cdots(QJP) & = & QJ^nP \end{array}$$

Then the infinite series simplifies:

$$\begin{aligned} e^{At} &= \sum_{n=0}^{\infty} A^n \frac{t^n}{n!} \\ &= \sum_{n=0}^{\infty} QJ^nP\frac{t^n}{n!} \\ &= Q\left(\sum_{n=0}^{\infty} J^n\frac{t^n}{n!}\right)P \\ &= Qe^{Jt}P \\ &= P^{-1}e^{Jt}P \end{aligned}$$

■

Verify $e^{At} = \sum_{n=0}^{\infty} A^n \frac{t^n}{n!}$.

The idea of the proof is to apply Picard iteration. By definition, the columns of e^{At} are vector solutions $\vec{\mathbf{w}}_1(t), \ldots, \vec{\mathbf{w}}_n(t)$ whose values at $t = 0$ are the corresponding columns of the $n \times n$ identity matrix. According to the theory of Picard iterates, a particular iterate is defined by

$$\vec{\mathbf{y}}_{n+1}(t) = \vec{\mathbf{y}}_0 + \int_0^t A\vec{\mathbf{y}}_n(r)dr, \quad n \geq 0.$$

Vector $\vec{\mathbf{y}}_0$ equals some column k of the identity matrix. The Picard iterates can be found explicitly, as follows.

$$\begin{aligned}
\vec{\mathbf{y}}_1(t) &= \vec{\mathbf{y}}_0 + \int_0^t A\vec{\mathbf{y}}_0 dr \\
&= (I + At)\,\vec{\mathbf{y}}_0, \\
\vec{\mathbf{y}}_2(t) &= \vec{\mathbf{y}}_0 + \int_0^t A\vec{\mathbf{y}}_1(r) dr \\
&= \vec{\mathbf{y}}_0 + \int_0^t A\,(I + At)\,\vec{\mathbf{y}}_0 dr \\
&= \left(I + At + A^2t^2/2\right)\vec{\mathbf{y}}_0, \\
&\vdots \\
\vec{\mathbf{y}}_n(t) &= \left(I + At + A^2\tfrac{t^2}{2} + \cdots + A^n\tfrac{t^n}{n!}\right)\vec{\mathbf{y}}_0.
\end{aligned}$$

The Picard-Lindelöf theorem implies

$$\lim_{n\to\infty} \vec{\mathbf{y}}_n(t) = \vec{\mathbf{w}}_k(t).$$

This being valid for each index k, then the columns of the matrix converge as $N \to \infty$ to $\vec{\mathbf{w}}_1(t)$, ..., $\vec{\mathbf{w}}_n(t)$. The matrix limit is formally the infinite series

$$\sum_{m=0}^{\infty} A^m \frac{t^m}{m!} = \lim_{N\to\infty} \sum_{m=0}^{N} A^m \frac{t^m}{m!} = \left\langle \vec{\mathbf{w}}_1(t)|\ldots|\vec{\mathbf{w}}_n(t)\right\rangle$$

but also $e^{At} \equiv \left\langle \vec{\mathbf{w}}_1(t)|\ldots|\vec{\mathbf{w}}_n(t)\right\rangle$. This proves the matrix identity

$$e^{At} = \sum_{n=0}^{\infty} A^n \frac{t^n}{n!}. \quad \blacksquare$$

Proofs of Theorems 11.16–11.20

Theorem 11.16, Proof of Putzer's $n \times n$ Formula:

The Cayley-Hamilton formula $(A - \lambda_1 I)\cdots(A - \lambda_n I) = \vec{\mathbf{0}}$ is valid for any $n \times n$ matrix A and the n roots $r = \lambda_1, \ldots, \lambda_n$ of the determinant equality $\det(A - rI) = 0$. Two facts will be used: (1) The Cayley-Hamilton formula implies $AP_n = \lambda_n P_n$; (2) The definition of P_k implies $\lambda_k P_k + P_{k+1} = AP_k$ for $1 \le k \le n-1$. Compute as follows.

$$\begin{aligned}
&\boxed{1} \quad \vec{\mathbf{x}}'(t) = (r_1'(t)P_1 + \cdots + r_n'(t)P_n)\,\vec{\mathbf{x}}(0) \\
&\boxed{2} \qquad = \left(\sum_{k=1}^{n} \lambda_k r_k(t) P_k + \sum_{k=2}^{n} r_{k-1}P_k\right)\vec{\mathbf{x}}_0 \\
&\boxed{3} \qquad = \left(\sum_{k=1}^{n-1} \lambda_k r_k(t) P_k + r_n(t)\lambda_n P_n + \sum_{k=1}^{n-1} r_k P_{k+1}\right)\vec{\mathbf{x}}_0 \\
&\boxed{4} \qquad = \left(\sum_{k=1}^{n-1} r_k(t)(\lambda_k P_k + P_{k+1}) + r_n(t)\lambda_n P_n\right)\vec{\mathbf{x}}_0 \\
&\boxed{5} \qquad = \left(\sum_{k=1}^{n-1} r_k(t) A P_k + r_n(t) A P_n\right)\vec{\mathbf{x}}_0
\end{aligned}$$

$$\boxed{6} \qquad = A\left(\sum_{k=1}^{n} r_k(t)P_k\right)\vec{\mathbf{x}}_0$$

$$\boxed{7} \qquad = A\vec{\mathbf{x}}(t).$$

Details: $\boxed{1}$ Differentiate the formula for $\vec{\mathbf{x}}(t)$. $\boxed{2}$ Use the differential equations for r_1,...,r_n. $\boxed{3}$ Split off the last term from the first sum, then re-index the last sum. $\boxed{4}$ Combine the two sums. $\boxed{5}$ Use the recursion for P_k and the Cayley-Hamilton formula $(A - \lambda_n I)P_n = \vec{\mathbf{0}}$. $\boxed{6}$ Factor out A on the left. $\boxed{7}$ Apply the definition of $\vec{\mathbf{x}}(t)$.

Then $\vec{\mathbf{x}}(t)$ is a solution. Because $\Phi(t) \equiv \sum_{k=1}^{n} r_k(t)P_k$ satisfies $\Phi(0) = I$, then $\vec{\mathbf{x}}(t)$ satisfies $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$. ■

Proof of Theorem 11.17, Compute e^{Jt} for J Triangular:

The first statement computes the solution of the problem $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) =$ column j of I, $1 \le j \le n$. These are the columns of e^{At}, by definition.

Each such problem is known to be solvable by linear first order integrating factor methods, using the variable list in reverse order.

An example for such a scalar system:

$$\begin{aligned}
&x_1' = 2x_1 + x_3,\\
&x_2' = 3x_2 + x_3,\\
&x_3' = 4x_3,\\
&x_1(0) = 1, x_2(0) = x_3(0) = 0.
\end{aligned}$$

The variable list reversed is x_3, x_2, x_1. The solution starts with $x_3' = 4x_3$, $x_3(0) = 0$. The solution is $x_3 = 0$. Then the equation for x_2 becomes $x_2' = 3x_2 + 0, x_2(0) = 0$. Again the solution is $x_2 = 0$. The last equation is $x_1' = 2x_1 + 0$, $x_1(0) = 1$ with solution $x_1 = e^{2t}$. ■

Proof of Theorem 11.18, Exponential of a Diagonal Matrix:

It suffices to prove that $\Phi(t) = \textbf{diag}\left(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\right)$ satisfies $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$. Because $e^{0t} = 1$, then $\Phi(0) = I$. The differential equation is satisfied by the following steps:

$$\begin{aligned}
\Phi'(t) &= \begin{pmatrix} \lambda_1 e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n e^{\lambda_n t} \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} \\
&= A\Phi(t)
\end{aligned}$$

■

Proof of Theorem 11.19, Block Diagonal Matrix Exponential:

Let $\Phi(t) = \textbf{diag}\left(e^{B_1 t}, \ldots, e^{B_k t}\right)$. To prove $\Phi(t)$ equals e^{At}, it suffices to prove identities $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$, Already $\Phi(0) = I$. Details for identity $\Phi'(t) = A\Phi(t)$ will use the formula $\frac{d}{dt}\,e^{Ct} = C\,e^{Ct}$. Apply block differentiation to show $\Phi'(t) = A\Phi(t)$:

$$\begin{aligned}
\Phi'(t) &= \begin{pmatrix} B_1\,e^{B_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_k\,e^{B_k t} \end{pmatrix} \\
&= \begin{pmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_k \end{pmatrix} \begin{pmatrix} e^{B_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{B_k t} \end{pmatrix} \\
&= A\Phi(t)
\end{aligned}$$

■

Proof of Theorem 11.20, Complex Exponential:

Assume $A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ with $b > 0$. Then A has eigenvalues $a \pm bi$. Putzer's 2×2 formula will be used, page 867:

$$e^{At} = e^{at}\left(\cos bt\,I + \frac{\sin bt}{b}(A - aI)\right).$$

Simplify the matrix expression on the right:

$$\begin{aligned}
e^{At} &= e^{at}\left(\begin{pmatrix} \cos bt & 0 \\ 0 & \cos bt \end{pmatrix} + \frac{\sin bt}{b}\begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}\right) \\
&= e^{at}\begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}
\end{aligned}$$

■

Exercises 11.4

Matrix Exponential.

1. **(Picard)** Let A be real 2×2. Write out the two initial value problems which define the columns $\vec{\mathbf{w}}_1(t)$, $\vec{\mathbf{w}}_2(t)$ of e^{At}.

2. **(Picard)** Let A be real 3×3. Write out the three initial value problems which define the columns $\vec{\mathbf{w}}_1(t)$, $\vec{\mathbf{w}}_2(t)$, $\vec{\mathbf{w}}_3(t)$ of e^{At}.

3. Let A be real 2×2. Show that $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{u}}_0$ satisfies $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{u}}_0$.

4. Let A be real $n \times n$. Show that $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}_0$ satisfies $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$.

Matrix Exponential 2×2. Find e^{At} from representation $e^{At} = \left\langle \vec{\mathbf{w}}_1 | \vec{\mathbf{w}}_2 \right\rangle$. Use first-order scalar methods.

5. $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

6. $A = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$.

7. $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$.

8. $A = \begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix}$.

Matrix Exponential Identities. Verify from exponential identities.

9. $e^{A}\,e^{-A} = I$

10. $e^{-A} = \left(e^{A}\right)^{-1}$

11. $A = \frac{d}{dt}\,e^{At}$ evaluated at $t = 0$

12. If $A^3 = \mathbf{0}$, then $e^A = I + A + \frac{1}{2}A^2$.

13. Let $A = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$ and $N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Verify $N^2 = \mathbf{0}$ and $e^{At+Nt} = e^{At}(I + Nt)$.

14. Let A be 3×3 diagonal and $N = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. Prove $N^3 = \mathbf{0}$ and

$$e^{At+Nt} = e^{At}(I + Nt + N^2\frac{t^2}{2}).$$

15. $e^{\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} t} = \begin{pmatrix} e^t & e^{2t} - e^t \\ 0 & e^{2t} \end{pmatrix}$

16. $e^{\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} t} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}$

Putzer's Spectral Formula.

17. Apply Picard-Lindelöf theory to conclude that $r_1, \ldots, r_n$ are everywhere defined,

18. Prove that $P_1, \ldots, P_k$ commute.

Putzer's Formula 2×2 .

19. Find a formula for $\frac{d}{dt}e^{At}$ for a 2×2 matrix A with eigenvalues $1, 2$.

20. Let $2{\times}2$ matrix A have duplicate eigenvalues $0, 0$. Compute r_1, r_2 and then report e^{At}.

Putzer: Real Distinct.

Find the matrix exponential.

21. $A = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$

22. $A = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}$

Putzer: Real Equal.

Find the matrix exponential.

23. $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

24. $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$

Putzer: Complex Eigenvalues.

Find the matrix exponential.

25. $A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

26. $A = \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}$

How to Remember Putzer's 2×2 Formula.

27. Find $\lim_{\lambda \to \lambda_1} \dfrac{e^{\lambda t} - e^{\lambda_1 t}}{\lambda - \lambda_1}$.

28. Let matrix A be 2×2 real. Take the real part: $e^{At} = I + \dfrac{e^{it} - e^{-it}}{2i}A$.

Classical $n \times n$ Spectral Formula.

Find e^{At}.

29. $A = \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

30. $A = \begin{pmatrix} 0 & 0 & 2 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$

Proofs of Matrix Exponential Properties.

31. Let $A\vec{\mathbf{u}} = B\vec{\mathbf{u}}$ for all vectors $\vec{\mathbf{u}}$. Prove $A = B$.

32. Let $A = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$. Compute the first four Picard iterates for $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$.

Special Cases e^{At}.

33. Show the details to solve

$x_1' = 2x_1 + x_3$,
$x_2' = 3x_2 + x_3$,
$x_3' = 4x_3$,
$x_1(0) = 1, x_2(0) = x_3(0) = 0$.

34. Let $A = \mathbf{diag}(1, 2, 3, 4)$. Find e^{At}.

35. Let $B = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, $A = \mathbf{diag}(B, B)$. Find e^{At}.

36. Let $B = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$ and $A = \mathbf{diag}(B, B)$. Find e^{At}.

11.5 Cayley-Hamilton-Ziebur, Spectral and Eigenanalysis Methods

Established earlier in this chapter:

$$\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}_0 \quad \text{solves} \quad \vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t), \quad \vec{\mathbf{x}}(0) - \vec{\mathbf{x}}_0$$

Matrix e^{At} is the augmented matrix of solutions $\vec{\mathbf{w}}_i(t)$ to $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ with $\vec{\mathbf{w}}_i(0) =$ column i of the identity matrix, $1 \le i \le n$.

Presented in this section are three premier methods for finding e^{At}:

Eigenanalysis Method
Spectral Method
Cayley-Hamilton-Ziebur (CHZ) Method

Eigenanalysis Method Requirements. The $n \times n$ real matrix A is required to have n independent eigenvectors in its list of eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, …, $(\lambda_n, \vec{\mathbf{v}}_n)$. Briefly, matrix A is **diagonalizable**. It is not required that the eigenvalues λ_1, …, λ_n be distinct and eigenvalues can be real or complex. The method uses independence of the Euler substitution solutions $e^{\lambda_i t}\vec{\mathbf{v}}_i$, $1 \le i \le n$, which are assembled into augmented matrix $\Phi(t)$. The general solution is $\vec{\mathbf{x}}(t) = e^{At}\,\vec{\mathbf{x}}(0)$, using identity $e^{At} = \Phi(t)\,\Phi(0)^{-1}$. A negative of the method occurs with complex eigenvalues: real solutions are found with extra effort via opaque identities. The method works best on diagonalizable matrices with only real eigenvalues, e.g., symmetric matrices.

Spectral Method Requirements. The method applies to any real $n \times n$ matrix A. Classical spectral theory of A provides a formula for e^{At} similar to Putzer's formula, thereby finding the solution $\vec{\mathbf{x}} = e^{At}\,\vec{\mathbf{x}}(0)$ of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$. Emphasis is on theory. Computational details are left to computer algebra systems, which efficiently implement the formulas. Hand computation is possible for low dimensions $n = 2, 3$ with time impact similar to Putzer's algorithm for e^{At}.

Cayley-Hamilton-Ziebur Method Requirements. The method applies to any real $n \times n$ matrix A. It provides a basis of n real vector solutions to the system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, which are found from characteristic equation $|A - \lambda I| = 0$ and Euler solution atom theory developed for scalar differential equations. The connection to e^{At} is direct and simple: $e^{At} = \Phi(t)\,\Phi(0)^{-1}$ where $\Phi(t)$ is the $n \times n$ augmented matrix of the vector solutions. Hand computation is possible for low dimensional examples $(n = 2, 3)$ with the lowest time impact of the three methods. A feature of the Cayley-Hamilton-Ziebur method is minimization of encounters with complex numbers. One important consequence of the method:

Solutions of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ are vector linear combinations of Euler solution atoms.

Eigenanalysis Method: 2×2 Matrix

Theorem 11.21 (Eigenanalysis Method 2×2)
Let matrix A be 2×2 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$. Assume eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are independent.

Then the general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ can be written as

$$\vec{\mathbf{x}}(t) = c_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2.$$

Proof:

Eigenvalues λ_1, λ_2 are either both real or a complex conjugate pair $\lambda_1 = \overline{\lambda_2} = a + ib$ with $b > 0$. Derivatives and calculations below apply in both cases.

$$\begin{aligned}
\vec{\mathbf{x}}' &= c_1(e^{\lambda_1 t})'\vec{\mathbf{v}}_1 + c_2(e^{\lambda_2 t})'\vec{\mathbf{v}}_2 && \text{Differentiate the formula for } \vec{\mathbf{x}}. \\
&= c_1 e^{\lambda_1 t}\lambda_1\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\lambda_2\vec{\mathbf{v}}_2 && \\
&= c_1 e^{\lambda_1 t}A\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}A\vec{\mathbf{v}}_2 && \text{Use } \lambda_1\vec{\mathbf{v}}_1 = A\vec{\mathbf{v}}_1,\ \lambda_2\vec{\mathbf{v}}_2 = A\vec{\mathbf{v}}_2. \\
&= A\left(c_1 e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t}\vec{\mathbf{v}}_2\right) && \text{Factor } A \text{ left.} \\
&= A\vec{\mathbf{x}} && \text{Definition of } \vec{\mathbf{x}}.
\end{aligned}$$

Re-write the solution $\vec{\mathbf{x}}$ in the vector-matrix form

$$\vec{\mathbf{x}}(t) = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}. \tag{1}$$

Because eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are assumed independent, then $\left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle$ is invertible and setting $t = 0$ in (1) gives

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle^{-1} \vec{\mathbf{x}}(0). \tag{2}$$

Because c_1, c_2 can be chosen to produce any initial condition $\vec{\mathbf{x}}(0)$, then $\vec{\mathbf{x}}(t)$ is the *general solution* of the system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

Proposition 11.2 (Exponential Matrix: 2×2) Let matrix A be 2×2 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$. Assume eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are independent.

Then:

$$e^{At} = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 \right\rangle^{-1} \tag{3}$$

Proof: Combine (1) and (2). ■

Formula (3) is immediately useful when the eigenpairs are real. It is problematic when the eigenvalues are complex. The complex arithmetic inherited by complex eigenpairs can be minimized by applying results collected into a Proposition.

Proposition 11.3 (Exponential Matrix: Complex $\lambda_2 = \overline{\lambda_1}$)
Assume matrix A is 2×2 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$. Let eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ be independent.

Assume $\lambda_2 = \overline{\lambda}_1$ and λ_1 is not real. Define for eigenpair $(\lambda_1, \vec{\mathbf{v}}_1)$ symbols a, b, P:

$$\lambda_1 = a + ib, \quad b > 0, \quad P = \Big\langle \mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1) | \mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1) \Big\rangle$$

Then

$$e^{At} = e^{at} P \begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix} P^{-1} \tag{4}$$

Proof on page 886.

Eigenanalysis Method: 3×3 Matrix

Theorem 11.22 (Eigenanalysis Method: 3×3)
Let matrix A be 3×3 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1), (\lambda_2, \vec{\mathbf{v}}_2), (\lambda_3, \vec{\mathbf{v}}_3)$. Assume $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are independent.

Then the general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is:

$$\vec{\mathbf{x}}(t) = c_1 e^{\lambda_1 t} \vec{\mathbf{v}}_1 + c_2 e^{\lambda_2 t} \vec{\mathbf{v}}_2 + c_3 e^{\lambda_3 t} \vec{\mathbf{v}}_3.$$

Proof on page 887.

Proposition 11.4 (Exponential Matrix: 3×3 Complex Form)
Let matrix A be 3×3 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1), (\lambda_2, \vec{\mathbf{v}}_2), (\lambda_3, \vec{\mathbf{v}}_3)$. Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ be independent. Then:

$$e^{At} = \Big\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \Big\rangle \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{pmatrix} \Big\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 \Big\rangle^{-1}.$$

The formula applies when the eigenpairs are real and also when the eigenpairs are complex. Proof on page 887.

Proposition 11.5 (Exponential Matrix: 3×3 Real Form)
Let matrix A be 3×3 real with eigenpairs $(\lambda_1, \vec{\mathbf{v}}_1)$, $(\lambda_2, \vec{\mathbf{v}}_2)$, $(\lambda_3, \vec{\mathbf{v}}_3)$. Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ be independent. Assume one eigenvalue λ_3 is real and the other eigenvalues are a complex conjugate pair $\lambda_1 = \overline{\lambda}_2 = a + ib$, $b > 0$. Define matrix $P = \Big\langle \mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1) | \mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1) | \vec{\mathbf{v}}_3 \Big\rangle$. Then P is invertible and the exponential matrix is:

$$e^{At} = P \begin{pmatrix} e^{at}\cos bt & e^{at}\sin bt & 0 \\ -e^{at}\sin bt & e^{at}\cos bt & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{pmatrix} P^{-1} \tag{5}$$

Proof on page 887.

Eigenanalysis Method: $n \times n$ Matrix

The general solution formula and the formula for e^{At} generalize routinely from the 2×2 and 3×3 cases to the general case of an $n \times n$ matrix. Proofs are left as an exercise, guided by the 3×3 case.

Theorem 11.23 (The Eigenanalysis Method)
Let the $n \times n$ real matrix A have eigenpairs

$$(\lambda_1, \vec{\mathbf{v}}_1), \quad (\lambda_2, \vec{\mathbf{v}}_2), \quad \ldots, \quad (\lambda_n, \vec{\mathbf{v}}_n)$$

with n independent eigenvectors $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_n$. Then the general solution of the linear system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is given by

$$\vec{\mathbf{x}}(t) = c_1 \vec{\mathbf{v}}_1 e^{\lambda_1 t} + c_2 \vec{\mathbf{v}}_2 e^{\lambda_2 t} + \cdots + c_n \vec{\mathbf{v}}_n e^{\lambda_n t}. \tag{6}$$

Proposition 11.6 (General Solution: $n \times n$ Complex Matrix Form)
General solution (6) can be expressed as a matrix product:

$$\vec{\mathbf{x}}(t) = \left\langle \vec{\mathbf{v}}_1 | \cdots | \vec{\mathbf{v}}_n \right\rangle \textbf{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}) \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}.$$

Definition 11.3 (Real Diagonal Form)
Assume $n \times n$ matrix A is diagonalizable. List all complex eigenvalues of A in pairs λ_1, $\overline{\lambda}_1$, ..., λ_p, $\overline{\lambda}_p$. Then list the real eigenvalues $r_1, \ldots, r_q$, $2p + q = n$. List the eigenpairs as $(\lambda_i, \vec{\mathbf{v}}_i)$, $(\overline{\lambda}_i, \overline{\vec{\mathbf{v}}}_i)$, $1 \le i \le p$ and $(r_j, \vec{\mathbf{v}}_{2p+j})$, $1 \le j \le q$. Define

$$\begin{aligned} P &= \left\langle \mathcal{R}\text{e}(\vec{\mathbf{v}}_1) | \mathcal{I}\text{m}(\vec{\mathbf{v}}_1) | \cdots | \mathcal{R}\text{e}(\vec{\mathbf{v}}_{2p-1}) | \mathcal{I}\text{m}(\vec{\mathbf{v}}_{2p-1}) | \vec{\mathbf{v}}_{2p+1} | \cdots | \vec{\mathbf{v}}_n \right\rangle \\ J_\lambda &= \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, \quad \lambda = a + ib, \quad b > 0 \end{aligned}$$

The **real diagonal form**:

$$A = P\, \textbf{diag}\left(J_{\lambda_1}, \cdots, J_{\lambda_p}, r_1, \cdots, r_q \right) P^{-1}$$

Proposition 11.7 (Exponential Matrix: $n \times n$ Real Matrix Form)
Define

$$R_\lambda(t) = e^{at} \begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}, \quad \lambda + a + ib, \quad b > 0.$$

Let $A = P\, \textbf{diag}\left(J_{\lambda_1}, \cdots, J_{\lambda_p}, r_1, \cdots, r_q \right) P^{-1}$ be the **real diagonal form** of diagonalizable matrix A.

Then:

$$e^{At} = P\, \textbf{diag}(R_{\lambda_1}(t), \ldots, R_{\lambda_p}(t), e^{r_1 t}, \ldots, e^{r_q t}) P^{-1}.$$

Remark on Euler Solution Atoms. If the characteristic equation is $(\lambda-1)^3 = 0$ and there are three independent eigenvectors, then the general solution $\vec{\mathbf{x}}(t) = c_1e^{\lambda_1 t}\vec{\mathbf{v}}_1 + c_2e^{\lambda_2 t}\vec{\mathbf{v}}_2 + c_3e^{\lambda_3 t}\vec{\mathbf{v}}_3$ contains no terms with te^t nor t^2e^t. Intuition from $(\lambda-1)^3 = 0$ suggests that solution components should be linear combinations of e^t, te^t, t^2e^t. How is that possible? The answer is contained in the linear combination $2e^t + 0te^t + 0t^2e^t$: it is indeed a linear combination of Euler atoms e^t, te^t, t^2e^t.

Classical Spectral Theory Method

The simplicity of Putzer's spectral method for computing e^{At} is appreciated, but we also recognize that the literature has an algorithm to compute e^{At}, devoid of differential equations, which is of fundamental importance in linear algebra. The parallel algorithm computes e^{At} directly from the eigenvalues λ_j of A and certain products of the nilpotent matrices $A - \lambda_j I$. Called **spectral formulas**, they can be implemented in a numerical laboratory or computer algebra system, in order to efficiently compute e^{At}, even in the case of multiple eigenvalues.

Theorem 11.24 (Spectral Formula for e^{At}: Simple Eigenvalues)
Let the $n \times n$ matrix A have n simple eigenvalues $\lambda_1, \ldots, \lambda_n$ (possibly complex) and define constant matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ by the formulas

$$\boldsymbol{Q}_j = \Pi_{i\neq j}\frac{A-\lambda_i I}{\lambda_j - \lambda_i}, \quad j = 1, \ldots, n.$$

Then

$$e^{At} = e^{\lambda_1 t}\boldsymbol{Q}_1 + \cdots + e^{\lambda_n t}\boldsymbol{Q}_n.$$

Theorem 11.25 (Spectral Formula for e^{At}: Multiple Eigenvalues)
Let the $n \times n$ matrix A have k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ of algebraic multiplicities $m_1, \ldots, m_k$. Let $p(\lambda) = \det(A - \lambda I)$ and define polynomials $a_1(\lambda), \ldots, a_k(\lambda)$ by the partial fraction identity

$$\frac{1}{p(\lambda)} = \frac{a_1(\lambda)}{(\lambda-\lambda_1)^{m_1}} + \cdots + \frac{a_k(\lambda)}{(\lambda-\lambda_k)^{m_k}}.$$

Define constant matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_k$ by the formulas

$$\boldsymbol{Q}_j = a_j(A)\Pi_{i\neq j}(A-\lambda_i I)^{m_i}, \quad j = 1, \ldots, k.$$

Then

$$(7) \qquad e^{At} = \sum_{i=1}^{k} e^{\lambda_i t}\,\boldsymbol{Q}_i \sum_{j=0}^{m_i-1}(A-\lambda_i I)^j\,\frac{t^j}{j!}.$$

Proof: Let $\boldsymbol{N}_i = \boldsymbol{Q}_i(A - \lambda_i I)$, $1 \le i \le k$. First:

Lemma 11.1 (Properties)
1. $\boldsymbol{Q}_1 + \cdots + \boldsymbol{Q}_k = I$,
2. $\boldsymbol{Q}_i\boldsymbol{Q}_j = \mathbf{0}$ for $i \neq j$,
3. $\boldsymbol{Q}_i\boldsymbol{Q}_i = \boldsymbol{Q}_i$,
4. $\boldsymbol{N}_i\boldsymbol{N}_j = \mathbf{0}$ for $i \neq j$,
5. $\boldsymbol{N}_i^{m_i} = \mathbf{0}$,
6. $A = \sum_{i=1}^k (\lambda_i\boldsymbol{Q}_i + \boldsymbol{N}_i)$.

To prove exponential formula (7), use Lemma 11.1 as follows:

$$\begin{aligned}
e^{At} &= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{At} && \text{Lemma 11.1, item } \mathbf{1}\\
&= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{\lambda_i It + (A-\lambda_i I)t} &&\\
&= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{\lambda_i t} e^{(A-\lambda_i I)t} &&\\
&= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{\lambda_i t} e^{\boldsymbol{Q}_i(A-\lambda_i I)t} && \text{Lemma 11.1, items } \mathbf{2}, \mathbf{3}\\
&= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{\lambda_i t} e^{\boldsymbol{N}_i t} && \text{Definition of } \boldsymbol{N}_i\\
&= \textstyle\sum_{i=1}^k \boldsymbol{Q}_i e^{\lambda_i t} \sum_{j=0}^{m_1-1} (A-\lambda_i I)^j \frac{t^j}{j!} && \text{Lemma 11.1, item } \mathbf{6}
\end{aligned}$$

Proof of Lemma 11.1:
Identity 1: Clear fractions in the partial fraction expansion of $1/p(\lambda)$:

$$1 = \sum_{i=1}^k a_i(\lambda)\frac{p(\lambda)}{(\lambda-\lambda_i)^{m_i}}.$$

Identity 2: Observe that $\boldsymbol{Q}_i$ and $\boldsymbol{Q}_j$ together contain all the factors of $p(A)$, therefore $\boldsymbol{Q}_i\boldsymbol{Q}_j = q(A)p(A)$ for some polynomial q. The Cayley-Hamilton theorem $p(A) = \mathbf{0}$ finishes the details.

Identity 3: Multiply identity **1** by $\boldsymbol{Q}_i$ and then use **2**.

Identity 4: Write $\boldsymbol{N}_i\boldsymbol{N}_j = (A-\lambda_i I)(A-\lambda_j I)\boldsymbol{Q}_i\boldsymbol{Q}_j$ and apply **3**.

Identity 5: Identity **2** implies $\boldsymbol{Q}_i^{m_i} = \boldsymbol{Q}_i$, then $\boldsymbol{N}_i^{m_i} = (A-\lambda_i I)^{m_i}\boldsymbol{Q}_i = p(A) = \mathbf{0}$.

Identity 6: Multiply identity **1** by A and rearrange:

$$\begin{aligned}
A &= \textstyle\sum_{i=1}^k A\boldsymbol{Q}_i\\
&= \textstyle\sum_{i=1}^k \lambda_i\boldsymbol{Q}_i + (A-\lambda_i I)\boldsymbol{Q}_i\\
&= \textstyle\sum_{i=1}^k \lambda_i\boldsymbol{Q}_i + \boldsymbol{N}_i
\end{aligned}$$

■

Cayley-Hamilton-Ziebur for $\vec{x}'(t) = A\vec{x}(t)$

Given $n \times n$ matrix A, determinant $|A - rI|$ is formed by subtracting r from the diagonal of A. The **characteristic polynomial** is $p(r) = |A - rI|$ and $|A - rI| = 0$ is the **characteristic equation**.

The famous result of Cayley and Hamilton is restated in Theorem 11.26. An elementary proof appears in linear algebra Chapter 5, Theorem 5.20, page 357.

Theorem 11.26 (Cayley-Hamilton)
Every square matrix A satisfies its own characteristic equation.

Let $|A - r\boldsymbol{I}| = (-r)^n + a_{n-1}(-r)^{n-1} + \cdots + a_0$ be the characteristic polynomial of $n \times n$ matrix A. Let $\boldsymbol{I}$ and $\mathbf{0}$ denote the $n \times n$ identity and zero matrix. Then:

$$(-A)^n + a_{n-1}(-A)^{n-1} + \cdots + a_1(-A) + a_0\boldsymbol{I} = \mathbf{0}$$

Theorem 11.27 (Cayley-Hamilton-Ziebur Theorem: Scalar Form)
Let A be an $n \times n$ real matrix. Each of the components $x_1(t), \ldots, x_n(t)$ of a real vector solution $\vec{\mathbf{x}}(t)$ of system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$ is a solution of an nth order scalar linear homogeneous constant-coefficient differential equation with characteristic equation $|A - rI| = 0$. The result remains true for complex solutions $\vec{\mathbf{x}}(t)$ and complex A. Proof on page 888.

Theorem 11.28 (Cayley-Hamilton-Ziebur Theorem: Vector Form)
Let A be an $n \times n$ real matrix. Let $A_1(t), \ldots, A_n(t)$ be Euler solution atoms constructed from the roots of $|A - rI| = 0$. The solution of system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ is a **vector** linear combination of $A_1(t), \ldots, A_n(t)$:

$$\vec{\mathbf{x}}(t) = \vec{\mathbf{d}}_1 \, A_1(t) + \cdots + \vec{\mathbf{d}}_n \, A_n(t).$$

Constant vectors $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_n$ are determined by A and $\vec{\mathbf{x}}(0)$ (see identity (8) *infra*). The result holds for complex A provided $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_n$ are complex. Euler atoms may be replaced by complex exponentials times powers of t. Proof on page 888.

Theorem 11.29 (Cayley-Hamilton-Ziebur Identity: Real)
Let $W(t)$ be the Wronskian matrix of Euler solution atoms $\{A_j\}_{j=1}^n$ constructed from the roots of $|A - rI| = 0$. Let $V = W(0)^T$. Constant vectors $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_n$ in Cayley-Hamilton-Ziebur Theorem 11.28 are determined by:

$$\left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle = \left\langle \vec{\mathbf{x}}(0) | A\vec{\mathbf{x}}(0) | \cdots | A^{n-1}\vec{\mathbf{x}}(0) \right\rangle V^{-1}. \tag{8}$$

Proof on page 888.

Theorem 11.30 (Cayley-Hamilton-Ziebur Identity: Complex)
Identity (8) remains valid if set $\{A_j\}_{j=1}^n$ is replaced by complex independent linear combinations $\{B_j\}_{j=1}^n$ of $\{A_j\}_{j=1}^n$ with $\{\vec{\mathbf{d}}_j\}_{j=1}^n$ possibly complex and $W(t)$ is replaced by the Wronskian matrix of $\{B_j\}_{j=1}^n$. Proof on page 888.

Theorem 11.31 (Vandermonde Matrix and Identity (8))
Assume the results of Theorems 11.29 and 11.30. If roots $\lambda = \lambda_1, \ldots, \lambda_n$ of $|A - \lambda I| = 0$ are distinct, then matrix $V = W(0)^T$ is the Vandermonde matrix of the roots:

$$V = \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{pmatrix}. \tag{9}$$

Proof on page 889.

Theorem 11.32 (Eigenvectors and Identity (8))
Assume the results of Theorems 11.29, 11.30. If $|A-\lambda I| = 0$ has distinct roots $\lambda_1, \ldots, \lambda_n$, then vector $\vec{\mathbf{d}}_j$ is a scalar multiple of eigenvector $\vec{\mathbf{v}}_j$ for eigenvalue λ_j, $1 \le j \le n$ (Warning: $\vec{\mathbf{d}}_j = \vec{\mathbf{0}}$ is possible). Proof on page 889.

Theorem 11.33 (Eigenvectors by Matrix Multiply)
Let A have distinct eigenvalues $\{\lambda_j\}_{j=1}^n$ and define for any n-vector $\vec{\mathbf{U}}$

$$V = \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{pmatrix}, \quad P = \left\langle \vec{\mathbf{U}} | A\vec{\mathbf{U}} | \cdots | A^{n-1}\vec{\mathbf{U}} \right\rangle V^{-1}.$$

Then column j of P is either zero or else an eigenvector of A for λ_j.

> **Notation**: $\left\langle \vec{\mathbf{y}}_1 | \cdots | \vec{\mathbf{y}}_n \right\rangle$ = augmented matrix of $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n$. To determine all eigenvectors experimentally, start with all $\vec{\mathbf{U}}$-components one, then change some ones in $\vec{\mathbf{U}}$ to zero or minus one and repeat.

Proof on page 889.

Example 11.7 (Eigenvectors by Matrix Multiply)
Compute by Theorem 11.33 all eigenvectors of matrix $A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$.

Details for Example 11.7:
The matrix of eigenvectors is $P = \frac{1}{2}\begin{pmatrix} 1-i & 1+i \\ 1+i & 1-i \end{pmatrix}$. Solve $|A-\lambda I| = 0$ for complex eigenvalues $\lambda_1, \lambda_2 = 1 \pm 2i$, then define $V = \begin{pmatrix} 1 & \lambda_1 \\ 1 & \lambda_2 \end{pmatrix}$, $U = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Compute $B = \left\langle U | AU \right\rangle = \begin{pmatrix} 1 & 3 \\ -1 & 1 \end{pmatrix}$ and $V^{-1} = \frac{1}{|V|}\,\mathbf{adj}(V) = \frac{1}{-4i}\begin{pmatrix} \lambda_2 & -\lambda_1 \\ -1 & 1 \end{pmatrix}$. Multiply to find $P = BV^{-1} = \frac{1}{2}\begin{pmatrix} 1-i & 1+i \\ 1+i & 1-i \end{pmatrix}$. Maple code to check the computation:

```
with(LinearAlgebra):A:=<1,2|-2,1>^+;EV:=Eigenvalues(A);
U:=<1,1>;V:=VandermondeMatrix(EV);P:=<U|A.U>.(1/V);
J:=DiagonalMatrix(EV);A.P-P.J; # Check eigenvectors
```

Inverse of a Vandermonde Matrix

Notation: Symmetric function $e_k(r_1, \ldots, r_N) = \displaystyle\sum_{1 \le i_1 < \cdots < i_k \le N} r_{i_1} \cdots r_{i_k}$

Vieta's formulas[7] supply coefficients $a_k = (-1)^{N-k} e_k(r_1, \ldots, r_N)$ of degree N polynomial $\sum_{k=0}^{N-1} a_k y^k + y^N = \prod_{p=1}^{N} (y - r_p)$ with roots $r_1, \ldots, r_N$.

[7] See https://en.wikipedia.org/wiki/Vieta%27s_formulas.

Theorem 11.34 (Vandermode Inverse)

Let $A = \begin{pmatrix} 1 & \cdots & x_1^{n-1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n^{n-1} \end{pmatrix}$ with $x_1, \ldots, x_n$ distinct. Then $A^{-1} = B = (b_{ij})$:

$$b_{ij} = \frac{(-1)^{n-j} e_{n-j}(\{x_1, \ldots, x_n\} \setminus \{x_i\})}{\displaystyle\prod_{p=1, p\neq i}^{n} (x_j - x_p)}, \quad 1 \leq i, j \leq n. \tag{10}$$

Proof on page 890.

Vieta's Formulas: maple

```
# Vieta's formulas, monic polynomial: maple library
n:=3;q:=expand(product((y-x[i]),i=1..n));
ListTools[Reverse]([coeffs(q,y)]);

# Vieta's formulas: basic algorithm, no library
# Monic polynomial, roots x[1] to x[n]
F:=proc(r) local A,n,i,j;
n:=nops(r);A:=[seq(0,i=1..n+2)];A[n+1]:=1;
for i from 1 to n do  for j from n+1-i to n+1 do
   A[j]:=A[j]+(-1)*r[i]*A[j+1]; od; od;
return simplify([seq(A[i],i=1..n+1)]); end proc:
F([seq(x[i],i=1..3)]); # Test n=3
```

Solving Planar Systems $\vec{x}'(t) = A\vec{x}(t)$

A 2×2 real system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$ can be solved in terms of matrix A and the two roots of the characteristic equation $\det(A - \lambda I) = 0$.

Two distinct methods are explored below, both with minimal use of complex numbers.

The most-used method on paper is the Cayley-Hamilton-Ziebur **Scalar Shortcut**. Implementations for embedded systems might use the formulas obtained from the **Matrix Shortcut**. The only requirement on matrix A is that it **not** be a diagonal matrix.

Theorem 11.35 (Cayley-Hamilton-Ziebur Scalar 2×2 Shortcut)
Let $b \neq 0$ in the scalar system

$$\begin{aligned} x_1' &= a\,x_1 + b\,x_2 \\ x_2' &= c\,x_1 + d\,x_2 \end{aligned} \tag{11}$$

Define $x_1(t) = c_1 y_1(t) + c_2 y_2(t)$. Solve for $x_2(t)$ in the first equation, then replace x_1 by $c_1 y_1 + c_2 y_2$ on the right of $bx_2 = x_1' - ax_1$ and simplify to find $x_2 = k_1 y_1 + k_2 y_2$. Proof on page 890.

Theorem 11.36 (Cayley-Hamilton-Ziebur Matrix 2×2 Shortcut)

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $b \neq 0$. Let $y_1(t), y_2(t)$ be the Euler solution atoms found from the roots of $|A - \lambda I| = 0$. Define constant matrix B by identity $\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = B \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$. Then the general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ with arbitrary constants c_1, c_2 is

$$\begin{cases} x_1(t) = c_1 y_1(t) + c_2 y_2(t), \\ x_2(t) = k_1 y_1(t) + k_2 y_2(t), \end{cases} \quad \text{where} \quad \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \frac{1}{b}(B^T - aI)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Proof on page 890.

> **Remark**. Theorems 11.35, 11.36 solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ when A is not a diagonal matrix (meaning either $b \neq 0$ or $c \neq 0$). The case $b = 0$ and $c \neq 0$ is treated by swapping b, c and x_1, x_2 in both of Theorems 11.35, 11.36.

Example 11.8 ()
(Scalar and Matrix 2×2 Shortcuts for Real Roots)
Solve the system

$$\begin{cases} x_1'(t) = x_1(t) + 2x_2(t), \\ x_2'(t) = 2x_1(t) + x_2(t), \end{cases} \quad A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \vec{\mathbf{x}}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},$$

verifying the general solution $\begin{cases} x_1(t) = c_1 e^{-t} + c_2 e^{3t}, \\ x_2(t) = -c_1 e^{-t} + c_2 e^{3t}. \end{cases}$

Details Example 11.8:

The characteristic polynomial $\begin{vmatrix} 1-r & 2 \\ 2 & 1-r \end{vmatrix} = (1-r)^2 - 4 = (r+1)(r-3)$ has roots $r = -1, r = 3$ and Euler solution atoms e^{-t}, e^{3t}.

Scalar Shortcut Details. To apply Theorem 11.35, define $x_1 = c_1 e^{-t} + c_2 e^{3t}$. Solve the first differential equation $x_1' = x_1 + 2x_2$ for $2x_2 = x_1' - x_1 = (c_1 e^{-t} + c_2 e^{3t})' - x_1 = -2c_1 e^{-t} + 2e^{3t}$. Then $x_2 = -e^{-t} + e^{3t}$.

Matrix Shortcut Details. To apply Theorem 11.36, first compute matrix $B = \begin{pmatrix} -1 & 0 \\ 0 & 3 \end{pmatrix}$ from $\frac{d}{dt}\begin{pmatrix} e^{-t} \\ e^{3t} \end{pmatrix} = \begin{pmatrix} -e^{-t} \\ e^{3t} \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} e^{-t} \\ e^{3t} \end{pmatrix}$.
Theorem 11.36 implies

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \frac{1}{b}(B^T - aI)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{2}(B^T - I)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Then $x_2(t) = -c_1 y_1 + c_2 y_2 = -c_1 e^{-t} + c_2 e^{3t}$.

Example 11.9 ()
(Scalar and Matrix 2×2 Shortcuts for Complex Roots)
Solve the system

$$\begin{cases} x_1'(t) = x_1(t) + 2x_2(t), \\ x_2'(t) = -2x_1(t) + x_2(t), \end{cases} \quad A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}, \quad \vec{\mathbf{x}}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},$$

verifying the general solution $\begin{cases} x_1(t) = c_1 e^t \cos(2t) + c_2 e^t \sin(2t), \\ x_2(t) = c_2 e^t \cos(2t) - c_1 e^t \sin(2t). \end{cases}$

Details Example 11.9: Characteristic polynomial $\begin{vmatrix} 1-r & 2 \\ -2 & 1-r \end{vmatrix} = (1-r)^2 + 4$ has roots $r = 1 \pm 2i$ and Euler solution atoms $e^t \cos(2t)$, $e^t \sin(2t)$.

Scalar Shortcut Details. To apply Theorem 11.35, let $x_1 = c_1 e^t \cos(2t) + c_2 e^t \sin(2t)$, then solve the first differential equation $x_1' = x_1 + 2x_2$ for $2x_2 = x_1' - x_1 = (c_1 e^t \cos(2t) + c_2 e^t \sin(2t))' - x_1 = 2c_2 e^t \cos(2t) - 2c_1 e^t \sin(2t)$. Then $x_2 = c_2 e^t \cos(2t) - c_1 e^t \sin(2t)$.

Matrix Shortcut Details. To apply Theorem 11.36, first compute matrix $B = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$:

$$\frac{d}{dt}\begin{pmatrix} e^t \cos(2t) \\ e^t \sin(2t) \end{pmatrix} = \begin{pmatrix} e^t \cos(2t) - 2e^t \sin(2t) \\ e^t \sin(2t) + 2e^t \cos(2t) \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} e^t \cos(2t) \\ e^t \sin(2t) \end{pmatrix}$$

Theorem 11.36 implies

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \tfrac{1}{b}(B^T - aI)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \tfrac{1}{2}(B^T - I)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Then $x_2(t) = c_2 y_1 - c_1 y_2 = c_2 e^t \cos(2t) - c_1 e^t \sin(2t)$.

Theorem 11.37 (Putzer's Spectral Formula: 2×2)

Consider the real planar system $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t)$. Let λ_1, λ_2 be the roots of the characteristic equation $\det(A - \lambda I) = 0$. The real general solution is $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}(0)$ where the 2×2 exponential matrix e^{At} is given by

Real $\lambda_1 \neq \lambda_2$	$e^{At} = e^{\lambda_1 t} I + \dfrac{e^{\lambda_2 t} - e^{\lambda_1 t}}{\lambda_2 - \lambda_1}(A - \lambda_1 I)$.
Real $\lambda_1 = \lambda_2$	$e^{At} = e^{\lambda_1 t} I + t e^{\lambda_1 t}(A - \lambda_1 I)$.
Complex $\lambda_1 = \overline{\lambda_2}$, $\lambda_1 = a + bi$, $b > 0$	$e^{At} = e^{at} \cos bt\, I + \dfrac{e^{at} \sin(bt)}{b}(A - aI)$.

Proof: The formulas are from Putzer's algorithm page 868 or equivalently from the spectral formulas with rearranged terms. The complex case is formally the real part of the distinct root case when $\lambda_2 = \overline{\lambda_1}$. The three formulas are analogous to the second order equation formulas Chapter 6 Section 1, Theorem 6.1. ■

Example 11.10 (Classical and Putzer Spectral Formulas)

Typical cases are represented by the following 2×2 matrices A, which correspond to roots λ_1, λ_2 of the characteristic equation $\det(A - \lambda I) = 0$ which are real distinct, real double or complex conjugate. The solution $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}(0)$ is given here in two forms, by writing e^{At} using $\boxed{1}$ a **classical spectral formula** from Theorems 11.24–11.25 and $\boxed{2}$ Putzer's **spectral formula** from Theorem 11.37.

$\lambda_1 = 5$, $\lambda_2 = 2$ Real distinct roots.

$$A = \begin{pmatrix} -1 & 3 \\ -6 & 8 \end{pmatrix}$$

$$\boxed{1}\ e^{At} = \frac{e^{5t}}{3}\begin{pmatrix} -3 & 3 \\ -6 & 6 \end{pmatrix} + \frac{e^{2t}}{-3}\begin{pmatrix} -6 & 3 \\ -6 & 3 \end{pmatrix}$$

$$\boxed{2}\ e^{At} = e^{5t} I + \frac{e^{2t} - e^{5t}}{2 - 5}\begin{pmatrix} -6 & 3 \\ -6 & 3 \end{pmatrix}$$

$\lambda_1 = \lambda_2 = 3$ Real double root.

$A = \begin{pmatrix} 2 & 1 \\ -1 & 4 \end{pmatrix}$ $\boxed{1}$ $e^{At} = e^{3t}\left(I + t\begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}\right)$

$\boxed{2}$ $e^{At} = e^{3t}I + te^{3t}\begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$

$\lambda_1 = \overline{\lambda}_2 = 2 + 3i$ Complex conjugate roots.

$A = \begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix}$ $\boxed{1}$ $e^{At} = 2\,\mathcal{R}\mathrm{e}\left(\dfrac{e^{2t+3it}}{2(3i)}\begin{pmatrix} 3i & 3 \\ -3 & 3i \end{pmatrix}\right)$

$\boxed{2}$ $e^{At}{=}e^{2t}\cos(3t)I{+}\dfrac{e^{2t}\sin(3t)}{3}\begin{pmatrix} 0 & 3 \\ -3 & 0 \end{pmatrix}$

The complex eigenvalue example is typical for real $n \times n$ matrices A with a complex conjugate pair of eigenvalues $\lambda_1 = \overline{\lambda}_2$. Then $\boldsymbol{Q}_2 = \overline{\boldsymbol{Q}}_1$ for $\boxed{1}$. The result is that λ_2 is not used and a simpler expression results by using the college algebra equality $z + \overline{z} = 2\,\mathcal{R}\mathrm{e}(z)$:

$$e^{\lambda_1 t}\boldsymbol{Q}_1 + e^{\lambda_2 t}\boldsymbol{Q}_2 = 2\,\mathcal{R}\mathrm{e}\left(e^{\lambda_1 t}\boldsymbol{Q}_1\right).$$

This observation explains why e^{At} is real when A is real, by pairing complex conjugate eigenvalues in Theorems 11.24–11.25,

Proofs and Methods

Proof of Proposition 11.3:

Eigenpair $(\lambda_2, \vec{\mathbf{v}}_2)$ is never computed or used, because $A\vec{\mathbf{v}}_1 = \lambda_1\vec{\mathbf{v}}_1$ implies $A\overline{\vec{\mathbf{v}}}_1 = \overline{\lambda}_1\overline{\vec{\mathbf{v}}}_1$, which implies λ_2 $(= \overline{\lambda}_1)$ has eigenvector $\vec{\mathbf{v}}_2 = \overline{\vec{\mathbf{v}}}_1$.

If A is real, then e^{At} is real. Take real parts across the formula for e^{At} to give a real formula. Due to the unpleasantness of the complex algebra, we will justify the answer with minimal use of complex numbers.

The formula is established by showing that the matrix $\Phi(t)$ on the right of equation (4) satisfies $\Phi(0) = I$ and $\Phi' = A\Phi$. Then by definition, $e^{At} = \Phi(t)$. For exposition, let

$$R(t) = e^{at}\begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}, \quad \Phi(t) = PR(t)P^{-1}.$$

Identity $\Phi(0) = I$ is verified as follows.

$$\begin{aligned} \Phi(0) &= PR(0)P^{-1} \\ &= Pe^0\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}P^{-1} \\ &= I \end{aligned}$$

Express $\vec{\mathbf{v}}_1 = \mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1) + i\,\mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1)$. Expand eigenpair relation $A\vec{\mathbf{v}}_1 = \lambda_1\vec{\mathbf{v}}_1$ into real and imaginary parts:

$$A\left(\mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1) + i\,\mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1)\right) = (a + ib)\left(\mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1) + i\,\mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1)\right)$$

Match real and imaginary parts left and right in this equation to obtain:

$$AP = P\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

Then:

$$\begin{aligned} \Phi'(t)\Phi^{-1}(t) &= PR'(t)P^{-1}PR^{-1}(t)P^{-1} \\ &= PR'(t)R^{-1}(t)P^{-1} \\ &= P\left(aI + \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}\right)P^{-1} \\ &= P\begin{pmatrix} a & b \\ -b & a \end{pmatrix}P^{-1} \\ &= A \end{aligned}$$

Because $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$, then $\Phi(t) = e^{At}$. The general solution is $\vec{\mathbf{x}}(t) = \Phi(t)\vec{\mathbf{x}}(0)$. Then

$$\vec{\mathbf{x}}(t) = e^{at}\Big\langle \mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1)|\mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1)\Big\rangle \begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

where values c_1, c_2 are related to the initial condition $\vec{\mathbf{x}}(0)$ by identity

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \Big\langle \mathcal{R}\mathrm{e}(\vec{\mathbf{v}}_1)|\mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1)\Big\rangle^{-1}\vec{\mathbf{x}}(0)$$

Proof of Theorem 11.22:

The eigenvalues $\lambda_1, \lambda_2, \lambda_3$ can be all real or eigenvalue λ_3 is real and the other eigenvalues are complex: $\lambda_1 = \overline{\lambda_2} = a + ib$ with $b > 0$.

The proposed solution $\vec{\mathbf{x}}$ can be written in vector-matrix form:

$$\vec{\mathbf{x}}(t) = \Big\langle \vec{\mathbf{v}}_1|\vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3\Big\rangle \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

Because the three eigenvectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$ are assumed independent, then $\Big\langle \vec{\mathbf{v}}_1|\vec{\mathbf{v}}_2|\vec{\mathbf{v}}_3\Big\rangle$ is invertible. Setting $t = 0$ in the previous display gives

$$\begin{pmatrix} c_1 \\ c_2 \\ c_2 \end{pmatrix} = \Big\langle \vec{\mathbf{v}}_1|\vec{\mathbf{v}}_2|\vec{\mathbf{v}}_3\Big\rangle^{-1}\vec{\mathbf{x}}(0).$$

Constants c_1, c_2, c_3 can be chosen to produce any initial condition $\vec{\mathbf{x}}(0)$, therefore $\vec{\mathbf{x}}(t)$ is the *general solution* of the 3×3 system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

Proofs of Propositions 11.4 and 11.5:

The proof of Theorem 11.22 supplies the proof details for Proposition 11.4.

Proposition 11.5 is proved in two steps: (1) Show P has independent columns, hence P is invertible; (2) The exponential matrix is given by equation (5).

(1) Let $\vec{\mathbf{v}}_2 = \overline{\vec{\mathbf{v}}}_1$. Replace the first two column vectors in P by

$$Re(\vec{\mathbf{v}}_1) = \frac{1}{2}(\vec{\mathbf{v}}_1 + \vec{\mathbf{v}}_2), \quad \mathcal{I}\mathrm{m}(\vec{\mathbf{v}}_1) = -\frac{i}{2}(\vec{\mathbf{v}}_1 - \vec{\mathbf{v}}_2).$$

Let d_1, d_2, d_3 be constants. Assume dependency relation $P\begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \vec{\mathbf{0}}$. Then:

$$\frac{1}{2}d_1(\vec{\mathbf{v}}_1 + \vec{\mathbf{v}}_2) - \frac{i}{2}d_2(\vec{\mathbf{v}}_1 - \vec{\mathbf{v}}_2) + d_3\vec{\mathbf{v}}_3 = \vec{\mathbf{0}}.$$

Independence of $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3$ implies all linear combination weights are zero:

$$\frac{1}{2}d_1 - \frac{i}{2}d_2 = 0, \quad \frac{1}{2}d_1 + \frac{i}{2}d_2 = 0, \quad d_3 = 0.$$

Solve this system to prove $d_1 = d_2 = d_3 = 0$. Conclude that the columns of P are independent.

(2) Let $B = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$. Define block matrix $J = \begin{pmatrix} B & 0 \\ \mathbf{0} & \lambda_3 \end{pmatrix}$. Diagonalization theory for matrices implies $AP = PJ$. Then:

$e^{Jt} = \begin{pmatrix} e^{Bt} & \vec{\mathbf{0}} \\ \mathbf{0} & e^{\lambda_3 t} \end{pmatrix}$	Theorem 11.19, page 869
$e^{Bt} = \begin{pmatrix} e^{at}\cos bt & e^{at}\sin bt \\ -e^{at}\sin bt & e^{at}\cos bt \end{pmatrix}$	Theorem 11.20, page 869
$e^{At} = P\,e^{Jt}\,P^{-1}$	Identities page 865
$e^{At} = P\begin{pmatrix} e^{at}\cos bt & e^{at}\sin bt & 0 \\ -e^{at}\sin bt & e^{at}\cos bt & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{pmatrix}P^{-1}$	

Proof of Theorem 11.27. Consider first the case $n = 2$, which has a routine generalization to higher dimensions.

$r^2 + a_1 r + a_0 = 0$	Expanded characteristic equation
$A^2 + a_1 A + a_0 I = 0$	Cayley-Hamilton matrix equation, where I and 0 are the identity and zero matrix.
$A^2\vec{\mathbf{x}} + a_1 A\vec{\mathbf{x}} + a_0\vec{\mathbf{x}} = \vec{\mathbf{0}}$	Right-multiply by $\vec{\mathbf{x}} = \vec{\mathbf{x}}(t)$
$\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}' = A^2\vec{\mathbf{x}}$	Differentiate $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$
$\vec{\mathbf{x}}'' + a_1\vec{\mathbf{x}}' + a_0\vec{\mathbf{x}} = \vec{\mathbf{0}}$	Replace $A^2\vec{\mathbf{x}} \to \vec{\mathbf{x}}''$, $A\vec{\mathbf{x}} \to \vec{\mathbf{x}}'$

Multiply the vector relation by the rows of the identity matrix to show the components $x_1(t)$, $x_2(t)$ of $\vec{\mathbf{x}}(t)$ satisfy the two differential equations

$$\begin{aligned} x_1''(t) + a_1 x_1'(t) + a_0 x_1(t) &= 0, \\ x_2''(t) + a_1 x_2'(t) + a_0 x_2(t) &= 0. \end{aligned}$$

This system implies that the components of $\vec{\mathbf{x}}(t)$ are solutions of the second order differential equation with characteristic equation $|A - rI| = 0$.

The proof remains valid if real solution $\vec{\mathbf{x}}(t)$ is replaced by a complex solution, no changes required in the above text. Because the Cayley-Hamilton theorem is valid for complex A, the proof is complete for $n = 2$. Details for any n are left to the reader.

Proofs of Theorems 11.27, 11.28, 11.29 and 11.30. The scalar form Theorem 11.27 can be written

$$x_i(t) = c_{i1}A_1 + \cdots + c_{in}A_n, \quad i = 1, \ldots, n.$$

In matrix form:

$$\vec{\mathbf{x}}(t) = \begin{pmatrix} c_{11} & \cdots & c_{1,n} \\ \vdots & \vdots & \vdots \\ c_{11} & \cdots & c_{1,n} \end{pmatrix} \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix}.$$

Then $\vec{\mathbf{d}}_1$ is the first column of matrix (c_{ij}) above, and so on, which proves

$$\vec{\mathbf{x}}(t) = \vec{\mathbf{d}}_1 A_1 + \cdots + \vec{\mathbf{d}}_n A_n \tag{12}$$

Left to prove is that column vectors $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_n$ depend only on A and initial data $\vec{\mathbf{x}}(0)$. We proceed as in the theory of Wronskian determinants by differentiation $n-1$ times of equation (12), then replace t by zero to obtain these formulas:

$$\left\{ \begin{array}{lcrcccr} \vec{\mathbf{x}}(0) & = & A_1(0)\,\vec{\mathbf{d}}_1 & + & \cdots & + & A_n(0)\,\vec{\mathbf{d}}_n \\ \vec{\mathbf{x}}'(0) & = & A_1'(0)\,\vec{\mathbf{d}}_1 & + & \cdots & + & A_n'(0)\,\vec{\mathbf{d}}_n \\ & \vdots & & & & & \\ \vec{\mathbf{x}}^{(n-1)}(0) & = & A_1^{(n-1)}(0)\,\vec{\mathbf{d}}_1 & + & \cdots & + & A_n^{(n-1)}(0)\,\vec{\mathbf{d}}_n \end{array} \right. \tag{13}$$

The derivatives on the left in equation (13) can be cleverly rewritten as $\vec{\mathbf{x}}(0)$, $A\vec{\mathbf{x}}(0)$, …, $A^{n-1}\vec{\mathbf{x}}(0)$ by successive differentiation of $\vec{\mathbf{x}}'(t) = Xx(t)$. For instance, $\vec{\mathbf{x}}''(t) = (A\vec{\mathbf{x}}(t))' = A\vec{\mathbf{x}}'(t) = AA\vec{\mathbf{x}}(t) = A^2\vec{\mathbf{x}}(t)$, then $t = 0$ gives $\vec{\mathbf{x}}''(0) = A^2\vec{\mathbf{x}}(0)$. The result in matrix form:

$$\left\langle \vec{\mathbf{x}}(0) | \cdots | A^{n-1}\vec{\mathbf{x}}(0) \right\rangle = \left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle \begin{pmatrix} A_1(0) \cdots A_1^{(n-1)}(0) \\ \vdots \quad \ldots \quad \vdots \\ A_n(0) \cdots A_n^{(n-1)}(0) \end{pmatrix} \tag{14}$$

The augmented matrix $\left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle$ of vectors $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_n$ is then obtained by matrix inversion: $\left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle = \left\langle \vec{\mathbf{x}}(0) | \cdots | A^{n-1}\vec{\mathbf{x}}(0) \right\rangle (W(0)^T)^{-1}$, where $W(t)$ is the Wronskian matrix of the n Euler solution atoms.

Suppose A_j is replaced by B_j which are independent linear combinations of atoms A_j with complex coefficients. Assume given a solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, then $\vec{\mathbf{x}}(t) = \sum_{j=1}^{n} \vec{\mathbf{d}}_j B_j(t)$ for some column vectors $\vec{\mathbf{d}}_j$. Let's differentiate this relation $n-1$ times and substitute $t = 0$, as before. The same analysis with matrix multiply, Wronskians and inverses applies, therefore identity (8) remains valid. Related details appear in the proof of Theorem 11.31.

Proof of Theorem 11.31. If all roots are real distinct, then the Euler solution atoms are $e^{\lambda_1 t}$, …, $e^{\lambda_n t}$. Find the Wronskian matrix of these functions, then let $t = 0$, which makes all exponentials equal to one. The first row is all ones, therefore the transpose matrix has first column all ones. If complex roots $a \pm bi$ appear, then in affected atoms $\cos bt = \frac{1}{2}(e^{ibt} + e^{-ibt})$, $\sin bt = \frac{1}{2i}(e^{ibt} - e^{-ibt})$. Collect terms into complex exponentials $e^{(a+ib)t}$ multiplied by vectors (complex entries allowed). Identity (8) is unchanged except for replacement of atoms by exponentials. Proceed to identify $W(t)$, $W(0)$ and $W(0)^T$ in the same manner as for real roots.

Proofs of Theorems 11.32, 11.33. Assume the CHZ results of previous theorems and that A has distinct eigenvalues, complex numbers allowed. Let $(\lambda_j, \vec{\mathbf{v}}_j)$, $1 \le j \le n$ be a list of eigenpairs of A. Let $A_j = e^{\lambda_j t}$, $1 \le j \le n$: they are independent functions with invertible Wronskian matrix $W(t)$ (see the Exercises). The Eigenanalysis method supplies solution $\vec{\mathbf{x}}(t) = \sum_{j=1}^{n} c_j \vec{\mathbf{v}}_j\, A_j(t)$ whereas CHZ supplies $\vec{\mathbf{x}}(t) = \sum_{j=1}^{n} \vec{\mathbf{d}}_j\, A_j(t)$. Independence of list $\{A_j(t)\}_{j=1}^{n}$ implies $\vec{\mathbf{d}}_j = c_j \vec{\mathbf{v}}_j$. However, $c_j = 0$ is possible, therefore $\vec{\mathbf{d}}_j = \vec{\mathbf{0}}$ or else $\vec{\mathbf{d}}_j$ is a nonzero multiple of eigenvector $\vec{\mathbf{v}}_j$, $1 \le j \le n$.

Theorem 11.33 directly applies Theorem 11.32, which implies the columns of augmented matrix $P = \left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle$ are either zero or a nonzero multiple of an eigenvector of A. Examples choose $\vec{\mathbf{U}}$ initially to be the column vector of ones, then ones are modified to zero or minus one: then re-apply the formula to find all eigenvectors.

Proof of Theorem 11.34, Vandermonde Inverse:

Case for $n = 3$. The inverse matrix $B = \begin{pmatrix} a_0 & \cdot & \cdot \\ a_1 & \cdot & \cdot \\ a_2 & \cdot & \cdot \end{pmatrix}$ of Vandermonde matrix $A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix}$ satisfies $AB = I$. Match column one on both sides of $AB = I$ using matrix multiply, then for polynomial $p_1(y) = a_0 + a_1 y + a_2 y^2$ there are three interpolation equations to be satisfied:

$$a_0 + a_1 x_1 + a_2 x_1^2 = 1,\ a_0 + a_1 x_2 + a_2 x_2^2 = 0,\ a_0 + a_1 x_3 + a_2 x_3^2 = 0.$$

Degree 2 polynomial $q_1(y) = \frac{1}{y - x_1} \prod_{i=1}^{3} (y - x_i)$ constructs the interpolation problem unique solution $p_1(y) = \frac{q_1(y)}{q_1(x_1)}$. Coefficients a_0, a_1, a_2 are found by matching y-coefficients after expanding equation $a_0 + a_1 y + a_2 y^2 = \frac{q_1(y)}{q_1(x_1)}$. Define q_2, q_3, p_2, p_3 analogously and repeat for columns 2, 3. Then inverse B equals:

$$\begin{pmatrix} x_2 x_3 & x_1 x_3 & x_1 x_2 \\ -x_2 - x_3 & -x_1 - x_3 & -x_1 - x_2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{(x_1 - x_2)(x_1 - x_3)} & 0 & 0 \\ 0 & \frac{1}{(x_2 - x_1)(x_2 - x_3)} & 0 \\ 0 & 0 & \frac{1}{(x_3 - x_1)(x_3 - x_2)} \end{pmatrix}$$

Case for General n. For i from 1 to n, define degree $n - 1$ polynomials $q_i(y) = \frac{1}{y - x_i} \prod_{p=1}^{n} (y - x_p)$, then expand $q_i(y) = \sum_{j=1}^{n} q_{ij} y^{j-1}$ to obtain

$$B = (b_{ij}) = \begin{pmatrix} q_{11} & \cdots & q_{n1} \\ \vdots & \ldots & \vdots \\ q_{1n} & \cdots & q_{nn} \end{pmatrix} \begin{pmatrix} \frac{1}{q_1(x_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{q_n(x_n)} \end{pmatrix}.$$

Formula for $q_i(x_i)$. Cancel factor $y - x_i$, then $q_i(x_i) = \prod_{p=1, p \ne i}^{n} (x_i - x_p)$.

Formula for q_{ij}. Let $N = n - 1$. Vieta's formulas applied to degree N polynomial $q_i(y) = \sum_{j=1}^{n} q_{i\,j} y^{j-1}$ give $q_{i\,j} = (-1)^{N-j+1} e_{N-j+1}(\{x_1, \ldots, x_n\} \setminus \{x_i\})$, for $j = 1, \ldots, n - 1$. Equality $b_{ij} = \frac{q_{ij}}{q_i(x_i)}$ then establishes equation (10).

Proof of Theorem 11.35. The details are in the proof of Theorem 11.36, which discusses the application of Theorem 11.27 and solving the first differential equation for variable x_2. This is the preferred shortcut on paper.

Proof of Theorem (11.36). The formula for $x_1(t)$ follows directly from Cayley-Hamilton-Ziebur Theorem 11.27. Equation $x_2(t)$=$k_1 y_1(t) + k_2 y_2(t)$ follows from the same theorem, for some constants k_1, k_2. It remains to prove that the constants are $\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \frac{1}{b}(B^T - aI) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$. Details:

$x_2 = \frac{1}{b} x_1' - \frac{a}{b} x_1$ Solve x_1'=$a x_1 + b x_2$ for x_2.

$x_2 = \frac{1}{b}(c_1 y_1' + c_2 y_2') - \frac{a}{b}(c_1 y_1 + c_2 y_2)$	Replace $x_1 = c_1 y_1 + c_2 y_2$.
$x_2 = \frac{1}{b}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T \begin{pmatrix}y_1'\\y_2'\end{pmatrix} - \frac{a}{b}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T \begin{pmatrix}y_1\\y_2\end{pmatrix}$	Rewrite as matrix multiply.
$x_2 = \frac{1}{b}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T B \begin{pmatrix}y_1\\y_2\end{pmatrix} - \frac{a}{b}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T \begin{pmatrix}y_1\\y_2\end{pmatrix}$	Definition of B.
$x_2 = \frac{1}{b}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T (B - aI) \begin{pmatrix}y_1\\y_2\end{pmatrix}$	Factor out $\begin{pmatrix}y_1\\y_2\end{pmatrix}$ right, $\begin{pmatrix}c_1\\c_2\end{pmatrix}^T$ left.
$x_2 = \frac{1}{b}\left((B^T - aI)\begin{pmatrix}c_1\\c_2\end{pmatrix}\right)^T \begin{pmatrix}y_1\\y_2\end{pmatrix}$	Matrix transpose properties $(CD)^T = D^T C^T$ and $(C+D)^T = C^T + D^T$.
$x_2 = \begin{pmatrix}k_1\\k_2\end{pmatrix}^T \begin{pmatrix}y_1\\y_2\end{pmatrix}$	Theorem's definition of k_1, k_2.
$x_2 = k_1 y_1 + k_2 y_2$	Verification complete.

Exercises 11.5

Determinant $|A - rI|$

Justify these statements.

1. Subtract r from the diagonal of A to form $|A - rI|$.

2. If A is 2×2, then $|A - rI|$ is a quadratic.

3. If A is 3×3, then $|A - rI|$ is a cubic.

4. Expansion of $|A - rI|$ by the cofactor rule often preserves factorizations.

5. If A is triangular, then $|A - rI|$ is the product of diagonal entries.

6. The *combo*, *mult* and *swap* rules for determinants are generally counter-productive for expansion of $|A - rI|$.

Characteristic Polynomial

Show expansion details for $|A - rI|$.

7. $A = \begin{pmatrix}2&3\\0&4\end{pmatrix}$.

Ans: $(2 - r)(4 - r)$

8. $A = \begin{pmatrix}2&3&4\\0&5&6\\0&0&7\end{pmatrix}$.

Ans: $(2 - r)(5 - r)(7 - r)$

Eigenanalysis Method: 2×2

Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

9. $A = \begin{pmatrix}1&0\\0&2\end{pmatrix}$

10. $A = \begin{pmatrix}1&1\\2&2\end{pmatrix}$

Eigenanalysis Method: 3×3

Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

11. $A = \begin{pmatrix}1&1&0\\2&2&0\\0&0&1\end{pmatrix}$

12. $A = \begin{pmatrix}1&1&0\\2&2&1\\0&0&1\end{pmatrix}$

Eigenanalysis Method: $n \times n$

Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$.

13. $A = \begin{pmatrix}1&1&0&0\\2&2&1&0\\0&0&1&0\\0&0&0&1\end{pmatrix}$

14. $A = \begin{pmatrix}1&1&0&0&1\\2&2&1&0&1\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}$

e^{At} for Simple Eigenvalues

Find a^{At} using classical spectral theory. Check by computer.

15. $A = \begin{pmatrix}1&1\\2&2\end{pmatrix}$

16. $A = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

e^{At} for Multiple Eigenvalues

Find a^{At} using classical spectral theory. Check by computer.

17. $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

18. $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$

Cayley-Hamilton Theorem

Prove the identity by applying the Cayley-Hamilton Theorem.

19. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $a_0 = |A| = ad - bc$, $a_1 = \mathbf{trace}(A) = a + d$. Then
$$A^2 + a_1(-A) + a_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

20. Let $A = \begin{pmatrix} 2 & 3 & 4 \\ 0 & 5 & 6 \\ 0 & 0 & 7 \end{pmatrix}$. Then:
$$(2I - A)(5I - A)(7I - A) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

CHZ Theorem: Scalar Form

21. Write Theorem 11.27 proof missing details for $n = 3$.

22. Write Theorem 11.27 proof missing details for any n.

CHZ Theorem: Vector Form

23. Write Theorem 11.28 proof details for $n = 2$.

24. Write Theorem 11.28 proof details for $n = 3$.

CHZ Identity: Vandermonde

Find matrix $D = \left\langle \vec{\mathbf{d}}_1 | \cdots | \vec{\mathbf{d}}_n \right\rangle$ using Theorems 11.29, 11.31, given $\vec{\mathbf{x}}(0) = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$.

25. $A = \begin{pmatrix} 1 & 0 \\ 2 & 2 \end{pmatrix}$. Ans: $W(0)^T, D =$ $\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$, $\begin{pmatrix} 0 & c_1 \\ 2c_1 + c_2 & -2c_1 \end{pmatrix}$

26. $A = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$. Ans: $W(0)^T, D =$ $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\begin{pmatrix} c_1 & 0 & 0 \\ -2c_1 & 2c_1 + c_2 & 0 \\ 0 & 0 & c_3 \end{pmatrix}$

CHZ and Eigenvectors

Supply details for the following.

27. Find a scalar 3rd order linear differential equation that has e^t, e^{2it}, e^{-2it} as solutions. Apply theorems to conclude that the Wronskian of the exponentials is invertible for every t.

28. Assume $e^{\lambda_1 t}, \ldots, e^{\lambda_n t}$ are independent exponentials . Apply theorems to conclude that the Wronskian of the exponentials is invertible for every t.

29. If $\vec{\mathbf{d}}_1 e^t + \vec{\mathbf{d}}_2 e^{-t} + \vec{\mathbf{d}}_3 e^{2t} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, then $\vec{\mathbf{d}}_1 = \vec{\mathbf{d}}_2 = \vec{\mathbf{d}}_3 = \vec{\mathbf{0}}$.

30. Independence of atoms applied to the n-vector equation $\vec{\mathbf{d}}_1 e^t + \vec{\mathbf{d}}_2 e^{-t} = c_1 \vec{\mathbf{v}}_1 e^t + c_2 \vec{\mathbf{v}}_2 e^{-t}$ implies $\vec{\mathbf{d}}_1 = c_1 \vec{\mathbf{v}}_1$ and $\vec{\mathbf{d}}_2 = c_2 \vec{\mathbf{v}}_2$.

31. There is a 2 × 2 system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ for which CHZ vectors $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2$ are not eigenvectors of A.

32. Let A be the 3 × 3 identity matrix. For $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$, two of the CHZ vectors $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2, \vec{\mathbf{d}}_3$ are zero.

Eigenvectors by Matrix Multiply

Find the eigenvectors of A by Theorem 11.33. Report the choice of $\vec{\mathbf{U}}$.

33. $A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$. Ans: $\vec{\mathbf{U}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

34. $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}$. Ans: $\vec{\mathbf{U}} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$.

CHZ 2×2 Matrix Shortcut Find the general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ using Theorem 11.36.

35. $A = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$, $r = -2, 4$

36. $A = \begin{pmatrix} 1 & 3 \\ -3 & 1 \end{pmatrix}$, $r = 1 \pm 3i$

CHZ Scalar 2×2 Shortcut Find the general solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$ using Theorem 11.35.

37. $A = \begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix}$, $r = -3, 5$

38. $A = \begin{pmatrix} 1 & 4 \\ -4 & 1 \end{pmatrix}$, $r = 1 \pm 4i$

Putzer's 2×2 Spectral Formula Verify the identity.

39. $A = \begin{pmatrix} -1 & 3 \\ -6 & 8 \end{pmatrix}$

$$e^{At} = e^{5t} I + \frac{e^{5t} - e^{2t}}{3} \begin{pmatrix} -6 & 3 \\ -6 & 3 \end{pmatrix}$$

40. $A = \begin{pmatrix} 0 & 1 \\ 6 & 1 \end{pmatrix}$

$$e^{At} = e^{-2t} I + \frac{e^{3t} - e^{-2t}}{5} \begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}$$

41. $A = \begin{pmatrix} 0 & 1 \\ -16 & 8 \end{pmatrix}$

$$e^{At} = e^{4t} I + t e^{4t} \begin{pmatrix} -4 & 1 \\ -16 & 4 \end{pmatrix}$$

42. $A = \begin{pmatrix} 3 & 2 \\ -2 & 3 \end{pmatrix}$, $e^{At} =$

$$e^{3t} \cos(2t) I + e^{3t} \sin(2t) \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}$$

11.6 Jordan Form and Eigenanalysis

Generalized Eigenanalysis

The main result is **Jordan's decomposition**

$$A = PJP^{-1},$$

valid for any real or complex square matrix A. Described here is how to compute the invertible matrix P of generalized eigenvectors and the upper triangular matrix J, called a **Jordan form** of A.

Jordan Block

An $m \times m$ upper triangular matrix $B(\lambda, m)$ is called a **Jordan block** provided all m diagonal elements are the same eigenvalue λ and all super-diagonal elements are one:

$$B(\lambda, m) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \quad (m \times m \text{ matrix})$$

Jordan Form

Given an $n \times n$ matrix A, a **Jordan form** J for A is a block diagonal matrix

$$J = \textbf{diag}(B(\lambda_1, m_1), B(\lambda_2, m_2), \ldots, B(\lambda_k, m_k)),$$

where λ_1, …, λ_k are eigenvalues of A (duplicates possible) and $m_1 + \cdots + m_k = n$. The eigenvalues of J are on the diagonal of J and J has exactly k eigenpairs. If $k < n$, then J is non-diagonalizable. Relation $AP = PJ$ implies A has exactly k eigenpairs and A fails to be diagonalizable for $k < n$.

The relation $A = PJP^{-1}$ is called a **Jordan decomposition** of A. The $n \times n$ matrix P is an augmented matrix of column vectors, i.e., $P = \left\langle \vec{\mathbf{v}}_1 | \ldots | \vec{\mathbf{v}}_n \right\rangle$, which is called the **matrix of generalized eigenvectors** of A. It defines a coordinate system $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$ in which the vector function $\vec{\mathbf{x}} \to A\vec{\mathbf{x}}$ is transformed to the simpler vector function $\vec{\mathbf{y}} \to J\vec{\mathbf{y}}$.

If equal eigenvalues are adjacent in J, then Jordan blocks with equal diagonal entries can be adjacent. Zeros can appear on the super-diagonal of J, because adjacent Jordan blocks join on the super-diagonal with a zero. A complete specification of how to build J from A appears below.

Organizing a Jordan Form

One scheme to organize J first lists distinct eigenvalues low to high $\lambda_1, \ldots, \lambda_k$. Then the Jordan blocks appear in J in that order, with block size high to low for those blocks with the same eigenvalue.

For instance, suppose $\lambda_1 = -4$, $\lambda_1 = 2$, $\lambda_1 = 7$ with respective multiplicities 5, 1 and 3. Then one possible Jordan form is:

$$\begin{aligned} J &= \textbf{diag}(B(\lambda_1, 3), B(\lambda_1, 2), B(\lambda_2, 1), B(\lambda 3, 3)) \\ &= \begin{pmatrix} -4 & 1 & 0 & & & & & & \\ 0 & -4 & 1 & & & & & & \\ 0 & 0 & -4 & & & & & & \\ & & & -4 & 1 & & & & \\ & & & 0 & -4 & & & & \\ & & & & & 2 & & & \\ & & & & & & 7 & 1 & 0 \\ & & & & & & 0 & 7 & 1 \\ & & & & & & 0 & 0 & 7 \end{pmatrix} \end{aligned}$$

Decoding a Jordan Decomposition $A = PJP^{-1}$

If J contains $m \times m$ Jordan block $B(\lambda, m)$, consuming rows 1 to m of J, then $P = \left\langle \vec{\mathbf{v}}_1 | \ldots | \vec{\mathbf{v}}_n \right\rangle$ and $AP = PJ$ implies m vector equations:

$$\begin{aligned} A\vec{\mathbf{v}}_1 &= \lambda\vec{\mathbf{v}}_1, \\ A\vec{\mathbf{v}}_2 &= \lambda\vec{\mathbf{v}}_2 + \vec{\mathbf{v}}_1, \\ \vdots \quad & \vdots \qquad \vdots \\ A\vec{\mathbf{v}}_m &= \lambda\vec{\mathbf{v}}_m + \vec{\mathbf{v}}_{m-1}. \end{aligned}$$

To justify this, start with $AP = PJ$. Expand $AP = \left\langle A\vec{\mathbf{v}}_1 | \ldots | A\vec{\mathbf{v}}_n \right\rangle$ and match its first m columns to those of PJ. This exploded view of the relation $AP = PJ$ according to the Jordan block $B(\lambda, m)$ is called a **Jordan chain**. The formulas can be compacted via matrix $N = A - \lambda I$ into the **Jordan chain relations**

$$(1) \qquad N\vec{\mathbf{v}}_1 = \vec{\mathbf{0}}, \quad N\vec{\mathbf{v}}_2 = \vec{\mathbf{v}}_1, \quad \ldots, \quad N\vec{\mathbf{v}}_m = \vec{\mathbf{v}}_{m-1}.$$

The first vector $\vec{\mathbf{v}}_1$ is an eigenvector. The remaining vectors $\vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_m$ are **not eigenvectors**, they are called **generalized eigenvectors**. Similar formulas can be written for each Jordan block in matrix J. A given eigenvalue may appear multiple times in the chain relations, due to the appearance of two or more Jordan blocks with the same eigenvalue. It is known that the vectors $\{\vec{\mathbf{v}}_i\}|_{i=1}^m$ in a Jordan chain are independent from vectors appearing in a different chain.

Theorem 11.38 (Jordan Decomposition)
Every $n \times n$ matrix A has a Jordan decomposition $A = PJP^{-1}$.
Induction proof on page 907.

Proposition 11.8 (Jordan's Extension) Any $n \times n$ matrix A can be represented in the block triangular form

$$A = PTP^{-1}, \quad T = \mathbf{diag}(T_1, \ldots, T_k),$$

where P is invertible and each matrix T_i is upper triangular with diagonal entries equal to a single eigenvalue of A.

See also Theorem 9.14 page 720. The theorem is proved from the Jordan decomposition theorem by defining $T_i = J_i$, a Jordan Block. A shorter, simpler induction proof exists for Jordan's extension, but the structure of the blocks T_i is unknown with no practical algorithm for their construction.

Geometric and Algebraic Multiplicity

Symbol $\mathbf{GeoMult}(\lambda) = \dim(\mathbf{kernel}(A-\lambda I))$ is called the **geometric multiplicity**. It is defined as the number of basis vectors in a solution to $(A-\lambda I)\vec{\mathbf{x}} = \vec{\mathbf{0}}$, or, equivalently, the number of free variables for this homogeneous problem.

The integer $k = \mathbf{AlgMult}(\lambda)$ is called the **algebraic multiplicity**, defined by the condition that $(r-\lambda)^k$ divides the characteristic polynomial $\det(A-rI)$, but larger powers do not.

Eigenvalue λ is called a **defective eigenvalue** provided inequality $\mathbf{GeoMult}(\lambda) < \mathbf{AlgMult}(\lambda)$ holds. If matrix A has a defective eigenvalue, then is called a **defective matrix**. Defective matrices are not diagonalizable, but they do admit a Jordan decomposition $A = PJP^{-1}$.

Theorem 11.39 (Algebraic and Geometric Multiplicity)
Let A be a square real or complex matrix. Then

$$1 \le \mathbf{GeoMult}(\lambda) \le \mathbf{AlgMult}(\lambda). \tag{2}$$

In addition, there are the following relationships between the Jordan form J and algebraic and geometric multiplicities.

$\mathbf{GeoMult}(\lambda)$	Equals the number of Jordan blocks in J with eigenvalue λ,
$\mathbf{AlgMult}(\lambda)$	Equals the number of times λ is repeated along the diagonal of J.

Proof: Let $d = \mathbf{GeoMult}(\lambda_0)$. Construct a basis $v_1, \ldots, v_n$ of $\mathcal{R}^n$ such that $v_1, \ldots, v_d$ is a basis for $\mathbf{kernel}(A-\lambda_0 I)$. Define $S = \left\langle v_1|\ldots|v_n \right\rangle$ and $B = S^{-1}AS$. The first d columns of AS are $\lambda_0 v_1, \ldots, \lambda_0 v_d$. Then $B = \left(\begin{array}{c|c} \lambda_0 I & C \\ \hline 0 & D \end{array}\right)$ for some matrices C and D. Cofactor expansion implies some polynomial g satisfies

$$\det(A-\lambda I) = \det(S(B-\lambda I)S^{-1}) = \det(B-\lambda I) = (\lambda-\lambda_0)^d g(\lambda)$$

and therefore $d \le \mathbf{AlgMult}(\lambda_0)$. Other details of proof are omitted. ■

Number of Jordan Blocks

Calculation of generalized eigenvectors of matrix A for eigenvalue λ is organized by computing only the Jordan chains of a certain size k. The sizes are found by computing ranks of the powers N^j of the nilpotent matrix $N = A - \lambda I$.

Theorem 11.40 (Number of Jordan Blocks)
Let matrix A have eigenvalue λ. Define $N = A - \lambda I$. Let p be the least integer such that $N^p = N^{p+1}$. Then the number $M(j)$ of Jordan blocks $B(\lambda, j)$ is given by

$$M(j) = \mathbf{rank}(N^{j+1}) + \mathbf{rank}(N^{j-1}) - 2\,\mathbf{rank}(N^j), \quad j = 2, \ldots, p.$$

The proof of the theorem[8] is in the exercises, where more detail appears for $p = 1$ and $p = 2$.

Chains of Generalized Eigenvectors

Given an eigenvalue λ of the matrix A, the topic of generalized eigenanalysis determines all Jordan blocks $B(\lambda, m)$ in J and the corresponding columns of P. The ordering of the blocks in J is not unique. The corresponding columns of P are not unique.

Let $N = A = \lambda I$. Suppose an m-chain is known to exist because of Theorem 11.40, $m \le$ **AlgMult**(λ). How exactly do we find $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_m$ in Jordan chain relations (1)?

A first step might be to write the chain relations (1) in reverse order using a new symbol $\vec{\mathbf{w}}$ that stands for $\vec{\mathbf{v}}_m$:

$$\text{(3)} \qquad \vec{\mathbf{v}}_1 = N^{m-1}\vec{\mathbf{w}}, \quad \ldots, \quad \vec{\mathbf{v}}_{m-1} = N\vec{\mathbf{w}}, \vec{\mathbf{v}}_m = \vec{\mathbf{w}}$$

For instance, if $m = 3$ then the equations are $\vec{\mathbf{v}}_1 = N^2\vec{\mathbf{w}}$, $\vec{\mathbf{v}}_2 = N\vec{\mathbf{w}}$, $\vec{\mathbf{v}}_3 = \vec{\mathbf{w}}$. The impact of (3) is to change the problem of finding an m-chain into finding a suitable vector $\vec{\mathbf{w}}$. Clearly $\vec{\mathbf{w}}$ is not unique. Generally, $\vec{\mathbf{w}}$ is not an eigenvector.

How to Choose Vector $\vec{\mathbf{w}}$

The requirements on $\vec{\mathbf{w}}$ are:

(1) $N^{m-1}\vec{\mathbf{x}} = \vec{\mathbf{w}}$ has no solution $\vec{\mathbf{x}}$.
(2) $N^m\vec{\mathbf{w}} = \vec{\mathbf{0}}$ or $\vec{\mathbf{w}} \in \mathbf{nullspace}(N^m)$
(3) $N^{m-1}\vec{\mathbf{w}} \ne \vec{\mathbf{0}}$ or $\vec{\mathbf{w}} \notin \mathbf{nullspace}(N^{m-1})$

[8] Jordan matrix. Encyclopedia of Mathematics. URL:
`https://encyclopediaofmath.org/index.php?title=Jordan_matrix&oldid=17628`
An equivalent formula is $M(j) = 2\,\mathbf{nullity}(N^j) - \mathbf{nullity}(N^{j+1}) - \mathbf{nullity}(N^{j-1})$.

Proposition 11.9 (Choosing Vector $\vec{\mathbf{w}}$)
Let $\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$ belong to the nullspace of $(N^{m-1})^T$. Then $N^{m-1}\vec{\mathbf{x}} = \vec{\mathbf{w}}$ has no solution $\vec{\mathbf{x}}$.

Proof: Assume a solution $\vec{\mathbf{x}}$ exists. Let $B = N^{m-1}$ and $S = \textbf{nullspace}(B^T)$. Given: $\vec{\mathbf{w}} \in S$. Equation $B\vec{\mathbf{x}} = \vec{\mathbf{w}}$ implies $\vec{\mathbf{w}} \in \textbf{Image}(B)$. The Fundamental Theorem of Linear Algebra (FTLA) gives $\textbf{Image}(B) = \textbf{nullspace}((B^T)^{\perp} = S^{\perp}$. Then $\vec{\mathbf{w}} \in S \cap S^{\perp}$. The intersection of S and $S^{\perp}$ is the zero vector. Then $\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$ and $\vec{\mathbf{w}} = \vec{\mathbf{0}}$, a contradiction. ■

Because of the **chain relations** of equation (1) the very first vector $\vec{\mathbf{v}}_1$ of the chain is an eigenvector: $(A - \lambda I)\vec{\mathbf{v}}_1 = \vec{\mathbf{0}}$. The others $\vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_k$ are not eigenvectors.

Table 2. Shortcut: How to Choose $\vec{\mathbf{w}}$

1. Let $B = (N^{m-1})^T$. Choose a nonzero vector $\vec{\mathbf{w}}$ in the nullspace of B which also satisfies $N^m\vec{\mathbf{w}} = \vec{\mathbf{0}}$. See Proposition 11.9.
2. Require vector $\vec{\mathbf{w}}$ to satisfy $B\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$, it is not in the nullspace of N^{m-1}.

Jordan Decomposition using `maple`

The matrix

$$A = \begin{pmatrix} 4 & -2 & 5 \\ -2 & 4 & -3 \\ 0 & 0 & 2 \end{pmatrix}$$

has a Jordan decomposition

$$A = PJP^{-1} = \begin{pmatrix} 1 & 4 & -7 \\ -1 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 6 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 1 & -\frac{7}{4} \\ -\frac{1}{4} & 1 & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}$$

```
# Maple, Find Jordan Form of matrix A
 A := Matrix([[4, -2, 5], [-2, 4, -3], [0, 0, 2]]);
 factor(LinearAlgebra[CharacteristicPolynomial](A,lambda));
 # Answer == (lambda-6)*(lambda-2)^2
 J,P:=LinearAlgebra[JordanForm](A,output=[’J’,’Q’]);
 zero:=A.P-P.J; # zero matrix expected
```

The `maple` algorithm for the Jordan Form employs the Frobenius Normal Form, which in 2022 differs from Wikipedia and Wolfram references in the ordering of the diagonal blocks. Expect `maple` and `mathematica` to deliver Jordan forms for a given matrix A with different ordering of Jordan blocks.

Examples: Jordan Form and m-Chain

Calculation of generalized eigenvectors of matrix A for eigenvalue λ is organized by computing only the Jordan chains of a certain size k. The sizes are found by rank computation of the powers N^j of the nilpotent matrix $N = A - \lambda I$.

Example 11.11 (Number of Jordan Blocks)
Let A be the 5×5 matrix in equation (4), which has one eigenvalue $\lambda = 2$ of multiplicity 5. Verify that a Jordan form of A has two Jordan blocks, one of size 2 and one of size 3, e.g., $J = \textbf{diag}(B(\lambda, 3), B(\lambda, 2))$.

$$A = \begin{pmatrix} 3 & -1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 \\ 1 & -1 & 2 & 1 & 0 \\ -1 & 1 & 0 & 2 & 1 \\ -3 & 3 & 0 & -2 & 3 \end{pmatrix} \tag{4}$$

Details.
First form the nilpotent matrix $N = A - \lambda I$, then compute N^2 and N^3:

$$N = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 2 & -2 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \\ -3 & 3 & 0 & -2 & 1 \end{pmatrix}, \quad N^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -2 & 2 & 0 & -1 & 1 \\ -2 & 2 & 0 & -1 & 1 \\ -2 & 2 & 0 & -1 & 1 \\ 2 & -2 & 0 & 1 & -1 \end{pmatrix}, \quad N^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Computer assist finds $\textbf{rank}(N) = 3$ and $\textbf{rank}(N^2) = 2$. Identity $N^3 = 0$ implies nilpotency $p = 3$.

Theorem 11.40 applied to Jordan block $B(\lambda, j)$ provides the equation $M(j) = \textbf{rank}(N^{j+1}) + \textbf{rank}(N^{j-1}) - 2\,\textbf{rank}(N^j), \quad j = 2, \ldots, p$. Then $M(1) = 0, M(2) = 1, M(3) = 1, M(4) = M(5) = 0$. There are only two Jordan blocks, size 2 and 3. One possible Jordan form:

$$J = \textbf{diag}(B(\lambda, 3), B(\lambda, 2)) = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

```
with(LinearAlgebra):
getBlockCounts:=proc(A,lambda) local m,N,j,r,p,txt;
 m:=RowDimension(A);
 N:=A-lambda*IdentityMatrix(m);
 for j from 1 to m do r[j]:=Rank(N^j); od:
 for p from m to 2 by -1 do
 if(r[p]<>r[p-1])then break;fi:od;
 printf("lambda=%d, nilpotency=p=%d\n",lambda,p);
 txt:=(j,x)-> printf("Blocks B(%a,%d):%d\n",lambda,j,x):
 for j from p to 2 by -1 do txt(j,-2*r[j]+r[j-1]+r[j+1]):
 od:
end proc:
#
A := Matrix([[3,-1,1,0,0],[2,0,1,1,0],[1,-1,2,1,0],
 [-1,1,0,2,1],[-3,3,0,-2,3]]);
getBlockCounts(A,2);
```

The results: $\lambda = 2$, nilpotency=3, Blocks $B(2,3) : 1$, Blocks $B(2,2) : 1$.

The maple answer for J is obtained by the single line JordanForm(A). Also possible is JordanForm(A,output=['J','Q']) to print J and Q for identity $AQ = QJ$. The maple answers:

$$(5) \qquad J = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}, \quad Q = \frac{1}{2}\begin{pmatrix} 0 & 1 & 2 & -1 & 0 \\ -4 & 2 & 2 & -2 & 2 \\ -4 & 1 & 1 & -1 & 1 \\ -4 & -3 & 1 & -1 & 1 \\ 4 & -5 & -3 & 1 & -3 \end{pmatrix}$$

Example 11.12 (Generalized Eigenvectors)
Let A be the 5×5 matrix in equation (4), which has one eigenvalue $\lambda = 2$ of multiplicity 5. Find the generalized eigenvectors of A as columns of a matrix P, verifying the answer satisfies $AP = PJ$.

Details: Duplicate matrices A, $N = A - 2I$ and J from the preceding example:

$$A = \begin{pmatrix} 3 & -1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 \\ 1 & -1 & 2 & 1 & 0 \\ -1 & 1 & 0 & 2 & 1 \\ -3 & 3 & 0 & -2 & 3 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 2 & -2 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \\ -3 & 3 & 0 & -2 & 1 \end{pmatrix}, \quad J = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

Jordan form J shows there is a 3-chain and a 2-chain of generalized eigenvectors for eigenvalue $\lambda = 2$. We will find the two chains.

The 3-chain. The plan is to find a vector $\vec{\mathbf{w}}$ with $N^3\vec{\mathbf{w}} = \vec{\mathbf{0}}$, $N^2\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$ and $N^2\vec{\mathbf{x}} = \vec{\mathbf{w}}$ has no solution $\vec{\mathbf{x}}$. Then $\vec{\mathbf{v}}_1 = N^2\vec{\mathbf{w}}$, $\vec{\mathbf{v}}_2 = N\vec{\mathbf{w}}$, $\vec{\mathbf{v}}_3 = \vec{\mathbf{w}}$ are the columns of P corresponding to Jordan block $B(\lambda, 3)$, to wit: columns 1,2,3 of P. Computer assist provides

$$N = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 2 & -2 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \\ -3 & 3 & 0 & -2 & 1 \end{pmatrix}, N^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -2 & 2 & 0 & -1 & 1 \\ -2 & 2 & 0 & -1 & 1 \\ -2 & 2 & 0 & -1 & 1 \\ 2 & -2 & 0 & 1 & -1 \end{pmatrix}, N^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We will choose $\vec{\mathbf{w}}$ to be a basis element for the nullspace of $(N^2)^T$, following Table 2 and Proposition 11.9. This clever choice works because $N^m = 0$. We still have to check $N^2\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$, as in Table 2. Employ `maple` to find the nullspace basis:

$$\mathbf{nullspace}((N^2)^T) = \mathbf{span}\left\{\begin{pmatrix}0\\1\\0\\0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\-1\\1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}\right\}$$

Choose vector $\vec{\mathbf{w}}$ to be the last basis vector above, that is, the vector with components $1, 0, 0, 0, 0$. Then (1) equation $N^2\vec{\mathbf{x}} = \vec{\mathbf{w}}$ is insolvable for $\vec{\mathbf{x}}$, (2) $N^2\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$, (3) $N^3\vec{\mathbf{w}} = \vec{\mathbf{0}}$.

Columns 1,2,3 of P will be defined by equations

$$\vec{\mathbf{v}}_1 = N^2\vec{\mathbf{w}} = \begin{pmatrix}0\\-2\\-2\\-2\\2\end{pmatrix}, \quad \vec{\mathbf{v}}_2 = N\vec{\mathbf{w}} = \begin{pmatrix}1\\2\\1\\-1\\-3\end{pmatrix}, \quad \vec{\mathbf{v}}_3 = \vec{\mathbf{w}} = \begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}$$

The computation means that $AP = PJ$[9] where

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 | \vec{\mathbf{0}} | \vec{\mathbf{0}} \right\rangle = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ -2 & 2 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -2 & -1 & 0 & 0 & 0 \\ 2 & -3 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{cases} N\vec{\mathbf{v}}_1 = \vec{\mathbf{0}} \\ N\vec{\mathbf{v}}_2 = \vec{\mathbf{v}}_1 \\ N\vec{\mathbf{v}}_3 = \vec{\mathbf{v}}_2 \end{cases}$$

The 2-chain. Let $m = 2$ (find a 2-chain). The plan is to find a vector $\vec{\mathbf{w}}$ with $N^2\vec{\mathbf{w}} = \vec{\mathbf{0}}$, $N\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$ and $N\vec{\mathbf{x}} = \vec{\mathbf{w}}$ has no solution $\vec{\mathbf{x}}$. Then $\vec{\mathbf{v}}_4 = N\vec{\mathbf{w}}$, $\vec{\mathbf{v}}_5 = \vec{\mathbf{w}}$ are the columns of P corresponding to Jordan block $B(\lambda, 2)$, to wit: columns 4,5 of P.

We will choose $\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$ to be a vector in the nullspace of N^T, following Table 2 and Proposition 11.9. First, find a basis for the nullspace of N^T, as in Proposition 11.9. Then write $\vec{\mathbf{w}}$ in terms of this basis:

$$\mathbf{nullspace}(N^T) = \mathbf{span}\left\{\begin{pmatrix}-2\\2\\0\\-1\\1\end{pmatrix}, \begin{pmatrix}1\\-1\\1\\0\\0\end{pmatrix}\right\},$$

$$\vec{\mathbf{w}} = c_1\begin{pmatrix}-2\\2\\0\\-1\\1\end{pmatrix} + c_2\begin{pmatrix}1\\-1\\1\\0\\0\end{pmatrix}.$$

[9] Zero columns in P allow rapid testing of $AP = PJ$.

Next, we force $\vec{\mathbf{w}}$ to belong to the nullspace of $N^m = N^2$. Equation

$$N^2\vec{\mathbf{w}} = \begin{pmatrix} 0 \\ 10c_1 - 4c_2 \\ 10c_1 - 4c_2 \\ 10c_1 - 4c_2 \\ -10c_1 + 4c_2 \end{pmatrix} = \vec{\mathbf{0}}$$

holds if and only if $5c_1 - 2c_2 = 0$. Choose $c_1 = 2$, $c_2 = 5$ to make it so, then compute

$$\vec{\mathbf{w}} = 2\begin{pmatrix} -2 \\ 2 \\ 0 \\ -1 \\ 1 \end{pmatrix} + 5\begin{pmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 5 \\ -2 \\ 2 \end{pmatrix}, \quad N\vec{\mathbf{w}} = \begin{pmatrix} 7 \\ 7 \\ 0 \\ 0 \\ 0 \end{pmatrix} \ne \vec{\mathbf{0}}$$

Conclusions: (1) equation $N\vec{\mathbf{x}} = \vec{\mathbf{w}}$ is insolvable for $\vec{\mathbf{x}}$, (2) $N\vec{\mathbf{w}} \ne \vec{\mathbf{0}}$ and (3) $N^2\vec{\mathbf{w}} = \vec{\mathbf{0}}$. Define

$$\vec{\mathbf{v}}_4 = N\vec{\mathbf{w}} = \begin{pmatrix} 7 \\ 7 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{\mathbf{v}}_5 = \vec{\mathbf{w}} = \begin{pmatrix} 1 \\ -1 \\ 5 \\ -2 \\ 2 \end{pmatrix}$$

Then

$$P = \left\langle \vec{\mathbf{v}}_1 | \vec{\mathbf{v}}_2 | \vec{\mathbf{v}}_3 | \vec{\mathbf{v}}_4 | \vec{\mathbf{v}}_5 \right\rangle = \begin{pmatrix} 0 & 1 & 1 & 7 & 1 \\ -2 & 2 & 0 & 7 & -1 \\ -2 & 1 & 0 & 0 & 5 \\ -2 & -1 & 0 & 0 & -2 \\ 2 & 3 & 0 & 0 & 2 \end{pmatrix}$$

Matrix multiply verifies $AP = PJ$, which means P is a matrix of generalized eigenvectors for A. The answer for P is not unique, as illustrated by `maple`'s answer in equation (5). ■

Direct Sum Decomposition

The **generalized eigenspace** of eigenvalue λ of an $n \times n$ matrix A is the subspace **kernel**$((A - \lambda I)^p)$ where $p =$ **AlgMult**(λ). We state without proof the main result for generalized eigenspace bases, because details appear in the exercises. A proof is included for the direct sum decomposition, even though Putzer's spectral theory independently produces the same decomposition.

Theorem 11.41 (Generalized Eigenspace Basis)
The subspace **kernel**$((A - \lambda I)^k)$, $k =$ **AlgMult**(λ) has a k-dimensional basis whose vectors are the columns of P corresponding to blocks $B(\lambda, j)$ of J, in Jordan decomposition $A = PJP^{-1}$.

Theorem 11.42 (Direct Sum Decomposition)
Given $n \times n$ matrix A with distinct eigenvalues λ_1, ..., λ_k, let $n_1 = \textbf{AlgMult}(\lambda_i)$, ..., $n_k = \textbf{AlgMult}(\lambda_k)$. Then A induces a direct sum decomposition

$$\mathcal{C}^n = \mathbf{kernel}((A - \lambda_1 I)^{n_1} \oplus \cdots \oplus \mathbf{kernel}((A - \lambda_k I)^{n_k}.$$

This equation means that each complex vector $\vec{\mathbf{x}}$ in $\mathcal{C}^n$ can be uniquely written as

$$\vec{\mathbf{x}} = \vec{\mathbf{x}}_1 + \cdots + \vec{\mathbf{x}}_k$$

where each $\vec{\mathbf{x}}_i$ belongs to $\mathbf{kernel}\,((A - \lambda_i)^{n_i})$, $i = 1, \ldots, k$.

Proof: The previous theorem implies there is a basis of dimension n_i for eigenspace $E_i \equiv \mathbf{kernel}((A - \lambda_i I)^{n_i})$, $i = 1, \ldots, k$. Because $n_1 + \cdots + n_k = n$, then there are n vectors in the union of these bases. The independence test for these n vectors amounts to showing that $\vec{\mathbf{x}}_1 + \cdots + \vec{\mathbf{x}}_k = \vec{\mathbf{0}}$ with $\vec{\mathbf{x}}_i$ in E_i, $i = 1, \ldots, k$, implies all $\vec{\mathbf{x}}_i = \vec{\mathbf{0}}$. This will be true provided $E_i \cap E_j = \{\vec{\mathbf{0}}\}$ for $i \neq j$.

Let's assume a Jordan decomposition $A = PJP^{-1}$. If $\vec{\mathbf{x}}$ is common to both E_i and E_j, then basis expansion of $\vec{\mathbf{x}}$ in both subspaces implies a linear combination of the columns of P is zero, which by independence of the columns of P implies $\vec{\mathbf{x}} = \vec{\mathbf{0}}$. ■

Remark. If A is real with real eigenvalues, then generalized eigenspaces have real bases and the decomposition $\vec{\mathbf{x}} = \vec{\mathbf{x}}_1 + \cdots + \vec{\mathbf{x}}_k$ uses real vectors.

The Real Jordan Form of A

Given a real matrix A, generalized eigenanalysis seeks to find a *real* invertible matrix $\mathcal{P}$ and a *real* upper triangular block matrix $\mathcal{J}$ such that $A = \mathcal{P}\mathcal{J}\mathcal{P}^{-1}$.

If λ is a real eigenvalue of A, then a **real Jordan block** is a matrix

$$B = \textbf{diag}(\lambda, \ldots, \lambda) + N, \quad N = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

If $\lambda = a + ib$ is a complex eigenvalue of A, then symbols λ, 1 and 0 are replaced respectively by 2×2 real matrices $\Lambda = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, $\mathcal{I} = \textbf{diag}(1,1)$ and $\mathcal{O} = \textbf{diag}(0,0)$. The corresponding $2m \times 2m$ real Jordan block matrix is given by the formula

$$B = \textbf{diag}(\Lambda, \ldots, \Lambda) + \mathcal{N}, \quad \mathcal{N} = \begin{pmatrix} \mathcal{O} & \mathcal{I} & \mathcal{O} & \cdots & \mathcal{O} & \mathcal{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \mathcal{O} & \mathcal{I} \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \mathcal{O} & \mathcal{O} \end{pmatrix}.$$

Real Jordan Decomposition

The ideas are best communicated by example. Let

$$A = \begin{pmatrix} -3 & 4 & 1 \\ 0 & -4 & 10 \\ 0 & -5 & 6 \end{pmatrix}.$$

The eigenpairs are

$$\left(-3, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\right), \quad \left(1+5i, \begin{pmatrix} -i \\ 1-i \\ 1 \end{pmatrix}\right), \quad \left(1-5i, \begin{pmatrix} i \\ 1+i \\ 1 \end{pmatrix}\right).$$

The complex Jordan decomposition of matrix A is $A\mathcal{P} = \mathcal{P}\mathcal{J}$ where

$$\mathcal{J} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1+5i & 0 \\ 0 & 0 & 1-5i \end{pmatrix}, \quad \mathcal{P} = \begin{pmatrix} 1 & -i & i \\ 0 & 1-i & 1+i \\ 0 & 1 & 1 \end{pmatrix}$$

The **Real Jordan Decomposition** of matrix A is $AP = PJ$ where

$$J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 5 \\ 0 & -1 & 5 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

The rules:

Replace $\begin{pmatrix} 1+5i & 0 \\ 0 & 1-5i \end{pmatrix}$ by $\begin{pmatrix} 1 & 5 \\ -1 & 5 \end{pmatrix}$

Replace the pair of complex eigenvector columns by the real and imaginary parts of the first eigenvector (the second is not used):

$$\begin{pmatrix} -i \\ 1-i \\ 1 \end{pmatrix}, \begin{pmatrix} i \\ 1+i \\ 1 \end{pmatrix} \text{ replaced by } \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 0 \end{pmatrix}.$$

The method for $n \times n$ real matrices with n eigenpairs is similar. Each pair of complex conjugate eigenvalues $a + ib$, $a - ib$ produces in J a real Jordan block $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$. The corresponding complex eigenvector pair $\vec{\mathbf{u}} + i\vec{\mathbf{v}}$, $\vec{\mathbf{u}} - i\vec{\mathbf{v}}$ is accounted for by inserting into P the real and imaginary parts $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$. A real eigenpair $(\lambda, \vec{\mathbf{x}})$ creates λ on the diagonal of J and real eigenvector $\vec{\mathbf{x}}$ is copied to the corresponding column of P.

Computing Real Exponential Matrices

Discussed here are methods for finding a real exponential matrix e^{At} when A is real. Two formulas are given, one for a real eigenvalue and one for a complex eigenvalue. These formulas supplement the spectral formulas given earlier in the text.

Nilpotent Matrices

A matrix N which satisfies $N^p = 0$ for some integer p is called **nilpotent**. The least integer p for which $N^p = 0$ is called the **nilpotency** of N. A nilpotent matrix N has a finite exponential series:

$$e^{Nt} = I + Nt + N^2\frac{t^2}{2!} + \cdots + N^{p-1}\frac{t^{p-1}}{(p-1)!}.$$

If $N = B(\lambda, p) - \lambda I$, then the finite sum has a splendidly simple expression due to $e^{\lambda t I + Nt} = e^{\lambda t}e^{Nt}$. These remarks motivate the following result.

Theorem 11.43 (Exponential of a Jordan Block Matrix)
If λ is real and

$$B = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \quad (p \times p \text{ matrix})$$

then

$$e^{Bt} = e^{\lambda t}\begin{pmatrix} 1 & t & \frac{t^2}{2} & \cdots & \frac{t^{p-2}}{(p-2)!} & \frac{t^{p-1}}{(p-1)!} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The equality also holds if λ is a complex number, in which case both sides of the equation are complex.

Proof: Let matrix $\Phi(t)$ be either the left side or the right side of the matrix equality. A computation shows that $\Phi'(t) = B\Phi(t)$, $\Phi(0) = I$. Apply uniqueness in the Picard-Lindelöf theorem. ■

Real Exponentials for Complex λ

A Jordan decomposition $A = \mathcal{P}J\mathcal{P}^{-1}$ in which A has only real eigenvalues has real generalized eigenvectors appearing as columns in the matrix $\mathcal{P}$, in the order matching Jordan blocks in J. When $\lambda = a + ib$ is complex, $b > 0$, then the real and imaginary parts of each generalized eigenvector are entered pairwise into $\mathcal{P}$; the conjugate eigenvalue $\overline{\lambda} = a - ib$ is skipped. The complex entry along the diagonal of J and the ones on the superdiagonal of J are each changed into a 2×2 matrix under the correspondence

$$a + ib \leftrightarrow \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

The result is a *real* matrix $\mathcal{P}$ and a *real* block upper triangular matrix $\mathcal{J}$ which satisfy $A = \mathcal{P}\mathcal{J}\mathcal{P}^{-1}$.

Theorem 11.44 (Real Block Diagonal Matrix, Eigenvalue $a+ib$)

Let $\Lambda = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, $\mathcal{I} = \mathbf{diag}(1,1)$ and $\mathcal{O} = \mathbf{diag}(0,0)$. Consider a real Jordan block matrix of dimension $2m \times 2m$ given by the formula

$$B = \begin{pmatrix} \Lambda & \mathcal{I} & \mathcal{O} & \cdots & \mathcal{O} & \mathcal{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \Lambda & \mathcal{I} \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \mathcal{O} & \Lambda \end{pmatrix}.$$

If $\mathcal{R} = \begin{pmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{pmatrix}$, then

$$e^{Bt} = e^{at} \begin{pmatrix} \mathcal{R} & t\mathcal{R} & \frac{t^2}{2}\mathcal{R} & \cdots & \frac{t^{m-2}}{(m-2)!}\mathcal{R} & \frac{t^{m-1}}{(m-1)!}\mathcal{R} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \mathcal{R} & t\mathcal{R} \\ \mathcal{O} & \mathcal{O} & \mathcal{O} & \cdots & \mathcal{O} & \mathcal{R} \end{pmatrix}.$$

Proof: Details are similar to the proof of Theorem 11.43. ■

Solving $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

The solution $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}(0)$ must be real if A is real. The real solution can be expressed as $\vec{\mathbf{x}}(t) = \mathcal{P}\vec{\mathbf{y}}(t)$ where $\vec{\mathbf{y}}'(t) = \mathcal{J}\vec{\mathbf{y}}(t)$ and $\mathcal{J}$ is a real Jordan form of A, containing real Jordan blocks $B_1, \ldots, B_k$ down its diagonal. Theorems above provide explicit formulas for the block matrices $e^{B_i t}$ in the relation

$$e^{\mathcal{J}t} = \mathbf{diag}\left(e^{B_1 t}, \ldots, e^{B_k t}\right).$$

The resulting formula

$$\vec{\mathbf{x}}(t) = \mathcal{P}e^{\mathcal{J}t}\mathcal{P}^{-1}\vec{\mathbf{x}}(0)$$

contains only real numbers, real exponentials, plus sine and cosine terms, which are possibly multiplied by polynomials in t.

Numerical Instability

The matrix $A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}$ has two possible Jordan forms

$$J(\varepsilon) = \begin{cases} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} & \varepsilon = 0, \\ \\ \begin{pmatrix} 1+\sqrt{\varepsilon} & 0 \\ 0 & 1-\sqrt{\varepsilon} \end{pmatrix} & \varepsilon > 0. \end{cases}$$

When $\varepsilon \approx 0$, then numerical algorithms become unstable, unable to lock onto the correct Jordan form. Briefly, $\lim_{\varepsilon\to 0} J(\varepsilon) \neq J(0)$.

Details and Proofs

Proof of Theorem 11.38 (Jordan Decomposition) The result holds by default for 1×1 matrices. Assume the result holds for all $k \times k$ matrices, $k < n$. The proof proceeds by induction on n.

The induction assumes, for any $k \times k$ matrix A, that there is a Jordan decomposition $A = PJP^{-1}$. Then the columns of P satisfy Jordan chain relations

$$A\vec{\mathbf{x}}_i^j = \lambda_i \vec{\mathbf{x}}_i^j + \vec{\mathbf{x}}_i^{j-1}, \quad j > 1, \quad A\vec{\mathbf{x}}_i^1 = \lambda_i \vec{\mathbf{x}}_i^1.$$

Conversely, if the Jordan chain relations are satisfied for k independent vectors $\{\vec{\mathbf{x}}_i^j\}$, then the vectors form the columns of an invertible matrix P such that $A = PJP^{-1}$ with J in Jordan form. The induction step centers upon producing the chain relations and proving that the n vectors are independent.

Let B be $n \times n$ and λ_0 an eigenvalue of B. The Jordan chain relations hold for $A = B$ if and only if they hold for $A = B - \lambda_0 I$. Without loss of generality, we can assume 0 is an eigenvalue of B.

Because B has 0 as an eigenvalue, then inequalities $p = \dim(\mathbf{kernel}(B)) > 0$ and $k = \dim(\mathbf{Image}(B)) < n$ hold, with $p + k = n$. If $k = 0$, then $B = \mathbf{0}$, which is a Jordan form, and there is nothing to prove. Assume henceforth p and k positive. Let $S = \left\langle\, \mathbf{col}(B, i_1)| \cdots | \mathbf{col}(B, i_k)\,\right\rangle$ denote the matrix of pivot columns $i_1, \ldots, i_k$ of B. The pivot columns are known to span $\mathbf{Image}(B)$. Let A be the $k \times k$ basis representation matrix defined by the equation $BS = SA$, or equivalently, $B\,\mathbf{col}(S, j) = \sum_{i=1}^k a_{ij}\,\mathbf{col}(S, i)$. The induction hypothesis applied to A implies there is a basis of k-vectors satisfying Jordan chain relations

$$A\vec{\mathbf{x}}_i^j = \lambda_i \vec{\mathbf{x}}_i^j + \vec{\mathbf{x}}_i^{j-1}, \quad j > 1, \quad A\vec{\mathbf{x}}_i^1 = \lambda_i \vec{\mathbf{x}}_i^1.$$

The values λ_i, $i = 1, \ldots, p$, are the distinct eigenvalues of A. Apply S to these equations to obtain for the n-vectors $\vec{\mathbf{y}}_i^j = S\vec{\mathbf{x}}_i^j$ the Jordan chain relations

$$B\vec{\mathbf{y}}_i^j = \lambda_i \vec{\mathbf{y}}_i^j + \vec{\mathbf{y}}_i^{j-1}, \quad j > 1, \quad B\vec{\mathbf{y}}_i^1 = \lambda_i \vec{\mathbf{y}}_i^1.$$

Because S has independent columns and the k-vectors $\vec{\mathbf{x}}_i^j$ are independent, then the n-vectors $\vec{\mathbf{y}}_i^j$ are independent.

The **plan** is to isolate the chains for eigenvalue zero, then extend these chains by one vector. Then 1-chains will be constructed from eigenpairs for eigenvalue zero to make n generalized eigenvectors.

Suppose q values of i satisfy $\lambda_i = 0$. We allow $q = 0$. For simplicity, assume such values i are $i = 1, \ldots, q$. The key formula $\vec{\mathbf{y}}_i^j = S\vec{\mathbf{x}}_i^j$ implies $\vec{\mathbf{y}}_i^j$ is in $\mathbf{Image}(B)$, while $B\vec{\mathbf{y}}_i^1 = \lambda_i \vec{\mathbf{y}}_i^1$ implies $\vec{\mathbf{y}}_1^1, \ldots, \vec{\mathbf{y}}_q^1$ are in $\mathbf{kernel}(B)$. Each eigenvector $\vec{\mathbf{y}}_i^1$ starts a Jordan chain ending in $\vec{\mathbf{y}}_i^{m(i)}$. Then[10] the equation $B\vec{\mathbf{u}} = \vec{\mathbf{y}}_i^{m(i)}$ has an n-vector solution $\vec{\mathbf{u}}$. We label $\vec{\mathbf{u}} = \vec{\mathbf{y}}_i^{m(i)+1}$. Because $\lambda_i = 0$, then $B\vec{\mathbf{u}} = \lambda_i \vec{\mathbf{u}} + \vec{\mathbf{y}}_i^{m(i)}$ results in an extended Jordan chain

$$\begin{array}{lcll}
B\vec{\mathbf{y}}_i^1 & = & \lambda_i \vec{\mathbf{y}}_i^1 & \\
B\vec{\mathbf{y}}_i^2 & = & \lambda_i \vec{\mathbf{y}}_i^2 & + \quad \vec{\mathbf{y}}_i^1 \\
 & \vdots & & \\
B\vec{\mathbf{y}}_i^{m(i)} & = & \lambda_i \vec{\mathbf{y}}_i^{m(i)} & + \quad \vec{\mathbf{y}}_i^{m(i)-1} \\
B\vec{\mathbf{y}}_i^{m(i)+1} & = & \lambda_i \vec{\mathbf{y}}_i^{m(i)+1} & + \quad \vec{\mathbf{y}}_i^{m(i)}
\end{array}$$

[10] The n-vector $\vec{\mathbf{u}}$ is constructed by setting $\vec{\mathbf{u}} = \vec{\mathbf{0}}$, then copy components of k-vector $\vec{\mathbf{x}}_i^{m(i)}$ into pivot locations: $\mathbf{row}(\vec{\mathbf{u}}, i_j) = \mathbf{row}(\vec{\mathbf{x}}_i^{m(i)}, j)$, $j = 1, \ldots, k$.

Extend the independent set $\{\vec{\mathbf{y}}_i^1\}_{i=1}^q$ to a basis of **kernel**(B) by adding $s = n - k - q$ additional independent vectors $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_s$. This basis consists of eigenvectors of B for eigenvalue 0. Then the set of n vectors $\vec{\mathbf{v}}_r$, $\vec{\mathbf{y}}_i^j$ for $1 \le r \le s$, $1 \le i \le p$, $1 \le j \le m(i)+1$ consists of eigenvectors of B and vectors that satisfy Jordan chain relations. These vectors are columns of a matrix $\mathcal{P}$ that satisfies $B\mathcal{P} = \mathcal{P}\mathcal{J}$ where $\mathcal{J}$ is a Jordan form.

To prove $\mathcal{P}$ invertible, assume a linear combination of the columns of $\mathcal{P}$ is zero:

$$\sum_{i=q+1}^{p} \sum_{j=1}^{m(i)} b_i^j \vec{\mathbf{y}}_i^j + \sum_{i=1}^{q} \sum_{j=1}^{m(i)+1} b_i^j \vec{\mathbf{y}}_i^j + \sum_{i=1}^{s} c_i \vec{\mathbf{v}}_i = \vec{\mathbf{0}}.$$

Apply B to this equation. Because $B\vec{\mathbf{w}} = \vec{\mathbf{0}}$ for any $\vec{\mathbf{w}}$ in **kernel**(B), then

$$\sum_{i=q+1}^{p} \sum_{j=1}^{m(i)} b_i^j B\vec{\mathbf{y}}_i^j + \sum_{i=1}^{q} \sum_{j=2}^{m(i)+1} b_i^j B\vec{\mathbf{y}}_i^j = \vec{\mathbf{0}}.$$

The Jordan chain relations imply that the k vectors $B\vec{\mathbf{y}}_i^j$ in the linear combination consist of $\lambda_i \vec{\mathbf{y}}_i^j + \vec{\mathbf{y}}_i^{j-1}$, $\lambda_i \vec{\mathbf{y}}_i^1$, $i = q+1, \ldots, p$, $j = 2, \ldots, m(i)$, plus the vectors $\vec{\mathbf{y}}_i^j$, $1 \le i \le q$, $1 \le j \le m(i)$. Independence of the original k vectors $\{\vec{\mathbf{y}}_i^j\}$ plus $\lambda_i \ne 0$ for $i > q$ implies this new set is independent. Then all coefficients in the linear combination are zero.

The first linear combination then reduces to $\sum_{i=1}^q b_i^1 \vec{\mathbf{y}}_i^1 + \sum_{i=1}^s c_i \vec{\mathbf{v}}_i = \vec{\mathbf{0}}$. Independence of the constructed basis for **kernel**(B) implies $b_i^1 = 0$ for $1 \le i \le q$ and $c_i = 0$ for $1 \le i \le s$. Therefore, the columns of $\mathcal{P}$ are independent. The induction is complete. ■

Exercises 11.6

Jordan block definition. Write out the Jordan form matrix explicitly.

1. **diag**$(B(7,2), B(5,3))$

   Answer: $\begin{pmatrix} 7 & 1 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$

2. **diag**$(B(0,2), B(4,3))$

3. **diag**$(B(-1,1), B(-1,2), B(5,3))$

4. **diag**$(B(1,1), B(5,2), B(5,3))$

Jordan form definition. Which are Jordan forms and which are not? Explain.

5. $\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$

6. $\begin{pmatrix} 5 & 1 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 5 \end{pmatrix}$

7. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 5 & 1 \end{pmatrix}$

8. $\begin{pmatrix} 5 & 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$

Decoding $A = PJP^{-1}$. Decode $A = PJP^{-1}$ in each case, displaying explicitly the Jordan chain relations and their solutions.

9. $A = \begin{pmatrix} 4 & 8 & 0 & 0 & -8 \\ 0 & 4 & 0 & 0 & 0 \\ 2 & 8 & 2 & 0 & -8 \\ 0 & 20 & 0 & 2 & -12 \\ 0 & 8 & 0 & 0 & -4 \end{pmatrix}$,

   $J = \textbf{diag}(-4, 2, 2, 4, 4)$

10. $A = \begin{pmatrix} -4 & -4 & -12 & 12 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ -8 & 4 & -12 & 16 & 0 \\ -8 & 4 & -16 & 20 & 0 \\ 0 & 0 & -4 & 4 & 0 \end{pmatrix}$,

$J = \textsf{diag}(-4, 4, 4, 0, 0)$

Geometric and algebraic multiplicity. Determine $\textsf{GeoMult}(\lambda)$ and $\textsf{AlgMult}(\lambda)$.

11. $A = \begin{pmatrix} 4 & 8 & 0 & 0 & -8 \\ 0 & 4 & 0 & 0 & 0 \\ 2 & 8 & 2 & 0 & -8 \\ 0 & 20 & 0 & 2 & -12 \\ 0 & 8 & 0 & 0 & -4 \end{pmatrix}$, $\lambda = 4$

12. $A = \begin{pmatrix} -4 & -4 & -12 & 12 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ -8 & 4 & -12 & 16 & 0 \\ -8 & 4 & -16 & 20 & 0 \\ 0 & 0 & -4 & 4 & 0 \end{pmatrix}$, $\lambda = 4$

Generalized eigenvectors. Find all generalized eigenvectors and represent $A = PJP^{-1}$. Check the answer in a computer algebra system.

13. $A = \begin{pmatrix} 4 & 8 & 0 & 0 & -8 \\ 0 & 4 & 0 & 0 & 0 \\ 2 & 8 & 2 & 0 & -8 \\ 0 & 20 & 0 & 2 & -12 \\ 0 & 8 & 0 & 0 & -4 \end{pmatrix}$,

Answer: $J = \textsf{diag}(-4, 4, 4, 2, 2)$,

$P = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 & 4 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$

14. $A = \begin{pmatrix} -4 & -4 & -12 & 12 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ -8 & 4 & -12 & 16 & 0 \\ -8 & 4 & -16 & 20 & 0 \\ 0 & 0 & -4 & 4 & 0 \end{pmatrix}$,

Answer: $J = \textsf{diag}(-4, 4, 4, 0, 0)$,

$P = \begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & -1 \\ 1 & -1 & 1 & 0 & 3 \\ 1 & 0 & 1 & 0 & 3 \\ 0 & 2 & 0 & 3 & 0 \end{pmatrix}$

15. $A = \begin{pmatrix} 0 & 2 & -2 & -2 \\ 2 & 0 & -2 & -4 \\ 2 & 2 & -4 & -2 \\ 0 & 0 & 0 & -4 \end{pmatrix}$,

Ans: $J = \textsf{diag}(0, -4, -2, -2)$,

$P = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 1 & 1 & -4 & 0 \\ 1 & 0 & -3 & -1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$

16. $A = \begin{pmatrix} -2 & 2 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$,

Ans: $J = \textsf{diag}(2, 2, B(2, 3))$,

$P = \begin{pmatrix} 1 & 1 & 1 & -2 & 3 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$

17. $A = \begin{pmatrix} 2 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$,

Ans: $J = \textsf{diag}(B(2, 3), B(2, 2))$,

$P = \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 0 & 0 & 2 & 0 & 2 \\ 0 & 2 & 1 & 2 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

18. $A = \begin{pmatrix} 2 & 0 & 0 & 1 & 0 \\ 1 & 3 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$,

Ans: $J = \textsf{diag}(B(2, 4), 2)$,

$P = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

Number of Jordan Blocks. Outlined here is the derivation of

$$s(j) = 2k(j-1) - k(j-2) - k(j).$$

Definitions:

- $s(j)$= number of blocks $B(\lambda, j)$
- $N = A - \lambda I$
- $k(j) = \dim(\mathbf{kernel}(N^j))$
- $L_j = \mathbf{kernel}(N^{j-1})^\perp$ relative to $\mathbf{kernel}(N^j)$
- $\ell(j) = \dim(L_j)$
- p minimizes $\mathbf{kernel}(N^p) = \mathbf{kernel}(N^{p+1})$

19. Verify $k(j) \le k(j+1)$ from

$$\mathbf{kernel}(N^j) \subset \mathbf{kernel}(N^{j+1}).$$

20. Verify the direct sum formula

$$\mathbf{kernel}(N^j) = \mathbf{kernel}(N^{j-1}) \oplus L_j.$$

Then $k(j) = k(j-1) + \ell(j)$.

21. Given $N^m\vec{\mathbf{w}} = \vec{\mathbf{0}}$, $N^{m-1}\vec{\mathbf{w}} \neq \vec{\mathbf{0}}$, define $\vec{\mathbf{v}}_i = N^{m-i}\vec{\mathbf{w}}$, $i = 1, \ldots, m$. Prove $\{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_m\}$ is independent and they satisfy Jordan chain relations $N\vec{\mathbf{v}}_1 = \vec{\mathbf{0}}$, $N\vec{\mathbf{v}}_{i+i} = \vec{\mathbf{v}}_i$.

22. A block $B(\lambda, p)$ corresponds to a Jordan chain $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_p$ constructed from the Jordan decomposition. Use $N^{p-1}\vec{\mathbf{v}}_p = \vec{\mathbf{v}}_1$ and $\mathbf{kernel}(N^p) = \mathbf{kernel}(N^{p+1})$ to show that the number of such blocks $B(\lambda, p)$ is $\ell(p)$. Then for $p > 1$, $s(p) = k(p) - k(p-1)$.

23. Show that $\ell(j-1) - \ell(j)$ is the number of blocks $B(\lambda, j)$ for $2 < j < p$. Then

$$s(j) = 2k(j-1) - k(j) - k(j-2).$$

24. Test the formulas above on the special matrices

$$A = \mathbf{diag}(B(\lambda,1), B(\lambda,1), B(\lambda,1)),$$

$$A = \mathbf{diag}(B(\lambda,1), B(\lambda,2), B(\lambda,3)),$$

$$A = \mathbf{diag}(B(\lambda,1), B(\lambda,3), B(\lambda,3)),$$

$$A = \mathbf{diag}(B(\lambda,1), B(\lambda,1), B(\lambda,3)),$$

Computing Jordan m-chains.

Find the Jordan m-chain formulas for the given eigenvalue. Then solve them to find the generalized eigenvectors.

25. $A = \begin{pmatrix} 1&0&1&0&1 \\ 0&1&0&1&1 \\ 0&0&1&0&0 \\ 0&0&0&1&1 \\ 0&0&0&0&1 \end{pmatrix}$

26. $A = \begin{pmatrix} 2&0&0&1&0 \\ 1&3&-1&0&0 \\ 1&1&1&0&0 \\ 0&0&0&2&1 \\ 0&0&0&0&2 \end{pmatrix}$, $\lambda = 2$

Generalized Eigenspace Basis.

Let A be $n \times n$ with distinct eigenvalues λ_i, $n_i = \mathbf{AlgMult}(\lambda_i)$ and $E_i = \mathbf{kernel}((A - \lambda_i I)^{n_i})$, $i = 1, \ldots, k$. Assume a Jordan decomposition $A = PJP^{-1}$.

27. Let Jordan block $B(\lambda, m)$ appear in J. Prove that a Jordan chain corresponding to this block is a set of m independent columns of P.

28. Let $\mathcal{B}_\lambda$ be the union of all columns of P originating from Jordan chains associated with Jordan blocks $B(\lambda, j)$. Prove that $\mathcal{B}_\lambda$ is an independent set.

29. Verify that $\mathcal{B}_\lambda$ has $\mathbf{AlgMult}(\lambda)$ basis elements.

30. Prove that $E_i = \mathbf{span}\,(\mathcal{B}_{\lambda_i})$ and $\dim(E_i) = n_i$, $i = 1, \ldots, k$.

Direct Sum Decomposition.

31. Let $A = \begin{pmatrix} 2&1&0 \\ 0&2&1 \\ 0&0&2 \end{pmatrix}$. Let $\lambda = 2$. Compute $k = \mathbf{AlgMult}(\lambda)$ and a basis of generalized eigenvectors for the subspace $\mathbf{kernel}((A - \lambda I)^k)$.

32. Let $A = \begin{pmatrix} 2&1&0&0 \\ 0&2&1&0 \\ 0&0&2&0 \\ 0&0&0&4 \end{pmatrix}$, $\vec{\mathbf{y}} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 9 \end{pmatrix}$.

Find $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2$ in decomposition $\vec{\mathbf{y}} = \vec{\mathbf{x}}_1 + \vec{\mathbf{x}}_2$ in Theorem 11.42.

Exponential Matrices.

Compute the exponential matrix e^{At} on paper. Check the answer using `maple`.

33. $A = \begin{pmatrix} 2&0&0 \\ 0&3&0 \\ 0&0&0 \end{pmatrix}$

34. $A = \begin{pmatrix} 2&1&0 \\ 0&2&0 \\ 0&0&4 \end{pmatrix}$

Nilpotent matrices.

Find the nilpotency of N.

35. $N = \begin{pmatrix} 0&1&0 \\ 0&0&0 \\ 0&0&0 \end{pmatrix}$

36. $N = \begin{pmatrix} 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \\ 0&0&0&0&0 \end{pmatrix}$

Real Jordan Decomposition

Find Jordan decomposition $A = PJP^{-1}$ where J and P are real matrices.

37. $A = \begin{pmatrix} -2 & 6 & -1 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix}$. Answer:

$\lambda = -2, 4 \pm i$,

$J = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & -1 & 4 \end{pmatrix}$, $P = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

38. $A = \begin{pmatrix} -31 & -10 & 18 \\ -15 & -5 & 10 \\ -54 & -20 & 32 \end{pmatrix}$. Answer:

$\lambda = -4, \pm 5i$

$J = \begin{pmatrix} -4 & 0 & 0 \\ 0 & 0 & 5 \\ 0 & -5 & 0 \end{pmatrix}$, $P = \begin{pmatrix} 2 & 2 & 0 \\ 0 & 1 & -1 \\ 3 & 4 & 0 \end{pmatrix}$

Solving $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$

Solve for $\vec{\mathbf{x}}$ in the differential equation.

39. $\vec{\mathbf{x}}' = \begin{pmatrix} -2 & 6 & -1 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \vec{\mathbf{x}}$.

40. $\vec{\mathbf{x}}' = \begin{pmatrix} -31 & -10 & 18 \\ -15 & -5 & 10 \\ -54 & -20 & 32 \end{pmatrix} \vec{\mathbf{x}}$.

Numerical Instability

Show directly that Jordan form J of A satisfies $\lim_{\epsilon \to 0+} J(\epsilon) \neq J(0)$.

41. $A = \begin{pmatrix} 1 & 1 \\ \epsilon & 1 \end{pmatrix}$

42. $A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & \epsilon & 1 \\ 0 & 0 & 0 \end{pmatrix}$

11.7 Nonhomogeneous Linear Systems

Variation of Parameters

The **Method of Variation of Parameters** solves a linear nonhomogeneous system

$$\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t).$$

Historically, it is substitution method which solves the nonhomogeneous system using a trial solution of the form

$$\vec{\mathbf{x}}(t) = e^{At}\,\vec{\mathbf{x}}_0(t).$$

The vector function $\vec{\mathbf{x}}_0(t)$ is to be determined. The method is imagined to originate by varying $\vec{\mathbf{x}}_0$ in the general solution $\vec{\mathbf{x}}(t) = e^{At}\,\vec{\mathbf{x}}_0$ of the linear homogenous system $\vec{\mathbf{x}}' = A\vec{\mathbf{x}}$. The names coined from this idea are *variation of parameters* and *variation of constants.*

Modern use of variation of parameters is through a formula, memorized for routine use.

Theorem 11.45 (Variation of Parameters: Constant Linear System)
Let A be a constant $n \times n$ matrix and $\vec{\mathbf{F}}(t)$ a continuous function near $t = t_0$. The unique solution $\vec{\mathbf{x}}(t)$ of the matrix initial value problem

$$\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t), \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0,$$

is given by the **Variation of Parameters formula**

$$\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{x}}_0 + e^{At}\int_{t_0}^{t} e^{-sA}\vec{\mathbf{F}}(s)ds. \tag{1}$$

Proof of Theorem 11.45. Define

$$\vec{\mathbf{u}}(t) = \vec{\mathbf{x}}_0 + \int_{t_0}^{t} e^{-sA}\vec{\mathbf{F}}(s)ds.$$

To show (1) holds, we must verify $\vec{\mathbf{x}}(t) = e^{At}\vec{\mathbf{u}}(t)$. First, the function $\vec{\mathbf{u}}(t)$ is differentiable with continuous derivative $e^{-tA}\vec{\mathbf{F}}(t)$, by the fundamental theorem of calculus applied to each of its components. The product rule of calculus applies to give

$$\begin{array}{rcl} \vec{\mathbf{x}}'(t) & = & \left(e^{At}\right)'\vec{\mathbf{u}}(t) + e^{At}\vec{\mathbf{u}}'(t) \\ & = & Ae^{At}\vec{\mathbf{u}}(t) + e^{At}e^{-At}\vec{\mathbf{F}}(t) \\ & = & A\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t). \end{array}$$

Therefore, $\vec{\mathbf{x}}(t)$ satisfies the differential equation $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$. Because $\vec{\mathbf{u}}(t_0) = \vec{\mathbf{x}}_0$, then $\vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$, which shows the initial condition is also satisfied. ■

Theorem 11.46 (Variation of Parameters: General Linear System)
Let $A(t)$ be an $n \times n$ matrix and $\vec{\mathbf{F}}(t)$ a vector function, both with continuous entries near $t = t_0$. Let $\Phi(t)$ be the $n \times n$ matrix solution of $\Phi'(t) = A(t)\Phi(t)$, $\Phi(t_0) = I$, established by the Picard-Lindelöf Theorem.

Then the unique solution $\vec{\mathbf{x}}(t)$ of the matrix initial value problem

$$\vec{\mathbf{x}}'(t) = A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t), \quad \vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$$

is given by

$$\vec{\mathbf{x}}(t) = \Phi(t)\vec{\mathbf{x}}_0 + \Phi(t)\int_{t_0}^{t} \Phi^{-1}(s)\vec{\mathbf{F}}(s)ds. \tag{2}$$

Proof of Theorem 11.46. Define

$$\vec{\mathbf{u}}(t) = \vec{\mathbf{x}}_0 + \int_{t_0}^{t} \Phi^{-1}(s)\vec{\mathbf{F}}(s)ds.$$

Equation (2) holds provided $\vec{\mathbf{x}}(t) = \Phi(t)\vec{\mathbf{u}}(t)$. First, the function $\vec{\mathbf{u}}(t)$ is differentiable with continuous derivative $\Phi(t)\vec{\mathbf{F}}(t)$, by the fundamental theorem of calculus applied to each of its components. The product rule of calculus implies

$$\begin{aligned}
\vec{\mathbf{x}}'(t) &= (\Phi(t))'\,\vec{\mathbf{u}}(t) + \Phi(t)\vec{\mathbf{u}}'(t) \\
&= A(t)\Phi(t)\vec{\mathbf{u}}(t) + \Phi(t)\Phi^{-1}(t)\vec{\mathbf{F}}(t) \\
&= A(t)\vec{\mathbf{x}}(t) + \vec{\mathbf{F}}(t).
\end{aligned}$$

Therefore, $\vec{\mathbf{x}}(t)$ satisfies the differential equation $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$. Because $\vec{\mathbf{u}}(t_0) = \vec{\mathbf{x}}_0$, then $\vec{\mathbf{x}}(t_0) = \vec{\mathbf{x}}_0$ and the initial condition is satisfied. ■

Example 11.13 (Variation of Parameters: 2×2 System)

Let $A = \begin{pmatrix} 4 & 0 \\ 0 & 5 \end{pmatrix}$ and $\vec{\mathbf{F}}(t) = e^t \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ using the formula $\vec{\mathbf{x}}_p = \int_0^t e^{A(t-s)}\vec{\mathbf{F}}(s)ds$ and find the shortest expression

$$\vec{\mathbf{x}}_p(t) = \begin{pmatrix} -\frac{2}{3}\,\mathrm{e}^t \\ -\frac{1}{4}\,\mathrm{e}^t \end{pmatrix}$$

Details for Example 11.13: Because A is diagonal, then $e^{At} = \begin{pmatrix} e^{4t} & 0 \\ 0 & e^{5t} \end{pmatrix}$. The integration problem:

$$\begin{aligned}
\vec{\mathbf{x}}_p(t) &= \int_0^t \mathrm{e}^{A(t-s)}\vec{\mathbf{F}}(s)ds \\
&= \int_0^t \begin{pmatrix} \mathrm{e}^{4t-4s} & 0 \\ 0 & \mathrm{e}^{5t-5s} \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} \mathrm{e}^s\,ds \\
&= \begin{pmatrix} \frac{2}{3}\,\mathrm{e}^{4\,t} - \frac{2}{3}\,\mathrm{e}^t \\ \frac{1}{4}\,\mathrm{e}^{5\,t} - \frac{1}{4}\,\mathrm{e}^t \end{pmatrix}
\end{aligned}$$

The integration was by CAS `maple`:

```
with(LinearAlgebra):A:=Matrix([[4,0],[0,5]]);
V:=t->MatrixExponential(A,t);
F:=t->Vector([2*exp(t),1*exp(t)]);
k:=(t,s)->V(t). V(s)^(-1) . F(s);# integrand=k(t,s)
w:=map(u->int(u,s=0..t),k(t,s));
```

Shortening the expression depends on superposition: $\vec{\mathbf{x}} = \vec{\mathbf{x}}_h + \vec{\mathbf{x}}_p$. The homogeneous terms for removal have the form

$$\begin{aligned} \vec{\mathbf{x}}_h(t) &= \begin{pmatrix} e^{4t} & 0 \\ 0 & e^{5t} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \\ &= \begin{pmatrix} c_1 e^{4t} \\ c_2 e^{5t} \end{pmatrix} \end{aligned}$$

Choose $c_1 = -\frac{2}{3}$, $c_2 = -\frac{1}{4}$, then add this $\vec{\mathbf{x}}_h(t)$ to the integration result:

$$\vec{\mathbf{x}}_p(t) = \begin{pmatrix} -\frac{2}{3}\,\mathrm{e}^t \\ -\frac{1}{4}\,\mathrm{e}^t \end{pmatrix}$$

Theorem 11.47 (Variation of Parameters: Scalar 2nd Order)

Let $a \neq 0$, b, c, f be continuous functions defined near $t = t_0$. Let $x_1(t)$, $x_2(t)$ be two linearly independent solutions of the homogeneous differential equation $a(t)x''(t) + b(t)x'(t) + c(t)x(t) = 0$. Then the unique solution $x_p(t)$ of the second order initial value problem

$$(3) \qquad a(t)x''(t) + b(t)x'(t) + c(t)x(t) = f(t), \quad x(t_0) = 0, \quad x'(t_0) = 0$$

is given by the **Variation of Parameters formula**

$$(4) \qquad x_p(t) = \int_{t_0}^{t} k(t,s)\frac{f(s)}{a(s)}ds, \quad k(t,r) = \frac{\begin{vmatrix} x_1(s) & x_2(s) \\ x_1(t) & x_2(t) \end{vmatrix}}{\begin{vmatrix} x_1(s) & x_2(s) \\ x_1'(s) & x_2'(s) \end{vmatrix}}$$

Proof of Theorem 11.47. Formula (4) is discovered via Theorem 11.46 using the **companion matrix** for scalar equation (3) on 846, which is $A(t) = \dfrac{1}{a(t)}\begin{pmatrix} 0 & 1 \\ -c(t) & -b(t) \end{pmatrix}$, and $\vec{\mathbf{F}}(t) = \dfrac{1}{a(t)}\begin{pmatrix} 0 \\ f(t) \end{pmatrix}$. This proof path is pursued in the exercises. A direct proof will be given which requires fewer background topics.

Verify Solution. To begin, expand $k(t,s) = u_1(s)x_1(t) + u_2(t)x)2(t)$ where $u_1(s) = -x_2(s)/W(s)$, $u_2(s) = x_1(s)/W(s)$ and $W(s) = \begin{vmatrix} x_1(s) & x_2(s) \\ x_1'(s) & x_2'(s) \end{vmatrix}$. Then

$$x_p(t) = x_1(t)\int_{t_0}^{t} u_1(s)\frac{f(s)}{a(s)}ds + x_2(t)\int_{t_0}^{t} u_2(s)\frac{f(s)}{a(s)}ds.$$

Expression $x_p(t)$ is expected to satisfy the differential equation, verified by the following steps.

$$
\begin{array}{lll}
\boxed{1} & \text{LHS} = ax_p'' + bx_p' + cx_p & \text{Define } x_p \text{ by (4)} \\
\boxed{2} & = \int_{t_0}^t (a(t)x_1'' + b(t)x_1' + c(t)x_1)\,\frac{f(s)}{a(s)}ds + & \\
 & \quad \int_{t_0}^t (a(t)x_2'' + b(t)x_2' + c(t)x_2)\,\frac{f(s)}{a(s)}ds + & \\
 & \quad a(t)\frac{f(t)}{a(t)} & \\
\boxed{3} & = 0 + 0 + f(t) & \text{Solution } x_p \text{ verified.}
\end{array}
$$

$\boxed{1}$: Symbol LHS is the left hand side of (3).

$\boxed{2}$: Product rule of calculus and the Fundamental Theorem of Calculus. In particular, due to cancellations:

$$
\begin{aligned}
x_p'(t) &= x_1'(t)\int_{t_0}^t u_1(s)\frac{f(s)}{a(s)}ds + x_2'(t)\int_{t_0}^t u_2(s)\frac{f(s)}{a(s)}ds, \\
x_p''(t) &= x_1''(t)\int_{t_0}^t u_1(s)\frac{f(s)}{a(s)}ds + x_2''(t)\int_{t_0}^t u_2(s)\frac{f(s)}{a(s)}ds + a(t)\frac{f(t)}{a(t)}.
\end{aligned}
$$

$\boxed{3}$: The homogeneous equation has solutions x_1, x_2.

Initial Conditions. Equation $x_p(t_0) = 0$ follows because the integral is taken over a zero-length interval. Equation $x_p'(t_0) = 0$ follows from $\boxed{2}$ details.

Example 11.14 (Scalar 2nd Order Euler Differential Equation)

Solve for the general solution:

$$x^2y'' + 3xy' + y = \ln(x^2), \quad x > 0$$

Details for Example 11.14:
The answer: $y_p(x) = \ln(x^2) - 4$, $y_h(t) = c_1x^{-1} + c_2x^{-1}\ln|x|$, details below.

Undetermined Coefficients is applied after a change of variables $x = e^t$ into the forced constant equation:

$$\frac{d^2y(e^t)}{dt^2} + 2\frac{dy(e^t)}{dt} + y(e^t) = 2t$$

It's characteristic equation is $r^2 + 2r + 1 = 0$. Then undetermined coefficient solution $2t - 4$ implies

$$y(e^t) = c_1e^{-t} + c_2te^{-t} + 2t - 4$$

$$y(x) = c_1\frac{1}{x} + c_2\frac{\ln|x|}{x} + 2\ln|x| - 4$$

Variation of Parameters directly finds $y(x)$ by integration. To use the formulas, change symbols: $x \to t$ and $y \to x$. Then the original differential equation becomes:

$$t^2x''(t) + 3tx'(t) + x(t) = \ln(t^2)$$

Euler differential equation theory finds a basis $x_1(t) = \frac{1}{t}$, $x_2(t) = \frac{\ln|t|}{t}$ for the homogeneous problem $t^2x''(t) + 3tx'(t) + x(t) = 0$. Then $\dfrac{f(s)}{a(s)} = s^{-2}\ln(s^2)$ and

$$k(t,s) = \frac{\begin{vmatrix} \dfrac{1}{s} & \dfrac{\ln|s|}{s} \\ \dfrac{1}{t} & \dfrac{\ln|t|}{t} \end{vmatrix}}{\begin{vmatrix} \dfrac{1}{s} & \dfrac{\ln|s|}{s} \\ \dfrac{-1}{s^2} & \dfrac{1}{s^2} - \dfrac{\ln|s|}{s^2} \end{vmatrix}} = \frac{s^2\ln|t/s|}{t}$$

Choose $t_0 = 1$ in the variation of parameters formula. Then for $t > 0$:

$$\begin{aligned} x_p(t) &= \int_1^t k(t,s)\frac{f(s)}{a(s)}ds \\ &= \int_1^t \left(\frac{\ln|t/s|\ln|s^2|}{t}\right) ds \\ &= \ln(t^2) - 4 + \frac{2\ln|t|}{t} + \frac{4}{t} \end{aligned}$$

The last two terms of x_p are homogeneous solutions, discarded to give the shortest particular solution $x_p(t) = \ln(t^2) - 4$.

Example 11.15 (Nonhomogeneous 2×2 System in CAS `maple`)

Solve $x' = 2x + y + t^2$, $y' = 2x + y$, $x(0) = y(0) = 0$ by computer algebra.

Details for Example 11.15:

```
f:=(x,y)->2*x+y; g:=(x,y)->2*x+y;
F:=t->t^2; G:=t->0;
des:=diff(x(t),t)=f(x(t),y(t))+F(t),
     diff(y(t),t)=g(x(t),y(t))+G(t);
dsolve({des,x(0)=0,y(0)=0},[x(t),y(t)]);
```

The reported answer:

$$x\,(t) = -\frac{2}{9}\,t^2 - \frac{4\,t}{27} + \frac{4\,e^{3\,t}}{81} - \frac{4}{81} + 1/9\,t^3$$

$$y\,(t) = -\frac{2}{9}\,t^3 - 2/9\,t^2 + \frac{4\,e^{3\,t}}{81} - \frac{4\,t}{27} - \frac{4}{81}$$

Undetermined Coefficients

The trial solution method known as the method of undetermined coefficients can be applied to vector-matrix systems $\vec{\mathbf{x}}\,' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}\,(t)$ when the components of $\vec{\mathbf{F}}$ are linear combinations of terms of the form

$$t^k e^{at}\cos(bt) \quad \text{or} \quad t^k e^{at}\sin(bt),$$

called **Euler solution atoms**. It is efficient for exposition to write $\vec{\mathbf{F}}$ in terms of the columns $\vec{\mathbf{e}}_1, \ldots, \vec{\mathbf{e}}_n$ of the $n \times n$ identity matrix I:

$$\vec{\mathbf{F}}(t) = \sum_{j=1}^{n} F_j(t)\vec{\mathbf{e}}_j.$$

Then a particular solution of $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ is given by

$$\vec{\mathbf{x}}(t) = \sum_{j=1}^{n} \vec{\mathbf{x}}_j(t)$$

where $\vec{\mathbf{x}}_j(t)$ for $1 \le j \le n$ is a particular solution of the simpler equation

$$\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t) + f(t)\vec{\mathbf{c}}, \quad f = F_j, \quad \vec{\mathbf{c}} = \vec{\mathbf{e}}_j.$$

A trial solution $\vec{\mathbf{x}}(t)$ for non-homogeneous equation $\vec{\mathbf{x}}'(t) = A\vec{\mathbf{x}}(t) + f(t)\vec{\mathbf{c}}$ can be determined from the following **Initial Trial Solution Rule**:

> Let $f(t)$ be a linear combination of Euler solution atoms. Identify independent Euler atoms $A_j(t)$ whose linear combinations include all derivatives of $f(t)$. The initial trial solution is expression $\vec{\mathbf{x}}(t) = \sum_j A_j(t)\vec{\mathbf{d}}_j$, a linear combination of atoms with undetermined vector coefficients $\left\{\vec{\mathbf{d}}_j\right\}$.

In the scalar case, the trial solution must be modified if it has an Euler solution atom which is a solution to the homogeneous equation. In the vector case, if $f(t)$ is a polynomial, then this *correction rule* for the initial trial solution is avoided by assuming the matrix A is invertible. This assumption means that $r = 0$ is not a root of $\det(A - rI) = 0$, which prevents the homogenous solution from having any polynomial terms.

The method substitutes the initial vector trial solution into the differential equation to find the undetermined coefficients $\left\{\vec{\mathbf{d}}_j\right\}$. The answers $\left\{\vec{\mathbf{d}}_j\right\}$ replaced in the trial solution determine a particular solution to the non-homogeneous vector differential equation.

Example 11.16 (Undetermined Coefficients: Polynomial Solution)

Solve by undetermined coefficients:

$$\frac{d\vec{\mathbf{x}}}{dt} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \vec{\mathbf{x}} + \begin{pmatrix} 1+t \\ t^2 \end{pmatrix}$$

Details Example 11.16:

Solution $\vec{\mathbf{x}}_h$:
Let $A = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$. Find e^{At}:

$$\begin{aligned}
e^{At} &= e^{\lambda_1 t}I + \frac{e^{\lambda_2 t}-e^{\lambda_1 t}}{\lambda_2-\lambda_1}(A-\lambda_1 I) && \text{Putzer's formula page 885.}\\
&= e^t I + \frac{e^{-t}-e^t}{-1-1}(A-I) && \text{Because } \lambda_1 = 1,\ \lambda_2 = -1.\\
&= e^t\begin{pmatrix}1&0\\0&1\end{pmatrix} + \frac{e^t-e^{-t}}{2}\begin{pmatrix}0&2\\0&-2\end{pmatrix} && \text{Because } A = \begin{pmatrix}1&2\\0&-1\end{pmatrix}.\\
&= \begin{pmatrix}e^t & e^t-e^{-t}\\ 0 & e^{-t}\end{pmatrix} && \text{Verified in } \texttt{maple}.
\end{aligned}$$

Then

$$\vec{\mathbf{x}}_h(t) = e^{At}\begin{pmatrix}c_1\\c_2\end{pmatrix} = (c_1+c_2)e^t\begin{pmatrix}1\\0\end{pmatrix} + c_2e^{-t}\begin{pmatrix}-1\\1\end{pmatrix}$$

The constant vectors in $\vec{\mathbf{x}}_h(t)$ are eigenvectors of A. The eigenanalysis method produces an equivalent formula.

Solution $\vec{\mathbf{x}}_p$:

The desired shortest particular solution is $\vec{\mathbf{x}}_p(t) = \begin{pmatrix}-2t^2-t-6\\t^2-2t+2\end{pmatrix}$, obtained by the method of undetermined coefficients.

The forcing term is a vector linear combination of Euler atoms $1, t, t^2$:

$$\vec{\mathbf{F}}(t) = \begin{pmatrix}1+t\\t^2\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix} + t\begin{pmatrix}1\\0\end{pmatrix} + t^2\begin{pmatrix}0\\1\end{pmatrix}$$

Select trial solution[11] $\vec{\mathbf{x}}(t) = \vec{\mathbf{d}}_1 + t\vec{\mathbf{d}}_2 + t^2\vec{\mathbf{d}}_3$. Substitute it into $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$:

$$\begin{aligned}
\vec{\mathbf{d}}_2 + 2t\vec{\mathbf{d}}_3 &= A\vec{\mathbf{d}}_1 + tA\vec{\mathbf{d}}_2 + t^2A\vec{\mathbf{d}}_3 + \vec{\mathbf{F}}(t)\\
\vec{\mathbf{d}}_2 + 2t\vec{\mathbf{d}}_3 &= A\vec{\mathbf{d}}_1 + tA\vec{\mathbf{d}}_2 + t^2A\vec{\mathbf{d}}_3 + \begin{pmatrix}1\\0\end{pmatrix} + t\begin{pmatrix}1\\0\end{pmatrix} + t^2\begin{pmatrix}0\\1\end{pmatrix}
\end{aligned}$$

Collect left on Euler atoms $1, t, t^2$:

$$\begin{aligned}
(1)\left(\vec{\mathbf{d}}_2 - A\vec{\mathbf{d}}_1 - \begin{pmatrix}1\\0\end{pmatrix}\right) &+ (t)\left(2\vec{\mathbf{d}}_3 - A\vec{\mathbf{d}}_2 - \begin{pmatrix}1\\0\end{pmatrix}\right)\\
&+ (t^2)\left(-A\vec{\mathbf{d}}_3 - \begin{pmatrix}0\\1\end{pmatrix}\right) = \vec{\mathbf{0}}
\end{aligned}$$

Independence of Euler atoms implies the vector coefficients are zero:

$$\begin{aligned}
\vec{\mathbf{d}}_2 - A\vec{\mathbf{d}}_1 - \begin{pmatrix}1\\0\end{pmatrix} &= \vec{\mathbf{0}}\\
2\vec{\mathbf{d}}_3 - A\vec{\mathbf{d}}_2 - \begin{pmatrix}1\\0\end{pmatrix} &= \vec{\mathbf{0}}\\
-A\vec{\mathbf{d}}_3 - \begin{pmatrix}0\\1\end{pmatrix} &= \vec{\mathbf{0}}
\end{aligned}$$

[11] Derivatives of $1, t, t^2$ are spanned by $1, t, t^2$.

Let $B = A^{-1}$. Solve as a triangular system, variables reversed:

$$\begin{aligned}\vec{\mathbf{d}}_3 &= -B\begin{pmatrix}0\\1\end{pmatrix} &&= \begin{pmatrix}-2\\1\end{pmatrix}\\ \vec{\mathbf{d}}_2 &= B\left(2\vec{\mathbf{d}}_3 - \begin{pmatrix}1\\0\end{pmatrix}\right) &&= \begin{pmatrix}-1\\-2\end{pmatrix}\\ \vec{\mathbf{d}}_1 &= B\left(\vec{\mathbf{d}}_2 - \begin{pmatrix}1\\0\end{pmatrix}\right) &&= \begin{pmatrix}-6\\2\end{pmatrix}\end{aligned}$$

Replace answers $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2, \vec{\mathbf{d}}_3$ in the trial solution to find particular solution:

$$\vec{\mathbf{x}}_p(t) = \begin{pmatrix}-6\\2\end{pmatrix} + t\begin{pmatrix}-1\\-2\end{pmatrix} + t^2\begin{pmatrix}-2\\1\end{pmatrix}$$

Example 11.17 (Undetermined Coefficients: Polynomial-Exponential)

Solve by undetermined coefficients:

$$\frac{d\vec{\mathbf{x}}}{dt} = \begin{pmatrix}1 & 2\\0 & -1\end{pmatrix}\vec{\mathbf{x}} + e^{2t}\begin{pmatrix}t\\3\end{pmatrix}$$

Details Example 11.17:

Solution $\vec{\mathbf{x}}_h$:

Let $A = \begin{pmatrix}1 & 2\\0 & -1\end{pmatrix}$. The homogenous solution from Example 11.16:

$$\vec{\mathbf{x}}_h(t) = e^{At}\begin{pmatrix}c_1\\c_2\end{pmatrix} = (c_1 + c_2)e^t\begin{pmatrix}1\\0\end{pmatrix} + c_2e^{-t}\begin{pmatrix}-1\\1\end{pmatrix}$$

Solution $\vec{\mathbf{x}}_p$:

The desired shortest particular solution is $\vec{\mathbf{x}}_p(t) = \begin{pmatrix}e^{2t} + te^{2t}\\e^{2t}\end{pmatrix}$, obtained by the method of undetermined coefficients.

The forcing term is a vector linear combination of Euler atoms e^{2t}, te^{2t}:

$$\vec{\mathbf{F}}(t) = \begin{pmatrix}te^{2t}\\3e^{2t}\end{pmatrix} = e^{2t}\begin{pmatrix}0\\3\end{pmatrix} + te^{2t}\begin{pmatrix}1\\0\end{pmatrix}$$

Select trial solution $\vec{\mathbf{x}}(t) = e^{2t}\vec{\mathbf{d}}_1 + te^{2t}\vec{\mathbf{d}}_2$.[12] Substitute it into $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$:

$$2e^{2t}\vec{\mathbf{d}}_1 + e^{2t}\vec{\mathbf{d}}_2 + 2te^{2t}\vec{\mathbf{d}}_2 = e^{2t}A\vec{\mathbf{d}}_1 + te^{2t}A\vec{\mathbf{d}}_2 + e^{2t}\begin{pmatrix}0\\3\end{pmatrix} + te^{2t}\begin{pmatrix}1\\0\end{pmatrix}$$

Cancel e^{2t}. Then collect left on Euler atoms $1, t$:

$$(1)\left(2\vec{\mathbf{d}}_1 - A\vec{\mathbf{d}}_1 - \begin{pmatrix}0\\3\end{pmatrix} + \vec{\mathbf{d}}_2\right) + (t)\left(2\vec{\mathbf{d}}_2 - A\vec{\mathbf{d}}_2 - \begin{pmatrix}1\\0\end{pmatrix}\right) = \vec{\mathbf{0}}$$

[12] Derivatives of e^{2t}, te^{2t} are spanned by e^{2t}, te^{2t}.

Independence of Euler atoms implies the vector coefficients are zero:

$$\begin{aligned} 2\vec{\mathbf{d}}_1 - A\vec{\mathbf{d}}_1 - \begin{pmatrix}0\\3\end{pmatrix} + \vec{\mathbf{d}}_2 &= \vec{\mathbf{0}} \\ 2\vec{\mathbf{d}}_2 - A\vec{\mathbf{d}}_2 - \begin{pmatrix}1\\0\end{pmatrix} \qquad &= \vec{\mathbf{0}} \end{aligned}$$

Factor $2I - A$ from each equation. Let $B = (2I - A)^{-1}$. Solve as a triangular system, variables reversed:

$$\begin{aligned} \vec{\mathbf{d}}_2 &= B\begin{pmatrix}1\\0\end{pmatrix} &&= \begin{pmatrix}1\\0\end{pmatrix} \\ \vec{\mathbf{d}}_1 &= B\left(\begin{pmatrix}0\\3\end{pmatrix} - \vec{\mathbf{d}}_2\right) &&= \begin{pmatrix}1\\1\end{pmatrix} \end{aligned}$$

Replace $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2$ in the trial solution to find particular solution

$$\vec{\mathbf{x}}_p(t) = e^{2t}\begin{pmatrix}1\\1\end{pmatrix} + te^{2t}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}e^{2t} + te^{2t}\\ e^{2t}\end{pmatrix}$$

There are nuances in the algorithm not revealed in the preceding two examples. Two theorems formalize the methods.

Theorem 11.48 (Polynomial Solutions)
Let $f(t) = \sum_{j=0}^{k} p_j \frac{t^j}{j!}$ be a polynomial of degree k. Assume A is an $n \times n$ constant invertible matrix. Then $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + f(t)\vec{\mathbf{c}}$ has a polynomial solution $\vec{\mathbf{u}}(t) = \sum_{j=0}^{k} \vec{\mathbf{d}}_j \frac{t^j}{j!}$ of degree k with vector coefficients $\left\{\vec{\mathbf{d}}_j\right\}$ given by the relations

$$\vec{\mathbf{d}}_j = -\sum_{i=j}^{k} p_i A^{j-i-1}\vec{\mathbf{c}}, \quad 0 \le j \le k.$$

Theorem 11.49 (Polynomial × Exponential Solutions)
Let $g(t) = \sum_{j=0}^{k} p_j \frac{t^j}{j!}$ be a polynomial of degree k. Assume A is an $n \times n$ constant matrix and $B = A - aI$ is invertible. Then $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + e^{at}g(t)\vec{\mathbf{c}}$ has a polynomial-exponential solution $\vec{\mathbf{u}}(t) = e^{at}\sum_{j=0}^{k} \vec{\mathbf{d}}_j \frac{t^j}{j!}$ with vector coefficients $\left\{\vec{\mathbf{d}}_j\right\}$ given by the relations

$$\vec{\mathbf{d}}_j = -\sum_{i=j}^{k} p_i B^{j-i-1}\vec{\mathbf{c}}, \quad 0 \le j \le k.$$

Proof of Theorem 11.48. Substitute $\vec{\mathbf{u}}(t) = \sum_{j=0}^{k} \vec{\mathbf{d}}_j \frac{t^j}{j!}$ into the differential equation, then

$$\sum_{j=0}^{k-1} \vec{\mathbf{d}}_{j+1}\frac{t^j}{j!} = A\sum_{j=0}^{k} \vec{\mathbf{d}}_j\frac{t^j}{j!} + \sum_{j=0}^{k} p_j\frac{t^j}{j!}\vec{\mathbf{c}}.$$

Terms on the right for $j = k$ must add to zero and the others must match the left side coefficients of $t^j/j!$, giving the relations

$$A\vec{\mathbf{d}}_k + p_k\vec{\mathbf{c}} = \vec{\mathbf{0}}, \quad \vec{\mathbf{d}}_{j+1} = A\vec{\mathbf{d}}_j + p_j\vec{\mathbf{c}}.$$

Solve the relations recursively to give the formulas

$$\begin{aligned}\vec{\mathbf{d}}_k &= -p_k A^{-1}\vec{\mathbf{c}},\\ \vec{\mathbf{d}}_{k-1} &= -\left(p_{k-1}A^{-1}+p_kA^{-2}\right)\vec{\mathbf{c}},\\ &\vdots\\ \vec{\mathbf{d}}_0 &= -\left(p_0A^{-1}+\cdots+p_kA^{-k-1}\right)\vec{\mathbf{c}}.\end{aligned}$$

The relations above can be summarized by the formula

$$\vec{\mathbf{d}}_j = -\sum_{i=j}^{k} p_i A^{j-i-1}\vec{\mathbf{c}}, \quad 0 \le j \le k.$$

The calculation shows that if $\vec{\mathbf{u}}(t) = \sum_{j=0}^{k}\vec{\mathbf{d}}_j \frac{t^j}{j!}$ and $\vec{\mathbf{d}}_j$ is given by the last formula, then $\vec{\mathbf{u}}(t)$ substituted into the differential equation gives matching LHS and RHS. ■

Proof of Theorem 11.49. Let $\vec{\mathbf{u}}(t) = e^{at}\vec{\mathbf{v}}(t)$. Then $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + e^{at}g(t)\vec{\mathbf{c}}$ implies $\vec{\mathbf{v}}' = (A-aI)\vec{\mathbf{v}} + g(t)\vec{\mathbf{c}}$. Apply Theorem 11.48 to $\vec{\mathbf{v}}' = B\vec{\mathbf{v}} + g(t)\vec{\mathbf{c}}$. ■

Exercises 11.7

Variation of Parameters

Let $A(t) = \begin{pmatrix} 0 & 1 \\ -c(t)/a(t) & -b(t)/a(t)\end{pmatrix}$, $\vec{\mathbf{F}}(t) = \frac{1}{a(t)}\begin{pmatrix}0\\ f(t)\end{pmatrix}$, $\vec{\mathbf{x}} = \begin{pmatrix}u(t)\\ u'(t)\end{pmatrix}$.

1. Verify equivalence of $a(t)u'' + b(t)u' + c(t)u = f(t)$ and $\vec{\mathbf{x}}' = A(t)\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$.

2. For $u'' + 100u = \sin(t)$, find $A(t)$ and $\vec{\mathbf{F}}(t)$.

3. For $u'' = f(t)$, find $A(t)$ and $\vec{\mathbf{F}}(t)$.

4. For $u'' = f(t)$, let $u_1 = 1$, $u_2 = t$, $\Phi(t) = \begin{pmatrix}u_1 & u_2\\ u_1' & u_2'\end{pmatrix}$. Verify $|\Phi(t)| = 1$, then find $A(t) = \Phi'(t)\Phi^{-1}(t)$.

5. State Theorem 11.46 for $n = 2$, then explain how it applies to this special case.

6. Prove Theorem 11.47 using the previous exercise.

Variation of Parameters: Scalar 2nd Order

Let $a(t)u'' + b(t)u' + c(t)u = 0$ have two independent solutions u_1, u_2.

Define $\Psi(t) = \begin{pmatrix}u_1 & u_2\\ u_1' & u_2'\end{pmatrix}$. Then:

7. Matrix $\Psi(t)$ has an inverse.

8. Matrix $\Phi(t) = \Psi(t)\Psi^{-1}(t_0)$ is invertible and $\Phi(t_0) = I$.

9. Let $\Psi(t) = \begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}$. Define $\begin{pmatrix}u\\ v\end{pmatrix} = \Psi(t)\int_0^t \Psi^{-1}(s)f(s)ds$. Then u is a particular solution of $u'' = f(t)$.

10. Let $\Psi(t) = \begin{pmatrix}e^t & e^{-t}\\ e^t & -e^{-t}\end{pmatrix}$. Define $\begin{pmatrix}u\\ v\end{pmatrix} = \Psi(t)\int_0^t \Psi^{-1}(s)f(s)ds$. Then u is a particular solution of $u'' - u = f(t)$.

Variation of Parameters

Let $A = \begin{pmatrix}2 & 0\\ 0 & 3\end{pmatrix}$. Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ using $\vec{\mathbf{x}}_p = \int_0^t e^{A(t-s)}\vec{\mathbf{F}}(s)ds$ and computer assist.

11. $\vec{\mathbf{F}}(t) = e^t\begin{pmatrix}1\\ 2\end{pmatrix}$, $\vec{\mathbf{x}}_p = \begin{pmatrix}e^{2t}-e^t\\ e^{3t}-e^t\end{pmatrix}$

12. $\vec{\mathbf{F}}(t) = \begin{pmatrix}e^t\\ e^{-t}\end{pmatrix}$, $\vec{\mathbf{x}}_p = \begin{pmatrix}e^{2t}-e^t\\ \frac{1}{4}e^{3t}-\frac{1}{4}e^{-t}\end{pmatrix}$

Undetermined Coefficients

Let $A = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$. Solve $\vec{\mathbf{x}}'=A\vec{\mathbf{x}}+\vec{\mathbf{F}}(t)$ by undetermined coefficients. Assume $\vec{\mathbf{x}}_h(t)=c_1e^t\begin{pmatrix} 1 \\ 0 \end{pmatrix}+c_2e^{-t}\begin{pmatrix} -1 \\ 1 \end{pmatrix}$

13. $\vec{\mathbf{F}}(t) = e^t\begin{pmatrix} 1 \\ 2 \end{pmatrix}$,

$\vec{\mathbf{x}}_p=\begin{pmatrix} e^{-t}+3te^t-e^t \\ e^t-e^{-t} \end{pmatrix}$

14. $\vec{\mathbf{F}}(t) = 2\begin{pmatrix} \cos t \\ e^t \end{pmatrix}$,

$\vec{\mathbf{x}}_p = \begin{pmatrix} 2te^t+\sin(t)-\cos(t)+e^{-t} \\ e^t-e^{-t} \end{pmatrix}$

Undetermined Coefficients

Let $A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. Solve $\vec{\mathbf{x}}' = A\vec{\mathbf{x}} + \vec{\mathbf{F}}(t)$ by undetermined coefficients. Assume $\vec{\mathbf{x}}_h(t) = \begin{pmatrix} c_1e^{2t} \\ c_2e^{3t} \end{pmatrix}$.

15. $\vec{\mathbf{F}}(t) = e^t\begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\vec{\mathbf{x}}_p = e^t\begin{pmatrix} -1 \\ -1 \end{pmatrix}$

16. $\vec{\mathbf{F}}(t) = 4\begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}$, $\vec{\mathbf{x}}_p = e^{-t}\begin{pmatrix} -4 \\ -1 \end{pmatrix}$

17. $\vec{\mathbf{F}}(t) = 10\begin{pmatrix} \cos t \\ e^t \end{pmatrix}$,

$\vec{\mathbf{x}}_p = \begin{pmatrix} -4\cos(t) + 2\sin(t) \\ -5e^t \end{pmatrix}$

18. $\vec{\mathbf{F}}(t) = 2e^t\begin{pmatrix} \cos t \\ 1 \end{pmatrix}$,

$\vec{\mathbf{x}}_p = e^t\begin{pmatrix} -\cos(t) + \sin(t) \\ -1 \end{pmatrix}$

11.8 Second-order Systems

A model problem for second order systems is the system of three masses coupled by springs studied in section 11.1, equation (6):

$$\begin{aligned} m_1x_1''(t) &= -k_1x_1(t)+k_2[x_2(t)-x_1(t)], \\ m_2x_2''(t) &= -k_2[x_2(t)-x_1(t)]+k_3[x_3(t)-x_2(t)], \\ m_3x_3''(t) &= -k_3[x_3(t)-x_2(t)]-k_4x_3(t). \end{aligned} \tag{1}$$

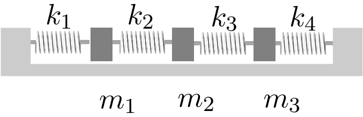


Figure 22. Three masses connected by springs. The masses slide on a frictionless surface.

In vector-matrix form, this system is a **second order system**

$$M\vec{\mathbf{x}}''(t) = K\vec{\mathbf{x}}(t)$$

where the **displacement** $\vec{\mathbf{x}}$, **mass matrix** M and **stiffness matrix** K are defined by the formulas

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \; M = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}, \; K = \begin{pmatrix} -k_1-k_2 & k_2 & 0 \\ k_2 & -k_2-k_3 & k_3 \\ 0 & k_3 & -k_3-k_4 \end{pmatrix}.$$

Because M is invertible, the system can always be written as

$$\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}, \quad A = M^{-1}K.$$

Euler's Substitution $\vec{x} = e^{\lambda t}\vec{v}$

Fundamental substitution $\vec{\mathbf{x}} = e^{\lambda t}\vec{\mathbf{v}}$ due to L. Euler applies to any vector-matrix differential system.

Euler's substitution $\vec{\mathbf{x}} = e^{\lambda t}\vec{\mathbf{v}}$ is perhaps the premier method for remembering the identities

$\lvert A-\lambda^2 I\rvert = 0$	**Characteristic equation** of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$
$\left(A-r^2I\right)\vec{\mathbf{v}} = \vec{\mathbf{0}}, \quad \vec{\mathbf{v}} \neq \vec{\mathbf{0}}$	**Eigenpair equation**

Theorem 11.50 (Properties of Euler's Substitution $\overset{\to}{\mathbf{x}}= e^{\lambda t}\overset{\to}{\mathbf{v}}$)
Equation $\vec{\mathbf{x}} = e^{rt}\vec{\mathbf{v}}$ defines a nonzero solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ if and only if $(r^2, \vec{\mathbf{v}})$ is an eigenpair of matrix A.

Proof: Assume $\vec{\mathbf{x}} = e^{rt}\vec{\mathbf{v}}$ is a solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$. Substitution gives $r^2e^{rt}\vec{\mathbf{v}} = A\vec{\mathbf{v}}e^{rt}$. Cancel the exponential, then $r^2\vec{\mathbf{v}} = A\vec{\mathbf{v}}$. Linear algebraic homogeneous system $\left(A-r^2I\right)\vec{\mathbf{v}} = \vec{\mathbf{0}}$ has a nonzero solution $\vec{\mathbf{v}}$ if and only if the determinant of coefficients vanishes: $|A-r^2I| = 0$.

Assume $(r^2, \vec{\mathbf{v}})$ is an eigenpair of A. The eigenpair equation: $r^2\vec{\mathbf{v}} = A\vec{\mathbf{v}}$. Multiply by e^{rt}: $r^2e^{rt}\vec{\mathbf{v}} = A\vec{\mathbf{v}}e^{rt}$. Then $\vec{\mathbf{x}} = e^{rt}\vec{\mathbf{v}} \neq \vec{\mathbf{0}}$ is a solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$.

Negative Eigenvalues of A

Suppose $(\lambda^2, \vec{\mathbf{v}})$ is an eigenpair of real $n \times n$ matrix A but λ^2 is negative or zero. What is the Euler solution $\vec{\mathbf{x}} = e^{\lambda t}\vec{\mathbf{v}}$ in this case?

For instance, if $\lambda^2 = -4$, then $\lambda = \pm 2i$. Nonzero eigenvector $\vec{\mathbf{v}}$ has real components, therefore Euler solution $\vec{\mathbf{x}}(t) = e^{\lambda t}\vec{\mathbf{v}}$ is a vector with complex entries: $\vec{\mathbf{x}}(t) = e^{2it}\vec{\mathbf{v}} = \cos(2t)\vec{\mathbf{v}} + i\sin(2t)\vec{\mathbf{v}}$. If A is real, then $\cos(2t)\vec{\mathbf{v}}$ and $\sin(2t)\vec{\mathbf{v}}$ are independent real solutions of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$. Formally, they are n-vectors times Euler solution atoms.

> To each negative root $\lambda = -\omega^2$ of $|A - \lambda I| = 0$ with associated eigenpair $(\lambda, \vec{\mathbf{v}})$ corresponds two independent real solutions $\cos(\omega t)\vec{\mathbf{v}}$ and $\sin(\omega t)\vec{\mathbf{v}}$ to the equation $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$.

Cayley-Hamilton-Ziebur Method for $\vec{x}'' = A\vec{x}$

The theory of Euler solution atoms impacts intuition for second order systems in an essential way. Acronym **CHZ** abbreviates *Cayley-Hamilton-Ziebur*. See page 841 for the history.

Theorem 11.51 (Cayley-Hamilton-Ziebur Structure for $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$)
The solution $\vec{\mathbf{x}}(t)$ of second order equation $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$ is a vector linear combination of Euler solution atoms corresponding to roots of the equation $\det(A - r^2 I) = 0$.

Remarks. The equation $|A - r^2 I| = 0$ is formed by substitution of $\lambda = r^2$ into the eigenanalysis characteristic equation $|A - \lambda I| = 0$. In symbols, the structure theorem says $\vec{\mathbf{x}} = \vec{\mathbf{d}}_1 A_1 + \cdots + \vec{\mathbf{d}}_k A_{2n}$, where $A_1, \ldots, A_{2n}$ are Euler solution atoms corresponding to roots r of the determining equation $|A - r^2 I| = 0$. Because Euler solution atoms are real, then all vectors in the relation have real entries. However, only $2n$ arbitrary real constants appear in the $2n^2$ components of $\vec{\mathbf{d}}_1, \ldots, \vec{\mathbf{d}}_{2n}$, the remaining components being dependent on them.

Proof of the CHZ Structure Theorem. Consider the case when A is 2×2 $(n = 2)$, because the proof details are similar in higher dimensions. Expand $|A - \lambda I| = 0$ to find the characteristic equation $\lambda^2 + c\lambda + d = 0$, for some constants c, d. The Cayley-Hamilton theorem says that $A^2 + cA + d\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. Let $\vec{\mathbf{x}}$ be a solution of $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$. Multiply the Cayley-Hamilton identity by vector $\vec{\mathbf{x}}$ and simplify to obtain

$$A^2\vec{\mathbf{x}} + cA\vec{\mathbf{x}} + d\vec{\mathbf{x}} = \vec{\mathbf{0}}.$$

Using equation $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$ backwards, we compute $A^2\vec{\mathbf{x}} = A\vec{\mathbf{x}}'' = \vec{\mathbf{x}}''''$. Replace the terms of the displayed equation to obtain the relation

$$\vec{\mathbf{x}}'''' + c\vec{\mathbf{x}}'' + d\vec{\mathbf{x}} = \vec{\mathbf{0}}.$$

Each component y of vector $\vec{\mathbf{x}}(t)$ then satisfies the 4th order linear homogeneous equation $y^{(4)} + cy^{(2)} + dy = 0$, which has characteristic equation $r^4 + cr^2 + d = 0$. This equation is the expansion of determinant equation $|A-r^2I|=0$. Therefore y is a linear combination of the Euler solution atoms found from the roots of this equation. It follows then that $\vec{\mathbf{x}}(t)$ is a vector linear combination of the Euler solution atoms so identified. ■

Theorem 11.52 (CHZ Method and Negative Eigenvalues)
Assume $n \times n$ matrix A has only negative eigenvalues. Then solution $\vec{\mathbf{x}}(t)$ of second order equation $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$ is a vector linear combination of Euler solution atoms of the form $\cos(\omega t)$, $\sin(\omega t)$, where $|A-\omega^2 I|=0$.

Proof: The result follows from Theorem 11.51, because negative roots of equation $|A-rI|=0$ have the form $r=-\omega^2$ for some positive number ω, which implies the Euler solution atoms for A are of the form $\cos(\omega t)$, $\sin(\omega t)$. ■

Euler Substitution and Solution Atoms

Euler's substitution $\vec{\mathbf{x}} = e^{kt}\vec{\mathbf{v}}$ has limited use for solving $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$. Advantages of the CHZ method will be illustrated.

Illustration 1. Assume A is 2×2 and $|A-\lambda I|=0$ has roots $\lambda = -4, -16$. Then $|A-r^2I|=0$ has four complex roots $\pm 2i, \pm 4i$ and Euler solution atom list $\cos(2t), \cos(4t), \sin(2t), \sin(4t)$. Because eigenvectors $\vec{\mathbf{v}}$ are real, then Euler substitutions are complex: $e^{2it}\vec{\mathbf{v}}$, $e^{-2it}\vec{\mathbf{v}}$, $e^{4it}\vec{\mathbf{v}}$ and $e^{-4it}\vec{\mathbf{v}}$.

The CHZ method is free of complex numbers. In the 2×2 example we have $\vec{\mathbf{x}} = \vec{\mathbf{d}}_1\cos(2t) + \vec{\mathbf{d}}_2\cos(4t) + \vec{\mathbf{d}}_3\sin(2t) + \vec{\mathbf{d}}_4\sin(4t)$, where $\vec{\mathbf{d}}_1$ to $\vec{\mathbf{d}}_4$ are *real* vectors.

Euler's Formula $e^{i\theta} = \cos\theta + i\sin\theta$ allows the switch between complex solutions and real solutions. Euler's substitution $\vec{\mathbf{x}} = e^{2it}\vec{\mathbf{v}}$ is a solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ provided $((2i)^2, \vec{\mathbf{v}})$ is an eigenpair of A. This means $\vec{\mathbf{v}}$ is a real eigenvector for eigenvalue -4 ($A\vec{\mathbf{v}} = -4\vec{\mathbf{v}}$ is required) and therefore $\vec{\mathbf{x}} = e^{2it}\vec{\mathbf{v}}$ is a *complex* solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$.

Illustration 2. Assume A is 2×2 and $|A-\lambda I|=0$ has roots $\lambda = 4, 16$. Then $|A-r^2I|=0$ has four real roots $2, 2, 4, 4$ and Euler atom list $e^{2t}, te^{2t}, e^{4t}, te^{4t}$.

The CHZ method implies the general solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ has the real form $\vec{\mathbf{d}}_1 e^{2t} + \vec{\mathbf{d}}_2 te^{2t} + \vec{\mathbf{d}}_3 e^{4t} + \vec{\mathbf{d}}_4 te^{4t}$.

Euler's substitution produces only two atoms e^{2t}, e^{4t} and we are left with the mystery of how atoms te^{2t}, te^{4t} were discovered to be part of the solution.

Converting $\vec{x}'' = A\vec{x}$ to $\vec{u}' = C\vec{u}$

Given a second order $n\times n$ system $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$, define the variable $\vec{\mathbf{u}}$ and the $2n\times 2n$ block matrix C as follows.

(2)
$$\vec{\mathbf{u}} = \begin{pmatrix} \vec{\mathbf{x}} \\ \vec{\mathbf{x}}' \end{pmatrix}, \quad C = \left(\begin{array}{c|c} \mathbf{0} & I \\ \hline A & \mathbf{0} \end{array}\right).$$

Then each solution $\vec{\mathbf{x}}$ of the second order system $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ produces a corresponding solution $\vec{\mathbf{u}}$ of the first order system $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$. Similarly, each solution $\vec{\mathbf{u}}$ of $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$ gives a solution $\vec{\mathbf{x}}$ of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ by the formula $\vec{\mathbf{x}} = \left\langle I|\mathbf{0}\right\rangle \vec{\mathbf{u}}$.

Characteristic Equation for $\vec{x}'' = A\vec{x}$

The characteristic equation for the $n \times n$ second order system $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ will be derived anew from the corresponding $2n \times 2n$ first order system $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$.

Theorem 11.53 (Characteristic Equation)
Let $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ be given with $n \times n$ constant matrix A. Let

$$\vec{\mathbf{u}} = \begin{pmatrix} \vec{\mathbf{x}}s \\ \vec{\mathbf{x}}' \end{pmatrix}, \quad C = \left(\begin{array}{c|c} \mathbf{0} & I \\ \hline A & \mathbf{0} \end{array}\right).$$

The first order system for $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ is $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$. Then:

$$\det(C - \lambda I) = (-1)^n \det(A - \lambda^2 I). \tag{3}$$

Proof: The method of proof is to verify the product formula

$$\left(\begin{array}{r|r} -\lambda I & I \\ \hline A & -\lambda I \end{array}\right) \left(\begin{array}{r|r} I & \mathbf{0} \\ \hline \lambda I & I \end{array}\right) = \left(\begin{array}{r|r} \mathbf{0} & I \\ \hline A - \lambda^2 I & -\lambda I \end{array}\right).$$

Then the determinant product formula applies to give

$$\det(C - \lambda I) \det \left(\begin{array}{r|r} I & \mathbf{0} \\ \hline \lambda I & I \end{array}\right) = \det \left(\begin{array}{r|r} \mathbf{0} & I \\ \hline A - \lambda^2 I & -\lambda I \end{array}\right). \tag{4}$$

Cofactor expansion is applied to give the two identities

$$\det \left(\begin{array}{r|r} I & \mathbf{0} \\ \hline \lambda I & I \end{array}\right) = 1, \quad \det \left(\begin{array}{r|r} \mathbf{0} & I \\ \hline A - \lambda^2 I & -\lambda I \end{array}\right) = (-1)^n \det(A - \lambda^2 I).$$

Then (4) implies (3). ■

Solving $\vec{u}' = C\vec{u}$ and $\vec{x}'' = A\vec{x}$

Theorem 11.54 (Eigenanalysis of A and C)
Consider the $n \times n$ second order system $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ and its corresponding $2n \times 2n$ first order system $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$ defined by

$$C = \left(\begin{array}{c|c} \mathbf{0} & I \\ \hline A & \mathbf{0} \end{array}\right), \quad \vec{\mathbf{u}} = \begin{pmatrix} \vec{\mathbf{x}} \\ \vec{\mathbf{x}}' \end{pmatrix}. \tag{5}$$

Then $(\lambda, \vec{\mathbf{y}})$ is an eigenpair of C if and only if $(\lambda^2, \vec{\mathbf{w}})$ is an eigenpair of A and $\vec{\mathbf{y}} = \begin{pmatrix} \vec{\mathbf{w}} \\ \lambda\vec{\mathbf{w}} \end{pmatrix}$.

Proof: The equivalent statement

$$(6) \qquad (C-\lambda I)\begin{pmatrix}\vec{\mathbf{w}}\\ \vec{\mathbf{z}}\end{pmatrix}=\vec{\mathbf{0}} \quad \text{if and only if} \quad \left\{\begin{array}{rcl} A\vec{\mathbf{w}} &=& \lambda^2\vec{\mathbf{w}},\\ \vec{\mathbf{z}} &=& \lambda\vec{\mathbf{w}}.\end{array}\right.$$

is proved from $C-\lambda I=\left(\begin{array}{c|c} -\lambda I & I\\ \hline A & -\lambda I\end{array}\right)$ and block multiply. ■

Theorem 11.55 (General Solutions of $\vec{\mathbf{u}}\,' = C\,\vec{\mathbf{u}}$ and $\vec{\mathbf{x}}\,'' = A\,\vec{\mathbf{x}}$)
Let A be a given $n\times n$ constant matrix and define the corresponding $2n\times 2n$ system by

$$\vec{\mathbf{u}}' = C\vec{\mathbf{u}}, \quad C=\left(\begin{array}{c|c} \mathbf{0} & I\\ \hline A & \mathbf{0}\end{array}\right), \quad \vec{\mathbf{u}}=\begin{pmatrix}\vec{\mathbf{x}}\\ \vec{\mathbf{x}}'\end{pmatrix}.$$

Assume C has eigenpairs $\{(\lambda_j,\vec{\mathbf{y}}_j)\}_{j=1}^{2n}$ and $\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_{2n}$ are independent. Let I and $\mathbf{0}$ denote the $n\times n$ identity and zero matrix. Define $\vec{\mathbf{w}}_j = \big\langle I|\mathbf{0}\big\rangle\vec{\mathbf{y}}_j$, $j=1,\ldots,2n$. Then $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$ and $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ have general solutions

$$\begin{array}{rcll} \vec{\mathbf{u}}(t) &=& c_1e^{\lambda_1 t}\vec{\mathbf{y}}_1+\cdots+c_{2n}e^{\lambda_{2n}t}\vec{\mathbf{y}}_{2n} & (2n\times 1),\\ \vec{\mathbf{x}}(t) &=& c_1e^{\lambda_1 t}\vec{\mathbf{w}}_1+\cdots+c_{2n}e^{\lambda_{2n}t}\vec{\mathbf{w}}_{2n} & (n\times 1).\end{array}$$

Proof:
General solution of $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$. Independence of vector Euler solutions $e^{\lambda_1 t}\vec{\mathbf{y}}_1$, ..., $e^{\lambda_{2n}t}\vec{\mathbf{y}}_{2n}$ will be verified. Assume a linear combination of these solutions is zero, then at $t=0$ the exponentials equal 1, which reduces to a linear combination of $\vec{\mathbf{y}}_1$, ..., $\vec{\mathbf{y}}_{2n}$. By independence of the latter, then all weights are zero: the Euler solutions are independent. Hence $\vec{\mathbf{u}}(t)$ is a general solution of $\vec{\mathbf{u}}' = C\vec{\mathbf{u}}$.

General solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$. Independence of vector Euler solution $e^{\lambda_1 t}\vec{\mathbf{w}}_1$, ..., $e^{\lambda_{2n}t}\vec{\mathbf{w}}_{2n}$ will be verified. Suppose constants $a_1,\ldots,a_{2n}$ are given with $\sum_{j=1}^{2n} a_j e^{\lambda_j t}\vec{\mathbf{w}}_j = \vec{\mathbf{0}}$. Replace $t=0$ in this relation to give (1) $\sum_{j=1}^{2n} a_j\vec{\mathbf{w}}_j = \vec{\mathbf{0}}$. Differentiate this relation on variable t to give $\sum_{j=1}^{2n} a_j\lambda_j\, e^{\lambda_j t}\vec{\mathbf{w}}_j = \vec{\mathbf{0}}$ for all t, then set $t=0$ to obtain (2) $\sum_{j=1}^{2n} a_j\lambda_j\,\vec{\mathbf{w}}_j = \vec{\mathbf{0}}$. Combine (1) and (2) using $\vec{\mathbf{y}}_j = \begin{pmatrix}\vec{\mathbf{w}}_j\\ \lambda_j\vec{\mathbf{w}}_j\end{pmatrix}$ from Theorem 11.54 into the vector equation $\sum_{j=1}^{2n} a_j\vec{\mathbf{y}}_j = \vec{\mathbf{0}}$. Independence of $\vec{\mathbf{y}}_1$, ..., $\vec{\mathbf{y}}_{2n}$ implies that the weights are zero: $a_1=\cdots=a_{2n}=0$. ■

Eigenanalysis for Non-positive Eigenvalues

Assume all eigenvalues μ of A are negative or zero. Eigenvalue μ of A is related to an eigenvalue λ of C by the relation $\mu=-\omega^2=\lambda^2$ for some real $\omega\ge 0$. Then $\lambda=\pm\omega i$ and $\omega=\sqrt{|\mu|}$.

Lemma 11.2 (Cosine and Sine Solutions)
Let $(-\omega^2,\vec{\mathbf{v}})$ be an eigenpair of the real $n\times n$ matrix A with $\omega\ge 0$. Define

$$u(t)=\left\{\begin{array}{ll} c_1\cos\omega t + c_2\sin\omega t & \omega>0,\\ c_1+c_2t & \omega=0.\end{array}\right.$$

Then $\vec{\mathbf{x}}(t)=u(t)\vec{\mathbf{v}}$ satisfies $\vec{\mathbf{x}}''(t)=A\vec{\mathbf{x}}(t)$.

Proof:
Then $u''(t) = -\omega^2 u(t)$ (both sides are zero for $\omega = 0$). Vector function $\vec{\mathbf{x}}(t) = u(t)\vec{\mathbf{v}}$ satisfies $\vec{\mathbf{x}}''(t) = -\omega^2\vec{\mathbf{x}}(t)$. Also, $A\vec{\mathbf{x}}(t) = u(t)A\vec{\mathbf{v}} = -\omega^2\vec{\mathbf{x}}(t)$. This proves $\vec{\mathbf{x}}(t) = u(t)\vec{\mathbf{v}}$ satisfies $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$. ■

Theorem 11.56 (Eigenanalysis Solution of $\vec{\mathbf{x}}\,'' = A\,\vec{\mathbf{x}}$)

Let real $n \times n$ matrix A have eigenpairs $\{(\mu_j, \vec{\mathbf{v}}_j)\}_{j=1}^n$. Assume A has distinct eigenvalues $\mu_j = -\omega_j^2$ with $\omega_j \ge 0$, $j = 1, \ldots, n$ and that $\vec{\mathbf{v}}_1$, ..., $\vec{\mathbf{v}}_n$ are linearly independent. Then the general solution of $\vec{\mathbf{x}}''(t) = A\vec{\mathbf{x}}(t)$ is given in terms of $2n$ arbitrary constants a_1, ..., a_n, b_1, ..., b_n by the formula

$$\vec{\mathbf{x}}(t) = \sum_{j=1}^{n}\left(a_j \cos\omega_j t + b_j \frac{\sin\omega_j t}{\omega_j}\right)\vec{\mathbf{v}}_j \tag{7}$$

This expression uses the limit convention $\left.\dfrac{\sin\omega t}{\omega}\right|_{\omega=0} = t$.

Proof:
Lemma 11.2 and superposition establish that $\vec{\mathbf{x}}(t)$ is a solution. It only remains to prove that it is the general solution, meaning that the arbitrary constants can be assigned to allow any possible initial condition $\vec{\mathbf{x}}(0) = \vec{\mathbf{x}}_0$, $\vec{\mathbf{x}}'(0) = \vec{\mathbf{y}}_0$. Define the constants uniquely by the relations

$$\begin{aligned} \vec{\mathbf{x}}_0 &= \textstyle\sum_{j=1}^n a_j\vec{\mathbf{v}}_j, \\ \vec{\mathbf{y}}_0 &= \textstyle\sum_{j=1}^n b_j\vec{\mathbf{v}}_j, \end{aligned}$$

which is possible by the assumed independence of the vectors $\{\vec{\mathbf{v}}_j\}_{j=1}^n$. Then equation (7) implies $\vec{\mathbf{x}}(0) = \sum_{j=1}^n a_j\vec{\mathbf{v}}_j = \vec{\mathbf{x}}_0$ and $\vec{\mathbf{x}}'(0) = \sum_{j=1}^n b_j\vec{\mathbf{v}}_j = \vec{\mathbf{y}}_0$. ■

Why doesn't equation (7) work for duplicate eigenvalues?

Consider $A = \begin{pmatrix} -4 & 0 \\ 0 & -4 \end{pmatrix}$ for which the characteristic equation $|A - r^2 I| = 0$ has duplicate complex roots $\pm 2i, \pm 2i$. Then CHZ predicts real solution $\vec{\mathbf{x}} = \vec{\mathbf{d}}_1\cos(2t) + \vec{\mathbf{d}}_2 t\cos(2t) + \vec{\mathbf{d}}_3\sin(2t) + \vec{\mathbf{d}}_4 t\sin(2t)$ whereas incorrect application of equation (7) would report $\vec{\mathbf{x}} = a_1\vec{\mathbf{v}}_1\cos(2t) + a_2\vec{\mathbf{v}}_2\cos(2t) + b_1\vec{\mathbf{v}}_1\sin(2t) + b_2\vec{\mathbf{v}}_2\sin(2t)$, the symbols $\vec{\mathbf{v}}_j$ being real eigenvectors of A for eigenvalues $-4, -4$.
Euler solution atoms $t\cos(2t), t\sin(2t)$ are missing in equation (7), but maybe the equation is correct anyway? The answer is **NO**, because differentiation across equation (7) on symbols a_1, a_2, b_1, b_2 reveals there are only two independent vector solutions represented, instead of the required four. The conclusion: equation (7) doesn't work for multiple eigenvalues.

Theorem 11.57 (CHZ and Eigenvectors: $\vec{\mathbf{x}}\,'' = A\,\vec{\mathbf{x}}$)

If the hypothesis of Theorem 11.56 holds, then in CHZ solution $\vec{\mathbf{x}} = \sum_{j=1}^{2n}\vec{\mathbf{d}}_j A_j(t)$ each $\vec{\mathbf{d}}_j$ is a scalar multiple of an eigenvector of A.[13]

Proof. Let $\vec{\mathbf{x}}$ be a solution of $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ and represent it in two ways, first by CHZ and second by eigenanalysis:

$$\vec{\mathbf{x}} = \sum_{j=1}^{2n}\vec{\mathbf{d}}_j A_j(t) = \sum_{j=1}^{n}\left(a_j\cos\omega_j t + b_j\frac{\sin\omega_j t}{\omega_j}\right)\vec{\mathbf{v}}_j$$

[13] **Warning**: A vector $\vec{\mathbf{d}}_j$ can be zero: $0\vec{\mathbf{v}}$ is a linear combination of eigenvector $\vec{\mathbf{v}}$.

Assume by re-labeling that the Euler atoms are $A_j(t) = \cos(\omega_j t)$ and $A_{j+n}(t) = \frac{\sin \omega_j t}{\omega_j}$, $1 \le j \le n$. Then $\sum_{j=1}^{2n} \vec{\mathbf{d}}_j A_j(t) = \sum_{j=1}^{n} a_j \vec{\mathbf{v}}_j A_j(t) + b_j \vec{\mathbf{v}}_j A_{j+n}(t)$. Independence of $\{A_j\}_{j=1}^{2n}$ implies vector coefficients of the atoms on each side of the equation must match: each $\vec{\mathbf{d}}_j$ is a scalar multiple of an eigenvector of A. ■

Earthquakes

Reproduced here are earthquake modeling formulas from page 833. The formulas are applied to 5-story buildings using the solution methods of this section.

A horizontal earthquake oscillation $F(t) = F_0 \cos \omega t$ affects each floor of a 5-floor building; see Figure 23. The effect of the earthquake depends upon the natural frequencies of oscillation of the floors.

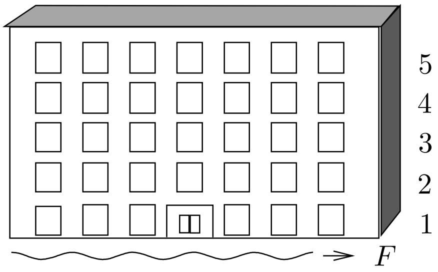


Figure 23. A 5-Floor Building.
A horizontal earthquake wave F affects every floor. A typical wave has wavelength many times larger than the illustration.

Assumptions and Symbols for a 5-Floor Building

- Each floor is considered a point mass located at its center-of-mass. The floors have masses $m_1, \ldots, m_5$.
- Each floor is restored to its equilibrium position by a linear restoring force or Hooke's force $-k$(elongation). The Hooke's constants are $k_1, \ldots, k_5$.
- The locations of masses representing the 5 floors are $x_1, \ldots, x_5$. The equilibrium position is $x_1 = \cdots = x_5 = 0$.
- Damping effects of the floors are ignored: it is a *frictionless* system.

Derivation Details

The differential equations for the model are obtained by **competition**: the Newton's second law force is set equal to the sum of the Hooke's forces and the external force due to the earthquake wave. This results in the following system, where $k_6 = 0$, $E_j = m_j F''$ for $j = 1, 2, 3, 4, 5$ and $F = F_0 \cos \omega t$.

$$\begin{array}{rcl} m_1 x_1'' &=& -(k_1 + k_2)x_1 + k_2 x_2 + E_1, \\ m_2 x_2'' &=& k_2 x_1 - (k_2 + k_3)x_2 + k_3 x_3 + E_2, \\ m_3 x_3'' &=& k_3 x_2 - (k_3 + k_4)x_3 + k_4 x_4 + E_3, \\ m_4 x_4'' &=& k_4 x_3 - (k_4 + k_5)x_4 + k_5 x_5 + E_4, \\ m_5 x_5'' &=& k_5 x_4 - (k_5 + k_6)x_5 + E_5. \end{array}$$

In particular, the equations for a floor depend only upon the neighboring floors. The bottom floor and the top floor are exceptions: they have just one neighboring floor.

Vector-Matrix 2nd Order System
Let:

$$M=\begin{pmatrix} m_1 & 0 & 0 & 0 & 0\\ 0 & m_2 & 0 & 0 & 0\\ 0 & 0 & m_3 & 0 & 0\\ 0 & 0 & 0 & m_4 & 0\\ 0 & 0 & 0 & 0 & m_5 \end{pmatrix}, \quad \vec{\mathbf{x}}=\begin{pmatrix} x_1\\ x_2\\ x_3\\ x_4\\ x_5\end{pmatrix}, \quad \vec{\mathbf{H}}=\begin{pmatrix} E_1\\ E_2\\ E_3\\ E_4\\ E_5\end{pmatrix},$$

$$K=\begin{pmatrix} -k_1-k_2 & k_2 & 0 & 0 & 0\\ k_2 & -k_2-k_3 & k_3 & 0 & 0\\ 0 & k_3 & -k_3-k_4 & k_4 & 0\\ 0 & 0 & k_4 & -k_4-k_5 & k_5\\ 0 & 0 & 0 & k_5 & -k_5-k_6 \end{pmatrix}$$

In the last row, $k_6=0$ reflects the absence of a floor above the fifth floor. The second order system:

$$M\vec{\mathbf{x}}''(t)=K\vec{\mathbf{x}}(t)+\vec{\mathbf{H}}(t)$$

Matrix M is called the **mass matrix** and matrix K is called the **Hooke's matrix**. The **external force** $\vec{\mathbf{H}}(t)$ can be written as a scalar function $E(t)=-F''(t)$ times a constant vector:

$$\vec{\mathbf{H}}(t)=-\omega^2F_0\cos\omega t\begin{pmatrix} m_1\\ m_2\\ m_3\\ m_4\\ m_5\end{pmatrix}.$$

Identical Floors
Assume that all floors have the same mass m and the same Hooke's constant k. Then $M=mI$ and $M\vec{\mathbf{x}}''(t)=K\vec{\mathbf{x}}(t)+\vec{\mathbf{H}}(t)$ becomes:

$$(8)\qquad \vec{\mathbf{x}}''=\frac{1}{m}\begin{pmatrix} -2k & k & 0 & 0 & 0\\ k & -2k & k & 0 & 0\\ 0 & k & -2k & k & 0\\ 0 & 0 & k & -2k & k\\ 0 & 0 & 0 & k & -k \end{pmatrix}\vec{\mathbf{x}}-F_0\omega^2\cos(\omega t)\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}$$

Hooke's matrix K is symmetric ($K^T=K$) with negative entries only on the diagonal. The last diagonal entry is $-k$ (a error to write $-2k$).

Particular Solution: Identical Floors
The method of undetermined coefficients predicts a trial solution $\vec{\mathbf{x}}(t)=\vec{\mathbf{c}}\,\cos\omega t$. Terms $\sin\omega t$ cannot appear in the trial solution because the $\vec{\mathbf{x}}'$ term is absent in equation (8).

Constant vector $\vec{\mathbf{c}}$ will be found by trial solution substitution. Let $\vec{\mathbf{b}}$ equal the column vector of ones in equation (8). Substitute the trial solution $\vec{\mathbf{x}}(t) = \vec{\mathbf{c}}\cos\omega t$ into (8). Cancel the common factor $\cos\omega t$. Then $\left(m^{-1}K + \omega^2 I\right)\vec{\mathbf{c}} = F_0\omega^2\vec{\mathbf{b}}$. Let $B = m^{-1}K + \omega^2 I$. Determinant formula $B^{-1} = \dfrac{\mathbf{adj}(B)}{\det(B)}$ gives:

$$\vec{\mathbf{c}} = F_0\omega^2\frac{\mathbf{adj}(B)}{\det(B)}\vec{\mathbf{b}}$$

Homogeneous Solution

Theorem 11.56 provides:

$$\vec{\mathbf{x}}_h(t) = \sum_{j=1}^{5}(a_j\cos\omega_j t + b_j\sin\omega_j t)\vec{\mathbf{v}}_j$$

where $r = \omega_j$ and $\vec{\mathbf{v}} = \vec{\mathbf{v}}_j \neq \vec{\mathbf{0}}$ satisfy the **eigenpair equation**:

$$\left(\frac{1}{m}K + r^2 I\right)\vec{\mathbf{v}} = \vec{\mathbf{0}}$$

Identical Floors $k/m = 10$

Then:

$$\frac{1}{m}K = \begin{pmatrix} -20 & 10 & 0 & 0 & 0 \\ 10 & -20 & 10 & 0 & 0 \\ 0 & 10 & -20 & 10 & 0 \\ 0 & 0 & 10 & -20 & 10 \\ 0 & 0 & 0 & 10 & -10 \end{pmatrix}$$

Let $B(\omega, k/m) = (1/m)K + \omega^2 I$. Natural frequency values $\omega_1, \ldots, \omega_5$ are found by solving for ω in determinant equation $|B(\omega, 10)| = 0$ to obtain Table 3.

Table 3. Natural Frequencies ω for the Special Case $k/m = 10$.

Frequency	Value
ω_1	0.900078068
ω_2	2.627315231
ω_3	4.141702938
ω_4	5.320554507
ω_5	6.068366391

Identical Floors: General Solution

Superposition provides the general solution $\vec{\mathbf{x}}(t) = \vec{\mathbf{x}}_h(t) + \vec{\mathbf{x}}_p(t)$. If the floors are at rest, then $\vec{\mathbf{x}}_h = \vec{\mathbf{0}}$. Term $\vec{\mathbf{x}}_p$ measures bounded oscillations of the center of mass of each floor due to the incoming earthquake wave.

Identical Floors: Resonance Effects for $k/m = 10$
Special solution $\vec{\mathbf{x}}_p(t)$ can be used to obtain some insight into practical resonance effects between the incoming earthquake wave and movement of the building floors.

Let ω be the incoming wave natural frequency. Solution $\vec{\mathbf{x}}_p$ has components $A_1\cos(\omega t), \ldots, A_5\cos(\omega t)$. Let I have columns $e_1, \ldots, e_5$. The amplitude formula for $1 \le j \le 5$:

$$A_j = e_j^T \vec{\mathbf{c}} \cos(0) = \frac{F_0\omega^2}{|B(\omega, 10)|} e_j^T\, \mathbf{adj}(B(\omega, 10))\vec{\mathbf{b}}$$

The fraction has bounded numerator. Determinant $|B(\omega, 10)|$ in the denominator can be near zero when ω is close to one of the natural frequencies $\omega_1, \ldots, \omega_5$. Then the amplitude of a component of $\vec{\mathbf{x}}_p$ can be very large, which means the floor takes an excursion that is too large to maintain structural integrity.

Physical Interpretation: An earthquake wave of proper frequency, lasting sufficiently long, can demolish a floor and hence demolish the entire building. Small amplitude earthquake waves can initiate destructive oscillation of structures having unlucky natural frequencies.

Coupled Spring-Mass Systems: Derivations

Reproduced here from page 813 are notation and assumptions for three masses attached to each other by four springs as in Figure 14.

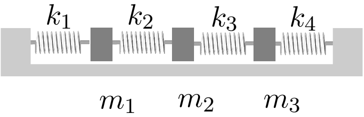


Figure 24. Three masses connected by springs. The masses slide along a frictionless track.

The analysis uses the following constants, variables and assumptions.

Mass Constants The boxcar masses m_1, m_2, m_3 are assumed to be point masses concentrated at their center of gravity.

Spring Constants The mass of each spring is negligible. The springs obey Hooke's law: *Force = k(elongation)*. The Hooke's constants are denoted k_1, k_2, k_3, k_4. The springs restore after compression and extension.

Position Variables Symbols $x_1(t)$, $x_2(t)$, $x_3(t)$ denote the mass positions along the horizontal surface, measured from their equilibrium positions, plus right and minus left.

Fixed Ends The first and last spring are attached to fixed walls.

The **competition method** is used to derive the equations of motion, using:

$$\text{Newton's Second Law Force} = \text{Sum of the Hooke's Forces}.$$

The model equations are

$$
(9)\qquad \begin{array}{rcl}
m_1x_1''(t) &=& -k_1x_1(t)+k_2[x_2(t)-x_1(t)],\\
m_2x_2''(t) &=& -k_2[x_2(t)-x_1(t)]+k_3[x_3(t)-x_2(t)],\\
m_3x_3''(t) &=& -k_3[x_3(t)-x_2(t)]-k_4x_3(t).
\end{array}
$$

The equations are justified in the case of all positive variables by observing that the first three springs are elongated by x_1, x_2-x_1, x_3-x_2, respectively. The last spring is compressed by x_3, which accounts for the minus sign.

Another way to justify the equations is through mirror-image symmetry: interchange $k_1 \longleftrightarrow k_4$, $k_2 \longleftrightarrow k_3$, $x_1 \longleftrightarrow x_3$, then equation 2 should be unchanged and equation 3 should become equation 1.

Matrix Formulation. System (9) can be written as a second order vector-matrix system

$$
\begin{pmatrix} m_1 & 0 & 0\\ 0 & m_2 & 0\\ 0 & 0 & m_3\end{pmatrix}\begin{pmatrix} x_1''\\ x_2''\\ x_3''\end{pmatrix} = \begin{pmatrix} -k_1-k_2 & k_2 & 0\\ k_2 & -k_2-k_3 & k_3\\ 0 & k_3 & -k_3-k_4\end{pmatrix}\begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix}.
$$

More succinctly, the system is written as

$$
M\vec{\mathbf{x}}''(t) = K\vec{\mathbf{x}}(t)
$$

where the **displacement** $\vec{\mathbf{x}}$, **mass matrix** M and **stiffness matrix** K are defined by the formulas

$$
\vec{\mathbf{x}} = \begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix},\quad M = \begin{pmatrix} m_1 & 0 & 0\\ 0 & m_2 & 0\\ 0 & 0 & m_3\end{pmatrix},\quad K = \begin{pmatrix} -k_1-k_2 & k_2 & 0\\ k_2 & -k_2-k_3 & k_3\\ 0 & k_3 & -k_3-k_4\end{pmatrix}.
$$

Two Masses

Modeling of two masses connected by springs uses ideas and methods from three-mass modeling equation (9).

Two Masses, Right End Free

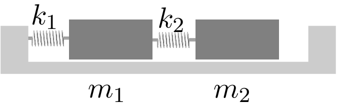


Figure 25. Two masses anchored left and connected by springs.

The model equations:

$$
(10)\qquad \begin{array}{rcl}
m_1x_1''(t) &=& -k_1x_1(t)+k_2[x_2(t)-x_1(t)]\\
m_2x_2''(t) &=& -k_2[x_2(t)-x_1(t)]
\end{array}
$$

Two Masses, Both Ends Free

Equations (10) modified with $k_1 = 0$ gives model equations:

$$(11)\qquad \begin{array}{rcl} m_1 x_1''(t) & = & k_2[x_2(t) - x_1(t)] \\ m_2 x_2''(t) & = & -k_2[x_2(t) - x_1(t)] \end{array}$$

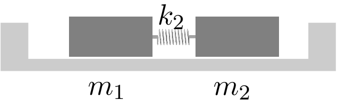


Figure 26. Two masses connected by one spring.

Example 11.18 (Two Masses with Free Right End)

Consider equation (10) with $m_1 = 2m_2$, $\dfrac{k_1}{m_1} = \dfrac{k_2}{m_2} = 50$:

$$\vec{\mathbf{x}}'' = \begin{pmatrix} -75 & 25 \\ 50 & -50 \end{pmatrix} \vec{\mathbf{x}}$$

Then the vector solution in terms of arbitrary constants a_1, a_2, b_1, b_2 is given by:

$$\vec{\mathbf{x}} \quad = \quad (a_1 \cos 5t + b_1 \sin 5t) \begin{pmatrix} 1 \\ 2 \end{pmatrix} + (a_2 \cos 10t + b_2 \sin 10t) \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Details Example 11.18:
Eigenpairs of $A = \begin{pmatrix} -75 & 25 \\ 50 & -50 \end{pmatrix}$ are $\left(-25, \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right)$, $\left(-100, \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)$. The example is completed by Theorem 11.56.

Three Rail Cars

A special case of the coupled spring-mass system is three rail cars on a level frictionless track connected by springs, as in Figure 28.[14]

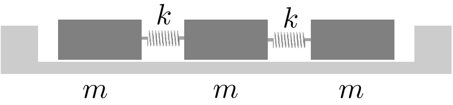


Figure 28. Three identical flatbed cars connected by identical springs.

Except for the springs on fixed ends, this problem is the same as the one in Figure 22. Let $k_1 = k_4 = 0$, $k_2 = k_3 = k$, $m_1 = m_2 = m_3 = m$ to give the system

$$(12)\qquad \begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & m \end{pmatrix} \begin{pmatrix} x_1'' \\ x_2'' \\ x_3'' \end{pmatrix} = \begin{pmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

[14] The cars are custom flatbed utility cars, **not boxcars**. Railway cars such as tankers, hoppers and boxcars are equipped with automatic Janney couplers, compression only dash-pots/bumpers and safety lanyards.

Figure 27. Railroad Boxcar Silhouette.

Example 11.19 (Identical Cars with $k = m$)

Consider equation (12) for $k = m$:

$$\vec{\mathbf{x}}'' = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix} \vec{\mathbf{x}}$$

Then the vector solution in terms of arbitrary constants a_1, a_2, a_3, b_1, b_2, b_3 is given by:

$$\begin{aligned} \vec{\mathbf{x}} \quad = \quad & (a_1 + b_1 t)\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + (a_2 \cos t + b_2 \sin t)\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \\ & + \left(a_3 \cos \sqrt{3}t + b_3 \sin \sqrt{3}t\right)\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \end{aligned} \tag{13}$$

Boxcars and Buffer Springs. Boxcars have buffer-spring shock absorbers which exert a force only under compression. Suppose one car moves along the track, then contacts two stationary cars, then transfers its momentum to the other cars, followed by disengagement. This situation could have a matrix model $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}} + B\vec{\mathbf{x}}'$. Matrix A contains Hooke's constants depending on $\vec{\mathbf{x}}$. Matrix B contains dashpot constants depending on $\vec{\mathbf{x}}$ and $\vec{\mathbf{x}}'$. The complexity seems suited for computer simulation.

Assume the dashpot constants are zero. The shock absorber springs act normally upon compression; the cars disengage upon full spring expansion. Model $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ has Hooke's constants in 3×3 matrix A. Solution expression (13) applies until a car disengages, measured by the first time $t = t_1 > 0$ at which $x_2(t) = x_1(t)$ or $x_3(t) = x_2(t)$. When a car contacts another car then the shock assembly compresses slightly but does not engage: the car making contact transfers momentum.

Analysis of one car moving into contact with two stationary cars uses equation (13) on $0 \le t \le t_1$. For $t > t_1$, model $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$ is discarded. One example is the first car transfers momentum and stops, while the other two cars travel at fixed speed. The model applies to determine both the time t_1 and the speed of the other two cars after $t = t_1$.

Dynamic Dashpot

A **dynamic dashpot** is a variable shock absorber, a component of **active suspension**.

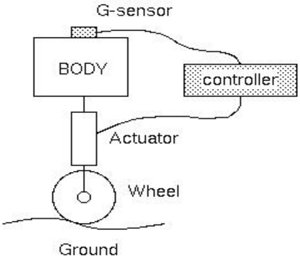


Figure 29. Active Suspension Components.

Bose Corporation (1980) designed variable shock absorbers for truck seats, especially 18-wheelers. Active suspension solutions for vehicles were designed by Toyota (1994), General Motors (2002), Volvo (2002), Range-Rover (2004) and Mercedes-Benz (2013). Camera and road-sensor devices have been implemented by Mercedes-Benz (2014).

An instance of Figure 29 is one wheel suspension with spring and shock absorber. Assumptions will fit the system to a damped spring-mass model.

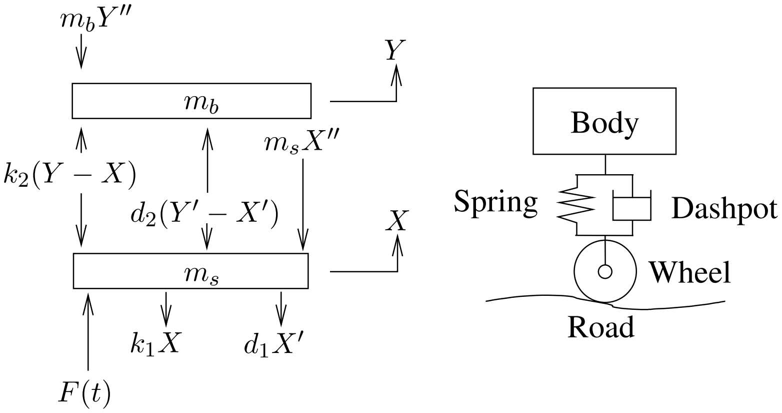


Figure 30. One wheel suspension with spring and shock absorber.

Assumptions for Figure 30.

Y = body mass displacement from equilibrium $Y = 0$
m_b = body mass
m_s = suspension system mass
X = suspension system mass displacement from equilibrium $X = 0$
k_1 = wheel and tire Hooke's constant
k_2 = suspension Hooke's constant
d_1 = wheel and tire dashpot constant
d_2 = suspension dashpot constant
$F(t)$ = roadway force on the wheel-suspension-body unit

$$(14) \qquad \begin{cases} m_sX'' = -k_1X - d_1X' - k_2(Y-X) - d_2(Y'-X') + F(t), \\ m_bY'' = k_2(Y-X) + d_2(Y'-X') \end{cases}$$

Ideal Suspension

Industrial solutions have used a tunable shock absorber, which means $c_2(t)$ is a function of time t defined in response to road data $F(t)$ and the current states $X(t), Y(t)$. **Is it realistic** to expect nearly motionless body vibration $Y \approx 0$ with suitable real-time changes in suspension dashpot constant $c_2(t)$? Manufacturers report yes, given suitably benign roadway data. Simulation uses the electrical-mechanical analogy to design an electrical circuit for model (14). Roadway data $F(t)$ from cameras and sensors is modeled by a variable input (emf) in the electrical circuit while mechanical displacements X, Y appear as electrical currents. Figure 31 shows an equivalent electrical network for computer simulation, extracted from a 2009 undergraduate Bachelor's Thesis project at Worcester Polytechnic Institute.[15]

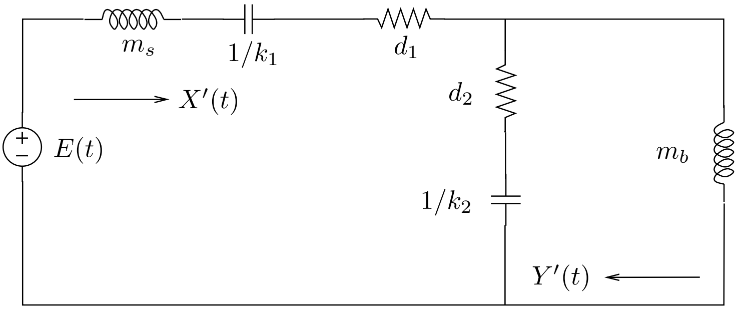


Figure 31. Dynamic shock absorber simulator circuit (2009)

Active Suspension Regulator

Added to Figure 30 is a regulator, which can be imagined as a linear electromagnetic motor that turns a shaft, one voltage input providing an upward force and the other input a downward force. Electrical supply voltages adjust the forces dynamically with sensor feedback. Standard suspension is $F(t) = 0$. Symbol $F(t)$ in Figures 30, 31 is a force, but each instance has a different meaning.

[15] Pashaj, B., Bermejo Calle, M. J., and Sebuwufu, P. (2009), **Dynamic Shock Absorber**, `https://digitalcommons.wpi.edu/mqp-all/2634`.

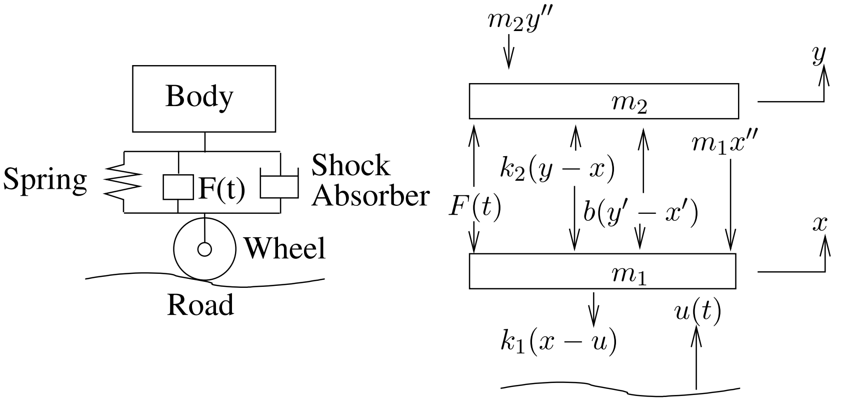


Figure 32. One wheel suspension with spring, shock absorber and regulator $F(t)$. Variables and forces are defined in the force diagram on the right. All units MKS.

Assumptions for Figure 32.

m_2 = body mass
y = body mass displacement from equilibrium $y = 0$
m_1 = suspension system mass
x = suspension system mass displacement from equilibrium $x = 0$
k_1 = wheel and tire Hooke's constant
k_2 = suspension Hooke's constant
b = shock absorber dashpot constant
$F(t)$ = regulator force between body and suspension system
$u(t)$ = roadway vertical displacement on the wheel-suspension-body unit

Equations (15) are derived from the force diagram in Figure 32.

$$(15)\qquad \begin{cases} m_1x'' = k_2(y-x) + b(y'-x') - k_1(x-u) - F(t), \\ m_2y'' = -k_2(y-x) - b(y'-x') + F(t) \end{cases}$$

Regulator. Assume system parameters in MKS units:

$$\begin{array}{ll} k_1 = 135000 & k_2 = 5700 \\ m_1 = 50 & m_2 = 465 \\ b = 290 & u(t) = 0.015\sin(t), \\ x(0) = x'(0) = 0 & \text{(suspension } m_1 \text{ at rest)} \\ y(t) = 0 & \text{(body } m_2 \text{ motionless)} \end{array}$$

The vertical roadway displacement $u(t) = 0.015\sin(t)$ fits a railroad track, zero to 1.5 cm deviation from perfectly flat. It is not suited for a highway. Period 2π is selected for simplicity. Equation (15) with values inserted implies equation (16):

$$(16)\qquad \begin{cases} 50\,x'' = 5700y - 140700x + 290y' - 290x' + 2025\sin(t) - F(t), \\ 465\,y'' = -5700y + 5700x - 290y' + 290x' + F(t), \\ x(0) = x'(0) = y(0) = y'(0) = 0 \end{cases}$$

Let's verify the **ideal suspension** regulator force:

$$(17)\qquad \begin{cases} F(t) &= \dfrac{2565\sqrt{3}}{2699}\sin\left(30\sqrt{3}t\right) - \dfrac{230850}{2699}\sin(t) \\ & +\dfrac{11745}{2699}\cos\left(30\sqrt{3}t\right) - \dfrac{11745}{2699}\cos(t) \end{cases}$$

if $y(t) = 0$, then $x(t)$ and $F(t)$ are determined by:

$$(18)\qquad \begin{cases} 50x'' = -140700x - 290x' + 2025\sin(t) - F(t), \\ 0 = 5700x + 290x' + F(t), \\ x(0) = x'(0) = 0 \end{cases}$$

Add equations (18):

$$50x'' = -135000x + 2025\sin(t), \quad x(0) = x'(0) = 0.$$

Then $x(t) = -\dfrac{9\sqrt{3}}{53980}\sin(30\sqrt{3}t) + \dfrac{81}{5398}\sin(t)$. Solve for $F(t)$ from the second equation in (18). Then equation (17) holds with approximation

$$F(t) \approx 4.35\cos(51.9\,t) + 1.64\sin(51.9\,t) - 4.35\cos(t) - 85.5\sin(t)$$

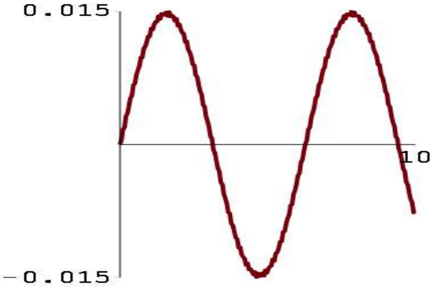


Figure 33. Suspension displacement

Solution $x(t)$ for a motionless body $y(t) = 0$ with roadway displacement $u(t) = 0.015\sin(t)$.

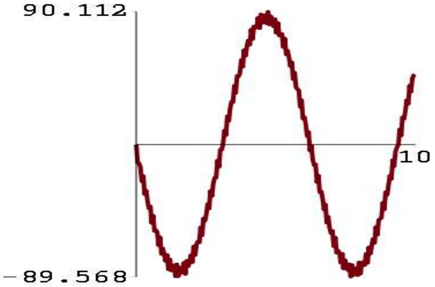


Figure 34. Regulator force

Force $F(t)$ for a motionless body $y(t) = 0$ with roadway displacement $u(t) = 0.015\sin(t)$.

Jagged edges in Figure 33 are caused by the high frequency term in $x(t) = -\dfrac{9\sqrt{3}}{53980}\sin(30\sqrt{3}t) + \dfrac{81}{5398}\sin(t)$. Similarly for Figure 34.

Exercises 11.8

Euler's Substitution: $\vec{\mathbf{u}}\,' = C\vec{\mathbf{u}}$

1. Change variables: $\vec{\mathbf{u}} = e^{rt}\vec{\mathbf{w}}$. Answer: $\vec{\mathbf{w}}\,' = (C - rI)\vec{\mathbf{w}}$

2. Prove: $(\lambda, \vec{\mathbf{v}})$ is an eigenpair of C if and only if $(0, \vec{\mathbf{v}})$ is an eigenpair of $C - \lambda I$.

3. Let $|C - \lambda I|$ have factor λ^2. Let $\vec{\mathbf{u}}\,' = C\vec{\mathbf{u}}$ have solution $\vec{\mathbf{u}} = \vec{\mathbf{d}}_1 + t\vec{\mathbf{d}}_2$. Prove: $C\vec{\mathbf{d}}_2 = \vec{\mathbf{0}}$, $C\vec{\mathbf{d}}_1 = \vec{\mathbf{d}}_2$. Are $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2$ eigenvectors of C? Discuss.

4. Let $C = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\vec{\mathbf{u}} = \vec{\mathbf{d}}_1 + t\vec{\mathbf{d}}_2$. Let $\vec{\mathbf{u}}$ solve $\vec{\mathbf{u}}\,' = C\vec{\mathbf{u}}$. Find $\vec{\mathbf{d}}_1, \vec{\mathbf{d}}_2$ in terms of arbitrary constants c_1, c_2.

Euler's Substitution: $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$

5. Change variables: $\vec{\mathbf{x}} = e^{rt}\vec{\mathbf{y}}$. Answer: $\vec{\mathbf{y}}\,'' + 2r\vec{\mathbf{y}}\,' = (A - r^2 I)\vec{\mathbf{y}}$

6. Prove: $\vec{\mathbf{x}} = e^{rt}\vec{\mathbf{v}}$ is a nonzero solution of $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$ if and only if $(r^2, \vec{\mathbf{v}})$ is an eigenpair of A.

Repeated Root: $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$

Let $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, eigenvalues $0, 0$.

7. Verify: Matrix A is a Jordan block with generalized eigenvectors the columns of I.

8. Prove: $x_1 = c_1 + c_2 t + c_3 \frac{t^2}{2} + c_4 \frac{t^3}{6}$, $x_2 = c_3 + c_4 t$ for arbitrary constants c_1 to c_4.

9. Prove: The solution of $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$ is a vector linear combination of atoms $1, t, t^2, t^3$.

10. Let $\vec{\mathbf{x}} = \vec{\mathbf{d}}_1 + \vec{\mathbf{d}}_2 t + \vec{\mathbf{d}}_3 \frac{t^2}{2} + \vec{\mathbf{d}}_4 \frac{t^3}{6}$. Assume $\vec{\mathbf{x}}$ solves $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$. Prove: $A\vec{\mathbf{d}}_3 = A\vec{\mathbf{d}}_4 = \vec{\mathbf{0}}$, $A\vec{\mathbf{d}}_1 = \vec{\mathbf{d}}_3$, $A\vec{\mathbf{d}}_2 = \vec{\mathbf{d}}_4$. These are generalized eigenvector chains for eigenvalue zero.

CHZ Method

11. Given a 3×3 matrix A, supply proof details for the Cayley-Hamilton-Ziebur structure theorem.

12. Invent a non-diagonal 3×3 example $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$ and solve it by CHZ.

13. Solve $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$ by CHZ for any 2×2 diagonal matrix with negative diagonal elements.

14. Solve $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$ by CHZ for any 3×3 diagonal matrix with negative diagonal elements.

Conversion

Given $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$, let $\vec{\mathbf{u}} = \begin{pmatrix} \vec{\mathbf{x}} \\ \vec{\mathbf{x}}\,' \end{pmatrix}$. Display system $\vec{\mathbf{u}}\,' = C\vec{\mathbf{u}}$.

15. $A = \begin{pmatrix} 1 & 3 \\ -1 & 2 \end{pmatrix}$

16. $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 2 & -1 & 2 \end{pmatrix}$

Eigenanalysis $\lambda \le 0$

Display the general solution of $\vec{\mathbf{x}}\,'' = A\vec{\mathbf{x}}$.

17. $A = \begin{pmatrix} -3 & 3 \\ 1 & -1 \end{pmatrix}$

18. $A = \begin{pmatrix} -3 & 3 & 0 \\ 1 & -1 & 0 \\ 5 & 0 & -1 \end{pmatrix}$

Earthquakes

Apply formulas from the *Earthquakes subsection* page 929 to find particular solution $\vec{\mathbf{x}}_p$, the natural frequencies ω_j and the amplitudes of $\vec{\mathbf{x}}_p(t)$ near the largest natural frequency. Assume $F(t) = F_0 \cos(\omega t)$.

19. Three-floor problem, $k/m = 10$.

20. Four-floor problem, $k/m = 10$.

Two Masses

Assume MKS units. Let $m_1 = 2$, $m_2 = 0.5$, $k_1 = 75$, $k_2 = 25$ in system:

$$\begin{aligned} m_1 x_1'' &= -k_1 x_1 + k_2[x_2 - x_1] \\ m_2 x_2'' &= -k_2[x_2 - x_1] \end{aligned}$$

21. Convert the system to the form $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$.

22. Show details for finding the vector solution $\vec{\mathbf{x}}(t)$.

Three Rail Cars: $k=2m$

Assume MKS units. Consider

$$\vec{\mathbf{x}}'' = \begin{pmatrix} -2 & 2 & 0 \\ 2 & -4 & 2 \\ 0 & 2 & -2 \end{pmatrix} \vec{\mathbf{x}}$$

23. Show eigenpair details for the 3×3 matrix.

24. Find the vector solution $\vec{\mathbf{x}}(t)$.

Three Rail Cars: Disengagement

For $\vec{\mathbf{x}}'' = A\vec{\mathbf{x}}$, assume FPS units and

$$A = \begin{pmatrix} -4 & 4 & 0 \\ 6 & -12 & 6 \\ 0 & 4 & -4 \end{pmatrix}$$

Suppose the springs disengage upon full expansion. Let the cars engage at $t=0$ with $x_1 = x_2 = x_3 = 0$.

25. Verify A has eigenvalues $\lambda = -16, 0, -4$ and corresponding eigenvectors
$$\begin{pmatrix} 1 \\ -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

26. For $x_1{=}x_2{=}x_3{=}0$ at $t{=}0$, verify:
$x_1(t){=}c_1t{+}c_2\sin(2t){-}c_3\sin(4t)$,
$x_2(t) = c_1t + 3c_3\sin(4t)$,
$x_3(t){=}c_1t{-}c_2\sin(2t){-}c_3\sin(4t)$

27. Let $x_1' = 48$, $x_2' = 0$, $x_3' = 0$ at $t = 0$. Verify disengagement time $t_1 = \pi/2$ and determine the car velocities thereafter.

28. Let $x_1'(0) = 144$, $x_2'(0) = 48$, $x_3'(0) = 48$. Verify disengagement time $t_1 = \pi/2$ and determine the car velocities thereafter.
Answer: Velocities 144, 48, 48 at $t = t_1$.

Dynamic Dashpot

Assume conventions for Figure 26 and dynamic dashpot system

$$\begin{aligned} m_sX'' &= -k_1X - d_1X' - k_2(Y-X) \\ &\quad - d_2(Y'-X') + F(t), \\ m_bY'' &= k_2(Y-X) + d_2(Y'-X') \end{aligned}$$

See page 936.

29. Assume $Y = 0$, ideal suspension. Derive:

$$\begin{aligned} m_sX'' &= -k_1X - d_1X' + F(t), \\ d_2X' + k_2X &= 0 \end{aligned}$$

30. Assume $Y = 0$, ideal suspension and $X(0) = 0.015$ meters. Find $X(t)$ and $F(t)$.

11.9 Numerical Methods for Systems

An initial value problem for a system of two differential equations is given by the equations

$$\begin{aligned} x'(t) &= f(t,x(t),y(t)),\\ y'(t) &= g(t,x(t),y(t)),\\ x(t_0) &= x_0,\\ y(t_0) &= y_0. \end{aligned}\tag{1}$$

A **numerical method** for (1) is an algorithm that computes an approximation table with first line t_0, x_0, y_0. Generally, the table has equally spaced t-values, two consecutive t-values differing by a constant value $h \neq 0$, called the **step size**. To illustrate, if $t_0 = 2$, $x_0 = 5$, $y_0 = 100$, then a typical approximation table for step size $h = 0.1$ might look like

t	x	y
2.0	5.00	100.00
2.1	5.57	103.07
2.2	5.62	104.10
2.3	5.77	102.15
2.4	5.82	101.88
2.5	5.96	100.55

Graphics

The approximation table represents the data needed to plot a solution curve to system (1) in three dimensions (t, x, y) or in two dimensions, using a tx-scene or a ty-scene. In all cases, the plot is a simple connect-the-dots graphic.

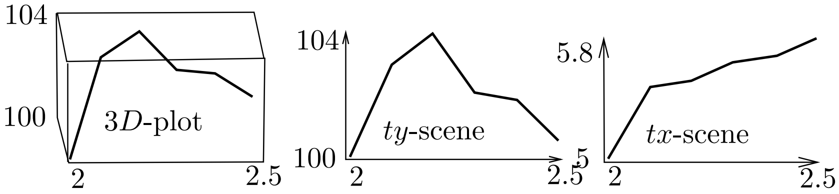


Figure 35. Dot table plots.
The three dimensional plot is a space curve made directly from the dot table. The tx-scene and the ty-scene are made from the same approximation table using corresponding data columns.

Near-Sighted Algorithms

All of the popular algorithms for numerical generation of an approximation table for system (1) are **near-sighted algorithm**, because they predict the next line in the table from the current table line, ignoring effects and errors for all other preceding table lines. Among such algorithms are **Euler's method**, **Heun's**

method and the **RK4 method**, which are showcased here for learning purposes. Computer production algorithms are available in `maple`, `mathematica` and `matlab`.

Numerical Algorithms: Planar Case

Stated here without proof are three numerical algorithms for solving planar initial value problems (1). Justification of the formulas is obtained from the vector relations in the next subsection.

Notation. Let t_0, x_0, y_0 denote the entries of the approximation table on a particular line. Let h be the increment for the table and let $t_0 + h$, x, y denote the table entries on the next line.

Planar Euler Method

$$\begin{aligned} x &= x_0 + hf(t_0, x_0, y_0), \\ y &= y_0 + hg(t_0, x_0, y_0). \end{aligned}$$

Planar Heun Method

$$\begin{aligned} x_1 &= x_0 + hf(t_0, x_0, y_0), \\ y_1 &= y_0 + hg(t_0, x_0, y_0), \\ x &= x_0 + h(f(t_0, x_0, y_0) + f(t_0 + h, x_1, y_1))/2 \\ y &= y_0 + h(g(t_0, x_0, y_0) + g(t_0 + h, x_1, y_1))/2. \end{aligned}$$

Planar RK4 Method

$$\begin{aligned} k_1 &= hf(t_0, x_0, y_0), \\ m_1 &= hg(t_0, x_0, y_0), \\ k_2 &= hf(t_0 + h/2, x_0 + k_1/2, y_0 + m_1/2), \\ m_2 &= hg(t_0 + h/2, x_0 + k_1/2, y_0 + m_1/2), \\ k_3 &= hf(t_0 + h/2, x_0 + k_2/2, y_0 + m_2/2), \\ m_3 &= hg(t_0 + h/2, x_0 + k_2/2, y_0 + m_2/2), \\ k_4 &= hf(t_0 + h, x_0 + k_3, y_0 + m_3), \\ m_4 &= hg(t_0 + h, x_0 + k_3, y_0 + m_3), \\ x &= x_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \\ y &= y_0 + \frac{1}{6}\left(m_1 + 2m_2 + 2m_3 + m_4\right). \end{aligned}$$

Example 11.20 (Planar Methods)
Solve $x' = x, y' = -2y$, $x(0) = y(0) = 2$ with step size $h = 0.1$ for 10 steps, using methods Euler, Heun and RK4 in computer algebra system MAPLE.

Details
Computer code for the three algorithms can be found in the solution to Exercise 1. Newer MAPLE versions have the algorithms available as documented below.

```
des:=diff(x(t),t)=x(t),diff(y(t),t)=-2*y(t);ics:=x(0)=2,y(0)=2;
args:=[des,ics],numeric,stepsize=0.1,output=listprocedure;
p:=dsolve(args,method=classical[foreuler]);# or: heunform, rk4
X:=eval(x(t),p); Y:=eval(y(t),p);
printf("Euler\n  t       X(t)      Y(t)\n");
seq(printf("%f  %f  %f\n",0.1*j,X(0.1*j),Y(0.1*j)),j=0..10);
```

The expected results are 1, 2, 4 digits of accuracy respectively for the computed values. At $t = 1$ the maple code for step size 0.1 computes $y(t)$ for Euler, Heun, RK4 as 0.214748, 0.274896, 0.270679 compared to exact value $y(1) = 2e^{-2} = 0.2706705664$.

Numerical Algorithms: General Case

Consider a vector initial value problem

$$\vec{\mathbf{u}}'(t) = \vec{\mathbf{F}}(t, \vec{\mathbf{u}}(t)), \quad \vec{\mathbf{u}}(t_0) = \vec{\mathbf{u}}_0.$$

Stated here are the vector formulas for Euler, Heun and RK4 methods. These myopic algorithms predict the next table entry $t_0 + h$, $\vec{\mathbf{u}}$ from the current entry t_0, $\vec{\mathbf{u}}_0$. The number of scalar values in a table row is $1 + n$, where n is the dimension of the vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{F}}$.

Vector Euler Method

$$\vec{\mathbf{u}} = \vec{\mathbf{u}}_0 + h\vec{\mathbf{F}}(t_0, \vec{\mathbf{u}}_0)$$

Vector Heun Method

$$\vec{\mathbf{w}} = \vec{\mathbf{u}}_0 + h\vec{\mathbf{F}}(t_0, \vec{\mathbf{u}}_0), \quad \vec{\mathbf{u}} = \vec{\mathbf{u}}_0 + \frac{h}{2}\left(\vec{\mathbf{F}}(t_0, \vec{\mathbf{u}}_0) + \vec{\mathbf{F}}(t_0 + h, \vec{\mathbf{w}})\right)$$

Vector RK4 Method

$$\begin{array}{rcl}
\vec{\mathbf{k}}_1 & = & h\vec{\mathbf{F}}(t_0, \vec{\mathbf{u}}_0), \\
\vec{\mathbf{k}}_1 & = & h\vec{\mathbf{F}}(t_0 + h/2, \vec{\mathbf{u}}_0 + \vec{\mathbf{k}}_1/2), \\
\vec{\mathbf{k}}_1 & = & h\vec{\mathbf{F}}(t_0 + h/2, \vec{\mathbf{u}}_0 + \vec{\mathbf{k}}_2/2), \\
\vec{\mathbf{k}}_1 & = & h\vec{\mathbf{F}}(t_0 + h, \vec{\mathbf{u}}_0 + \vec{\mathbf{k}}_3), \\
\vec{\mathbf{u}} & = & \vec{\mathbf{u}}_0 + \dfrac{1}{6}\left(\vec{\mathbf{k}}_1 + 2\vec{\mathbf{k}}_2 + 2\vec{\mathbf{k}}_3 + \vec{\mathbf{k}}_4\right).
\end{array}$$

Example 11.21 (Exact Solution $\vec{\mathbf{u}}\,' = A\,\vec{\mathbf{u}} + \vec{\mathbf{F}}\,(t)$)

Let $A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, $\vec{\mathbf{F}}(t) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$. Solve $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$.

Details

Handwritten method: find a fundamental matrix $\Phi(t)$ and then $e^{At} = \Phi(t)\Phi(0)^{-1}$. The homogeneous solution is $u_h(t) = e^{At}\vec{\mathbf{c}}$ for constant vector $\vec{\mathbf{c}}$. A particular solution $\vec{\mathbf{u}}_p(t)$ is computed from the variation of parameters formula page 912.

CAS method: One possible method uses MAPLE library DEtools:

```
A:=Matrix([[1, -1 , 0],[1 , 1 , 0],[0 , 0 , 2]]);
F:=Vector([1,1,0]);Sol:=DEtools[matrixDE](A,F,t);
Xh:=Sol[1].Vector([c1,c2,c3]);Xp:=Vector(convert(Sol[2],list));
U:=unapply(Xh+Xp,t);U(t);# General solution of u'=Au+F(t)
simplify(A.U(t)+F-map(diff,U(t),t));# Answer check
```

$$\vec{\mathbf{u}}(t) = \begin{bmatrix} e^t \cos(t)\, c_1 + e^t \sin(t)\, c_2 - 1 \\ e^t \sin(t)\, c_1 - e^t \cos(t)\, c_2 \\ e^{2\,t} c_3 \end{bmatrix}$$

Example 11.22 (Vector Euler Method)

Let $A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. $\vec{\mathbf{F}}(t) = \begin{pmatrix} e^t \\ 1 \\ 0 \end{pmatrix}$. Solve $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ with step size $h = 0.1$ for 10 steps, using the vector Euler method implemented in computer algebra system MAPLE.

Details

The vector algorithm uses MAPLE functions and basic vector-matrix algebra.

```
# Euler's method with vector notation
A:=Matrix([[1, -1 , 0],[1 , 1 , 0],[0 , 0 , 2]]);
F0:=unapply(A.<x,y,z>+Vector([exp(t),1,0]),(t,x,y,z)):
F0(t,x,y,z);# Scalar variables
F:=(t,X)->F0(t,X[1],X[2],X[3]);# Vector variables
U0:=<1,0,0>;n:=10;h:=0.1;t0:=0;Vals:=U0; # Initialize
for j from 1 to n do
U:=U0+h*F(t0,U0);U0:=U;t0:=t0+h;Vals:=Vals,U0;
od:
ValsEuler:=Vals[n+1];
```

$$\texttt{ValsEuler} = \begin{bmatrix} 3.1116983042 \\ 4.4649291918 \\ 0.0 \end{bmatrix}$$

Example 11.23 (Vector Heun Method)

Let $A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. $\vec{\mathbf{F}}(t) = \begin{pmatrix} e^t \\ 1 \\ 0 \end{pmatrix}$. Solve $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ with step size $h = 0.1$ for 10 steps, using the vector Heun method implemented in computer algebra system MAPLE.

Details

```
# Heun's method with vector notation
A:=Matrix([[1, -1 , 0],[1 , 1 , 0],[0 , 0 , 2]]);
F0:=unapply(A.<x,y,z>+Vector([exp(t),1,0]),(t,x,y,z)):
F0(t,x,y,z);# Scalar variables
F:=(t,X)->F0(t,X[1],X[2],X[3]);# Vector variables
U0:=<1,0,0>;n:=10;h:=0.1;t0:=0:Vals:=U0; # Initialize
for j from 1 to n do
w:=U0+h*F(t0,U0);
U:=U0+0.5*h*(F(t0,U0)+F(t0+h,w));U0:=U;t0:=t0+h;Vals:=Vals,U0;
od:
ValsHeun:=Vals[n+1];
```

$$\texttt{ValsHeun} = \begin{bmatrix} 2.8724813157 \\ 4.9105494201 \\ 0.0 \end{bmatrix}$$

Example 11.24 (Vector RK4 Method)

Let $A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. $\vec{\mathbf{F}}(t) = \begin{pmatrix} e^t \\ 1 \\ 0 \end{pmatrix}$. Solve $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ with step size $h = 0.1$ for 10 steps, using the vector RK4 method implemented in computer algebra system MAPLE.

Details

```
# RK4 method with vector notation
A:=Matrix([[1, -1 , 0],[1 , 1 , 0],[0 , 0 , 2]]);
F0:=unapply(A.<x,y,z>+Vector([exp(t),1,0]),(t,x,y,z)):
F0(t,x,y,z);# Scalar variables
F:=(t,X)->F0(t,X[1],X[2],X[3]);# Vector variables
U0:=<1,0,0>;n:=10;h:=0.1;t0:=0:Vals:=U0; # Initialize
for j from 1 to n do
k1:=h*F(t0,U0);
k2:=h*F(t0+h/2,U0+k1/2);
k3:=h*F(t0+h/2,U0+k2/2);
k4:=h*F(t0+h,U0+k3);
U:=U0+(k1+2*k2+2*k3+k4)/6;U0:=U;t0:=t0+h;Vals:=Vals,U0;od:
ValsRK4:=Vals[n+1];
```

$$\mathtt{ValsRK4} = \begin{bmatrix} 2.8467234249 \\ 4.9149919169 \\ 0.0 \end{bmatrix}$$

Example 11.25 (Compare Vector Methods Euler, Heun and RK4)

Let $A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, $\vec{\mathbf{F}}(t) = \begin{pmatrix} e^t \\ 1 \\ 0 \end{pmatrix}$. Solve $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ with step size $h = 0.1$ for 10 steps, using the vector methods Euler, Heun and RK4 in computer algebra system MAPLE. Compare to 6 digits computed values at $t = 1$ for the three methods.

Details

Refer to the previous three examples for `maple` values `ValsEuler, ValsHeun, ValsRK4, Exact`.

$$\begin{bmatrix} 2.872481 \\ 4.910549 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 2.872481 \\ 4.910549 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 2.846723 \\ 4.914992 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 2.846719 \\ 4.914968 \\ 0.0 \end{bmatrix}.$$

Exercises 11.9

Planar Methods

Apply the Euler, Heun and RK4 methods. Compare with the exact solution in a table.

1. $x' = x$, $y' = -y$, $x(0) = 2$, $y(0) = 2$. $h = 0.1$, 10 steps

2. $x' = -3x + y$, $y' = x - 3y$, $x(0) = 2$, $y(0) = 0$, $h = 0.1$, 10 steps

3. $x' = -x + y$, $y' = -x - y$, $x(0) = 0$, $y(0) = 3$, $h = 0.2$, 5 steps

4. $x' = 2x - 4y$, $y' = x - 3y$, $x(0) = 4$, $y(0) = 0$, $h = 0.1$, 10 steps

Vector Methods $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$, 2×2

Apply vector Euler, Heun and RK4 methods for 10 steps with $h = 0.1$.

5. $\vec{\mathbf{u}}' = \begin{pmatrix} u_1 + u_2 \\ -u_1 + u_2 \end{pmatrix}$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$.

6. $\vec{\mathbf{u}}' = \begin{pmatrix} -3u_1 + u_2 \\ u_1 - 3u_2 \end{pmatrix}$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$.

Vector Methods $\vec{\mathbf{u}}' = A\vec{\mathbf{u}} + \vec{\mathbf{F}}(t)$

Apply vector Euler, Heun and RK4 methods for 10 steps with $t_0 = 0$, $h = 0.1$. Compare results for the last step.

7. $A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$, $\vec{\mathbf{F}} = \begin{pmatrix} e^t \\ 0 \end{pmatrix}$,

$\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Ans Euler: $3.81, -5.33$

8. $A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$, $\vec{\mathbf{F}} = \begin{pmatrix} e^t \\ 0 \\ 0 \end{pmatrix}$,

$\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

Ans RK4: $2.576, -5.528, 0.0$

Vector Methods $\vec{\mathbf{u}}' = A\vec{\mathbf{u}}$, 3×3

Apply vector Euler, Heun and RK4 methods for 10 steps with $h = 0.1$.

9. $A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

Ans Heun: $1.36, -3.67, 0.00$

10. $A = \begin{pmatrix} 1 & 3 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\vec{\mathbf{u}}(0) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

Ans RK4: $-2.307, -3.075, 0.00$

Chapter 12

Series Methods

Contents

Introduction

The differential equation

$$(1) \qquad (1+x^2)y'' + (1+x+x^2+x^3)y' + (x^3-1)y = 0$$

has polynomial coefficients. It will be shown in this chapter that the solution $y(x)$ is **approximately a polynomial**, that is, the general solution y has an approximation formula

$$y(x) \approx c_1 f_1(x) + c_2 f_2(x),$$

where f_1 and f_2 are polynomials. Graphically, the polynomials depend on the graph window, the pixel resolution and a maximum value for $|c_1| + |c_2|$.

The approximation means that solution graphs can be made with a graphing hand calculator, a computer algebra system or a numerical laboratory by entering two polynomials f_1, f_2. For (1), the polynomials

$$f_1(x) = 1 + \frac{1}{2}x^2 - \frac{1}{6}x^3 - \frac{1}{12}x^4 - \frac{1}{60}x^5,$$

$$f_2(x) = x - \frac{1}{2}x^2 + \frac{1}{6}x^3 - \frac{1}{15}x^5$$

can be used to plot solutions within a reasonable range of initial conditions.

The theory will show that (1) has a basis of solutions $y_1(x)$, $y_2(x)$, each represented as a convergent power series

$$y(x) = \sum_{n=0}^{\infty} a_n x^n.$$

Truncation of power series y_1 to a polynomial f_1 and power series y_2 to a polynomial f_2 provide approximate solutions suitable for graphing and calculation.

12.1 Review of Calculus Topics

A **power series** in the variable x is a formal sum

$$\sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + \cdots. \tag{2}$$

It is called a **convergent series** at x provided the limit below exists:

$$\lim_{N\to\infty} \sum_{n=0}^{N} c_n x^n = L.$$

The value L is a finite number called the **sum** of the series, written usually as $L = \sum_{n=0}^{\infty} c_n x^n$. Otherwise, the power series is called **divergent**. Convergence of the power series for every x in some interval J is called **convergence on** J. Similarly, **divergence on** J means the power series fails to have a limit at each point x of J. The series is said to **converge absolutely** if the series of absolute values $\sum_{n=0}^{\infty} |c_n||x|^n$ converges.

Given a power series $\sum_{n=0}^{\infty} c_n x^n$, define the **radius of convergence** R by the equation

$$R = \lim_{n\to\infty} \left| \frac{c_n}{c_{n+1}} \right|. \tag{3}$$

The radius of convergence R is undefined if the limit does not exist. Radius $R = \infty$ is common (it does *not* mean undefined).

Theorem 12.1 (Maclaurin Expansion)
If $f(x) = \sum_{n=0}^{\infty} c_n x^n$ converges for $|x| < R$, and $R > 0$, then f has infinitely many derivatives on $|x| < R$ and its coefficients $\{c_n\}$ are given by the **Maclaurin formula**

$$c_n = \frac{f^{(n)}(0)}{n!}. \tag{4}$$

The example $f(x) = e^{-1/x^2}$ shows the theorem has no converse. The following basic result summarizes what typically appears in calculus texts.

Theorem 12.2 (Convergence of power series)
Let the power series $\sum_{n=0}^{\infty} c_n x^n$ have radius of convergence R. If $R = 0$, then the series converges for $x = 0$ only. If $R = \infty$, then the series converges for all x. If $0 < R < \infty$, then

1. The series $\sum_{n=0}^{\infty} c_n x^n$ converges absolutely if $|x| < R$.
2. The series $\sum_{n=0}^{\infty} c_n x^n$ diverges if $|x| > R$.
3. The series $\sum_{n=0}^{\infty} c_n x^n$ may converge or diverge if $|x| = R$. The **interval of convergence** may be of the form $-R < x < R$, $-R \le x < R$, $-R < x \le R$ or $-R \le x \le R$.

Library of Maclaurin Series

The key Maclaurin series formulas used in applications are recorded below.

Geometric Series: $\dfrac{1}{1-x} = \displaystyle\sum_{n=0}^{\infty} x^n$	Converges for $-1 < x < 1$.
Log Series: $\ln(1+x) = \displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}x^n}{n}$	Converges for $-1 < x \le 1$.
Exponential Series: $e^x = \displaystyle\sum_{n=0}^{\infty} \frac{x^n}{n!}$	Converges for all x.
Cosine Series: $\cos x = \displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$	Converges for all x.
Sine Series: $\sin x = \displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$	Converges for all x.

Theorem 12.3 (Properties of power series)
Given two power series $\sum_{n=0}^{\infty} b_n x^n$ and $\sum_{n=0}^{\infty} c_n x^n$ with radii of convergence R_1, R_2, respectively, define $R = \min(R_1, R_2)$, so that both series converge for $|x| < R$. The power series have these properties:

1. $\sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} c_n x^n$ for $|x| < R$ implies $b_n = c_n$ for all n.

3. $\sum_{n=0}^{\infty} b_n x^n + \sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty} (b_n + c_n) x^n$ for $|x| < R$.

4. $k \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} k b_n x^n$ for all constants k, $|x| < R_1$.

5. $\frac{d}{dx} \sum_{n=0}^{\infty} b_n x^n = \sum_{n=1}^{\infty} n b_n x^{n-1}$ for $|x| < R_1$.

6. $\int_a^b \left(\sum_{n=0}^{\infty} b_n x^n\right) dx = \sum_{n=0}^{\infty} b_n \int_a^b x^n dx$ for $-R_1 < a < b < R_1$.

Taylor Series

A series expansion of the form

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

is called a **Taylor series** expansion of $f(x)$ about $x = x_0$. If valid, then the series converges and represents $f(x)$ for an interval of convergence $|x - x_0| < R$. Taylor expansions are general-use extensions of Maclaurin expansions, obtained by translation $x \to x - x_0$. If a Taylor series exists, then $f(x)$ has infinitely many derivatives. Therefore, the examples $|x|$ and x^α $(0 < \alpha < 1)$ fail to have Taylor expansions about $x = 0$. On the other hand, e^{-1/x^2} has infinitely many derivatives, but no Taylor expansion at $x = 0$.

Exercises 12.1

Series Convergence

Find R, the radius of convergence.

1. $\sum_{k=2}^{\infty} \frac{x^k}{k\ \ln(k)}$

2. $\sum_{k=1}^{\infty} a_k\, x^k$, $a_{2n} = 2$, $a_{2n+1} = 4$.

Series Properties

Compute the series given by the indicated operation(s).

3. $\frac{d}{dx} \sum_{k=2}^{\infty} \frac{x^k}{k\ \ln(k)}$

4. $4 \sum_{k=1}^{\infty} \frac{1}{1+k}\, x^k + \sum_{k=2}^{\infty} \frac{1}{1+k^2}\, x^k$

Maclaurin Series

Find the Maclaurin series expansion.

5. $f(x) = \frac{1}{1+x^3}$ for $|x| < 1$.

6. $f(x) = \arctan(x)$, using $\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$.

7. $f(x) = \left(\frac{3}{2}\right)^x$ for all x.

8. $f(x) = \int_0^x \frac{\sin t}{t}\, dt$, called the **Sine Integral**.

9. $f(x)$ is the solution of $f' = 1 + xf$, $f(0) = 0$.

10. The first 4 terms, $f(x) = \tan x$.

Taylor Series

Find the series expansion about the given point.

11. $f(x) = \ln|1 - x|$, at $x = 0$.

12. $f(x) = \frac{1}{x^2}$, at $x = 1$.

12.2 Algebraic Techniques

Derivative Formulas

Differential equations are solved with series techniques by assuming a **trial solution** of the form

$$y(x) = \sum_{n=0}^{\infty} c_n (x - x_0)^n.$$

The trial solution is thought to have **undetermined coefficients** $\{c_n\}$, to be found explicitly by the method of undetermined coefficients, i.e., substitute the trial solution and its derivatives into the differential equation and resolve the constants. The various derivatives of $y(x)$ can be written as power series. Below are the mostly commonly used derivative formulas.

$$\begin{aligned}
y(x) &= \sum_{n=0}^{\infty} c_n (x - x_0)^n, \\
y'(x) &= \sum_{n=1}^{\infty} n c_n (x - x_0)^{n-1}, \\
y''(x) &= \sum_{n=2}^{\infty} n(n-1) c_n (x - x_0)^{n-2}, \\
y'''(x) &= \sum_{n=3}^{\infty} n(n-1)(n-2) c_n (x - x_0)^{n-3}.
\end{aligned}$$

The summations are over a different subscript range in each case, because differentiation eliminates the constant term each time it is applied.

Changing Subscripts

A change of variable $t = x - a$ changes an integral $\int_a^\infty f(x)dx$ into $\int_0^\infty f(t+a)dt$. This change of variable is indicated when several integrals are added, because then the interval of integration is $[0, \infty)$, allowing the various integrals to be collected on one integral sign. For instance,

$$\int_2^\infty f(x)dx + \int_\pi^\infty g(x)dx = \int_0^\infty (f(t+2) + g(t+\pi))dt.$$

A similar change of variable technique is possible for summations, allowing several summation signs with different limits of summation to be collected under one summation sign. The rule:

$$\sum_{n=a}^{n=a+h} x_n = \sum_{k=0}^{h} x_{k+a}.$$

It is remembered via the change of variable $k = n - a$, which is formally applied to the summation just as it is applied in integration theory. If $h = \infty$, then the rule reads as follows:

$$\sum_{n=a}^{\infty} x_n = \sum_{k=0}^{\infty} x_{k+a}.$$

An illustration, in which LHS refers to the substitution of a trial solution into the left hand side of some differential equation:

$$\begin{aligned}
\text{LHS} &= \textstyle\sum_{n=2}^{\infty} n(n-1)c_n x^{n-2} + 2x\sum_{n=0}^{\infty} c_n x^n & \boxed{\mathbf{1}}\\
&= \textstyle\sum_{k=0}^{\infty}(k+2)(k+1)c_{k+2}x^k + \sum_{n=0}^{\infty} 2c_n x^{n+1} & \boxed{\mathbf{2}}\\
&= 2c_0 + \textstyle\sum_{k=1}^{\infty}(k+2)(k+1)c_{k+2}x^k + \sum_{k=1}^{\infty} 2c_{k-1}x^k & \boxed{\mathbf{3}}\\
&= 2c_0 + \textstyle\sum_{k=1}^{\infty}((k+2)(k+1)c_{k+2} + 2c_{k-1})x^k. & \boxed{\mathbf{4}}
\end{aligned}$$

Step details:

$\boxed{\mathbf{1}}$ is the result of substitution of the trial solution into the differential equation $y'' + 2xy$;

$\boxed{\mathbf{2}}$ makes a change of index variable $k = n - 2$;

$\boxed{\mathbf{3}}$ makes a change of index variable $k = n + 1$;

$\boxed{\mathbf{4}}$ adds the two series, which now have the same range of summation and equal powers of x.

The change of index variable in each case was dictated by attempting to match the powers of x, e.g., $x^{n-2} = x^k$ in $\boxed{\mathbf{2}}$ and $x^{n+1} = x^k$ in $\boxed{\mathbf{3}}$.

The formulas for derivatives of a trial solution $y(x)$ can all be written with the same index of summation, if desired:

$$\begin{aligned}
y(x) &= \sum_{n=0}^{\infty} c_n(x-x_0)^n,\\
y'(x) &= \sum_{n=0}^{\infty}(n+1)c_{n+1}(x-x_0)^n,\\
y''(x) &= \sum_{n=0}^{\infty}(n+2)(n+1)c_{n+2}(x-x_0)^n,\\
y'''(x) &= \sum_{n=0}^{\infty}(n+3)(n+2)(n+1)c_{n+3}(x-x_0)^n.
\end{aligned}$$

Linearity and Power Series

The set of all power series convergent for $|x| < R$ forms a vector space under function addition and scalar multiplication. This means:

1. The sum of two power series is a power series.

2. A scalar multiple of a power series is a power series.

3. The zero power series is the zero function: all coefficients are zero.

4. The negative of a power series is (-1) times the power series.

Cauchy Product

Multiplication and division of power series is possible and the result is again a power series convergent on some interval $|x| < R$. The **Cauchy product** of two series is defined by the relations

$$\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{m=0}^{\infty} b_m x^m\right) = \sum_{k=0}^{\infty} c_k x^k, \quad c_k = \sum_{n=0}^{k} a_n b_{k-n}.$$

Division of two series can be defined by its equivalent Cauchy product formula, which determines the coefficients of the quotient series.

To illustrate, we compute the coefficients $\{c_n\}$ in the formula

$$\sum_{n=0}^{\infty} c_n x^n = \left(\sum_{k=0}^{\infty} \frac{x^k}{k+1}\right) / \left(\sum_{m=0}^{\infty} x^m\right).$$

Limitations exist: the division is allowed only when the denominator is nonzero. In the present example, the denominator sums to $1/(1-x)$, which is never zero. The equivalent Cauchy product relation is

$$\left(\sum_{n=0}^{\infty} c_n x^n\right)\left(\sum_{m=0}^{\infty} x^m\right) = \sum_{k=0}^{\infty} \frac{x^k}{k+1}.$$

This relation implies the formula

$$\sum_{n=0}^{k} (c_n)(1) = \frac{1}{k+1}.$$

Therefore, back-substitution implies $c_0 = 1$, $c_1 = -1/2$, $c_2 = -1/6$. More coefficients can be found and perhaps also a general formula can be written for c_n. A general formula is needed infrequently, so we spend no time discussing how to find it.

Power Series Expansions of Rational Functions

A rational function $f(x)$ is a quotient of two polynomials, therefore it is a quotient of two power series, hence also a power series. Sometimes the easiest method known to find the coefficients c_n of the power series of f is to apply Maclaurin's formula

$$c_n = \frac{f^{(n)}(0)}{n!}.$$

In a number of limited cases, in which the polynomials have low degree, it is possible to use Cauchy's product formula to find $\{c_n\}$. An illustration:

$$\frac{x+1}{x^2+1} = \sum_{n=0}^{\infty} c_n x^n, \quad c_{2k+1} = c_{2k} = (-1)^k.$$

To derive this formula, write the quotient as a Cauchy product:

$$\begin{aligned} x+1 &= (1+x^2)\sum_{n=0}^{\infty} c_n x^n \\ &= \sum_{n=0}^{\infty} c_n x^n + \sum_{m=0}^{\infty} c_m x^{m+2} \\ &= c_0 + c_1 x + \sum_{n=2}^{\infty} c_n x^n + \sum_{k=2}^{\infty} c_{k-2} x^k \\ &= c_0 + c_1 x + \sum_{k=2}^{\infty} (c_k + c_{k-2}) x^k \end{aligned}$$

The third step uses variable change $k = m + 2$. The terms on the right then have the same index range, allowing the addition of the final step. To match coefficients on each side of the equation, we require $c_0 = 1$, $c_1 = 1$, $c_k + c_{k-2} = 0$. Solving, $c_2 = -c_0$, $c_3 = -c_1$, $c_4 = -c_2 = (-1)^2 c_0$, $c_5 = -c_3 = (-1)^2 c_1$. By induction, $c_{2k} = (-1)^k$ and $c_{2k+1} = (-1)^k$. This gives the series reported earlier.

The same series expansion can be obtained in a more intuitive manner, as follows. The idea depends upon substitution of $r = -x^2$ into the geometric series expansion $(1-r)^{-1} = 1 + r + r^2 + \cdots$, which is valid for $|r| < 1$.

$$\begin{aligned} \frac{x+1}{x^2+1} &= (1+x)\sum_{n=0}^{\infty} r^n \quad \text{where } r = -x^2 \\ &= \sum_{n=0}^{\infty} (-x^2)^n + x\sum_{n=0}^{\infty} (-x^2)^n \\ &= \sum_{n=0}^{\infty} (-1)^n x^{2n} + \sum_{n=0}^{\infty} (-1)^n x^{2n+1} \\ &= \sum_{k=0}^{\infty} c_k x^k, \end{aligned}$$

where $c_{2k} = (-1)^k$ and $c_{2k+1} = (-1)^k$. The latter method is preferred to discover a useful formula. The method is a shortcut to the expansion of $1/(x^2 + 1)$ as a Maclaurin series, followed by series properties to write the indicated Cauchy product as a single power series.

Instances exist where neither the Cauchy product method nor other methods are easy, for instance, the expansion of $f(x) = 1/(x^2 + x + 1)$. Here, we might find a formula from $c_n = f^{(n)}(0)/n!$, or equally unpleasant, find $\{c_n\}$ from the formula $1 = (x^2 + x + 1)\sum_{n=0}^{\infty} c_n x^n$.

Recursion Relations

The relations

$$c_0 = 1, \quad c_1 = 1, \quad c_k + c_{k-2} = 0 \text{ for } k \geq 2$$

are called **recursion relations**. They are often solved by ad hoc algebraic methods. Developed here is a systematic method for solving such recursions.

First order recursions. Given x_0 and sequences of constants $\{a_n\}_{n=0}^{\infty}$, $\{b_n\}_{n=0}^{\infty}$, consider the abstract problem of finding a formula for x_k in the recursion relation

$$x_{k+1} = a_k x_k + b_k, \quad k \geq 0.$$

For $k = 0$ the formula gives $x_1 = a_0 x_0 + b_0$. Similarly, $x_2 = a_1 x_1 + b_1 = a_1 a_0 x_0 + a_1 b_0 + b_1$, $x_3 = a_2 x_2 + b_2 = a_2 a_1 a_0 x_0 + a_2 a_1 b_0 + a_2 b_1 + b_2$. By induction, the unique solution is

$$x_{k+1} = \left(\Pi_{r=0}^{k} a_r\right) x_0 + \sum_{n=0}^{k} \left(\Pi_{r=n+1}^{k} a_r\right) b_n.$$

Two-termed second order recursions. Given c_0, c_1 and sequences $\{a_k\}_{k=0}^{\infty}$, $\{b_k\}_{k=0}^{\infty}$, consider the problem of solving for c_{k+2} in the two-termed second order recursion

$$c_{k+2} = a_k c_k + b_k, \quad k \geq 0.$$

The idea to solve it comes from splitting the problem into even and odd subscripts. For even subscripts, let $k = 2n$. For odd subscripts, let $k = 2n+1$. Then the two-termed second order recursion splits into two first order recursions

$$\begin{aligned} c_{2n+2} &= a_{2n} c_{2n} + b_{2n}, \quad n \geq 0, \\ c_{2n+3} &= a_{2n+1} c_{2n+1} + b_{2n+1}, \quad n \geq 0. \end{aligned}$$

Define $x_n = c_{2n}$ or $x_n = c_{2n+1}$ and apply the general theory for first order recursions to solve the above recursions:

$$\begin{aligned} c_{2n+2} &= \left(\Pi_{r=0}^{n} a_{2r}\right) c_0 + \sum_{k=0}^{n} \left(\Pi_{r=k+1}^{n} a_{2r}\right) b_{2r}, \quad n \geq 0, \\ c_{2n+3} &= \left(\Pi_{r=0}^{n} a_{2r+1}\right) c_1 + \sum_{k=0}^{n} \left(\Pi_{r=k+1}^{n} a_{2r+1}\right) b_{2r+1}, \quad n \geq 0. \end{aligned}$$

Two-termed third order recursions. Given c_0, c_1, c_2, $\{a_k\}_{k=0}^{\infty}$, $\{b_k\}_{k=0}^{\infty}$, consider the problem of solving for c_{k+3} in the two-termed third order recursion

$$c_{k+3} = a_k c_k + b_k, \quad k \geq 0.$$

The subscripts are split into three groups by the equations $k = 3n$, $k = 3n + 1$, $k = 3n + 2$. Then the third order recursion splits into three first order recursions, each of which is solved by the theory of first order recursions. The solution for $n \ge 0$:

$$\begin{aligned}
c_{3n+3} &= \left(\Pi_{r=0}^{n} a_{3r}\right) c_0 + \sum_{k=0}^{n} \left(\Pi_{r=k+1}^{n} a_{3r}\right) b_{3r},\\
c_{3n+4} &= \left(\Pi_{r=0}^{n} a_{3r+1}\right) c_1 + \sum_{k=0}^{n} \left(\Pi_{r=k+1}^{n} a_{3r+1}\right) b_{3r+1},\\
c_{3n+5} &= \left(\Pi_{r=0}^{n} a_{3r+2}\right) c_2 + \sum_{k=0}^{n} \left(\Pi_{r=k+1}^{n} a_{3r+2}\right) b_{3r+2}.
\end{aligned}$$

Exercises 12.2

Differentiation

Verify using term–by–term differentiation. Document all series and calculus steps.

1. $\frac{d}{dx}\sum_{n=1}^{\infty}\frac{1}{n}\,x^n = \sum_{n=0}^{\infty} x^n$.
Is this valid for $x = -1$?

2. $\frac{d}{dx}\sum_{n=0}^{\infty}(-1)^n\,x^{2n+1} = \sum_{n=0}^{\infty}(-1)^n\,x^{2n}$.

Subscripts

Perform a change of variables to verify the identity.

3. $\sum_{n=0}^{\infty} c_n x^{n+2} = \sum_{k=2}^{\infty} c_{k-2}\,x^k$

4. $\sum_{n=2}^{\infty} n(n-1)c_n(x-x_0)^{n-2} = \sum_{k=0}^{\infty}(k+2)(k+1)c_{k+2}\,(x-x_0)^k$

5. $-1+x+\sum_{n=2}^{\infty}(-1)^{n+1}\,x^n = \sum_{k=0}^{\infty}(-1)^{k+1}\,x^k$

6. $\sum_{n=0}^{\infty}\frac{1}{n+1}\,x^n + \sum_{m=1}^{\infty}\frac{1}{m+2}\,x^m = 1+\sum_{k=1}^{\infty}\frac{2k+1}{(k+1)(k+2)}\,x^k$

Linearity

Find the power series of the given function.

7. $\cos(x) + 2\sin(x)$

8. $e^x + \sin(x)$

Cauchy Product

Find the power series.

9. $(1+x)\sin(x)$

10. $\frac{\sin(x)}{e^x}$

Recursion Relations

Solve the given recursion.

11. $x_{k+1} = 2x_k$

12. $x_{k+1} = 2x_k + 1$

13. $x_{k+2} = 2x_k + 1$

14. $x_{k+3} = 2x_k + 1$

12.3 Power Series Methods

Detailed below are trial solution methods for first and second order differential equations. A **trial solution** is an infinite series, a Maclaurin expansion or a Taylor series expansion about $x = x_0$. Techniques for trial solution methods involve series methods, undetermined coefficients and algebraic results to solve recursions. The Taylor series method employs the calculus Taylor polynomial formula and requires only a calculus background.

A Series Method for First Order

Illustrated here is a method to solve the differential equation $y' - 2y = 0$ for a power series solution. Assume a power series **trial solution**

$$y(x) = \sum_{n=0}^{\infty} c_n x^n.$$

Let LHS stand for the left hand side of $y' - 2y = 0$. Substitute the trial series solution into LHS to obtain:

(1)

$$\begin{aligned}
\text{LHS} &= y' - 2y \\
&= \sum_{n=1}^{\infty} n c_n x^{n-1} - 2\sum_{n=0}^{\infty} c_n x^n \\
&= \sum_{k=0}^{\infty}(k+1)c_{k+1}x^k + \sum_{n=0}^{\infty}(-2)c_n x^n \quad \boxed{\mathbf{1}} \\
&= \sum_{k=0}^{\infty}((k+1)c_{k+1} - 2c_k)x^k \quad \boxed{\mathbf{2}}
\end{aligned}$$

(2)

The change of variable $k = n-1$ was used in $\boxed{\mathbf{1}}$, the objective being to add on like powers of x in $\boxed{\mathbf{2}}$. Assume LHS = 0. The zero function is **uniquely** represented by the power series with all zero coefficients. By uniqueness, all coefficients in the series for LHS must be zero, which gives the recursion relation

$$(k+1)c_{k+1} - 2c_k = 0, \quad k \geq 0.$$

This first order two-termed recursion is solved by back-substitution or by using the general theory for first order recursions which is in the preceding section, page 957. Using the results, then

$$\begin{aligned}
c_{k+1} &= \left(\Pi_{r=0}^{k} \frac{2}{r+1}\right) c_0 \\
&= \frac{2^{k+1}}{(k+1)!} c_0.
\end{aligned}$$

The **trial solution** becomes a **power series solution**:

$$\begin{aligned}
y(x) &= c_0 + \sum_{k=0}^{\infty} c_{k+1} x^{k+1} && \text{Re-index the trial solution.} \\
&= c_0 + \sum_{k=0}^{\infty} \frac{2^{k+1}}{(k+1)!}\, c_0\, x^{k+1} && \text{Substitute the recursion answer.} \\
&= c_0 + \left(\sum_{n=1}^{\infty} \frac{2^n}{(n)!}\, x^n \right) c_0 && \text{Change index } n = k+1. \\
&= \left(\sum_{n=0}^{\infty} \frac{(2x)^n}{(n)!} \right) c_0 && \text{Factor out } c_0\text{, then reindex.} \\
&= e^{2x} c_0. && \text{Maclaurin expansion library.}
\end{aligned}$$

The **solution** $y(x) = c_0 e^{2x}$ agrees with the growth-decay theory formula for the first order differential equation $y' = ky$ ($k = 2$ in this case).

A Series Method for Second Order

Shown here are the details for finding two independent power series solutions

$$y_1(x) = 1 + \frac{1}{6}x^3 + \frac{1}{180}x^6 + \frac{1}{12960}x^9 + \frac{1}{1710720}x^{12} + \cdots$$
$$y_2(x) = x + \frac{1}{12}x^4 + \frac{1}{504}x^7 + \frac{1}{45360}x^{10} + \frac{1}{7076160}x^{13} + \cdots$$

for Airy's airfoil differential equation

$$y'' = xy.$$

The two independent solutions give the general solution as

$$y(x) = c_1 y_1(x) + c_2 y_2(x).$$

The solutions are related to the classical **Airy wave functions**, denoted `AiryAi` and `AiryBi` in the literature, and documented for example in the computer algebra system `maple`. The wave functions `AiryAi`, `AiryBi` are special linear combinations of y_1, y_2.

The **trial solution** in the second order power series method is generally a Taylor series. In this case, it is a Maclaurin series

$$y(x) = \sum_{n=0}^{\infty} c_n x^n.$$

Write Airy's differential equation in standard form $y'' - xy = 0$ and let LHS stand for the left hand side of this equation. Then substitution of the trial solution into LHS gives:

$$\begin{aligned}
\text{LHS} &= y'' - xy \\
&= \sum_{n=0}^{\infty}(n+2)(n+1)c_{n+2}x^n - x\sum_{k=0}^{\infty}c_kx^k \quad \boxed{\mathbf{1}} \\
&= \sum_{n=0}^{\infty}(n+2)(n+1)c_{n+2}x^n - \sum_{k=0}^{\infty}c_kx^{k+1} \quad \boxed{\mathbf{2}} \\
&= 2c_2 + \sum_{n=1}^{\infty}(n+2)(n+1)c_{n+2}x^n - \sum_{n=1}^{\infty}c_{n-1}x^n \quad \boxed{\mathbf{3}} \\
&= 2c_2 + \sum_{n=1}^{\infty}\left((n+2)(n+1)c_{n+2} - c_{n-1}\right)x^n \quad \boxed{\mathbf{4}}
\end{aligned}$$

The **steps**: $\boxed{\mathbf{1}}$ Substitute the trial solution into LHS using derivative formulas; $\boxed{\mathbf{2}}$ Move x inside the summation by linearity; $\boxed{\mathbf{3}}$ Index change $n = k+1$ to match powers of x; $\boxed{\mathbf{4}}$ Match summation index ranges and collect on powers of x.

Because LHS $= 0 =$ RHS and the power series for the zero function has zero coefficients, all coefficients in the series LHS must be zero. This implies the relations

$$c_2 = 0, \quad (n+2)(n+1)c_{n+2} - c_{n-1} = 0, \quad n \ge 1.$$

Replace n by $k+1$. Then the relations above become the two-termed third order recursion

$$c_{k+3} = \frac{1}{(k+2)(k+3)}c_k, \quad k \ge 0.$$

The answers are obtained from page 957, with appropriate definitions of a_k and b_k:

$$\begin{aligned}
c_{3n+3} &= \left(\Pi_{r=0}^{n}\frac{1}{(3r+2)(3r+3)}\right)c_0, \\
c_{3n+4} &= \left(\Pi_{r=0}^{n}\frac{1}{(3r+3)(3r+4)}\right)c_1, \\
c_{3n+5} &= \left(\Pi_{r=0}^{n}\frac{1}{(3r+4)(3r+5)}\right)c_2 \\
&= 0 \quad (\text{because } c_2 = 0).
\end{aligned}$$

Taking $c_0 = 1$, $c_1 = 0$ gives one solution

$$y_1(x) = 1 + \sum_{n=0}^{\infty}\left(\Pi_{r=0}^{n}\frac{1}{(3r+2)(3r+3)}\right)x^{3n+3}.$$

Taking $c_0 = 0$, $c_1 = 1$ gives a second independent solution

$$\begin{aligned}
y_2(x) &= x + \sum_{n=0}^{\infty}\left(\Pi_{r=0}^{n}\frac{1}{(3r+3)(3r+4)}\right)x^{3n+4} \\
&= x\left(1 + \sum_{n=0}^{\infty}\left(\Pi_{r=0}^{n}\frac{1}{(3r+3)(3r+4)}\right)x^{3n+3}\right).
\end{aligned}$$

Power Series Maple Code

It is possible to reproduce the first few terms (below, up to x^{20}) of the power series solutions y_1, y_2 using the computer algebra system `maple`. Here's how:

```
de1:=diff(y1(x),x,x)-x*y1(x)=0; Order:=20;
dsolve({de1,y1(0)=1,D(y1)(0)=0},y1(x),type=series);
de2:=diff(y2(x),x,x)-x*y2(x)=0;
dsolve({de2,y2(0)=0,D(y2)(0)=1},y2(x),type=series);
```

The `maple` global variable `Order` assigns the number of terms to compute in the series method for `dsolve()`.

The Airy wave functions are defined so that

$$\begin{array}{rcccl} \sqrt{3}\,\texttt{AiryAi}(0) & = & \texttt{AiryBi}(0) & \approx & 0.6149266276, \\ -\sqrt{3}\,\texttt{AiryAi}'(0) & = & \texttt{AiryBAi}'(0) & \approx & 0.4482883572. \end{array}$$

A warning: the Airy wave functions are *not identical* to y_1, y_2.

A Simple Taylor Polynomial Method

The first power series solution

$$y(x) = 1 + \frac{1}{6}x^3 + \frac{1}{180}x^6 + \frac{1}{12960}x^9 + \frac{1}{1710720}x^{12} + \cdots$$

for Airy's airfoil differential equation $y'' = xy$ can be found without knowing anything about recursion relations or properties of infinite series. Detailed here is a Taylor polynomial method which requires only a calculus background. The computation reproduces by hand the answer given by the `maple` code below.

```
    de:=diff(y(x),x,x)-x*y(x)=0; Order:=10;
    dsolve([de,y(0)=1,D(y)(0)=0],y(x),type=series);
```

The calculus background:

Theorem 12.4 (Taylor Polynomials)
Let $f(x)$ have $n+1$ continuous derivatives on $a < x < b$ and assume given x_0, $a < x_0 < b$. Then

$$(3) \qquad f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!} + R_n$$

where the remainder R_n has the form

$$R_n = f^{(n+1)}(x_1)\frac{(x - x_0)^{n+1}}{(n+1)!}$$

for some point x_1 between a and b.

The polynomial on the right in (3) is called the **Taylor polynomial** of degree n for $f(x)$ at $x = x_0$. If f is infinitely differentiable, then it has Taylor polynomials of all orders. The **Taylor series** of f is the infinite series obtained formally by letting $n = \infty$ and $R_n = 0$.

For the Airy differential equation problem, $x_0 = 0$. Let's assume that $y(x)$ is determined by initial conditions $y(0) = 1$, $y'(0) = 0$. The method is a simple one:

> Differentiate the differential equation formally several times, then set $x = x_0$ in all these equations. Resolve from the several equations the values of $y''(x_0)$, $y'''(x_0)$, $y^{iv}(x_0)$, …and then write out the Taylor polynomial approximation
>
> $$y(x) \approx y(x_0) + y'(x_0)(x - x_0) + y''(x_0)\frac{(x - x_0)^2}{2} + \cdots$$

The successive derivatives of Airy's differential equation are

$$\begin{aligned} y'' &= xy, \\ y''' &= y + xy', \\ y^{iv} &= 2y' + xy'', \\ y^{v} &= 3y'' + xy''', \\ &\vdots \end{aligned}$$

Set $x = x_0 = 0$ in the above equations. Then

$y(0) = 1$	Given.
$y'(0) = 0$	Given.
$y''(0) = xy\vert_{x=0}$ $= 0$	Use Airy's equation $y'' = xy$.
$y'''(0) = (y + xy')\vert_{x=0}$ $= 1$	Use $y''' = y + xy'$.
$y^{iv}(0) = (2y' + xy'')\vert_{x=0}$ $= 0$	Use $y^{iv} = 2y' + xy''$.
$y^{v}(0) = (3y'' + xy''')\vert_{x=0}$ $= 0$	Use $y^{v} = 3y'' + xy'''$.
$y^{vi}(0) = (4y''' + xy^{iv})\vert_{x=0}$ $= 4$	Use $y^{vi} = 4y''' + xy^{iv}$.

Finally, we write out the Taylor polynomial approximation of y:

$$\begin{aligned} y(x) &\approx y(0) + y'(0)x + y''(0)\frac{x^2}{2} + \cdots \\ &= 1 + 0 + 0 + \frac{x^3}{6} + 0 + 0 + \frac{4x^6}{6!} + \cdots \\ &= 1 + \frac{x^3}{6} + \frac{x^6}{180} + \cdots \end{aligned}$$

Computer algebra systems can replace the hand details, finding the Taylor polynomial directly.

Exercises 12.3 ↗

First Order Series Method

Solve by power series.

1. $y' - 4y = 0$

2. $y' - xy = 0$

Second Order Series Method

Solve by power series using the Airy equation example.

3. $y'' = 4y$

4. $y'' + y = 0$

Taylor Series Method

Solve by Taylor series about $x = 0$, finding the first 8 terms.

5. $y' = 16y$

6. $y'' = y$

7. $y' = (1 + x)y$

8. $y'' = (2 + x)y$

12.4 Ordinary Points

Developed here is the mathematical theory for 2nd order differential equations and their Taylor series solutions. Assume a differential equation

$$(1) \qquad a(x)y'' + b(x)y' + c(x)y = 0, \quad a(x) \neq 0.$$

Such an equation can always be converted by division of $a(x) \neq 0$ to the **standard form**

$$(2) \qquad y'' + p(x)y' + q(x)y = 0,$$

using formulas

$$p(x) = b(x)/a(x), \quad q(x) = c(x)/a(x).$$

A point $x = x_0$ is called an **Ordinary Point** of equation (2) provided both $p(x)$ and $q(x)$ have Taylor series expansions valid in an interval $|x - x_0| < R$, $R > 0$. Any point that is not an ordinary point is called a **Singular Point**. For equation (1), $x = x_0$ is an ordinary point provided $a(x) \neq 0$ at $x = x_0$ and each of $a(x)$, $b(x)$, $c(x)$ has a Taylor series expansion valid in some interval about $x = x_0$.

Theorem 12.5 (Power series solutions)
Let $a(x)y'' + b(x)y' + c(x)y = 0$, $a(x) \neq 0$, be given and assume that $x = x_0$ is an ordinary point. If the Taylor series of both $p(x) = b(x)/a(x)$ and $q(x) = c(x)/a(x)$ are convergent in $|x - x_0| < R$, then the differential equation has two independent Taylor series solutions

$$y_1(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n, \quad y_2(x) = \sum_{n=0}^{\infty} b_n (x - x_0)^n,$$

convergent in $|x - x_0| < R$. Any solution $y(x)$ defined in $|x - x_0| < R$ can be written as $y(x) = c_1 y_1(x) + c_2 y_2(x)$ for a unique set of constants c_2, c_2.

A proof of this result can be found in Birkhoff-Rota [BirkRota]. The maximum allowed value of R is the distance from x_0 to the nearest singular point.

Ordinary Point Illustration

Two independent solutions y_1, y_2 of Theorem 12.5 will be determined for the second order differential equation

$$y'' - 2xy' + y = 0.$$

Let LHS stand for the left side of the differential equation. Assume a trial solution $y = \sum_{n=0}^{\infty} c_n x^n$. Then formulas on pages 953 and 954 imply

$$\text{LHS} = y'' - 2xy' + y$$

$$
\begin{aligned}
&= \sum_{n=0}^{\infty}(n+1)(n+2)c_{n+2}x^n - 2x\sum_{n=1}^{\infty} nc_n x^{n-1} + \sum_{n=0}^{\infty} c_n x^n \\
&= \sum_{n=0}^{\infty}(n+1)(n+2)c_{n+2}x^n + \sum_{n=1}^{\infty}(-2)nc_n x^n + \sum_{n=0}^{\infty} c_n x^n \\
&= 2c_2 + c_0 + \sum_{n=1}^{\infty}((n+1)(n+2)c_{n+2} - 2nc_n + c_n)x^n \\
&= 2c_2 + c_0 + \sum_{n=1}^{\infty}((n+1)(n+2)c_{n+2} - (2n-1)c_n)x^n
\end{aligned}
$$

The power series LHS equals the zero power series, which gives rise to the recursion relations $2c_2 + c_0 = 0$, $(n+1)(n+2)c_{n+2} - (2n-1)c_n = 0$, $n \ge 1$, or more succinctly the two-termed second order recursion

$$c_{n+2} = \frac{2n-1}{(n+1)(n+2)}c_n, \quad n \ge 0.$$

Using the formulas on page 957, we obtain the recursion answers

$$
\begin{aligned}
c_{2k+2} &= \left(\Pi_{r=0}^{k}\frac{4r-1}{(2r+1)(2r+2)}\right)c_0, \\
c_{2k+3} &= \left(\Pi_{r=0}^{k}\frac{4r+1}{(2r+2)(2r+3)}\right)c_1.
\end{aligned}
$$

Taking $c_0 = 1$, $c_1 = 0$ gives y_1 and taking $c_0 = 0$, $c_1 = 1$ gives y_2:

$$
\begin{aligned}
y_1(x) &= 1 + \sum_{k=0}^{\infty}\left(\Pi_{r=0}^{k}\frac{4r-1}{(2r+1)(2r+2)}\right)x^{2k+2}, \\
y_2(x) &= x + \sum_{k=0}^{\infty}\left(\Pi_{r=0}^{k}\frac{4r+1}{(2r+2)(2r+3)}\right)x^{2k+3}.
\end{aligned}
$$

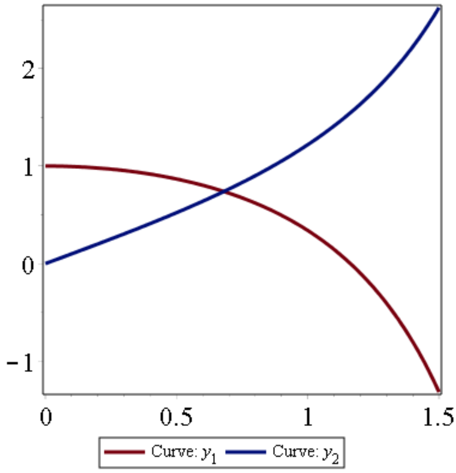
These solutions have Wronskian 1 at $x = 0$, hence they are independent and they form a basis for the solution space of the differential equation.

Plots and Computation in `maple`

It is possible to directly program the basis y_1, y_2 in `maple`, ready for plotting and computation of solutions to initial value problems. At the same time, we can check the series formulas against the `maple` engine, which is able to solve for the series solutions y_1, y_2 to any order of accuracy.

```
f:=t->(2*t-1)/((t+1)*(t+2)):
c1:=k->product(f(2*r),r=0..k):
c2:=k->product(f(2*r+1),r=0..k):
```

```
y1:=(x,N)->1+sum(c1(k)*x^(2*k+2),k=0..N);
y2:=(x,N)->x+sum(c2(k)*x^(2*k+3),k=0..N);
de:=diff(y(x),x,x)-2*x*diff(y(x),x)+y(x)=0: Order:=10:
dsolve({de,y(0)=1,D(y)(0)=0},y(x),type=series); # find y1
'y1'=y1(x,5);
dsolve({de,y(0)=0,D(y)(0)=1},y(x),type=series); # find y2
'y2'=y2(x,5);
opts:=font=[courier,18],axes=boxed,thickness=3;
plot(2*y1(x,infinity)+3*y2(x,infinity),x=0..3);
plot([y1(x,infinity),y2(x,infinity)],x=0..1.5,opts);
```



The `maple` `dsolve` formulas are

$$y_1(x) = 1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 - \frac{7}{240}x^6 - \frac{11}{1920}x^8 + \cdots$$

$$y_2(x) = x + \frac{1}{6}x^3 + \frac{1}{24}x^5 + \frac{1}{112}x^7 + \frac{13}{8064}x^9 + \cdots$$

Approximation of $2y_1+3y_2$ to order 20 agrees with the exact solution for the first 8 digits. Often the N =infinity required for the exact solution can be replaced by integer $N = 10$ to produce exactly the same plot.

Exercises 12.4

Standard Form

Convert to form $y'' + p(x)y' + q(x)y = 0$. Find the singular points and ordinary points.

1. $(x+1)y'' + xy' + y = 0$

2. $x^2y'' + 3xy' + 4y = 0$

3. $x(1+x)y'' + xy' + (1+x)y = 0$

4. $xy'' = (1+x)y' + e^x y$

Ordinary Point Method

Find a power series solution, following the method in the text for $y'' - 2xy' + y = 0$. Use a CAS or mathematical workbench to check the answer.

5. $y'' + xy' = 0$

6. $y'' + x^2y' + y = 0$

12.5 Regular Singular Points

The model differential equation for Frobenius singular point theory is the 2nd order **Cauchy-Euler differential equation**

$$ax^2y'' + bxy' + cy = 0. \tag{1}$$

The Frobenius theory treats a **perturbation** of the Cauchy-Euler equation obtained by replacement of the constants a, b, c by Maclaurin power series. A **Frobenius differential equation** has the special form

$$x^2a(x)y'' + xb(x)y' + c(x)y = 0$$

where $a(x) \neq 0$, $b(x)$, $c(x)$ have Maclaurin series expansions.

Intuition from the Cauchy-Euler Equation

The Cauchy-Euler differential equation (1) provides intuition about the possible kinds of solutions for Frobenius equations. It is known that equation (1) can be transformed to a constant-coefficient differential equation

$$a\frac{d^2z}{dt^2} + (b-a)\frac{dz}{dt} + cz = 0 \tag{2}$$

via the change of variables

$$z(t) = y(e^t), \quad x = e^t.$$

By constant-coefficient formulas from Chapter 6, Theorem 6.1 page 431, a Cauchy-Euler equation (1) has three kinds of possible solutions, organized by the character of the roots r_1, r_2 of the characteristic equation $ar^2 + (b-a)r + c = 0$ of (2). The three kinds are

Case 1:	Discriminant positive Real $r_1 \neq r_2$	$y = c_1x^{r_1} + c_2x^{r_2}$
Case 2:	Discriminant zero Real $r_1 = r_2$	$y = c_1x^{r_1} + c_2x^{r_1}\ln\lvert x\rvert$
Case 3:	Discriminant negative Complex $r_1 = \overline{r}_2 = \alpha + i\beta$	$y = c_1x^{\alpha}\cos(\beta\ln\lvert x\rvert)$ $+ c_2x^{\alpha}\sin(\beta\ln\lvert x\rvert)$

The last solution is **Singular** at $x = 0$, the location where the leading coefficient ax^2 in (1) is zero. The second solution is singular at $x = 0$ when $c_2 \neq 0$. The other solutions involve powers x^r; they can be singular solutions at $x = 0$ if $r < 0$.

Cauchy-Euler Conjecture

The conjecture about solutions of Frobenius equations is often made by differential equation rookies:

> Isn't it true that a Frobenius differential equation has a general solution obtained from the general solution of the Cauchy-Euler differential equation
>
> $$x^2a(0)y'' + xb(0)y' + c(0)y = 0$$
>
> by replacement of the constants c_1, c_2 by Maclaurin power series?

As a tribute to this intuitive conjecture, we can say in hindsight that the **Cauchy-Euler conjecture** is *almost correct*! Perhaps it is a good way to remember the results of the Frobenius theory which follows.

Frobenius theory

A **Frobenius differential equation** singular at $x = x_0$ has the form

$$(x - x_0)^2 A(x)y'' + (x - x_0)B(x)y' + C(x)y = 0 \tag{3}$$

where $A(x_0) \neq 0$ and $A(x)$, $B(x)$, $C(x)$ have Taylor series expansions at $x = x_0$ valid in an interval $|x - x_0| < R$, $R > 0$. Such a point $x = x_0$ is called a **regular singular point** of (3). Any other point $x = x_0$ is called an **irregular singular point**.

A Frobenius regular singular point differential equation generalizes the Cauchy-Euler differential equation, because if the Taylor series are constants and the translation $x \to x - x_0$ is made, then the Frobenius equation reduces to a Cauchy-Euler equation.

The **Indicial Equation** of (3) is defined to be the quadratic equation

$$A(x_0)r^2 + (B(x_0) - A(x_0))r + C(x_0) = 0.$$

Technically, the definition is a useful shortcut, because the indicial equation is obtained by calculation in two steps:

(1) Transform the Cauchy-Euler differential equation

$$(x - x_0)^2 A(x_0)y'' + (x - x_0)B(x_0)y' + C(x_0)y = 0$$

by the change of variables $x - x_0 = e^t$, $z(t) = y(x_0 + e^t)$ to obtain the constant-coefficient differential operator form

$$A(x_0)(D - 1)Dz + B(x_0)Dz + C(x_0)z = 0, \quad D = \frac{d}{dt}.$$

The expanded constant-coefficient equation is

$$A(x_0)\frac{d^2z}{dt^2} + (B(x_0) - A(x_0))\frac{dz}{dt} + C(x_0)z = 0$$

(2) The indicial equation is the characteristic equation of the constant-coefficient differential equation.

The indicial equation can be used to directly solve Cauchy-Euler differential equations. The roots of the indicial equation plus the constant-coefficient formulas in Theorem 6.1 provide answers which directly transcribe the general solution of the Cauchy-Euler equation.

The Frobenius theory analyzes the Frobenius differential equation only in the case when the roots of the indicial equation are real, which corresponds to the discriminant positive or zero in the discriminant table, page 968.

The cases in which the discriminant is non-negative have their own complications. Expected from the Cauchy-Euler conjecture is a so-called **Frobenius solution**

$$y(x) = (x - x_0)^r \left(c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + \cdots\right),$$

in which r is a root of the indicial equation. Two independent Frobenius solutions may or may not exist, therefore the Cauchy-Euler conjecture turns out to be partly true, but false in general.

The last case, in which the discriminant of the indicial equation is negative, is not treated here.

Theorem 12.6 (Frobenius Solutions)
Let $x = x_0$ be a regular singular point of the Frobenius equation

$$(x - x_0)^2 A(x)y'' + (x - x_0)B(x)y' + C(x)y = 0. \tag{4}$$

Let the indicial equation $A(x_0)r^2 + (B(x_0) - A(x_0))r + C(x_0) = 0$ have real roots r_1, r_2 with $r_1 \geq r_2$. Then equation (4) always has one Frobenius series solution y_1 of the form

$$y_1(x) = (x - x_0)^{r_1} \sum_{n=0}^{\infty} c_n (x - x_0)^n, \quad c_0 \neq 0.$$

The root r_1 has to be the larger root: the equation can fail for the smaller root r_2.

Equation (4) has a second independent solution y_2 in the following cases.

(a) If $r_1 \neq r_2$ and $r_1 - r_2$ is not an integer, then, for some coefficients $\{d_n\}$ with $d_0 \neq 0$,

$$y_2(x) = (x - x_0)^{r_2} \sum_{n=0}^{\infty} d_n (x - x_0)^n.$$

(b) If $r_1 \neq r_2$ and $r_1 - r_2$ is a positive integer, then, for some coefficients $\{d_n\}$ with $d_0 \neq 0$ and either $C = 0$ or $C = 1$,

$$y_2(x) = C y_1(x) \ln|x - x_0| + (x - x_0)^{r_2} \sum_{n=0}^{\infty} d_n (x - x_0)^n.$$

(c) If $r_1 = r_2$, then, for some coefficients $\{d_n\}$ with $d_0 = 0$,

$$y_2(x) = y_1(x)\ln|x - x_0| + (x - x_0)^{r_1}\sum_{n=0}^{\infty} d_n(x - x_0)^n.$$

Proof: A Frobenius theorem proof can be found in Birkhoff-Rota [BirkRota] 4th edition page 282. The method of proof, due to Frobenius, is a generalization of Cauchy's Method of Majorants [BirkRota] page 113. ■

Independence tests for y_1, y_2 plus calculation details for y_1, y_2 appear below in the examples. In part **(b)** of the theorem, the formula compresses two trial solutions into one, but the intent is that they be tried separately, in order $C = 0$, then $C = 1$. Sometimes it is possible to combine the two trials into one complicated computation, but that is not for the faint of heart.

The examples use symbol $L(y)$, defined by

$$L(y) = (x - x_0)^2 A(x)y'' + (x - x_0)B(x)y' + C(x)y,$$

which is the left hand side of the Frobenius equation (4). Implicit use is made of the **linearity property** $L(c_1y_1 + c_2y_2) = c_1L(y_1) + c_2L(y_2)$.

Example 12.1 (Frobenius Theorem Case (a))

Use the Frobenius theory to solve for y_1, y_2 in the differential equation $2x^2y'' + xy' + xy = 0$.

Solution: The indicial equation is $2r^2 + (1 - 2)r + 0 = 0$ with roots $r_1 = 1/2$, $r_2 = 0$. The roots do not differ by an integer, therefore two independent Frobenius solutions y_1, y_2 exist, according to Theorem 12.6(a). The answers are

$$\begin{aligned} y_1(x) &= x^{1/2}\left(1 - \frac{1}{3}x + \frac{1}{30}x^2 - \frac{1}{630}x^3 + \frac{1}{22680}x^4 + \cdots\right), \\ y_2(x) &= x^0\left(1 - x + \frac{1}{6}x^2 - \frac{1}{90}x^3 + \frac{1}{2520}x^4 + \cdots\right). \end{aligned}$$

The method. Let r be a variable, to eventually be set to either root $r = r_1$ or $r = r_2$. We expect to compute two solutions $y_1 = y(x, r_1)$, $y_2 = y(x, r_2)$ from

$$y(x, r) = x^r \sum_{n=0}^{\infty} c(n, r)x^n.$$

The symbol $c(n, r)$ plays the role of c_n during the computation, but emphasizes the dependence of the coefficient on the root r.

Independence of y_1, y_2. Assume $k_1y_1(x) + k_2y_2(x) = 0$ for all x. Proving $k_1 = k_2 = 0$ implies y_1, y_2 are independent. Divide the equation $k_1y_1 + k_2y_2 = 0$ by x^{r_2}. The series representations of y_1, y_2 contain factors x^{r_2}, x^{r_2}. The division by x^{r_2} leaves two Maclaurin series and a factor of $x^{r_1 - r_2}$ on the y_1-series. This factor equals zero at $x = 0$, because $r_1 - r_2 > 0$. Substitute $x = 0$ to show that $k_2 = 0$. Then $k_1y_1(x) + k_2y_2(x) = 0$ gives $k_1 = 0$ because $y_1 \neq 0$. The test of independence is complete.

A formula for $c(n,r)$. The method applied is substitution of the series $y(x,r)$ into the differential equation in order to resolve the coefficients. At certain steps, series indexed from zero to infinity are split into the $n=0$ term plus the rest of the series, in order to match summation ranges. Index changes are used to match powers of x. The details:

$$\begin{aligned}
x^2A(x)y'' &= 2x^2y''(x,r) \\
&= 2x^2\sum_{n=0}^{\infty}(n+r)(n+r-1)c(n,r)x^{n+r-2} \\
&= 2r(r-1)c(0,r)x^r + \sum_{n=1}^{\infty}2(n+r)(n+r-1)c(n,r)x^{n+r}, \\
xB(x)y' &= xy'(x,r) \\
&= \sum_{n=0}^{\infty}(n+r)c(n,r)x^{n+r} \\
&= rc(0,r)x^r + \sum_{n=1}^{\infty}(n+r)c(n,r)x^{n+r} \\
C(x)y &= xy(x,r) \\
&= \sum_{n=0}^{\infty}c(n,r)x^{n+r+1} \\
&= \sum_{n=1}^{\infty}c(n-1,r)x^{n+r}.
\end{aligned}$$

Recursion. Let $p(r) = 2r(r-1)+r+0$ be the indicial polynomial. Let LHS stand for the left hand side of the Frobenius differential equation. Add the preceding equations. Then

$$\begin{aligned}
\text{LHS} &= 2x^2y''(x,r) + xy'(x,r) + xy(x,r) \\
&= p(r)c(0,r)x^r + \sum_{n=1}^{\infty}\left(p(n+r)c(n,r)+c(n-1,r)\right)x^{n+r}.
\end{aligned}$$

Because LHS equals the zero series, all coefficients are zero, which implies $p(r)=0$, $c(0,r)\neq 0$, and the recursion relation

$$p(n+r)c(n,r)+c(n-1,r)=0, \quad n\geq 1.$$

Solution of the recursion. The recursion answers on page 957 imply for $c_0 = c(0,r) = 1$ the relations

$$\begin{aligned}
c(n+1,r) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{1}{p(k+1+r)}\right) \\
c(n+1,r_1) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{1}{p(k+3/2)}\right) \\
c(n+1,r_2) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{1}{p(k+1)}\right)
\end{aligned}$$

Then $y_1(x) = y(x, r_1)$, $y_2(x) = y(x, r_2)$ imply

$$\begin{aligned}
y_1(x) &= x^{1/2}\left(1+\sum_{n=0}^{\infty}(-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{1}{(2k+3)(k+1)}\right)x^{n+1}\right)\\
&= x^{1/2}\left(1+\sum_{n=0}^{\infty}(-1)^{n+1}\frac{2^{n+1}}{(2n+3)!}x^{n+1}\right),\\
y_2(x) &= x^{0}\left(1+\sum_{n=0}^{\infty}(-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{1}{(k+1)(2k+1)}\right)x^{n+1}\right)\\
&= x^{0}\left(1+\sum_{n=0}^{\infty}(-1)^{n+1}\frac{2^{n}}{(n+1)(2n+1)!}x^{n+1}\right).
\end{aligned}$$

Answer checks. It is possible to verify the answers using `maple`, as follows.

```
c:=n->(-1)^(n+1)*product(1/((2*k+3)*(k+1)),k=0..n);
d:=n->(-1)^(n+1)*product(1/((2*k+1)*(k+1)),k=0..n);
N:=6;1+sum(c(n)*x^(n+1),n=0..N);
1+sum((-1)^(n+1)*2^(n+1)/((2*n+3)!)*x^(n+1),n=0..N);
1+sum(d(n)*x^(n+1),n=0..N);
1+sum((-1)^(n+1)*2^(n)/((n+1)*(2*n+1)!)*x^(n+1),n=0..N);
```

Verified by `maple` is exact solution formula $y(x) = c_1\cos(\sqrt{2x}) + c_2\sin(\sqrt{2x})$ in terms of elementary functions. Details:

```
de:=2*x^2*diff(y(x),x,x)+x*diff(y(x),x)+x*y(x)=0;
dsolve(de,y(x));
```

Example 12.2 (Frobenius Theorem Case (b))

Use the Frobenius theory to solve for y_1, y_2 in the differential equation $x^2y'' + x(3 + x)y' - 3y = 0$.

Solution: The indicial equation is $r^2 + (3 - 1)r - 3 = 0$ with roots $r_1 = 1$ (the larger root) and $r_2 = -3$. The roots differ by an integer, therefore one Frobenius solution y_1 exists and the second independent solution y_2 must be computed according to Theorem 12.6 part **(b)**. The answers are

$$\begin{aligned}
y_1(x) &= x\left(1-\frac{1}{5}x+\frac{1}{30}x^2-\frac{1}{210}x^3+\frac{1}{1680}x^4+\cdots\right),\\
y_2(x) &= x^{-3}\left(1-x+\frac{1}{2}x^2-\frac{1}{6}x^3\right).
\end{aligned}$$

Let r denote either root r_1 or r_2. We expect to compute solutions y_1, y_2 by the following scheme.

$$\begin{aligned}
y(x,r) &= x^r\sum_{n=0}^{\infty}c(n,r)x^n,\\
y_1(x) &= y(x,r_1),\\
y_2(x) &= Cy_1(x)\ln(x)+x^{r_2}\sum_{n=0}^{\infty}d_nx^n.
\end{aligned}$$

The constant C is either zero or one, but the value cannot be decided until the end of the computation. Likewise, $d_0 \neq 0$ is known, but little else about the sequence $\{d_n\}$ is known.

Find a formula for $c(n,r)$. The method substitutes the series $y(x,r)$ into the differential equation and then solves for the undetermined coefficients. The details:

$$\begin{aligned}
x^2A(x)y'' &= x^2y''(x,r) \\
&= x^2\sum_{n=0}^{\infty}(n+r)(n+r-1)c(n,r)x^{n+r-2} \\
&= r(r-1)c(0,r)x^r + \sum_{n=1}^{\infty}(n+r)(n+r-1)c(n,r)x^{n+r} \\
xB(x)y' &= (3+x)xy'(x,r) \\
&= (3+x)xy'(x,r) \\
&= (3+x)x\sum_{n=0}^{\infty}(n+r)c(n,r)x^{n+r-1} \\
&= \sum_{n=0}^{\infty}3(n+r)c(n,r)x^{n+r} + \sum_{n=0}^{\infty}(n+r)c(n,r)x^{n+r+1} \\
&= 3rc(0,r)x^r + \sum_{n=1}^{\infty}3(n+r)c(n,r)x^{n+r} \\
&= +\sum_{n=1}^{\infty}(n+r-1)c(n-1,r)x^{n+r} \\
C(x)y &= -3y(x,r) \\
&= -3c(0,r)x^r + \sum_{n=1}^{\infty}-3c(n,r)x^{n+r}.
\end{aligned}$$

Find the recursions. Let $p(r) = r(r-1)+3r-3$ be the indicial polynomial. Let LHS denote the left hand side of $x^2y''+x(3+x)y'-3y=0$. Add the three equations above. Then

$$\begin{aligned}
\text{LHS} &= x^2y''(x,r)+(3+x)xy'(x,r)-3y(x,r) \\
&= p(r)c(0,r)x^r + \sum_{n=1}^{\infty}\left(p(n+r)c(n,r)+(n+r-1)c(n-1,r)\right)x^{n+r}.
\end{aligned}$$

Symbol LHS equals the zero series, therefore all the coefficients are zero. Given $c(0,r) \neq 0$, then $p(r)=0$ and we have the recursion relation

$$p(n+r)c(n,r)+(n+r-1)c(n-1,r)=0, \quad n \geq 1.$$

Solve the recursion. Using $c(0,r)=1$ and the recursion answers on page 957 gives

$$\begin{aligned}
c(n+1,r) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{k+r}{p(k+1+r)}\right) \\
c(n+1,1) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{k+1}{(k+1)(k+5)}\right) \\
&= (-1)^{n+1}\frac{24}{(n+5)!}
\end{aligned}$$

Therefore, the first few coefficients $c_n = c(n,1)$ of y_1 are given by

$$c_0 = 1, \quad c_1 = \frac{-1}{5}, \quad c_2 = \frac{1}{30}, \quad c_3 = \frac{-1}{210}, \quad c_4 = \frac{1}{1680}.$$

This agrees with the reported solution y_1, whose general definition is

$$y_1(x) = 1 + \sum_{n=0}^{\infty} (-1)^{n+1} \frac{24}{(n+5)!} x^{n+1}.$$

Find the second solution y_2. Assume that $C = 0$ in the trial solution y_2. Let $d_n = c(n, r_2)$. Then the preceding formulas give the recursion relations

$$p(r_2)d_0 = 0, \quad p(n+r_2)d_n + (n + r_2 - 1)d_{n-1} = 0, \quad n \geq 1.$$

We require $r_2 = -3$ and $d_0 \neq 0$. The recursions reduce to

$$p(n-3)d_n + (n-4)d_{n-1} = 0, \quad n \geq 1.$$

The solution for $0 \leq n \leq 3$ is found from $d_n = -\dfrac{n-4}{p(n-3)} d_{n-1}$:

$$d_0 \neq 0, \quad d_1 = -d_0, \quad d_2 = \frac{1}{2}d_0, \quad d_3 = -\frac{1}{6}d_0.$$

There is no condition at $n = 4$, leaving d_4 arbitrary. This gives the recursion

$$p(n+2)d_{n+5} + (n+1)d_{n+4} = 0, \quad n \geq 0.$$

The solution of this recursion is

$$\begin{aligned} d_{n+5} &= (-1)^{n+1} \left(\Pi_{k=0}^{n} \frac{k+1}{p(k+2)} \right) d_4 \\ &= (-1)^{n+1} \left(\Pi_{k=0}^{n} \frac{k+1}{(k+1)(k+5)} \right) d_4 \\ &= (-1)^{n+1} \frac{24}{(n+5)!} d_4. \end{aligned}$$

For the moment let $d_4 = 1$. Then

$$d_4 = 1, \quad d_5 = -\frac{1}{5}, \quad d_6 = \frac{1}{30}, \quad d_7 = -\frac{1}{210},$$

and then the series terms for $n = 4$ and higher equal

$$x^{-3} \left(x^4 - \frac{1}{5}x^5 + \frac{1}{30}x^6 - \frac{1}{210}x^7 + \cdots \right) = y_1(x).$$

This implies

$$\begin{aligned} y_2(x) &= x^{-3}\left(d_0 + d_1 x + d_2 x^2 + d_3 x^3\right) + d_4 y_1(x) \\ &= x^{-3}\left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3\right) d_0 + d_4 y_1(x). \end{aligned}$$

By superposition, y_1 can be dropped from the formula for y_2. The conclusion for case $C = 0$ is

$$y_2(x) = x^{-3}\left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3\right).$$

False path for $C = 1$. We take $C = 1$ and repeat the derivation of y_2, just to see why this path leads to no solution with a $\ln(x)$-term. We have a 50% chance in Frobenius series problems of taking the wrong path to the solution. We will see details for success and also the signal for failure.

Let $L(y) = x^2y'' + x(3 + x)y' - 3y$ denote the left hand side of the Frobenius differential equation.

Decompose $y_2 = A+B$ where $A = y_1(x)\ln(x)$ and $B = x^{r_2}\sum_{n=1}^{\infty} d_nx^n$. Then $L(y_2) = 0$ becomes $L(B) = -L(A)$.

Compute $L(B)$. The substitution of B into the differential equation to obtain LHS has been done above. Let $d_n = c(n, r_2)$, $r_2 = -3$. The equation $p(r_2) = 0$ eliminates the extra term $p(r_2)c(0, r_2)x^{r_2}$. Split the summation into $1 \le n \le 4$ and $5 \le n < \infty$. Change index $n = m + 4$ to obtain:

$$\begin{aligned} L(B) &= \sum_{n=1}^{\infty} (p(n + r_2)c(n, r_2) + (n + r_2 - 1)c(n - 1, r_2))\, x^{n+r_2} \\ &= \sum_{n=1}^{3} (p(n - 3)d_n + (n - 4)d_{n-1})\, x^{n-3} + (p(1)d_4 + (0)d_3)x \\ &\quad + \sum_{m=1}^{\infty} (p(m + 1)d_{m+4} + (m)d_{m+3})\, x^{m+1}. \end{aligned}$$

Compute $L(A)$. Use $L(y_1) = 0$ in the third step and $r_1 = 1$ in the last step, below.

$$\begin{aligned} L(A) &= x^2(y_1'' \ln(x) + 2x^{-1}y_1' - x^{-2}y_1) \\ &\quad +(3 + x)x(y_1' \ln(x) + x^{-1}y_1) - 3y_1 \ln(x) \\ &= L(y_1)\ln(x) + (2 + x)y_1 + 2xy_1' \\ &= (2 + x)y_1 + 2xy_1' \\ &= \sum_{n=0}^{\infty} 2c_nx^{n+r_1} + \sum_{n=1}^{\infty} c_{n-1}x^{n+r_1} + \sum_{n=0}^{\infty} 2(n + r_1)c_nx^{n+r_1} \\ &= 4c_0x + \sum_{n=1}^{\infty}((2n + 4)c_n + c_{n-1})x^{n+1}. \end{aligned}$$

Find $\{d_n\}$. The equation $L(B) = -L(A)$ produces recursion relations by matching corresponding powers of x on each side of the equality. We are given $d_0 \ne 0$. For $1 \le n \le 3$, the left side matches zero coefficients on the right side, therefore as we saw in the case $C = 0$,

$$d_0 \ne 0, \quad d_1 = -d_0, \quad d_2 = \frac{1}{2}d_0, \quad d_3 = -\frac{1}{6}d_0.$$

The term for $n = 4$ on the left is $(p(1)d_4 + (0)d_3)x$, which is always zero, regardless of the values of d_3, d_4. On the other hand, there is the **nonzero term** $4c_0x$ on the right. We can never match terms, therefore there is **no solution** with $C = 1$. This is the only *signal for failure.*

Independence of y_1**,** y_2. Two functions y_1, y_2 are called independent provided $k_1y_1(x)+ k_2y_2(x) = 0$ for all x implies $k_1 = k_2 = 0$. For the given solutions, test independence by solving for k_1, k_2 in the equation

$$k_1x\left(1 - \frac{1}{5}x + \frac{1}{30}x^2 - \frac{1}{210}x^3 + \cdots\right) + k_2x^{-3}\left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3\right) = 0.$$

Divide the equation by x^{r_2}, then set $x = 0$. We get $k_2 = 0$. Substitute $k_2 = 0$ in the above equation. Divide by x^{r_1}, then set $x = 0$ to obtain $k_1 = 0$. Therefore, $k_1 = k_2 = 0$ and the independence test is complete.

Answer checks. The simplest check uses `maple` as follows. It is interesting that both y_1 and y_2 are expressible in terms of elementary functions, seen by executing the code below, and detected as a matter of course by `maple dsolve()`.

```
de:=x^2*diff(y(x),x,x)+x*(3+x)*diff(y(x),x)+(-3)*y(x)=0;
Order:=5;dsolve({de},y(x),type=series);
c:=n->(-1)^(n+1)*product((k+1)/((k+5)*(k+1)),k=0..n);
y1:=x+sum(c(n)*x^(n+2),n=0..5);
x+sum(c(n)*x^(n+2),n=0..infinity);
y2:=x->x^(-3)*( 1-x + x^2/2 -(1/6)*x^3);
simplify(subs(y(x)=y2(x),de));
dsolve(de,y(x));
```

Example 12.3 (Frobenius Theorem Case (c))

Use the Frobenius theory to solve for y_1, y_2 in the differential equation $x^2y'' + x(3 + x)y' + y = 0$.

Solution: The indicial equation is $r^2+(3-1)r+1 = 0$ with roots $r_1 = -1$, $r_2 = -1$. The roots are equal, therefore one Frobenius solution y_1 exists and the second independent solution y_2 must be computed according to Theorem 12.6. The answers:

$$\begin{aligned} y_1(x) &= x^{-1}(1 + x), \\ y_2(x) &= x^{-1}\left(-3x - \frac{1}{4}x^2 + \frac{1}{36}x^3 - \frac{1}{288}x^4 + \frac{1}{2400}x^5 + \cdots\right) \end{aligned}$$

Trial solution formulas for y_1, y_2. Based upon statement **(c)** of the Frobenius theorem page 970, we expect to compute the two solutions as follows.

$$\begin{aligned} y(x,r) &= x^r\sum_{n=0}^{\infty} c(n,r)x^n, \\ y_1(x) &= y(x,r_1), \\ y_2(x) &= \left.\frac{\partial y(x,r)}{\partial r}\right|_{r=r_1} \\ &= \left.\left(y(x,r)\ln(x) + x^r\sum_{n=0}^{\infty}\frac{\partial c(n,r)}{\partial r}x^n\right)\right|_{r=r_1} \\ &= y(x,r_1)\ln(x) + x^{r_1}\sum_{n=1}^{\infty} d_n x^n \end{aligned}$$

for some constants d_1, d_2, d_3, ... In some applications, it seems easier to use the partial derivative formula, in others, the final expression in symbols $\{d_n\}$ is more tractable. Finally, we might reject both methods in favor of the reduction of order formula for y_2.

Independence of y_1, y_2. To test independence, let $k_1y_1(x) + k_2y_2(x) = 0$ for all x. Proving $k_1 = k_2 = 0$ implies y_1, y_2 are independent. Divide the equation $k_1y_1 + k_2y_2 = 0$ by x^{r_1}. The series representations of y_1, y_2 contain a factor x^{r_1} which divides out, leaving two Maclaurin series and a $\ln(x)$-term. Then $\ln(0) = -\infty$ and assumption $c(0, r_1) \neq 0$

together with finiteness of the series shows that $k_2 = 0$. Hence also $k_1 = 0$. This completes the independence test.

Find a formula for $c(n, r)$. The method is to substitute the series $y(x, r)$ into the differential equation and then resolve the coefficients. The details:

$$\begin{aligned}
x^2A(x)y'' &= x^2y''(x,r) \\
&= x^2\sum_{n=0}^{\infty}(n+r)(n+r-1)c(n,r)x^{n+r-2} \\
&= r(r-1)c(0,r)x^r + \sum_{n=1}^{\infty}(n+r)(n+r-1)c(n,r)x^{n+r} \\
xB(x)y' &= (3+x)xy'(x,r) \\
&= (3+x)x\sum_{n=0}^{\infty}(n+r)c(n,r)x^{n+r-1} \\
&= \sum_{n=0}^{\infty}3(n+r)c(n,r)x^{n+r} + \sum_{n=0}^{\infty}(n+r)c(n,r)x^{n+r+1} \\
&= 3rc(0,r)x^r + \sum_{n=1}^{\infty}3(n+r)c(n,r)x^{n+r} \\
&= +\sum_{n=1}^{\infty}(n+r-1)c(n-1,r)x^{n+r} \\
C(x)y &= y(x,r) \\
&= c(0,r)x^r + \sum_{n=1}^{\infty}c(n,r)x^{n+r}.
\end{aligned}$$

Find the recursions. Let $p(r) = r(r-1) + 3r + 1$ be the indicial polynomial. Let LHS stand for the left hand side of the Frobenius differential equation. Add the above equations. Then

$$\begin{aligned}
\text{LHS} &= x^2y''(x,r) + (3+x)xy'(x,r) + y(x,r) \\
&= p(r)c(0,r)x^r + \sum_{n=1}^{\infty}\left(p(n+r)c(n,r) + (n+r-1)c(n-1,r)\right)x^{n+r}.
\end{aligned}$$

Because LHS equals the zero series, all coefficients are zero, which implies $p(r) = 0$ for $c(0, r) \neq 0$, plus the recursion relation

$$p(n+r)c(n,r) + (n+r-1)c(n-1,r) = 0, \quad n \geq 1.$$

Solve the recursions. Using the recursion answers on page 957 gives

$$\begin{aligned}
c(n+1,r) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{k+r}{p(k+1+r)}\right)c(0,r) \\
c(n+1,-1) &= (-1)^{n+1}\left(\Pi_{k=0}^{n}\frac{k-1}{(k+1)^2}\right)c(0,r).
\end{aligned}$$

Therefore, $c(0, -1) \neq 0$, $c(1, -1) = c(0, -1)$, $c(n + 1, -1) = 0$ for $n \geq 1$.

A formula for y_1. Choose $c(0, -1) = 1$. Then the formula for $y(x, r)$ and the requirement $y_1(x) = y(x, r_1)$ gives

$$y_1(x) = x^{-1}(1+x).$$

A formula for y_2. Of the various expressions for the solution, we choose

$$y_2(x) = y_1(x)\ln(x) + x^{r_1}\sum_{n=1}^{\infty} d_n x^n.$$

Let us put the trial solution y_2 into the differential equation left hand side $L(y) = x^2y'' + x(3+x)y' + y$ in order to determine the undetermined coefficients $\{d_n\}$. Arrange the computation as $y_2 = A + B$ where $A = y_1(x)\ln(x)$ and $B = x^{r_1}\sum_{n=1}^{\infty} d_n x^n$. Then $L(y_2) = L(A) + L(B) = 0$, or $L(B) = -L(A)$. The work has already been done for series B, because of the work with $y(x,r)$ and LHS. We define $d_0 = c(0, r_1) = 0$, $d_n = c(n, r_1)$ for $n \ge 1$. Then

$$L(B) = 0 + \sum_{n=1}^{\infty} \left(p(n+r)d_n + (n+r-1)d_{n-1}\right) x^{n+r_1}.$$

A direct computation, tedious and routine, gives

$$L(A) = 3 + x.$$

Comparing terms in the equation $L(B) = -L(A)$ results in the recursion relations

$$d_1 = -3, \quad d_2 = -\frac{1}{4}, \quad d_{n+1} = -\frac{n-1}{(n+1)^2}\, d_n \quad (n \ge 2).$$

Solving for the first few terms duplicates the coefficients reported earlier:

$$d_1 = -3, \quad d_2 = -\frac{1}{4}, \quad d_3 = \frac{1}{36}, \quad d_4 = \frac{-1}{288}, \quad d_5 = \frac{1}{2400}.$$

A complete formula:

$$\begin{aligned} y_2(x) &= x^{-1}\left((1+x)\ln(x) - 3x - \frac{1}{4}x^2 + \frac{1}{4}\sum_{n=2}^{\infty}(-1)^n\left(\Pi_{k=2}^{n}\frac{k-1}{p(k)}\right)x^{n+1}\right) \\ &= x^{-1}\left((1+x)\ln(x) - 3x - \frac{1}{4}x^2 + \sum_{n=2}^{\infty}(-1)^n\frac{(n-1)!}{((n+1)!)^2}\,x^{n+1}\right) \\ &= x^{-1}\left((1+x)\ln(x) - 3x - \frac{1}{4}x^2 + \sum_{n=2}^{\infty}\frac{(-1)^n}{n(n+1)}\,\frac{x^{n+1}}{(n+1)!}\right). \end{aligned}$$

Answer check. The solutions displayed here can be checked in `maple` as follows.

```
de:=x^2*diff(y(x),x,x)+x*(3+x)*diff(y(x),x)+y(x);
y1:=((1+x)/x)*ln(x);
eqA:=simplify(subs(y(x)=y1,de));
dsolve(de=0,y(x),series);
d:=n->(-1)^(n-1)/((n-1)*n*(n!));
y2:=x^(-1)*((1+x)*ln(x)-3*x-x^2/4+sum(d(n+1)*x^(n+1),n=2..6));
```

Exercises 12.5

Regular Singular Point

Test the equation for regular singular points.

1. $x^2y'' + xy' + y = 0$

2. $x^2(x-1)y'' + \sin(x)y' + y = 0$

3. $x^3(x^2-1)y'' - x(x+1)y' + (1-x)y = 0$

4. $x^3(x-1)y'' + (x-1)y' + 2xy = 0$

Indicial Equation

Each equation is an Euler differential equation $ax^2y'' + bxy' + cy = 0$ with a, b, c replaced by power series. Find the Euler differential equation and the indicial equation.

5. $x^2y'' - 2x(x+1)y' + (x-1)y = 0$
Ans: $x^2y'' - 2xy' - y = 0$, $r(r-1) - 2r - 1 = 0$.

6. $x^2y'' - 2xy' + y = 0$
Ans: The same equation, $r(r-1) - 2r + 1 = 0$.

7. $xy'' + (1-x)y' + 2y = 0$

8. $x^2y'' - 2xy' + (2 + \sin x)y = 0$

Frobenius Solutions

Find two linearly independent solutions. Follow Examples 1, 2, 3 for cases (a), (b), (c) in the Frobenius Theorem page 970. Examples: **(a)** page 971, **(b)** page 973, **(c)** page 977.

9. $2x^2y'' + xy' - y = 0$

10. $4x^2y'' + (2x-7)y' + 6y = 0$

11. $4x^2(x+1)y'' + x(3x-1)y' + y = 0$

12. $3x^2y'' + xy' - (1+x)y = 0$

13. $x^2y'' + 3xy' + (1+x)y = 0$

14. $xy'' + (1-x)y' + 3y = 0$

15. $x^2y'' + x(x-1)y' + (1-x)y = 0$

16. $xy'' + (2x+3)y' + 4y = 0$

12.6 Bessel Functions

The work of Friedrich W. Bessel (1784-1846) on planetary orbits led to his 1824 derivation of the equation known in this century as the **Bessel differential equation** or order p:

$$x^2y'' + xy' + (x^2 - p^2)y = 0.$$

This equation appears in a 1733 work on hanging cables by Daniel Bernoulli (1700-1782). A particular solution y is called a **Bessel function**. While any real or complex value of p may be considered, we restrict the case here to $p \geq 0$ an integer.

Frobenius theory page 970 applies directly to Bessel's equation, which has a regular singular point at $x = 0$. The indicial equation is $r^2 - p^2 = 0$ with roots $r_1 = p$ and $r_2 = -p$. The assumptions imply that cases (b) and (c) of the Frobenius theorem apply: either $r_1 - r_2 =$ positive integer [case (b)] or else $r_1 = r_2 = 0$ and $p = 0$ [case (c)]. In both cases there is a Frobenius series solution for the larger root. This solution is referenced as $J_p(x)$ in the literature, and called a **Bessel function of nonnegative integral order** p. The formulas most often used appear below.

$$\begin{aligned}
J_p(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n (x/2)^{p+2n}}{n!(p+n)!}, \\
J_0(x) &= 1 - (x/2)^2 + \frac{(x/2)^4}{4^2} - \frac{(x/2)^6}{6^2} + \cdots \\
J_1(x) &= \frac{x}{2} - \frac{(x/2)^3}{(1)(2)} + \frac{(x/2)^5}{(2)(6)} - \frac{(x/2)^7}{(6)(24)} + \cdots
\end{aligned}$$

The derivation of the formula for J_p is obtained by substitution of the trial solution $y = x^r \sum_{n=0}^{\infty} c_n x^n$ into Bessel's equation. Let $Q(r) = r(r-1) - p^2$ be the indicial polynomial. The result is

$$\sum_{n=0}^{\infty} Q(n+r)c_n x^{n+r} + \sum_{n=0}^{\infty} c_n x^{n+p+2} = 0.$$

Matching terms on the left to the zero coefficients on the right gives the recursion relations

$$Q(r)c_0 = 0, \quad Q(r+1)c_1 = 0, \quad Q(n+r)c_n + c_{n-2} = 0, \quad n \geq 2.$$

To resolve the relations, let $r = p$ (the larger root), $c_0 = 1$, $c_1 = 0$ (because $Q(p+1) \neq 0$), and

$$c_{n+2} = \frac{-1}{Q(n+2+p)} c_n.$$

This is a two-termed second order recursion which can be solved with formulas developed on page 957 to give

$$\begin{aligned}
c_{2n+2} &= (-1)^{n+1}\left(\prod_{k=0}^{n} \frac{1}{(2k+2+p)^2-p^2}\right) c_0 \\
&= (-1)^{n+1}\prod_{k=0}^{n} \frac{1}{4(k+1)(k+1+p)} \\
&= \frac{(-1)^{n+1}}{4^{n+1}}\frac{1}{(n+1)!}\frac{p!}{(n+1+p)!} \\
&= (2^p p!)\frac{(-1)^{n+1}}{2^{2n+2+p}}\frac{1}{(n+1)!}\frac{1}{(n+1+p)!} \\
c_{2n+3} &= (-1)^{n+1}\left(\prod_{k=0}^{n} \frac{1}{(2k+3+p)^2-p^2}\right) c_1 \\
&= 0.
\end{aligned}$$

The common factor $(2^p p!)x^p$ can be factored out from each term except the first, which is $c_0 x^p$ or x^p. Dividing the answer so obtained by $(2^p p!)$ gives the series reported for J_p.

Properties of Bessel Functions

Sine and cosine identities from trigonometry have direct analogs for Bessel functions. We would like to say that $\cos(x) \leftrightarrow J_0(x)$, and $\sin(x) \leftrightarrow J_1(x)$, but that is not exactly correct. There are asymptotic formulas

$$\begin{aligned}
J_0(x) &\approx \sqrt{\tfrac{2}{\pi x}}\cos\left(x-\tfrac{\pi}{4}\right), \\
J_1(x) &\approx \sqrt{\tfrac{2}{\pi x}}\sin\left(x-\tfrac{\pi}{4}\right).
\end{aligned}$$

See the reference by G.N. Watson [Watson] for details about these asymptotic formulas. At a basic level, based upon the series expressions for J_0 and J_1, the following identities can be quickly checked.

Bessel Functions			Trig Functions		
$J_0(0)$	$=$	1	$\cos(0)$	$=$	1
$J_0'(0)$	$=$	0	$(\cos(x))'\big\vert_{x=0}$	$=$	0
$J_1(0)$	$=$	0	$\sin(0)$	$=$	0
$J_1'(0)$	$=$	$1/2$	$(\sin(x))'\big\vert_{x=0}$	$=$	1
$J_0(-x)$	$=$	$J_0(x)$	$\cos(-x)$	$=$	$\cos(x)$
$J_1(-x)$	$=$	$-J_1(x)$	$\sin(-x)$	$=$	$-\sin(x)$

Some deeper relations exist, obtained by series expansion of both sides of the identities. Suggestions for the derivations are in the exercises. Watson's basic reference [Watson] can be consulted to find complete details.

$$\begin{array}{lcl}
J_0'(x) & = & -J_1(x) \\
J_1'(x) & = & J_0(x) - \dfrac{1}{x}\,J_1(x) \\
(x^p J_p(x))' & = & x^p J_{p-1}(x), \quad p \geq 1, \\
\big(x^{-p} J_p(x)\big)' & = & -x^{-p} J_{p+1}(x), \quad p \geq 0, \\
J_{p+1} & = & \dfrac{2p}{x} J_{p+1}(x) - J_{p-1}(x), \quad p \geq 1, \\
J_{p+1}(x) & = & -2J_p'(x) + J_{p-1}(x), \quad p \geq 1.
\end{array}$$

The Zeros of Bessel Functions

It is a consequence of the second order differential equation for Bessel functions that these functions have infinitely many zeros on the positive x-axis. As seen from asymptotic expansions, the zeros of J_0 satisfy $x - \pi/4 \approx (2n-1)\pi/2$ and the zeros of J_1 satisfy $x - \pi/4 \approx n\pi$. These approximations are already accurate to one decimal digit for the first five zeros, as seen from the following table.

The positive zeros of J_0 and J_1

n	$J_0(x)$	$J_1(x)$	$\left(\dfrac{2n-1}{2} + \dfrac{1}{4}\right)\pi$	$n\pi + \dfrac{\pi}{4}$
1	2.40482556	3.83170597	2.35619449	3.92699082
2	5.52007811	7.01558667	5.49778714	7.06858347
3	8.65372791	10.17346813	8.63937980	10.21017613
4	11.79153444	13.32369194	11.78097245	13.35176878
5	14.93091771	16.47063005	14.92256511	16.49336143

The values are conveniently obtained by the following `maple` code.

```
seq(evalf(BesselJZeros(0,n)),n=1..5);
seq(evalf(BesselJZeros(1,n)),n=1..5);
seq(evalf((2*n-1)*Pi/2+Pi/4),n=1..5);
seq(evalf((n)*Pi+Pi/4),n=1..5);
```

The Sturm theory of oscillations of second order differential equations provides the theory which shows that Bessel functions oscillate on the positive x-axis. Part of that theory translates to the following theorem about the interlaced zero property. Trigonometric graphs verify the interlaced zero property for sine and cosine. The theorem for $p = 0$ says that the zeros of $J_0(x) \leftrightarrow \cos(x)$ and $J_1(x) \leftrightarrow \sin(x)$ are interlaced.

Theorem 12.7 (Interlaced Zeros)
Between pairs of zeros of J_p there is a zero of J_{p+1} and between zeros of J_{p+1} there is a zero of J_p. In short, the zeros of J_p and J_{p+1} are interlaced.

Proof: A complete proof including the basic Sturm theory can be found in the text by Kreider, Kuller, Ostberg and Perkins (1966), [KKOP] page 234. ■

Exercises 12.6

Values of J_0 and J_1

Use series representations and identities to find an identity for values of the following functions. Use a computer algebra system to compute the answers.

1. $J_0(1)$

2. $J_1(1)$

3. $J_0(1/2)$

4. $J_1(1/2)$

Bessel Function Properties

Prove the following relations by expanding LHS and RHS in series.

5. $J_0'(x) = -J_1(x)$

6. $J_1'(x) = J_0(x) - \dfrac{1}{x} J_1(x)$

7. $(x^p J_p(x))' = x^p J_{p-1}(x)$, $p \geq 1$

8. $\left(x^{-p} J_p(x)\right)' = -x^{-p} J_{p+1}(x)$, $p \geq 0$

Bessel Function Recursion Proofs

Add and subtract the expanded equations of the previous exercises.

9. $J_{p+1} = \dfrac{2p}{x} J_p(x) - J_{p-1}(x)$, $p \geq 1$

10. $J_{p+1}(x) = -2J_p'(x) + J_{p-1}(x)$, $p \geq 1$

Recurrence Relations

Use results of the previous exercises.

11. Express J_3 and J_4 in terms of J_0 and J_1.

12. Prove by induction that $J_p(x) = c_1(1/x)J_0(x) + c_2(1/x)J_1(x)$ where c_1 and c_2 are polynomials.

Laplace Transform

Assume Laplace identity $\mathcal{L}(J_n(t)) = \frac{\left(\sqrt{s^2+1}-s\right)^n}{\sqrt{s^2+1}}$ holds for $s \geq 0$. Prove the following results.

13. $\int_0^\infty J_{n+1}(x)dx = \int_0^\infty J_{n-1}(x)dx$ for integers $n > 0$.

14. $\displaystyle\int_0^\infty \frac{J_n(x)dx}{x} = \frac{1}{n}$ for integers $n > 0$

Bessel Function Bounds

Assume L. J. Landau's result $J_p(x) \leq c|x|^{-1/3}$ for all x and $p > 0$, where $c = 0.78574687\ldots$ is the best possible constant. Prove the following results.

15. $\lim_{x\to\infty} J_1(x) = 0$

16. $\lim_{x\to\infty} J_0'(x) = 0$

12.7 Legendre Polynomials

The differential equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0$$

is called the **Legendre differential equation** of order n, after the French mathematician Adrien Marie Legendre (1752-1833), because of his work on gravitation.[1] The value of n is a nonnegative integer. For each n, the corresponding Legendre equation is known to have a polynomial solution $P_n(x)$ of degree n, called the nth **Legendre polynomial**. The first few of these are recorded below.

$P_0(x)$	$=$	1	$P_4(x)$	$=$	$\frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}$
$P_1(x)$	$=$	x			
$P_2(x)$	$=$	$\frac{3}{2}x^2 - \frac{1}{2}$	$P_5(x)$	$=$	$\frac{63}{8}x^5 - \frac{35}{4}x^3 + \frac{15}{8}x,$
$P_3(x)$	$=$	$\frac{5}{2}x^3 - \frac{3}{2}x$	$P_6(x)$	$=$	$\frac{231}{16}x^6 - \frac{315}{16}x^4 + \frac{105}{16}x^2 - \frac{5}{16}.$

The general formula for $P_n(x)$ is obtained by using ordinary point theory on Legendre's differential equation. The polynomial is normalized to satisfy $P_n(1) = 1$. The **Legendre polynomial of order** n is defined by

$$(1) \qquad P_n(x) = \frac{1}{2^n}\sum_{k=0}^{N} \frac{(-1)^k(2n-2k)!}{k!(n-2k)!(n-k)!}\, x^{n-2k},$$

according to $n = 2N$ even or $n = 2N+1$ odd. Proof on page 989.

There are alternative formulas available from which to compute P_n. The most famous one is **Rodrigues' formula**, after the French economist and mathematician Olinde Rodrigues (1794-1851),

$$P_n(x) = \frac{1}{2^n\, n!}\frac{d^n}{dx^n}\left(x^2-1\right)^n,$$

proof on page 993. The classical generating function derivation is in Exercise 5. Equally famous is **Bonnet's recursion**

$$P_{n+1}(x) = \frac{2n+1}{n+1}xP_n(x) - \frac{n}{n+1}P_{n-1}(x),$$

which was used to produce the table of Legendre polynomials above. Bonnet's recursion is derived from Rodrigues' formula on page 993.

[1] Legendre is recognized more often for his 40 years of work on elliptic integrals.

Properties of Legendre Polynomials

The main relations known for Legendre polynomials P_n are recorded here.

$$\begin{aligned}
P_n(1) &= 1 \\
P_n(-1) &= (-1)^n \\
P_{2n+1}(0) &= 0 \\
P'_{2n}(0) &= 0 \\
P_n(-x) &= (-1)^n P_n(x) \\
(n+1)P_{n+1}(x) &= (2n+1)xP_n(x) - nP_{n-1}(x) \\
P'_{n+1}(x) - P'_{n-1}(x) &= (2n+1)P_n(x) \\
P'_{n+1}(x) - xP'_n(x) &= (n+1)P_n(x) \\
(1-2xt+t^2)^{-1/2} &= \sum_{n=0}^{\infty} P_n(x)t^n \\
\int_{-1}^{1} |P_n(x)|^2 dx &= \frac{2}{2n+1} \\
\int_{-1}^{1} P_n(x)P_m(x)dx &= 0 \quad (n \neq m)
\end{aligned}$$

Example 12.4 (Boundary Data for P_n)
The polynomial solution $P_n(x)$ of Legendre's equation $(1-x^2)y''-2xy'+n(n+1)y = 0$ satisfies $P_n(1) = 1$ and $P'_n(1) = \dfrac{n(n+1)}{2}$.

Details for Example 12.4
Identity $P_n(1) = 1$ is derived in the proof of the Legendre polynomial formula page 989. Used in calculations below are identities from algebra and calculus:

$$\begin{array}{lll}
(1) & (a+b)^k = \displaystyle\sum_{r=0}^{k} \binom{k}{r} a^r b^{k-r} & \text{Binomial theorem} \\
(2) & (uv)^{(n)} = \displaystyle\sum_{r=0}^{n} \binom{n}{r} u^{(r)} v^{(n-r)} & \text{Product theorem}
\end{array}$$

Identity $P'_n(1) = \dfrac{n(n+1)}{2}$ for $n > 1$ will be derived from Rodrigues' formula and identities (1), (2). For $n = 0, 1$, the identity follows from $P_0(x) = 1$, $P_1(x) = x$. Assume $n \geq 1$. Let $c = \dfrac{1}{2^n n!}$. Then Rodrigues' formula implies

$$\begin{aligned}
P'_n(x) &= c\frac{d}{dx}\left((x^2-1)^n\right)^{(n)} \\
&= c\left(\frac{d}{dx}\left(x^2-1)^n\right)\right)^{(n)} \\
&= c\left(2nx(x^2-1)^{n-1}\right)^{(n)}
\end{aligned}$$

$$\begin{aligned} &= 2nc\,(uv)^{(n)} \text{ where } u = x,\ v = (x^2-1)^{n-1} \\ &= 2nc\sum_{r=0}^{n}\binom{n}{r}u^{(r)}v^{(n-r)} \qquad \text{by identity (2)} \end{aligned}$$

Let $x = 1$ in the last display. Because $u(x) = x$, then $u(1) = u'(1) = 1$ and $u^{(r)} = 0$ for $r \geq 2$. The sum reduces to two terms:

$$P_n'(1) = 2nc\binom{n}{0}v^{(n)}(1) + 2nc\binom{n}{1}v^{(n-1)}(1)$$

Insert $\binom{n}{0} = 1$ and $\binom{n}{1} = n$, then:

$$P_n'(1) = 2ncv^{(n)}(1) + 2n^2cv^{(n-1)}(1)$$

Calculus with mathematical induction on formula $v = (x^2-1)^{n-1}$ gives these results:

$$v^{(n-1)}(1) = 2^{n-1}(n-1)!, \qquad v^{(n)}(1) = 2^{n-2}(n-1)\,n!$$

The details are aided by substitution $y = x - 1$. Then $v = (y^2+2y)^{n-1}$ is a polynomial in y obtained explicitly by expansion (1). Then $2^{n-1}\,n! = \dfrac{1}{2c}$ implies:

$$\begin{aligned} P_n'(1) &= 2ncv^{(n)}(1) + 2ncn\,v^{(n-1)}(1) \\ &= c(2^{n-2}(2)(n!)(n)(n-1)) + c(2^{n-1}(2n)(n)(n-1)!) \\ &= \frac{n(n+1)}{2} \end{aligned}$$

Gaussian Quadrature

A high-speed low overhead numerical procedure **Gaussian quadrature** is defined in terms of the zeros $\{x_k\}_{k=1}^n$ of $P_n(x) = 0$ in $-1 < x < 1$ and certain constants $\{a_k\}_{k=1}^n$ by the approximation formula

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} a_k f(x_k).$$

The approximation is exact when f is a polynomial of degree less than $2n$. This fact is enough to evaluate the sequence of numbers $\{a_k\}_{k=1}^n$, because we can replace f by the basis functions $1, x, \ldots, x^{n-1}$ to get an $n \times n$ system for the variables $a_1, \ldots, a_n$. The last critical element: the sequence $\{x_k\}_{k=1}^n$ is the set of n distinct roots of $P_n(x) = 0$ in $-1 < x < 1$. Here we need some theory, that says that these roots number n and are all real.

Theorem 12.8 (Roots of Legendre Polynomials)
The Legendre polynomial P_n has exactly n distinct real roots $x_1, \ldots, x_n$ located in the interval $-1 < x < 1$.

The importance of the Gaussian quadrature formula lies in the ability to make a table of values that generates the approximation, except for the evaluations $f(x_k)$. This makes Gaussian quadrature a very high speed method, because it is based upon function evaluation and a dot product for a fixed number of vector entries. Vector parallel computers are able to perform these operations at high speed.

A question: *How is Gaussian quadrature different than the rectangular rule*? They are similar methods in the arithmetic requirements of function evaluation and dot product. The answer: the rectangular rule has less accuracy than Gaussian quadrature.

Gaussian quadrature can be compared with Simpson's rule. For $n = 3$, which uses three function evaluations, Gaussian quadrature becomes

$$\int_{-1}^{1} f(x)dx \approx \frac{5f(\sqrt{.6}) + 8f(0) + 5f(-\sqrt{.6})}{9},$$

whereas Simpson's rule with one interval is

$$\int_{-1}^{1} f(x)dx \approx \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1).$$

Left as a puzzle is comparison of the two approximations using polynomials f of degree higher than 4, or perhaps a smooth positive function f on $-1 < x < 1$, say $f(x) = \cos(x)$.

Table generation. The pairs (x_j, a_j), $1 \le j \le n$, required for the right side of the Gaussian quadrature formula, can be generated just once for a given n by the following `maple` procedure.

```
GaussQuadPairs:=proc(n)
 local a,x,xx,ans,p,eqs;
 xx:=fsolve(orthopoly[P](n,x)=0,x);
 x:=array(1..n,[xx]);
 eqs:=seq(sum(a[j]*x[j]^k,j=1..n)=int(t^k,t=-1..1),
  k=0..n-1);
 ans:=solve({eqs},{seq(a[j],j=1..n)});
 assign(ans);
 p:=[seq([x[j],a[j]],j=1..n)];
 end proc;
```

For simple applications, the `maple` code above can be attached to the application to generate the table on-the-fly. To generate tables, such as the one below, run the procedure for a given n, e.g., to generate the table for $n = 5$, insert the above procedure, then use the command `GaussQuadPairs(5);`.

Table 1. Gaussian Quadrature Pairs for $n = 5$

j	x_j	a_j
1	-0.9061798459	0.2369268851
2	-0.5384693101	0.4786286705
3	0.0000000000	0.5688888887
4	0.5384693101	0.4786286705
5	0.9061798459	0.2369268851

Derivation: Legendre Polynomial Formula

Let's start with the differential equation

$$(1-x^2)y'' - 2xy' + \lambda y = 0$$

where λ is a real constant. It will be shown that the differential equation has a polynomial solution if and only if $\lambda = n(n+1)$ for some nonnegative integer n, in which case the polynomial solution is a scalar multiple of P_n, which is given by equation (1) page 985.

Proof: The trial solution is a Maclaurin series $y = \sum_{n=0}^{\infty} c_n x^n$. We will find two independent solutions y_1, y_2, a basis of solutions on an interval about $x = 0$. The background required is the theory of ordinary points. [2]

Substitute the trial solution into Legendre's equation:

$$\begin{aligned}
(1-x^2)y'' &= \sum_{k=0}^{\infty}(k+2)(k+1)c_{k+2}x^k - \sum_{n=2}^{\infty} n(n-1)c_n x^n,\\
-2xy' &= \sum_{n=1}^{\infty} -2nc_n x^n,\\
\lambda y &= \sum_{n=0}^{\infty} \lambda c_n x^n.
\end{aligned}$$

Let $L(y) = (1-x^2)y'' - 2xy' + \lambda y$, then, adding the above equations,

$$\begin{aligned}
L(y) &= (1-x^2)y'' - 2xy' + \lambda y\\
&= (2c_2 + \lambda c_0) + (6c_3 - 2c_1 + \lambda c_1)x\\
&\quad + \sum_{n=2}^{\infty}((n+2)(n+1)c_{n+2} + (-n(n-1) - 2n + \lambda)c_n)x^n.
\end{aligned}$$

The requirement $L(y) = 0$ makes the right side coefficients equal the coefficients of the zero series, giving the relations

$$\begin{aligned}
&2c_2 + \lambda c_0 = 0,\\
&6c_3 - 2c_1 + \lambda c_1 = 0,\\
&(n+2)(n+1)c_{n+2} + (-n(n-1) - 2n + \lambda)c_n = 0 \quad (n \geq 2).
\end{aligned}$$

[2] Legendre polynomials P_n are solutions of Legendre's equation for $n \geq 0$ an integer, known to be orthogonal on $[-1, 1]$. Legendre's equation has regular singular points at $x = \pm 1$ and $x = \infty$. Frobenius theory applies to find solutions when n in the factor $n(n+1)$ is a real number (not an integer). The solutions are called a **Legendre function of the first kind** and a **Legendre function of the second kind**, denoted `LegendreP(n,x)` and `LegendreQ(n,x)` in both `maple` and `mathematica` languages.

These compress to a single two-termed second order recursion

$$c_{n+2} = \frac{n^2+n-\lambda}{(n+2)(n+1)}c_n = 0, \quad (n \geq 0),$$

whose solution is

$$\begin{aligned} c_{2n+2} &= \left(\Pi_{k=0}^{n} \frac{2k(2k+1)-\lambda}{(2k+1)(2k+2)}\right) c_0, \\ c_{2n+3} &= \left(\Pi_{k=0}^{n} \frac{(2k+1)(2k+2)-\lambda}{(2k+2)(2k+3)}\right) c_1. \end{aligned}$$

Let y_1 be the series solution using $c_0 = 1$, $c_1 = 0$ and let y_2 be the series solution using $c_0 = 0$, $c_1 = 1$. Then

$$\begin{aligned} y_1 &= 1 + \textstyle\sum_{n=0}^{\infty} a_{2n+2}x^{2n+2}, \qquad & a_{2n+2} = \prod_{k=0}^{n} \frac{2k(2k+1)-\lambda}{(2k+1)(2k+2)} \\ y_2 &= x + \textstyle\sum_{n=0}^{\infty} b_{2n+3}x^{2n+3}, \qquad & b_{2n+3} = \prod_{k=0}^{n} \frac{(2k+1)(2k+2)-\lambda}{(2k+2)(2k+3)} \end{aligned}$$

The radius of convergence of y_1 and y_2 is 1, by the ratio test. They form a basis of solutions to Legendre's equation defined on $-1 < x < 1$.

Lemma A. If $\lambda = m(m+1)$ for some integer $m \geq 0$, then one of the two series solutions y_1, y_2 is a polynomial.

Proof of Lemma A: For $m = 2n + 2$ (m even), there is a zero factor in the product equation for a_{2n+2}, causing $a_{2j+2} = 0$ for $j \geq n$, which means y_1 is a polynomial. Similarly, if $m = 2n + 3$ (m is odd), then $b_{2j+3} = 0$ for $j \geq n$: y_2 is a polynomial. If $m = 0$, then $c_2 = 0$ from the recursion relations, giving polynomial solution $y_1 = 1$. If $m = 1$, then $c_0 = c_{2k+2} = 0$, $c_1 = 1$,$c_{2k+3} = 0$ for $k \geq 0$, giving polynomial solution $y_2 = x$. The proof of Lemma B is complete. ■

Lemma B. If some solution y is a nonzero polynomial, then $\lambda = n(n + 1)$ for some integer $n \geq 0$.

Proof of Lemma B: Assume some solution y is a nonzero polynomial. Assume the contrary, that λ does not equal $n(n+1)$ for any integer $n \geq 0$. Let's seek a contradiction to complete the proof.

Because y_1, y_2 are a basis of solutions, then $y = d_1y_1 + d_2y_2$ for some $|d_1| + |d_2| > 0$ and the derivative $y^{(m)}$ is identically zero for m equal to one plus the degree of polynomial y.

Differentiate $y = d_1y_1 + d_2y_2$ to obtain the two equations

$$\begin{array}{lcll} d_1y_1^m(0) & + & d_2y_2^m(0) & = & 0, \\ d_1y_1^{m+1}(0) & + & d_2y_2^{m+1}(0) & = & 0 \end{array}$$

Then d_1, d_2 satisfy the 2×2 linear system

$$\begin{pmatrix} y_1^{(m)}(0) & y_2^{(m)}(0) \\ y_1^{(m+1)}(0) & y_2^{(m+1)}(0) \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Because $|d_1| + |d_2| > 0$, then the 2×2 linear system has a nonzero solution, implying the determinant of coefficients must vanish:

$$D = \begin{vmatrix} y_1^{(m)}(0) & y_2^{(m)}(0) \\ y_1^{(m+1)}(0) & y_2^{(m+1)}(0) \end{vmatrix} = 0$$

Series y_1 and y_2 are Maclaurin expansions. The four derivatives in determinant D appear in the series expansions of y_1 and y_2. For instance, $y_1^{(m)}(0)/m!$ is the coefficient of x^m in series y_1. Assume $m > 1$ and $m = 2n + 2$ (m is even).

The odd case $m > 1$ and $m = 2n + 3$ is treated similarly, details omitted.

Then $y_1^{(m)}(0)/m! = a_{2n+2}$ by the definition of y_1, giving relation

$$D = \begin{vmatrix} (2n+2)!a_{2n+2} & y_2(2n+2)(0) \\ (2n+3)!a_{2n+3} & y_2^{(2n+3)}(0) \end{vmatrix} = 0$$

Even terms of y_2 are zero, therefore $y^{(2n+2)}(0) = 0$ and the determinant evaluates to $D = (2n+2)!a_{2n+2}y_2^{(2n+3)}(0)$. If λ is not the product of two consecutive integers, then product $a_{2n+2} \neq 0$, and $y_2^{(2n+3)}(0) \neq 0$ by a similar analysis, using the recursion product formulas for a_{2n+2} and b_{2n+3}, which contain nonzero factors of the form $j(j+1) - \lambda$. So $D \neq 0$. The contradiction: $D = 0$ and $D \neq 0$.

Two cases remain: (1) $m = 0$, (2) $m = 1$. Consider case (1), then

$$\begin{aligned} D &= \begin{vmatrix} y_1^{(m)}(0) & y_2^{(m)}(0) \\ y_1^{(m+1)}(0) & y_2^{(m+1)}(0) \end{vmatrix} \\ &= \begin{vmatrix} y_1(0) & y_2(0) \\ y_1'(0) & y_2'(0) \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \neq 0. \end{aligned}$$

Consider case (2), then

$$\begin{aligned} D &= \begin{vmatrix} y_1^{(m)}(0) & y_2^{(m)}(0) \\ y_1^{(m+1)}(0) & y_2^{(m+1)}(0) \end{vmatrix} \\ &= \begin{vmatrix} y_1'(0) & y_2'(0) \\ y_1''(0) & y_2''(0) \end{vmatrix} \\ &= \begin{vmatrix} 0 & 1 \\ a_2 & y_2''(0) \end{vmatrix} \\ &= -a_2 = -(-\lambda/2) \end{aligned}$$

Because λ is not of the form $n(n+1)$ then $\lambda \neq 0$ and again a contradiction: D is both zero and nonzero. The proof of Lemma B is complete. ■

Simplification of Coefficients.
Let $P_n = y_1$ for n even and $P_n = y_2$ for n odd. Only the case of n even, $n = 2N$, will be verified. The odd case is left as an easily-solved puzzle. The equation $2r(2r+1) - n(n+1) = (2r-n)(n+2r+1)$ implies the following relation for the coefficients of y_1:

$$\begin{aligned} c_{2p+2} &= c_0\Pi_{r=0}^{p}\frac{2r(2r+1)-n(n+1)}{(2r+1)(2r+2)} \\ &= c_0\Pi_{r=0}^{p}\frac{(2r-n)(n+2r+1)}{(2r+1)(2r+2)}. \end{aligned}$$

Choose

$$c_0 = \frac{(-1)^N}{2^n(N!)^2} \quad (n = 2N \text{ even}).$$

Let's match coefficients in the reported formula for P_n against the series solution. The constant terms match by the choice of c_0. To match powers x^{n-2k} and x^{2p+2}, we require $n-2k=2p+2$. To match coefficients, we must prove

$$c_{2p+2} = \frac{1}{2^n}\frac{(-1)^r(2n-2k)!}{k!(n-2k)!(n-k)!}.$$

Solving $n-2k=2p+2$ for p when $n=2N$ gives $p=N-k-1$ and then

$$\begin{aligned} c_{2p+2} &= c_0\Pi_{r=0}^{p}\frac{(-1)(n-2r)(n+2r+1)}{(2r+1)(2r+2)} \\ &= \frac{(-1)^{2N-k}}{2^n(N!)^2}\Pi_{r=0}^{N-k-1}\frac{(n-2r)(n+2r+1)}{(2r+1)(2r+2)}. \end{aligned}$$

The product factor will be converted to powers and factorials.

$$\begin{aligned} \boxed{\mathbf{1}} &= \Pi_{r=0}^{N-k-1}(n-2r) \\ &= (2N)(2N-2)\cdots(2k+2) \\ &= 2^{N-k}(N)(N-1)\cdots(k+1) \\ &= 2^{N-k}\frac{N!}{k!}. \\ \boxed{\mathbf{2}} &= \Pi_{r=0}^{N-k-1}(n+2r+1) \\ &= (2N+1)(2N+3)\cdots(4N-2k-1) \\ &= \frac{(2N+1)(2N+2)\cdots(4N-2k-1)(4N-2k)}{(2N+2)(2N+4)\cdots(4N-2k)} \\ &= \frac{(4N-2k)!}{(2N)!(2N)(4N)\cdots(4N-2k)} \\ &= \frac{(4N-2k)!}{(2N)!2^{N-k}(N+1)(N+2)\cdots(2N-k)} \\ &= \frac{(4N-2k)!N!}{(2N)!2^{N-k}(2N-k)!} \\ &= \frac{(2n-2k)!N!}{(n)!2^{N-k}(n-k)!} \quad \text{because } n=2N. \\ \boxed{\mathbf{3}} &= \Pi_{r=0}^{N-k-1}(2r+1)(2r+2) \\ &= [1\cdot 2][3\cdot 4]\cdots[(2N-2k-1)(2N-2k)] \\ &= (n-2k)! \quad \text{because } n=2N. \end{aligned}$$

Then

$$\begin{aligned} c_{2p+2} &= \frac{(-1)^{2N-k}}{2^n(N!)^2}\frac{\boxed{\mathbf{1}}\,\boxed{\mathbf{2}}}{\boxed{\mathbf{3}}} \\ &= \frac{(-1)^{2N-k}}{2^n(N!)^2}\,\frac{2^{N-k}\frac{N!}{k!}\,\frac{(2n-2k)!N!}{(n)!2^{N-k}(n-k)!}}{(n-2k)!} \\ &= \frac{(-1)^k}{2^nk!(n-2k)!(n-k)!}. \end{aligned}$$

This completes the derivation of the Legendre polynomial formula. ■

Derivation of Rodrigues' Formula

It must be shown that Legendre's polynomial formula

$$P_n(x) = \frac{1}{2^n}\sum_{k=0}^{N} \frac{(-1)^k(2n-2k)!}{k!(n-2k)!(n-k)!}\, x^{n-2k},$$

derived above from ordinary point theory applied to Legendre's differential equation, is also given by Rodrigues' formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left((x^2-1)^n\right).$$

Proof: Start with the binomial expansion $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$. Substitute $a=-1$, $b=x^2$, $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$ to obtain

$$(-1+x^2)^n = \sum_{k=0}^{n} \frac{(-1)^k n!}{k!(n-k)!} x^{2n-2k}.$$

The plan is to differentiate this equation n times. Calculus derivative $(d/du)^n u^m$ can be written as $\frac{m!}{(m-n)!}u^{m-n}$. Each differentiation annihilates the constant term. Therefore, there are $N=n/2$ terms for n even and $N=(n-1)/2$ terms for n odd, and we have

$$\begin{aligned}\frac{d^n}{dx^n}\left((-1+x^2)^n\right) &= \sum_{k=0}^{N}\frac{(-1)^k n!(2n-2k)!}{k!(n-k)!(n-2k)!}x^{n-2k}\\ &= n!\,2^n\,\frac{1}{2^n}\sum_{k=0}^{N}\frac{(-1)^k(2n-2k)!}{k!(n-k)!(n-2k)!}x^{n-2k}\\ &= 2^n n! P_n(x).\end{aligned}$$

■

Derivation of Bonnet's Recursion

Proof: Rodrigues' formula will be used to define P_n:

$$P_n(x) = c_n D^n(u^n), \quad u = x^2-1, \quad D = \frac{d}{dx}, \quad c_n = \frac{1}{n!2^n}$$

To be proved is **Bonnet's recursion**:

$$P_{n+1}(x) = \frac{2n+1}{n+1}xP_n(x) - \frac{n}{n+1}P_{n-1}(x)$$

Lemma A. $c_m = 2(m+1)c_{m+1}$

Lemma B. Bonnet's recursion is equivalent to the identity

$$D^{n+1}u^{n+1} = 2(2n+1)xD^n u^n - 4n^2 D^{n-1}u^{n-1} \qquad (2)$$

Lemma C. $D^{n+1}u^{n+1} = 2(n+1)(2n+1)D^{n-1}u^n + 4n(n+1)D^{n-1}u^{n-1}$

Lemma D. $(n+1)D^{n-1}u^n = xD^nu^n - 2nD^{n-1}u^{n-1}$

Proof of Bonnet's recursion: Let's verify that equation (2) in Lemma B is satisfied:

$$\begin{aligned}
\text{LHS} &= D^{n+1}u^{n+1} && \text{Left side of (2).}\\
&= \left\{\begin{array}{l} 2(n+1)(2n+1)D^{n-1}u^n \\ +4n(n+1)D^{n-1}u^{n-1}\end{array}\right. && \text{By Lemma C.}\\
&= \left\{\begin{array}{l} 2(2n+1)\ \left(\boldsymbol{xD^nu^n - 2nD^{n-1}u^{n-1}}\right) \\ \quad +4n(n+1)D^{n-1}u^{n-1}\end{array}\right. && \text{By Lemma D.}\\
&= \left\{\begin{array}{l} 2(2n+1)xD^nu^n \\ +4n\left(-(2n+1)+(n+1)\right)D^{n-1}u^{n-1}\end{array}\right. && \text{Expand.}\\
&= 2(2n+1)xD^nu^n - 4n^2D^{n-1}u^{n-1} && \text{Which equals the RHS of (2) in Lemma B.}
\end{aligned}$$

This completes the proof of Bonnet's recursion, except for proofs of the lemmas.

Proof of Lemma A: See the Exercise 3 solution. ■

Proof of Lemma B: Replace $P_k(x)$ by $c_kD^ku^k$ in Bonnet's recursion:

$$(n+1)c_{n+1}D^{n+1}u^{n+1} = (2n+1)xD^nu^n - nc_{n-1}D^{n-1}u^{n-1}$$

Divide by $(n+1)c_{n+1}$ and simplify using Lemma A:

$$\begin{aligned}
D^{n+1}u^{n+1} &= \frac{2n+1}{(n+1)c_{n+1}}xD^nu^n - \frac{nc_{n-1}}{(n+1)c_{n+1}}D^{n-1}u^{n-1}\\
&= 2(n+1)xD^nu^n - \frac{2n(n)(2)(n+1)c_{n+1}}{(n+1)c_{n+1}}D^{n-1}u^{n-1}\\
&= 2(n+1)xD^nu^n - 4n^2D^{n-1}u^{n-1}
\end{aligned}$$

All steps are reversible, so Bonnet's recursion is equivalent to equation (2). ■

Proof of Lemma C: Let's write $x^2 = (x^2-1)+1 = u+1$ to strip symbol x from the expansion. The calculus product rule and definition $u = x^2-1$ gives

$$\begin{aligned}
D^{n+1}u^{n+1} &= D^{n-1}\left(D\left(Du^{n+1}\right)\right)\\
&= D^{n-1}\left(D\left(2(n+1)xu^n\right)\right)\\
&= D^{n-1}\left(2n(n+1)u^n + 4n(n+1)x^2u^{n-1}\right)\\
&= D^{n-1}\left(2n(n+1)u^n + 4n(n+1)(u+1)u^{n-1}\right)\\
&= D^{n-1}\left(2n(n+1)(2n+1)u^n + 4n(n+1)u^{n-1}\right)\\
&= 2n(n+1)(2n+1)D^{n-1}u^n + 4n(n+1)D^{n-1}u^{n-1}
\end{aligned}$$

■

Proof of Lemma D: The Leibitz Rule for differentiation of a power $(fg)^k$ gives

$$D^n(xu^n) = xD^nu^n - 2nD^{n-1}u^n \tag{3}$$

because there are only two nonzero terms $\binom{n}{r}D^r(x)D^{n-r}(u^n)$ in the Leibnitz identity.

The calculus product rule gives

$$D^n(xu^n) = (2n+1)D^{n-1}u^n + 2nD^{n-1}u^{n-1} \tag{4}$$

because

$$\begin{aligned} D^n(D^n(xu^n) &= D^{n-1}(D(xu^n)) \\ &= D^{n-1}(u^n + 2nx^2u^{n-1} \\ &= D^{n-1}u^n + 2nD^{n-1}((u+1)u^{n-1}) \\ &= (2n+1)D^{n-1}u^n + 2nD^{n-1}u^{n-1} \end{aligned}$$

Subtract equation (4) from equation (3).

$$\begin{aligned} 0 &= xD^nu^n + nD^{n-1}u^n - (2n+1)D^{n-1}u^n - 2nD^{n-1}u^n - 2nD^{n-1}u^{n-1} \\ &= xD^nu^n - (n+1)D^{n-1}u^n - 2nD^{n-1}u^{n-1} \end{aligned}$$

Rearrange this equation to

$$(n+1)D^{n-1}u^n = xD^nu^n - 2nD^{n-1}u^{n-1}$$

$\blacksquare$

Exercises 12.7

Equivalent Legendre Equations

Prove the following are equivalent to $(1{-}x^2)y''{-}2xy'{+}n(n{+}1)y{=}0$

1. $((1-x^2)y')' + n(n+1)y = 0$

2. Let $x = \cos\theta$, $' = \frac{d}{d\theta}$, then $\sin\theta y''{+}\cos\theta y'{+}n(n{+}1)\sin\theta y{=}0$.

Proof of Bonnet's Recursion

3. Define $c_n = \frac{1}{n!2^n}$. Prove $c_m = 2(m+1)c_{m+1}$.

4. Let $D = \frac{d}{dx}$, $u = x^2 - 1$. Verify $D^2u^2 = 12x^2 - 4$ using D and the binomial theorem.

5. Prove Bonnet's recursion from the generating function equation

$$\frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{\infty} P_n(x)t^n$$

6. Prove that $P_n(1) = 1$ directly from Rodrigues' formula.

Boundary Data at $x = \pm 1$

Use these identities:

(1) $(a+b)^k{=}\sum_{r=0}^{k}\binom{k}{r}a^rb^{k-r}$
(2) $(uv)^{(n)}{=}\sum_{r=0}^{n}\binom{n}{r}u^{(r)}v^{(n-r)}$

7. In Rodrigues' formula, let Let $y = x - 1$ to prove

$$P_n(y+1){=}\frac{1}{n!2^n}\left(\frac{d}{dy}\right)^n\left(y^2+2y\right)^n$$

8. Verify from identity (1): $\left(y^2{+}2y\right)^n = \sum_{r=0}^{n}\binom{n}{r}2^ry^{2n-r}$

9. Prove $P_n(1) = 1$ from Bonnet's recursion.

10. Assume $P_n(-x){=}(-1)^nP_n(x)$ and $P_n'(1) = \dfrac{n(n+1)}{2}$. Prove $P_n(-1) = (-1)^n$ and $P_n'(-1) = (-1)^n\dfrac{n(n+1)}{2}$.

Legendre Integrals

Use Legendre properties page 986.

11. Use $(2n{+}1)P_n = P_{n+1}' - P_{n-1}'$ to prove $\int_0^1 P_n(x)dx = 0$ for $n > 0$ even.

12. Use Bonnet's recursion to show that $\int_0^1 P_n(x)dx = \frac{P_{n-1}(0)}{n+1}$ for $n > 0$.

12.8 Orthogonality

The notion of orthogonality originates in $\mathcal{R}^3$, where nonzero vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$ are defined to be **orthogonal**, written $\vec{\mathbf{v}}_1 \perp \vec{\mathbf{v}}_2$, provided $\vec{\mathbf{v}}_1 \cdot \vec{\mathbf{v}}_2 = 0$. The dot product in $\mathcal{R}^3$ is defined by

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1y_1 + x_2y_2 + x_3y_3.$$

Similarly, $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$ defines the dot product in $\mathcal{R}^n$. Literature uses the notation $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ as well as $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}$. Modern terminology uses **Inner Product** instead of **dot product**, to emphasize the use of functions and abstract properties. The inner product $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ satisfies the following properties.

$(\vec{\mathbf{x}}, \vec{\mathbf{x}}) \ge 0$	Non-negativity.
$(\vec{\mathbf{x}}, \vec{\mathbf{x}}) = 0$ implies $\vec{\mathbf{x}} = \vec{\mathbf{0}}$	Uniqueness.
$(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = (\vec{\mathbf{y}}, \vec{\mathbf{x}})$	Symmetry.
$k(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = (k\vec{\mathbf{x}}, \vec{\mathbf{y}})$	Homogeneity.
$(\vec{\mathbf{x}} + \vec{\mathbf{y}}, \vec{\mathbf{z}}) = (\vec{\mathbf{x}}, \vec{\mathbf{z}}) + (\vec{\mathbf{y}}, \vec{\mathbf{z}})$	Additivity.

The storage system of choice for answers to differential equations is a real vector space V of functions f. A **real inner product space** is a vector space V with real-valued inner product function $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ defined for each $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$ in V, satisfying the preceding rules.

Dot Product for Functions

The extension of the notion of dot product to functions replaces $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}$ by average value. Insight can be gained from the approximation

$$\frac{1}{b-a}\int_a^b F(x)dx \approx \frac{F(x_1) + F(x_2) + \cdots + F(x_n)}{n}$$

where $b - a = nh$ and $x_k = a + kh$. The left side of this approximation is called the **average value of** F **on** $[a, b]$. The right side is the classical average of F at n equally spaced values in $[a, b]$. If we replace F by a product fg, then the average value formula reveals that $\int_a^b fgdx$ acts like a dot product:

$$\frac{1}{b-a}\int_a^b f(x)g(x)dx \approx \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}}{n}, \quad \vec{\mathbf{x}} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}, \quad \vec{\mathbf{y}} = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}.$$

The formula says that $\int_a^b f(x)g(x)dx$ is approximately a constant multiple of the dot product of samples of f, g at n points of $[a, b]$.

Given functions f and g integrable on $[a, b]$, the formula

$$(f, g) = \int_a^b f(x)g(x)dx$$

defines a dot product satisfying the abstract properties cited above. When dealing with solutions to differential equations, this dot product, along with the abstract properties of a dot product, provide the notions of *distance* and *orthogonality* analogous to those in $\mathcal{R}^3$.

Orthogonality, Norm and Distance

Define nonzero functions f and g to be **orthogonal** on $[a, b]$ provided $(f, g) = 0$. Define the **norm** or the **distance** from f to 0 to be the number $\|f\| = \sqrt{(f, f)}$ and the distance from f to g to be $\|f - g\|$. The basic properties of the norm $\|\cdot\|$ are as follows.

$\|f\| \geq 0$	Non-negativity.
$\|f\| = 0$ implies $f = 0$	Uniqueness.
$\|cf\| = \|c\|\|f\|$	Homogeneity.
$\|f\| = \sqrt{(f, f)}$	Norm and the inner product.
$\|f + g\| \leq \|f\| + \|g\|$	The triangle inequality.
$\|(f, g)\| \leq \|f\|\,\|g\|$	Cauchy-Schwartz inequality.

Weighted Dot Product

In applications of Bessel functions, use is made of the **weighted dot product**

$$(f, g) = \int_a^b f(x)g(x)\rho(x)dx,$$

where $\rho(x) > 0$ on $a < x < b$.

The possibility that $\rho(x) = 0$ at some set of points in (a, b) has been considered by researchers, as well as the possibility of a singularity at $x = a$ or $x = b$, or $a = -\infty$ and/or $b = \infty$. Properties advertised here mostly hold in these extended cases, provided appropriate additional assumptions are invoked.

Theorem 12.9 (Orthogonality of Legendre Polynomials)
The Legendre polynomials $\{P_n\}_{n=0}^{\infty}$ satisfy the orthogonality relation

$$\int_{-1}^{1} P_n(x)P_m(x)dx = 0, \quad n \neq m.$$

The relation means that P_n and P_m $(n \neq m)$ are orthogonal on $[-1, 1]$ relative to the dot product $(f, g) = \int_{-1}^{1} f(x)g(x)dx$.

Proof: The details use only the Legendre differential equation $(1-x^2)y''-2xy'+n(n+1)y=0$ in the form $((1-x^2)y')'+n(n+1)y=0$ and the fact that $a(x)=1-x^2$ is zero at $x=\pm 1$. From the definition of the Legendre polynomials, the following differential equations are satisfied:

$$\begin{aligned}(aP_n')'+n(n+1)P_n &= 0,\\ (aP_m')'+m(m+1)P_m &= 0.\end{aligned}$$

Multiply the first by P_m and the second by P_n, then subtract to obtain

$$(m(m+1)-n(n+1))P_nP_m=(aP_n')'P_m-(aP_m')'P_n.$$

Re-write the right side of this equation as $(aP_n'P_m-aP_m'P_n)'$. Then integrate over $-1<x<1$ to obtain

$$\begin{aligned}\text{LHS} &= (m(m+1)-n(n+1))\int_{-1}^{1}P_n(x)P_m(x)dx\\ &= (a(x)P_n'(x)P_m(x)-a(x)P_m'(x)P_n(x))|_{x=-1}^{x=1}\\ &= 0.\end{aligned}$$

The result is zero because $a(x)=1-x^2$ is zero at $x=1$ and $x=-1$. The dot product of P_n and P_m is zero, because $m(m+1)-n(n+1)\neq 0$. ■

Theorem 12.10 (Orthogonality of Bessel Functions)
Let the Bessel function J_n have positive zeros $\{b_{mn}\}_{m=1}^{\infty}$. Given $R>0$, define $f_m(r)=J_n(b_{mn}r/R)$. Then the following weighted orthogonality relation holds.

$$\int_0^R f_i(r)f_j(r)rdr=0,\quad i\neq j.$$

The relation means that f_i and f_j $(i\neq j)$ are orthogonal on $[0,R]$ relative to the weighted dot product $(f,g)=\int_0^R f(r)g(r)\rho(r)dr$, where $\rho(r)=r$.

Proof: The details depend entirely upon the Bessel differential equation of order n, $x^2y''+xy'+(x^2-n^2)y=0$, and the condition $y(b_{mn})=0$, valid when $y=J_n$. Let $\lambda=b_{mn}/R$ and change variables by $x=\lambda r$, $w(r)=y(\lambda r)$. Then w satisfies $dw/dr=y'(x)\lambda$, $d^2w/dr^2=y''(x)\lambda^2$ and the differential equation for y implies the equation

$$r^2\frac{d^2w}{dr^2}(r)+r\frac{dw}{dr}(r)+(\lambda^2r^2-n^2)w(r)=0.$$

Apply this change of variables to Bessel's equation of orders i and j. Then

$$\begin{aligned}r^2f_i''(r)+rf_i'(r)+(b_{in}^2r^2R^{-2}-n^2)f_i(r) &= 0,\\ r^2f_j''(r)+rf_j'(r)+(b_{jn}^2r^2R^{-2}-n^2)f_j(r) &= 0.\end{aligned}$$

Multiply the first equation by $f_j(r)$ and the second by $f_i(r)$, then subtract and divide by r to obtain

$$rf_i''f_j-rf_j''f_i+f_i'f_j-f_j'f_i+(b_{in}^2-b_{jn}^2)rR^{-2}f_if_j=0.$$

Because of the calculus identities $rw'' + w' = (rw')'$ and $(rw_1'w_2 - rw_2'w_2)' = (rw_1')'w_2 - (rw_2')'w_1$, this equation can be re-written in the form

$$(b_{jn}^2 - b_{in}^2)R^{-2}rf_if_j = (rf_i'f_j - rf_j'f_i)'.$$

Integrate this equation over $0 < r < R$. Then the right side evaluates to zero, because of the conditions $f_i(R) = f_j(R) = 0$. The left side evaluates to a nonzero multiple of $\int_0^R f_i(r)f_j(r)rdr$. Therefore, the weighted dot product of f_i and f_j is zero. ■

Series of Orthogonal Functions

Let (f, g) denote a dot product defined for functions f, g. Especially, we include $(f, g) = \int_a^b fgdx$ and a weighted dot product $(f, g) = \int_a^b fg\rho dx$. Let $\{f_n\}$ be a sequence of nonzero functions orthogonal with respect to the dot product (f, g), that is, a system $\{f_n\}_{n=1}^{\infty}$ satisfying the **orthogonality relations**

$$(f_i, f_j) = 0, \quad i \neq j, \quad (f_i, f_i) > 0, \quad i = 1, 2, \ldots.$$

A **Generalized Fourier series** is a convergent series of such orthogonal functions

$$F(x) = \sum_{n=1}^{\infty} c_n f_n(x).$$

The coefficients $\{c_n\}$ are called the **Generalized Fourier Coefficients** of F. Convergence is taken in the sense of the norm $\|g\| = \sqrt{(g, g)}$, defined as follows:

$$F = \sum_{n=1}^{\infty} c_n f_n \quad \text{means} \quad \lim_{N\to\infty} \left\| \sum_{n=1}^{N} c_n f_n - F \right\| = 0.$$

For example, when $\|g\| = \sqrt{(g, g)}$ and $(f, g) = \int_a^b fgdx$, then series convergence is called **Mean-Square convergence**, defined by

$$\lim_{N\to\infty} \sqrt{\int_a^b \left| \sum_{n=1}^{N} c_n f_n(x) - F(x) \right|^2 dx} = 0.$$

Orthogonal Series Method. The coefficients $\{c_n\}$ in an orthogonal series are determined by a technique called the **Orthogonal series method**, described in words as follows.

> The coefficient c_n in an orthogonal series is found by taking the dot product of the equation with the orthogonal function that multiplies c_n.

The details of the method:

$$(F, f_n) = \left(\sum_{k=1}^{\infty} c_k f_k, f_n \right) \qquad \text{Dot product the equation with } f_n.$$

$(F, f_n) = \sum_{k=1}^{\infty} c_k (f_k, f_n)$ Apply dot product properties.

$(F, f_n) = c_n (f_n, f_n)$ By orthogonality, just one term remains from the series on the right.

Division after the last step leads to the **Fourier Coefficient Formula**

$$c_n = \frac{(F, f_n)}{(f_n, f_n)}.$$

Orthogonal Projection

The **shadow projection** of vector $\vec{X}$ onto the direction of vector $\vec{Y}$ is the number d defined by

$$d = \frac{\vec{X} \cdot \vec{Y}}{|\vec{Y}|}.$$

The triangle determined by $\vec{X}$ and $d\dfrac{\vec{Y}}{|\vec{Y}|}$ is a right triangle.

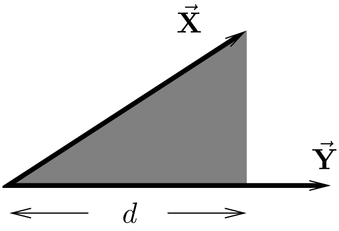


Figure 1. Shadow projection d of vector $\vec{\mathbf{X}}$ onto the direction of vector $\vec{\mathbf{Y}}$.

The **vector shadow projection** of $\vec{X}$ onto the line L through the origin in the direction of $\vec{Y}$ is defined by

$$\mathbf{proj}_{\vec{Y}}(\vec{X}) = d\frac{\vec{Y}}{|\vec{Y}|} = \frac{\vec{X} \cdot \vec{Y}}{\vec{Y} \cdot \vec{Y}}\vec{Y}.$$

Shadow Projection and Fourier Coefficients

The term $c_n f_n$ in a generalized Fourier series can be expanded as

$$c_n f_n = \frac{(F, f_n)}{(f_n, f_n)} f_n = \text{ Shadow projection of } F \text{ onto } f_n.$$

This formula appears in Gram-Schmidt formulas and in Least Squares formulas, because those formulas also involve orthogonal projections. The complexity of such formulas is removed by thinking of the results as sums of shadow projections or as subtractions of shadow projections.

Bessel inequality and Parseval equality

Assume given a dot product (f,g) for an orthogonal series expansion

$$F(x) = \sum_{n=1}^{\infty} c_n f_n(x).$$

Bessel's inequality

$$\sum_{n=1}^{N} \frac{|(F,f_n)|^2}{\|f_n\|^2} \le \|F\|^2$$

is proved as follows. Let $N \ge 1$ be given and let $S_N = \sum_{n=1}^{N} c_n f_n$. Then

$$(S_N, S_N) = \left(\sum_{n=1}^{N} c_n f_n, \sum_{k=1}^{N} c_k f_k\right)$$ Definition of S_N.

$$= \sum_{n=1}^{N}\sum_{k=1}^{N} c_n c_k (f_n, f_k)$$ Linearity properties of the dot product.

$$= \sum_{n=1}^{N} c_n c_n (f_n, f_n)$$ Because $(f_n, f_k) = 0$ for $n \ne k$.

$$= \sum_{n=1}^{N} |c_n|^2 \|f_n\|^2$$ Because $\|g\|^2 = (g,g)$.

$$(F, S_N) = \sum_{n=1}^{N} c_n (F, f_n)$$ Linearity of the dot product.

$$= \sum_{n=1}^{N} |c_n|^2 \|f_n\|^2$$ Fourier coefficient formula.

Then

$$0 \le \|F - S_N\|^2$$ The norm is non-negative.

$$= (F - S_N, F - S_N)$$ Use $\|g\|^2 = (g,g)$.

$$= (F,F) + (S_N, S_N) - 2(F, S_N)$$ Dot product properties.

$$= (F,F) - \textstyle\sum_{n=1}^{N} |c_n|^2 \|f_n\|^2$$ Apply previous formulas.

This proves

$$\sum_{n=1}^{N} |c_n|^2 \|f_n\|^2 \le (F,F),$$

or what is the same, because of the Fourier coefficient formula,

$$\sum_{n=1}^{N} \frac{|(F,f_n)|^2}{\|f_n\|^2} \le (F,F).$$

Letting $N \to \infty$ gives Bessel's inequality $\sum_{n=1}^{\infty} \frac{|(F,f_n)|^2}{\|f_n\|^2} \le (F,F)$.

Parseval's equality is equality in Bessel's inequality:

$$\|F\|^2 = \sum_{n=1}^{N} \frac{|(F,f_n)|^2}{\|f_n\|^2}.$$

There is a fundamental relationship between Parseval's equality and the possibility to expand a function F as an infinite orthogonal series in the functions $\{f_n\}$. In literature, the relationship is known as **completeness** of the orthogonal sequence $\{f_n\}$. The definition: $\{f_n\}$ is complete if and only if each function F has a series expansion $F = \sum_{n=1}^{\infty} c_n f_n$ for some set of coefficients $\{c_n\}$. When equality holds, the coefficients c_n are given by Fourier's coefficient formula.

Theorem 12.11 (Parseval)
A sequence $\{f_n\}$ is a complete orthogonal sequence if and only if Parseval's equality holds.

Therefore, the equation $F = \sum_{n=1}^{\infty} \frac{(F,f_n)}{(f_n,f_n)} f_n$ holds for every F if and only if Parseval's equality holds for every F.

Legendre series

A convergent series of the form

$$F(x) = \sum_{n=0}^{\infty} c_n P_n(x)$$

is called a **Legendre series**. The orthogonal system $\{P_n\}$ on $[-1,1]$ under the dot product $(f,g) = \int_{-1}^{1} f(x)g(x)dx$ together with Fourier's coefficient formula gives

$$c_n = \frac{\int_{-1}^{1} F(x)P_n(x)dx}{\int_{-1}^{1} |P_n(x)|^2 dx}.$$

The denominator in this fraction can be evaluated for all values of n:

$$\int_{-1}^{1} |P_n(x)|^2 dx = \frac{2}{2n+1}.$$

Theorem 12.12 (Legendre expansion)
Let F be defined on $-1 \le x \le 1$ and assume F and F' are piecewise continuous. Then the Legendre series expansion of F converges and equals $F(x)$ at each point of continuity of F. At other points, the series converges to $\frac{1}{2}(F(x+) + F(x-))$.

Bessel series

A convergent infinite series of the form

$$F(r) = \sum_{n=1}^{\infty} c_n J_m(b_{nm} r/R), \quad 0 < r < R,$$

is called a **Bessel series**. The index m, assumed here to be a non-negative integer, is fixed throughout the series terms. The sequence $\{b_{nm}\}_{n=1}^{\infty}$ is an ordered list of the positive zeros of J_m.

The weighted dot product $(f, g) = \int_0^R f(r)g(r)rdr$ is used. It is known that the sequence of functions $f_n(r) = J_m(b_{nm}r/R)$ is orthogonal relative to the weighted dot product $(\cdot, \cdot)$. Then Fourier's coefficient formula implies

$$c_n = \frac{\int_0^R F(r)J_m(b_{nm}r/R)rdr}{\int_0^R |J_m(b_{nm}r/R)|^2 rdr}.$$

To evaluate the denominator of the above fraction, let's denote $' = d/dr$, $y(r) = f_n(r) = J_m(b_{nm}r/R)$. Use $r(ry')' + (b_{nm}^2 r^2 R^{-2} - n^2)y = 0$, the equation used to prove orthogonality of Bessel functions. Multiply this equation by $2y'$. Re-write the resulting equation as

$$[(ry')^2]' + (b_{nm}^2 r^2 R^{-2} - n^2)[y^2]' = 0.$$

Integrate this last equation over $[0, R]$. Use parts on the term involving $r^2[y^2]'$. Then use $J_m(0) = 0$, $y' = (b_{nm}/R)J_m'(b_{nm}r/R)$ and $xJ_m'(x) = mJ_m(x) - xJ_{m+1}(x)$ to obtain

$$\int_0^R |J_m(b_{nm}r/R)|^2 rdr = \frac{R^2}{2}|J_{m+1}(b_{nm})|^2.$$

Theorem 12.13 (Bessel expansion)

Let F be defined on $0 \le x \le R$ and assume F and F' are piecewise continuous.

Then the Bessel series expansion of F converges and equals $F(x)$ at each point of continuity of F. At other points, the series converges to the average $\frac{1}{2}(F(x+) + F(x-))$ of left-hand and right-hand limits.

Exercises 12.8

Legendre series. Establish the following results.

1. Prove using orthogonality that $\int_{-1}^{1} P_n(x)F(x)dx = 0$ for any polynomial $F(x)$ of degree less than n.

2. Use identity
$xP_n'(x) - P_{n-1}'(x) = nP_n(x)$
to prove $\int_{-1}^{1} |P_n(x)|^2 dx = \frac{2}{2n+1}$.

3. Let $\left\langle f, g \right\rangle = \int_0^{\pi} f(x)g(x)\sin(x)dx$. Show that the sequence $\{P_n(\cos x)\}$ is orthogonal on $0 \le x \le \pi$ with respect to inner product $\left\langle f, g \right\rangle$.

4. Let $F(x) = \sin^3(x) - \sin(x)\cos(x)$. Expand F as a Legendre series $F(x) = \sum_{n=0}^{\infty} c_n P_n(\cos x)$.

Chebyshev Series. The **Chebyshev polynomials** are $T_n(x) = \cos(n \arccos(x))$ with inner product $(f, g) = \int_{-1}^{1} f(x)g(x)(1-x^2)^{-1/2}dx$.

5. Show that $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$.

6. Show that $T_3(x) = 4x^3 - 3x$.

7. Prove that (f, g) satisfies the abstract properties of an inner product.

8. Show that T_n is a solution of the **Chebyshev equation** $(1-x^2)y'' - xy' + n^2y = 0$.

9. Prove that $\{T_n\}$ is orthogonal relative to the weighted inner product (f, g).

10. Prove: $T_n(x)$ is an even function for n even and an odd function for n odd.

Hermite Polynomials. Define the Hermite polynomials by $H_0(x) = 1$,
$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\left(e^{-x^2}\right)$.
Define the inner product
$(f, g) = \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2}dx$.

11. Verify: $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$, $H_3(x) = 8x^3 - 12x$, $H_4(x) = 16x^4 - 48x^2 + 12$.

12. Prove: $H_n(-x) = (-1)^n H_n(x)$.

13. Prove $H_n'(x) = 2xH_n(x) - H_{n+1}(x)$. Then use recursion $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$ to show $H_n'(x) = 2nH_{n-1}(x)$.

14. Let $y = H_5 = 32x^5 - 160x^3 + 120x$. Show y satisfies **Hermite's equation** $y'' - 2xy' + 2ny = 0$ with $n = 5$.

15. Prove recursion
$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$.

16. Show that the sequence $\{H_n(x)\}$ is orthogonal with respect to (f, g).

Alternate Laguerre Polynomials. Define the alternate Laguerre polynomials by $L_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x})$. Define $(f, g) = \int_0^{\infty} f(x)g(x)e^{-x}dx$. A **warning**: Laguerre polynomials in the literature are $\frac{1}{n!}L_n$.

17. Prove: $L_1(x) = 1 - x$ and
$L_2(x) = 2 - 4x + x^2$.

18. Prove:
$L_3(x) = 6 - 18x + 9x^2 - x^3$.

19. Prove that (f, g) satisfies the abstract properties for an inner product.

20. Show that L_0, L_1, L_2, L_3 are orthogonal with respect to the inner product (f, g), using direct integration methods.

21. Prove:
$L_n(x) = \sum_{k=0}^{n} \frac{(-1)^k\,(n!)^2}{(n-k)!(k!)^2} x^k$.

22. Show that $\{L_n\}$ is an orthogonal sequence with respect to (f, g).

23. Find an expression for a polynomial solution to **Laguerre's equation** $xy'' + (1-x)y' + ny = 0$ using Frobenius theory.

24. Show that $y = e^x \frac{d^n}{dx^n}(x^n e^{-x})$ satisfies **Laguerre's equation**: $xy'' + (1-x)y' + ny = 0$.

25. Verify by computer the **Laguerre** formulas

$$\begin{aligned}
L_0(x) &= 1\\
L_1(x) &= -x + 1\\
L_2(x) &= x^2 - 4x + 2\\
L_3(x) &= -x^3 + 9x^2 - 18x + 6
\end{aligned}$$

26. Find to 6 digits by computer the roots of $L_4(x)$.

27. Prove: Up to a constant, L_n is the only polynomial solution of $xy'' + (1-x)y' + ny = 0$, $n \ge 0$ an integer.

28. Assume standard Laguerre polynomials $\{\mathcal{L}_n\}$ satisfy recurrence

$$\begin{aligned}
(n+1)\mathcal{L}_{n+1}(x) = {} & (2n+1-x)\mathcal{L}_n(x)\\
& -n\mathcal{L}_{n-1}(x).
\end{aligned}$$

Prove: The alternate Laguerre polynomials $\{L_n\}$ satisfy recurrence

$$\begin{aligned}
L_{n+1}(x) = {} & (2n+1-x)L_n(x)\\
& -n^2L_{n-1}(x).
\end{aligned}$$

Chapter A

Background Topics

Introduction

The appendices to follow contain a short list of topics extracted from pre-calculus and calculus courses.

Contents

A.1 Calculus

The selected topics from differential and integral calculus are used in differential equations. The special notation of differential equations is introduced, along with some ideas of Isaac Newton concerning the elementary kinetics formula $D = RT$, which has the physical interpretation $\mathsf{Distance} = \mathsf{Rate} \times \mathsf{Time}$.

Derivative

The calculus derivative $f'(x_0) = \lim_{h \to 0} \dfrac{f(x_0 + h) - f(x_0)}{h}$ makes sense provided the indicated limit exists. Implicit in the formula is the *assumption* that f is defined in an open interval of the form $|x - x_0| < H$. Differential equations use this standard notation, plus the **Leibniz notation**

$$\frac{df}{dx} = f'(x).$$

Variable names used in science and engineering often follow this standard:

$$\begin{aligned} y &= \text{dependent variable},\\ x &= \text{independent variable}. \end{aligned}$$

Within certain disciplines, such as kinetics, the variable names change, and the following standard exists:

$$\begin{aligned} x &= \text{displacement, dependent variable},\\ t &= \text{time, independent variable}, \end{aligned}$$

$$\begin{aligned} \frac{dx}{dt} &= \text{velocity} & \frac{d^2x}{dt^2} &= \text{acceleration}\\ &= x'(t) & &= x''(t)\\ &= \dot{x}(t) & &= \ddot{x}(t)\\ &= Dx(t), & &= D^2x(t). \end{aligned}$$

The functional notation $y(x)$ means y is a dependent variable which depends on the independent variable x. For example, $x(t)$ means displacement x depends on time t. In a graphic, it is expected that x is the vertical axis and t is the horizontal axis. The dot-notation $\dot{x}(t)$ and $\ddot{x}(t)$, instead of $x'(t)$ and $x''(t)$, is common in literature on statics and dynamics. Operator notation Dx, D^2x appears in differential equations literature and in computer algebra systems, e.g., `maple` and `mathematica`.

Slope, Rates and Averages

The derivative can be interpreted geometrically as the **slope** of the line tangent to a curve at a point; see Figure 1.

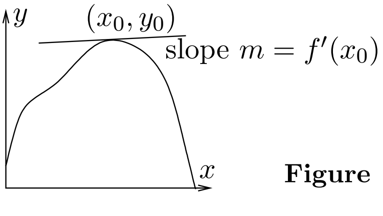


Figure 1. Slope of the tangent line.

The **tangent line** itself can be viewed as the **linearization** of the curve. For example, if the curve is the path of an automobile which at speedometer reading v instantly skids off the road, then the car follows the tangent line with constant speed v. Travel along the tangent line is **linear motion** at constant speed.

The line equation tangent to $y = f(x)$ at $x = x_0$ is given by the **point-slope form** of a line

$$\begin{aligned} y - y_0 &= m(x - x_0),\\ y_0 &= f(x_0), \quad m = f'(x_0). \end{aligned}$$

The notation $y(x)$, usual in differential equations, conflicts with the notation from geometry. In handwritten and blackboard work it is recommended to change x and y to capital letters X and Y, then replace f by y, as follows:

$$\begin{aligned} &Y - y_0 = m(X - x_0), \\ &y_0 = y(x_0), \quad m = y'(x_0). \end{aligned}$$

Other forms of a straight line in coordinate geometry are the **slope-intercept form** $y = mx + b$, the **standard form** $Ax + By + C = 0$ and the **parametric form**

$$\begin{cases} x = x_0 + at, \\ y = y_0 + bt, \quad -\infty < t < \infty. \end{cases}$$

In the parametric form, the vector $a\vec{\mathbf{i}} + b\vec{\mathbf{j}}$ is tangent to the line. For example, $a = 0$ and $b = 1$ gives a vertical line through (x_0, y_0).

Applied sciences interpret the derivative $f'(x)$ as the **rate of change** of $y = f(x)$ with respect to x. Typical interpretations appear below.

$\dot{x}(t)$	$\approx$	change in displacement x for a unit change in t
$\dfrac{dQ}{dt}$	$\approx$	change in charge Q for a unit change in t
$\ddot{Q}(t)$	$\approx$	change in current $I = \dot{Q}$ for a unit change in t
$A'(t)$	$\approx$	expected decrease in the amount A of radioactive material for time interval $[t, t+1]$

The **average** of n samples $y_1, \ldots, y_n$ is defined to be

$$\frac{y_1 + y_2 + \cdots + y_n}{n}.$$

The term **simple average** is sometimes used. The **average value** $\overline{f}$ of a continuous function $f(x)$ on $[a, b]$ is defined by

$$\overline{f} = \frac{\int_a^b f(x)dx}{b - a}.$$

This abstract notion has connections with the simple average. The theory of the integral $\int_a^b f(x)dx$ includes the **rectangular rule** for numerical integration. For step size $h = (b - a)/n$ and sample values $y_1 = f(a)$, $y_2 = f(a + h)$, $\ldots$, $y_n = f(a + nh - h)$ it gives the approximation formula

$$\int_a^b f(x)dx \approx h(y_1 + y_2 + \cdots + y_n).$$

Multiply this relation by $1/(b - a)$ and replace the left side by the average value $\overline{f}$. Then

$$\overline{f} \approx \frac{y_1 + y_2 + \cdots + y_n}{n},$$

or in words,

> The average value $\overline{f}$ is approximately a simple average of n samples of f, taken at equi-spaced points in $[a, b]$.

In the language of kinetics, f is **velocity** and $\overline{f}$ is the **average velocity** or the **speed**.

The language of kinetics agrees with common public notions of speed. For example, the average of various speedometer reading samples during an automobile trip give a good indication of the average speed of the car on the trip. The average speed $R = \overline{f}$ is related to the trip time $T = b - a$ and the trip mileage D by the classical formula $D = RT$, which is taught in elementary school.

The expression for the trip mileage D in terms of the instantaneous velocity f,

$$D = \int_a^b f(x)dx,$$

is due to the creative genius of Isaac Newton. This relation of Newton today appears in texts as the **fundamental theorem of calculus**.

Fundamental Theorem of Calculus

The foundations of the study of differential equations rests with Newton's discovery of a way to state the relation $D = RT$ using instantaneous velocities instead of speed averages.

Theorem A.1 (Fundamental theorem of calculus)
Let G be continuous and let F be continuously differentiable on $[a, b]$. Then

(a) $F(b) - F(a) = \displaystyle\int_a^b F'(x)dx,$

(b) $\displaystyle\frac{d}{dx}\int_a^x G(t)dt = G(x).$

Part (a) of the fundamental theorem is used by calculus students to evaluate integrals. In differential equations, it is applied to find solutions.

Part (b) of the fundamental theorem computes the instantaneous rate of an averaging process. Calculus students use it to check answers to integration problems. In differential equations it is used to verify solutions.

The justification of $D = RT$ for instantaneous rates $f(x) = F'(x)$ is contained in part (a): divide both sides by $b - a$ and interpret the right side as *the average velocity* or *speed* to get the formula $D/T = R$.

Example A.1 (Leibniz Notation)
Change $y''(x) + y(x)$ into Leibniz notation.

Solution:

$$\begin{aligned} y''(x) &= \frac{d}{dx} y'(x) && \text{Definition of second derivative.} \\ &= \frac{d}{dx}\frac{dy}{dx} && \text{Leibniz notation for the first derivative.} \\ &= \frac{d^2y}{dx^2} && \text{Leibniz notation.} \end{aligned}$$

Therefore, the converted expression is $\dfrac{d^2y}{dx^2} + y$.

Example A.2 (Notation Conversion)
Convert the equation $\dfrac{du}{dt} = u + e^t \sin t$ to dot notation.

Solution: By convention, $\dfrac{du}{dt} = \dot{u}(t)$ and $u = u(t)$. Therefore, the converted equation is $\dot{u}(t) = u(t) + e^t \sin t$.

Example A.3 (Slope of the Tangent Line)
Compute the slope m of the line tangent to $y = x \sin x$ at $x = \pi/2$.

Solution:

$$\begin{aligned} m &= y' && \text{Definition of slope and derivative.} \\ &= (x \sin x)' && \text{Definition of } y. \\ &= \sin x + x \cos x && \text{Product rule and derivative tables. Variable } x \text{ to be replaced by } \pi/2. \\ &= \sin(\pi/2) + \frac{1}{2}\pi \cos(\pi/2) && \text{Replacement } x = \pi/2. \\ &= 1 && \text{Identities } \cos(\pi/2) = 0,\ \sin(\pi/2) = 1 \text{ applied.} \end{aligned}$$

Example A.4 (Tangent Line Equation)
Find the tangent line equation at $x = \pi/2$ for $y = x \sin x$ in point-slope form and in slope-intercept form.

Solution: The point-slope equation in an XY-system is $Y - y_0 = m(X - x_0)$. In this formula, $x_0 = \pi/2$, $y_0 = x_0 \sin x_0 = \pi/2$. Example A.3 gives $m = 1$. The tangent line equation in point-slope form is $Y - \pi/2 = (1)(X - \pi/2)$, which simplifies to the slope-intercept form $Y = X$.

Example A.5 (Line Equations)
Convert the line equation $y - 2 = 5(x - 3)$ to slope-intercept and parametric forms.

Solution: The *slope-intercept* form $y = 5x - 13$ is found by expansion to an explicit equation for y. A *parametric* form can be found by setting $x = t$ and then $y = 5x - 13 = 5t - 13$. The vector form is

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} t \\ 5t - 13 \end{pmatrix}, \quad -\infty < t < \infty.$$

Example A.6 (Decay Law Derivation)

Derive the decay law $\dfrac{dA}{dt} = kA(t)$ from the sentence

> *Radioactive material decays at a rate proportional to the amount present.*

Solution: The sentence is first dissected into English phrases **1** to **4**.

1: *Radioactive material*	The phrase causes the invention of a symbol A for the amount present at time t.
2: *decays at a rate*	It means A undergoes decay. Then A changes. Calculus conventions imply the rate of change is dA/dt.
3: *proportional to*	Literally, it means *equal to a constant multiple of*. Let k be the proportionality constant.
4: *the amount present*	The amount of radioactive material present is $A(t)$.

The four phrases are translated into mathematical notation as follows.

Phrases 1 and 2	Symbol dA/dt.
Phrase 3	Equal sign '$=$' and a constant k.
Phrase 4	Symbol $A(t)$.

Let $A(t)$ be the amount present at time t. The translation is $\dfrac{dA}{dt} = kA(t)$.

Example A.7 (Average Value)

Given $f(x) = xe^x + \sin^2(\pi x)$, find the average value on $0 \le x \le 2$.

Solution: The value is $\frac{1}{2}e^2 + 1$. The details:

$$\begin{aligned}
\overline{f} &= \frac{1}{2}\int_0^2 f(x)dx && \text{Definition of average value, page 1007.} \\
&= \frac{1}{2}\int_0^2 [xe^x + \sin^2(\pi x)]dx && \text{Substitute for } f(x). \\
&= \frac{1}{2}(x-1)e^x\Big|_{x=0}^{x=2} && \text{Integral tables.} \\
&\quad + \frac{1}{4\pi}\left(-\cos \pi x \sin \pi x + \pi x\right)\Big|_{x=0}^{x=2} && \\
&= \frac{1}{2}e^2 + 1 && \text{Use } \sin(n\pi) = 0.
\end{aligned}$$

Example A.8 (Speed)

Find the speed for a car trip of 2 hours, given the velocity profile

$$\dot{x}(t) = \begin{cases} 1200t & 0 \le t \le 0.05, \\ 60 & 0.05 \le t \le 2. \end{cases}$$

Solution: The speed R is given by

$$\begin{aligned}
R &= \frac{1}{2}\int_0^2 \dot{x}(t)dt && \text{Average value of } \dot{x}\text{, page 1007.}\\
&= \frac{1}{2}\left(\int_0^{0.05} 1200t dt + \int_{0.05}^{2} 60dt\right) && \text{Use } \textstyle\int_a^b f = \int_a^c f + \int_c^b f.\\
&= \frac{1}{2}\left(600(0.05)^2 + 60(2-0.05)\right) && \text{Evaluate integrals.}\\
&= \frac{237}{4}. && \text{About 59.25 mph.}
\end{aligned}$$

The unrealistic 3-minute acceleration to 60 mph can be replaced by a more realistic 18-second acceleration to give 59.925 mph.

Example A.9 (Speed Estimation)

Estimate the average speed of a car which accelerates from 0 to 65 miles per hour in 12 seconds.

Solution: The purpose of this example is to explain the layman's answer of 65/2 mph. The answer must be justified in the context of calculus.

If the acceleration is constant, then $\ddot{x}(t) = a = \text{constant}$. Therefore, $\dot{x}(t) = at$, since $\dot{x}(0) = 0$. Let $t_0 = 12/3600$ hours. The average speed R for time interval $0 \le t \le t_0$ is

$$\begin{aligned}
R &= \frac{1}{t_0}\int_0^{t_0} \dot{x}(t)dt && \text{Definition of average speed, page 1007.}\\
&= \frac{a}{t_0}\frac{t_0^2}{2} && \text{Evaluate integral with } \dot{x} = at.\\
&= \frac{65}{2} && \text{Because } 65 = \dot{x}(t_0) = at_0.
\end{aligned}$$

It can be argued on physical grounds that no car has constant acceleration, so the answer 65/2 is merely an estimate. The layman's answer can be obtained by averaging the two speeds 0 and 65.

Example A.10 (Integral Identity)

Verify the integral evaluation $\displaystyle\int_0^1 xe^x dx = 1$.

Solution:

$$\begin{aligned}
I &= \int_0^1 xe^x dx && \text{Integral } I \text{ to be evaluated.}\\
&= \int_0^1 (xe^x - e^x)'\, dx && \text{Identity } xe^x = (xe^x - e^x)' \text{ derived below.}\\
&= (xe^x - e^x)|_{x=0}^{x=1} && \text{Apply the fundamental theorem of calculus, part (a). See page 1008.}\\
&= 1 && \text{Use } e^0 = 1.
\end{aligned}$$

The identity $xe^x = (xe^x - e^x)'$ applied in the solution above is obtained by experiment, as follows.

$$(xe^x)' = (1)e^x + xe^x \qquad \text{Product rule } (uv)' = u'v + uv'.$$

$$= (e^x)' + xe^x \qquad \text{Term } xe^x \text{ isolated on the right.}$$

Solving the last equation for xe^x gives the identity $xe^x = (xe^x - e^x)'$. A more systematic method for finding such identities is *integration by parts*.

Example A.11 (Integral Answer Check)
Verify the identity

$$\int_0^x t\ln(1+t)dt = \frac{1}{2}\left(x^2-1\right)\ln(1+x) + \frac{x}{2} - \frac{x^2}{4}.$$

Solution: Both sides evaluate to zero at $x = 0$, because $\ln(1) = 0$. According to the fundamental theorem of calculus, part (b), page 1008, it is sufficient to differentiate the answer on the right and verify that the derivative so obtained matches the integrand on the left. Let RHS denote the right hand side. Then

$$\begin{aligned}\text{RHS}' &= \left(\frac{x^2-1}{2}\ln(1+x) + \frac{x}{2} - \frac{x^2}{4}\right)' && \text{The \textit{Right Hand Side} of the identity, to be differentiated.}\\ &= x\ln(1+x) + \frac{x^2-1}{2x+2} + \frac{1}{2} - \frac{x}{2} && \text{Product rule, power rule and the identity } (\ln(u))' = u'/u.\\ &= x\ln(1+x). && \text{Simplified derivative of the RHS.}\end{aligned}$$

The derivative of RHS matches the integrand of the left side, which completes the verification.

Example A.12 (Distance Estimate)
Estimate the distance D traveled by an automobile in two hours, and its average speed R, given that for $t = 20$ to $t = 120$ the speedometer readings every 20 minutes are 55, 70, 66, 71, 72, 65 miles per hour.

Solution: The answers are 133 miles and 66.5 mph. To estimate the values of R and D, it will be assumed that the speed was constant during the 20-minute period before the reading. The actual velocity $\dot{x}(t)$ of the automobile is related to the average velocity R by the formula

$$R = \frac{1}{120}\int_0^{120} \dot{x}(t)dt.$$

The samples are used to find the average R as follows.

$$\begin{aligned}R &\approx \frac{55+70+66+71+72+65}{6} && \text{Used } \overline{f} \approx \frac{y_1+\cdots+y_n}{n}\text{, page 1007.}\\ &= \frac{399}{6} && \text{About 66.5 miles per hour.}\end{aligned}$$

Then $D = RT$ implies $D \approx \frac{399}{6}\frac{120}{60} = 133$ miles.

Exercises A.1

Derivative notation Convert from the given notation, prime, dot, Leibniz or operator, to the other three forms.

1. $\dfrac{du}{dt}$

2. $\dot{u}(t_0)$

3. $\ddot{u}(1+t)$

4. $\dfrac{dx}{dt} = 1 + x(t)$

5. $D^2w(x) = 1 + w(x) + x$

6. $Dy(x) = y^{-2}(x)$

7. $\ln(w(r)) = \dfrac{dw}{dr}$

8. $e^{-y(x)} = y'(x)$

9. $\dot{y}(t) = 1 + t$

10. $\dot{x}(t) = e^{-2x(t)}$

Slope Compute the slope of the line tangent to the curve at the given point.

11. $y = x^2 - 3x + 1, \quad x = 0.$

12. $y = x^5 - x + 2, \quad x = 2.$

13. $y = \sin x + x, \quad x = \pi/4.$

14. $y = \cos x - x, \quad x = \pi/4.$

15. $y = \tan^{-1} x + e^{-x}\ln(1+x), \quad x = 1.$

16. $y = \sin^{-1} x + e^{x}\ln(2+x), \quad x = 1.$

Tangent line equation Find the tangent line equation in the three possible forms, point-slope, slope-intercept and parametric.

17. $y = x^3 - x, \quad x = 1.$

18. $y = x^3 + x + 1, \quad x = 0.$

19. $y = \sin^{-1}(x), \quad x = 1/2.$

20. $y = \tan^{-1}(x), \quad x = 1.$

21. $y = e^{-x}, \quad x = \ln(2).$

22. $y = \ln(1+x), \quad x = 0.$

23. $y = \dfrac{1+x}{1-x}, \quad x = 0.$

24. $y = \dfrac{1-x^2}{1+x^2}, \quad x = 0.$

Rates Model as a rate of change equation.

25. The expected change in charge Q is equal to the electromotive force $\sin(\omega t)$.

26. The damping force F is proportional to the instantaneous change in $x(t)$.

27. The angular rate of change is proportional to the external force $\cos(\omega t)$.

28. The amount in a bank account changes at a rate proportional to the current balance.

29. The expected population change is proportional to the present population P.

30. The temperature flux and the temperature difference from the surrounding medium are proportional.

Average value Find the average value of f on $[a,b]$,

$$\overline{f} = \frac{1}{b-a}\int_a^b f(x)dx.$$

31. $xe^{-x}, \quad 0 \le x \le 1.$

32. $\frac{1}{2}e^{x} - \frac{1}{2}e^{-x}, \quad 0 \le x \le 2.$

33. $\ln x, \quad 1 \le x \le 3.$

34. $\sec x, \quad 0 \le x \le \pi/4.$

35. $x^3 - x, \quad 0 \le x \le 2.$

36. $\dfrac{x-1}{x+1}, \quad 0 \le x \le 1.$

37. $\dfrac{\sin x}{1+\cos x}, \quad 0 \le x \le \pi/4.$

38. $\sin^3 x \cos x, \quad 0 \le x \le \pi.$

39. $\dfrac{1}{1+x^2}$ on $0 \le x \le 1/2$, $4/5$ on $1/2 \le x \le 1$.

40. $\dfrac{1}{x}$ on $1 \le x \le 2$, $\dfrac{5}{8}\dfrac{x^2}{1+x^2}$ on $2 \le x \le 3$.

41. $\tan x$ on $0 \le x \le \pi/4$, and $1+(x-\pi/4)$ on $\pi/4 \le \pi/3$.

42. $\cot x$ on $\pi/4 \le x \le \pi/2$, and $x - \pi/2$ on $\pi/2 \le x \le \pi$.

Integral identities Verify the given integration identity by applying the fundamental theorem of calculus.

43. $\displaystyle\int_0^1 \frac{1+t}{2+t}dt = 1 + \ln\frac{2}{3}.$

44. $\displaystyle\int_0^1 \frac{1+t^2}{2+t}dt = 5\ln\frac{3}{2} - \frac{3}{2}.$

45. $\displaystyle\int_0^\pi t\sin(2t)dt = \frac{\pi-2}{4}.$

46. $\displaystyle\int_0^{\pi/2} t\cos(2t)dt = -\frac{1}{2}.$

47. $\displaystyle\int_0^1 te^{-t}dt = 1 - \frac{2}{e}.$

48. $\displaystyle\int_0^1 t^2e^{-t}dt = 2 - \frac{5}{e}.$

49. $\displaystyle\int_0^x \sin^4(t)\cos(t)dt = \frac{\sin^5(x)}{5}.$

50. $\displaystyle\int_0^x \tan(t)dt = -\ln(\cos x).$

Car trip Estimate the average speed R and the distance traveled D on a car trip, given the velocity samples.

51. Every 10 minutes from $t = 10$ to $t = 120$ minutes, 51, 62, 55, 53, 60, 67, 61, 67, 55, 70, 71, 66 miles per hour.

52. Every 15 minutes from $t = 15$ to $t = 225$ minutes, 90, 92, 110, 112, 120, 113, 109, 90, 95, 97, 60, 90, 100, 105, 103 kilometers per hour.

53. Every 5 minutes from $t = 5$ to $t = 75$ minutes, 45, 60, 61, 63, 60, 58, 61, 65, 25, 40, 45, 60, 65, 59, 60 miles per hour.

54. Every 5 minutes from $t = 5$ to $t = 100$ minutes, 50, 90, 100, 120, 110, 112, 130, 120, 110, 40, 60, 100, 90, 80, 20, 55, 130, 130, 120, 125 kilometers per hour.

A.2 Graphics

Engineers and scientists prefer a computer approach to graphing solutions to differential equations $y' = f(x, y)$. In special cases, the fastest graph generation is by hand or by calculator. Experience with hand calculations and calculators is useful for judging the accuracy of a computer graphic.

Small *numeric data sets* may be graphed by hand using graph paper or engineering paper. Large data sets are best graphed using a computer spreadsheet program, e.g., Microsoft `Excel`, a computer algebra system, e.g., `maple` or `mathematica`, a numerical laboratory, e.g., `matlab`, `octave` or `scilab`, or a freely available graphing program, e.g., `gnuplot`.

The Standard Curve Library

Feasibility for hand graphing of equations in *explicit form* is tested using the **standard curve library**, which includes the following equation types.

$y = mx + b$ Equation of a line in slope-intercept form. Includes constant equations. Increases for $m > 0$, decreases for $m < 0$.

$y = x^n$, $y = \dfrac{1}{x^n}$ Power curves. Even for n even, odd for n odd. Reciprocal powers have asymptote at $x = 0$. Special cases $y = x^2$, $y = 1/x$ occur often.

$y = \sin x$, $y = \tan x$ The sine is 2π-periodic and the cosine graph is a translation by $\pi/2$. The tangent is a π-periodic curve with asymptotes at odd multiples of $\pi/2$.

$y = e^x$, $y = \ln x$ All exponential and logarithmic curves are obtained from these basic graphs.

Four Transformations

The *standard curve library* is modified for use by allowing four transformations. The first two transformations are **rigid motions**, that is, the shape is unchanged. The last two are not rigid motions.

(1) Replace x by $x - x_0$ and y by $y - y_0$. The effect is to change the origin of coordinates in the graph from $(0, 0)$ to (x_0, y_0).

(2) Replace y by $-y$. The effect on a paper graphic is to turn the paper over, swapping horizontal edges. Examples are $y = x^2$ and $y = -x^2$.

(3) Replace y by $1/y$. The effect is to swap the roles of 0 and ∞ in the original graph. Examples are $y = x^2$ and $y = 1/x^2$.

(4) Replace y by ky with $k > 0$. The effect is to change the y-axis scale. Examples are $y = x^2$ and $y = 4x^2$ $(k = 1/4)$.

Special Equations

The standard curve library includes some special equations in *implicit* form $F(x, y) = c$. Recognized from the subject of **analytic geometry** are the following.

$(x - x_0)^2 + (y - y_0)^2 = r^2$	**Circle**. Radius r and center (x_0, y_0).
$\dfrac{(x - x_0)^2}{a^2} + \dfrac{(y - y_0)^2}{b^2} = 1$	**Ellipse**. Semiaxes a, b and center (x_0, y_0).
$\dfrac{(x - x_0)^2}{a^2} - \dfrac{(y - y_0)^2}{b^2} = 1$	**Hyperbola**. Center (x_0, y_0).

Polynomial Quotients

Rough graphs of polynomials and rational functions can be made with curve library methods. The graphs are accurate for the sign of y and the general shape. The following equation types are suited for use with the library.

$y = (x - a_1)^{n_1} \cdots (x - a_k)^{n_k}$	Factored polynomial curves. Roots at a_1, $\ldots$, a_k.
$y = \dfrac{(x - a_1)^{n_1} \cdots (x - a_k)^{n_k}}{(x - b_1)^{m_1} \cdots (x - b_\ell)^{m_\ell}}$	Factored rational curve. The roots are at a_1, $\ldots$, a_k and the vertical asymptotes are at b_1, $\ldots$, b_ℓ.

The curve graphic at each root looks like a power curve $Y = X^n$ and at each vertical asymptote it looks like an inverse power curve $Y = 1/X^n$, subject to the four transformations on page 1015.

Example A.13 (Curve Library Graphing Methods)
Apply curve library methods to graph on one set of axes for $-2 \le x \le 2$ the equations

$$y = 2x - 1, \quad y = (x - 1)^2, \quad y = -(x + 1)^4, \quad y = -1/x.$$

Solution: The curve library templates for the given graphs are

$$Y = X, \quad Y = X^2, \quad Y = -X^4, \quad Y = -1/X.$$

Transformation of the four types described on page 1015 are applied to change the templates into the correct figures. They are:

$Y = X$	Replace X by $2x$ and Y by $y+1$ to obtain $y = 2x - 1$.
$Y = X^2$	Replace X by $x-1$ and Y by y to obtain $y = (x-1)^2$.
$Y = -X^4$	Replace X by $x+1$ and Y by y to get $y = -(x+1)^4$.
$Y = -1/X$	Replace X by x and Y by $-y$ to obtain $y = -1/x$.

The transformations amount to changing the origin and/or flipping the curve compared to the template graphic from the curve library. The graphics to be assembled onto one set of axes appear in Figure 2.

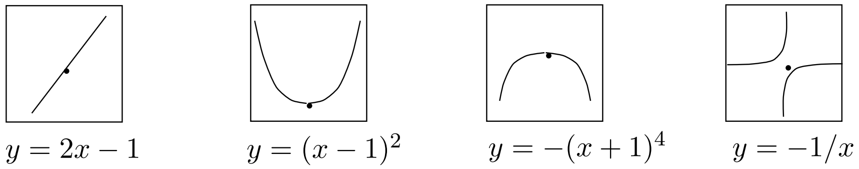


Figure 2. Transformed template graphics with centers $(0,-1)$, $(1,0)$, $(-1,0)$, $(0,0)$.

The final graphic is completed by assembling the transformed template graphics onto a single set of axes, locating each template onto its center. In the intermediate stage of completion, Figure 3, some portions of the graphic are left incomplete. The final graphic is Figure 4.

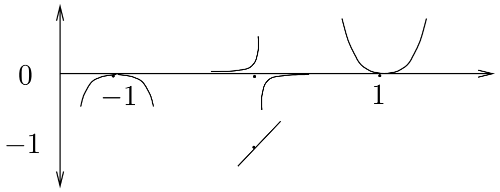


Figure 3. Combined graphic made from the four templates.

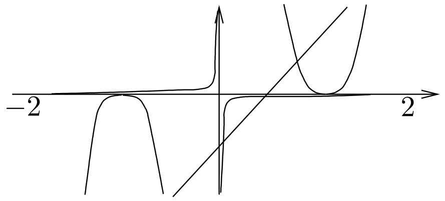


Figure 4. Combined graphic on $-2 \le x \le 2$ for curves $y = 2x - 1$, $y = (x-1)^2$, $y = -(x+1)^4$ and $y = -1/x$.

Example A.14 (Factored Polynomial Graphs)

Apply curve library methods to make a rough graph of the factored polynomial

$$y = -3x(x-1)^2(x-2)^3(x-3), \quad 0 \le x \le 4.$$

Solution: The distinct factors correspond to the template graphics to be used in the assembly of the final graphic:

$$y = c_1 x, \quad y = c_2(x-1)^2, \quad y = c_3(x-2)^3, \quad y = c_3(x-3).$$

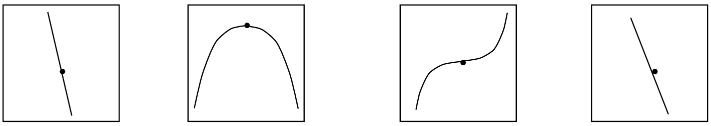
The constants c_1, c_2, c_3, c_4 are evaluated from the original curve equation $y = -3x(x-1)^2(x-2)^3(x-3)$ by arguing that, for (x,y) *close* to the center of each template graphic, the template and the original should be graphically the same. For example, if x is close to $x = 2$, then

$$\begin{aligned}
y &= -3x(x-1)^2(x-2)^3(x-3) && \text{Original equation. Analysis near } x = 2. \\
&= \left(-3x(x-1)^2(x-3)\right)(x-2)^3 && \text{Isolate all factors not containing the factor } (x-2). \\
&\approx \left[-3x(x-1)^2(x-3)\big|_{x=2}\right](x-2)^3 && \text{Isolated factors are nearly constant close to } x = 2. \\
&= 6(x-2)^3 && \text{Found template } y = 6(x-3)^3.
\end{aligned}$$

By this process, $y = -3x(x-1)^2(x-2)^3(x-3)$ has template equations

$$y = -72x, \quad y = -3(x-1)^2, \quad y = 6(x-2)^3, \quad y = -36(x-3).$$

As will be seen, the equalities $c_1 = -72$, $c_2 = -3$, $c_3 = 6$, $c_4 = -36$ are not actually used, *only the signs matter.* Therefore, knowing $c_1 < 0$, $c_2 < 0$, $c_3 > 0$, $c_4 < 0$ is enough for a rough graphic. This information is rapidly obtained by counting signs of the factors involved. The graphics for the templates, taken from the *standard curve library*, appear in Figure 5.



$y = -72x$ $\quad y = -3(x-1)^2$ $\quad y = 6(x+1)^3$ $\quad y = -36(x-3)$

Figure 5. Template graphics for $y = -3x(x-1)^2(x-2)^3(x-3)$.

To make the final graphic in Figure 6, the templates are located at their respective centers on one set of axes, then they are connected with a smooth curve (boldface). The connections stay either in the upper or the lower half-plane, because all zeros of y are accounted for by the template graphics.

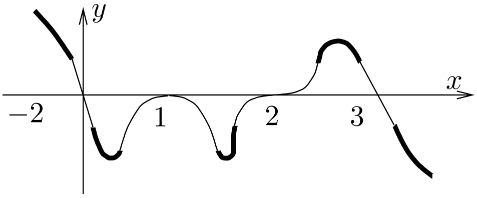


Figure 6. Final graph for the polynomial $y = -3x(x-1)^2(x-2)^3(x-3)$.

The graphic is accurate for the sign of y. The general shape is correct, but details like maxima, minima and slopes are flawed. Nevertheless, the hand graphic is perhaps more useful than a computer graphic.

Polynomial graphs exit the paper at $x = \pm\infty$ in the same way as their leading term, which could be called the **horizontal asymptote**. In the present example, the leading term is $y = -3x^7$, which is a curve from the standard curve library. This information can be used to detect fundamental graphing errors.

Example A.15 (Graphing Polynomial Fractions)

Apply curve library methods to make a rough graph of the rational function

$$y = \frac{-3x(x-1)^2(x-4)}{(x+1)^2(x-2)(x-3)}.$$

Solution: The rational function is a the quotient of two quadrics. By long division of the polynomials it follows that $y = -3 + r/q$ where the degree of r is less than the degree of q. Therefore, $r/q \approx 0$ at $\pm\infty$, and this means $y = -3$ is the horizontal asymptote.

The effect of these remarks about asymptotes is to declare that the graph exits the paper left and right along the line $y = -3$.

The vertical asymptotes $x = -1$, $x = 2$, $x = 3$ similarly cause the graph to exit the paper top and/or bottom. The curve library method uses this information implicitly.

A rough graphic can be drawn immediately from this basic information. The portions of the graph that appear in Figure 7 are approximate only, valid for values of x very close to $x = -\infty$ or $x = \infty$. In Figure 7, vertical asymptotes are shown, even though they already appear in library templates. It is a matter of taste to add vertical asymptotes to this figure. If added, then just portions of the vertical lines near $|y| = \infty$ are valid.

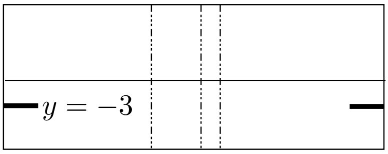


Figure 7. Rough graphic near $x = \pm\infty$ and $y = \pm\infty$.

The remainder of the graphic is obtained from the assembly of curve library templates. The distinct factors of the numerator and denominator of the rational function become the templates:

$$y = c_1 x, \qquad y = c_2(x-1)^2, \quad y = c_3(x-4),$$

$$y = \frac{d_1}{(x+1)^2}, \quad y = \frac{d_2}{x-2}, \qquad y = \frac{d_3}{x-3}.$$

Calculation of the constants c_1, c_2, c_3, d_1, d_2, d_3 is unnecessary, only the signs matter for template selection. For example, to compute d_1:

$$y = \frac{-3x(x-1)^2(x-4)}{(x+1)^2(x-2)(x-3)}$$

Given rational function. Analysis near $x = -1$.

$$= \left[\frac{-3x(x-1)^2(x-4)}{(x-2)(x-3)}\right]\frac{1}{(x+1)^2}$$

Isolate all factors not containing $(x+1)$.

$$\approx \frac{-3(-1)(-1-1)^2(-1-4)}{(-1-2)(-1-3)}\frac{1}{(x+1)^2}$$

Substitute $x = -1$ into the isolated factors.

$$= -5\frac{1}{(x+1)^2}$$

Template equation found.

The logic of the substitution $x = -1$ into the *isolated factors* is that they are nearly constant for $x \approx -1$. The template equation $y = -5/(x+1)^2$ at center $(-1, 0)$ will be used to plot the final graphic. Just the sign of $d_1 = -5$ is needed, which can be obtained by counting signs, the actual value 5 being irrelevant for the graphic.

By similar methods, the signs of the constants are found to be $c_1 > 0$, $c_2 > 0$, $c_3 < 0$, $d_1 < 0$, $d_2 < 0$, $d_3 > 0$. The six templates arise from four different library curves.

The six templates are placed at their centers and joined by a smooth curve (boldface) to produce the final graphic. See Figure 8. The plot needs some explanation. First, nothing is to scale, although the signs are correct for y and the general shape is valid. The curve goes off the paper left and right at $|x| = \infty$, the exit curve being $y = -3$. The curve also goes off the paper on the bottom edge at $x = -1, 2, 3$ and on the top edge at $x = 2, 3$, in the manner shown. The maxima and minima of the curve have not been computed, so this information is not to scale either.

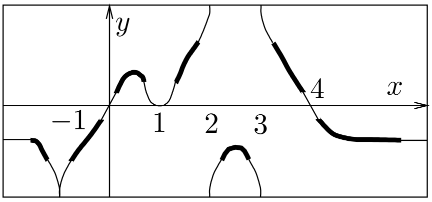


Figure 8. Graphic for $y = \dfrac{-3x(x-1)^2(x-4)}{(x+1)^2(x-2)(x-3)}$.

Computer algebra systems like `maple` and `mathematica` can produce similar plots, with limitations. Below is the `maple` code for the resulting plot in Figure 9. The unwanted and incorrect vertical lines in the plot are an artifact of the discontinuities. It is especially difficult to see some of the fine features, e.g., the double zero of y at $x = 1$ and the horizontal asymptote values.

```
# Maple V 5.1
F:=x-> -3*x*(x-1)^2*(x-4)/((x+1)^2*(x-2)*(x-3));
plot(F(x),x=-infinity..infinity);
```

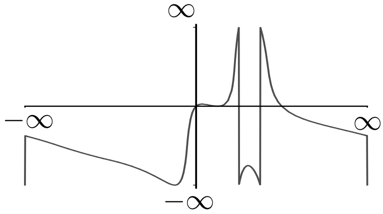


Figure 9. **Graphic by `maple` for** $y = \dfrac{-3x(x-1)^2(x-4)}{(x+1)^2(x-2)(x-3)}$.

Example A.16 (Computer Graphing)
Graph the 20 data points generated by the approximation formula

$$y(x + 0.05) \approx y(x) + 0.05(x + y(x)), \quad y(0) = 1,$$

from $x = 0$ to $x = 1$ in uniform steps of 0.05, using a computer.

Solution: The formula is applied as a **recursion formula**, which details how to generate from a *given table pair* x, y the *next table pair* X, Y via the formulas

$$X = x + 0.05, \quad Y = y + 0.05(x + y).$$

Mathematical translation includes elimination of the approximation symbol ($\approx$) and the use of equal signs ($=$) in the final formulas.

The first step is to generate a table of values. Then the table is plotted by a standard method. The process of determining the data pairs can be done by hand as follows.

$x = 0, \quad y = 1$	The first data pair arises from $y(0) = 1$, which means $y = 1$ at $x = 0$.
$X = x + 0.05$	The next x-value is the old one plus 0.05.
$Y = y + 0.05(x + y)$	Approximation formula for the next y-value.
$= 1 + 0.05(0 + 1)$	Use $x = 0$, $y = 1$.
$= 1.05.$	The second data pair is $X = 0.05$, $Y = 1.05$.

The first three pairs of values are verified to be

$$(0.00, 1.000), \quad (0.05, 1.050), \quad (0.10, 1.105).$$

A `maple` plot of the data uses the following code, resulting in Figure 10. Similar mechanisms for plotting data points are available in `matlab`, `mathematica`, `gnuplot` and `scilab`. Libreoffice `CALC` and Microsoft `Excel` can be used for such graphics.

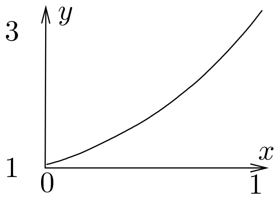


Figure 10. Sample computer graphic for the approximation formula $y(x + 0.05) \approx y(x) + 0.05(x + y(x))$, $y(0) = 1$.

```
# Maple V,  plot data points in list L
L:=[0.00,1.000],[0.05,1.050],[0.10,1.105],
   [0.15,1.165],[0.20,1.231],[0.25,1.303],
   [0.30,1.380],[0.35,1.464],[0.40,1.555],
   [0.45,1.653],[0.50,1.758],[0.55,1.870],
   [0.60,1.992],[0.65,2.121],[0.70,2.260],
   [0.75,2.408],[0.80,2.566],[0.85,2.734],
   [0.90,2.913],[0.95,3.104],[1.00,3.307]:
plot([L]);
```

In computer algebra systems, it is possible to avoid typing the numeric data, because of the formulas $X = x + 0.05$, $Y = y + 0.05(x + y)$. To generate the list L in maple, execute the two code groups below.

```
# Execute the first group once
X:=0:Y:=1:L:=[X,Y]:
# Execute the second group 20 times
Y:=Y+0.05*(X+Y):X:=X+0.05:L:=L,[X,Y]:
```

Example A.17 (Computer Plotting)

Graph by computer the explicit equation $y = e^{-x}\sin(x)$ on $0 \le x \le 2\pi$.

Solution: Plot commands for five plotting systems are given below. The graphic in Figure 11 represents the maple output.

```
plot(exp(-x)*sin(x),x=0..2*Pi);                  maple
Plot[{exp(-x) sin(x)}, {x,0,2 Pi} ];             mathematica
plot [0:2*pi] exp(-x)*sin(x)                     gnuplot
x=0:0.1:2*PI; y=exp(-x).*sin(x); plot(x,y)       matlab and scilab
```

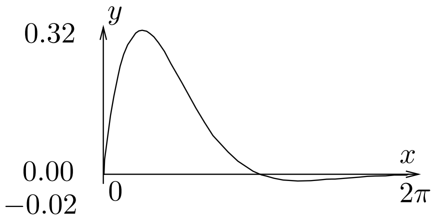


Figure 11. Computer plot of $y = e^{-x}\sin(x)$ on $0 \le x \le 2\pi$.

Example A.18 (Computer Plotting)

Plot by computer the implicit equation $x^2 + 2y^2 + xy = 10$.

Solution: Some background will be reviewed, which provides the source of intuition for plotting similar implicit equations. Quadratic forms $Ax^2 + 2Bxy + Cy^2 = D$ are studied in *analytic geometry*, and it is known how to classify the graphic based upon the sign of $B^2 - AC$. A change of variables to eliminate the cross term xy would result in a hand solution to this example, an ellipse with semiaxes $a \approx 3.55$ and $b \approx 2.13$ rotated about -22.5 degrees with major axis along the line $y = (1 - \sqrt{2})x$. The exact semiaxis lengths are given by

$$\frac{1}{a^2} = \frac{3 - \sqrt{2}}{20}, \quad \frac{1}{b^2} = \frac{3 + \sqrt{2}}{20}.$$

The graphic in Figure 12 is the result of the `maple` code below. Plots of *implicit equations* require tweaking of the domain and various plot parameters. The feature is not available in some programs, e.g., `gnuplot`.

```
# Maple V
with(plots):
eq:=x^2+2*y^2 + x*y = 10:
opt:=scaling=constrained,grid=[40,40]:
implicitplot(eq,x=-4..4,y=-4..4,opt);
```

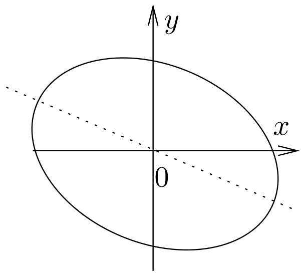


Figure 12. Implicit plot of $x^2 + 2y^2 + xy = 10$.

Exercises A.2

Curve library graphics Apply the curve library method to construct by hand a graphic of the given equations on one set of axes.

1. $y = 2x + 1$, $y = 3(x + 1)^2$

2. $y = \dfrac{-1}{x+1}$, $y = -2x - 1$

3. $y = \dfrac{2}{(x+1)^2}$

4. $y = \dfrac{-1}{(x+1)^3}$

5. $y = x^2$, $y = (x - 1)^4$, $y = (x - 2)^6$

6. $y = \dfrac{1}{x+1}$, $y = \dfrac{1}{(x-1)^2}$

Factored polynomial graphics Apply the curve library method to construct by hand a graphic of the given factored polynomial on one set of axes.

7. $y = -2x(x - 1)^2$

8. $y = 2x(x + 1)^3$

9. $y = -(x + 1)^2(x - 1)^3$

10. $y = (x + 1)^3(x - 1)^4$

11. $y = (x + 1)(x - 1)^3(x + 2)$

12. $y = -x^3(1 - x)(1 + x)$

13. $y = \pi(x + 1)(x - 1)(x + 2)^2$

14. $y = \pi^2(x + 1)(x - 1)(x + 2)^3$

Factored rational graphics Apply the curve library method to construct by hand a graphic of the given factored rational function on one set of axes.

15. $y = \dfrac{x-1}{x+1}$

16. $y = \dfrac{2x+1}{x+2}$

17. $y = \dfrac{x(x+1)}{(x+2)(x-2)}$

18. $y = \dfrac{x(2x+1)}{(x+2)(x-2)}$

19. $y = \dfrac{-x(1-x)}{(x+1)(x-2)}$

20. $y = \dfrac{5x(x+1)}{(x-1)(x-2)}$

Computer plotting of tables Make a table of values $x = 0$ to $x = 1$ in steps of 0.05 for the given approximate equation and plot the table of values. Cite the *recursion* formulas applied to obtain the next table pair from the previous table pair.

21. $y(x+0.05) \approx y(x) + 0.05(1 - y(x))$, $y(0) = 1$

22. $y(x+0.05) \approx y(x) + 0.05(1 + y(x))$, $y(0) = 1$

23. $y(x+0.05) \approx y(x) + 0.05(x - y(x))$, $y(0) = 0$

24. $y(x+0.05) \approx y(x) + 0.05(2x + y(x))$, $y(0) = 0$

25. $y(x+0.05) \approx y(x) + 0.05(\sin x + xy(x))$, $y(0) = 2$

26. $y(x+0.05) \approx y(x) + 0.05(\sin x - x^2 y(x))$, $y(0) = 2$

Computer plots of explicit equations Plot by computer the given explicit equation over $0 \le x \le 1$.

27. $y = e^{-x} \sin \pi x$

28. $y = e^{-x} \cos \pi x$

29. $y = e^{-x} \ln(1 + x)$

30. $y = e^{-x} \ln(1 + x^2)$

31. $y = \sin(\pi x) \sin^2(2\pi x)$

32. $y = \sin(\pi x) \cos^2(\pi x)$

Implicit plots Plot by computer or by hand the given implicit equation.

33. $x^2 + y^2 + 3xy = 10$

34. $x^2 + y^2 - 3xy = 10$

35. $x^2 - (y+1)^2 = 1$

36. $x^2 - y^2 + xy = 10$

37. $x(x-1)y = 5$

38. $xy(1 + y^2) = 10$

A.3 Explicit and Implicit Answers

Important to engineers and scientists are methods by which an existing *answer* can be tested for correctness. Given here are tests for explicit and implicit equations, as applied to the initial value problem

$$\begin{cases} y' = f(x,y), \\ y(x_0) = y_0. \end{cases} \tag{1}$$

It is possible to test mathematical equations of the form $y = y(x)$ and $F(x,y) = 0$, to see if they represent a solution to the problem (1). Both methods rely upon the expansion of the left side (LHS) and the right side (RHS) of equations. The two sides are compared for equality, either symbolically or else as constants. A proposed *answer* passes the test if the two sides are equal, that is, LHS = RHS.

Explicit Equations

An *explicit equation* $y = y(x)$ represents a solution of (1) provided checkpoints (a), (b) hold below.

(a) The equation $y' = f(x,y)$ is expanded using $y = y(x)$ to produce a LHS and a RHS that depend on x. The expressions LHS and RHS are tested for symbolic equality at each x in the domain of $y(x)$.

(b) The equation $y(x_0) = y_0$ has a constant LHS, evaluated using the given expression for $y(x)$ and the value $x = x_0$. The constant RHS is y_0. The expressions LHS and RHS are tested for numerical equality.

Implicit Equations

A given *implicit equation* $F(x,y) = 0$ represents a solution of (1) provided checkpoints (c), (d) hold below.

(c) Briefly, *implicit differentiation* of $F(x,y) = 0$ reproduces (1). Technically, the equation $f(x,y) = -F_x(x,y)/F_y(x,y)$ is expanded using the formulas for F and its partial derivatives F_x and F_y, to produce a LHS and a RHS which are expressions in the two symbols x, y. The symbolic equality LHS = RHS must hold for all (x,y) satisfying $F(x,y) = 0$.

(d) Initial condition $y(x_0) = y_0$ is tested by expansion of the equation $F(x_0,y_0) = 0$ into LHS and RHS. The constant expressions must be equal, LHS = RHS.

The equation $F(x,y) = 0$ can be viewed as a **conservation law**, e.g., if F is energy, then $F = 0$ says the energy is constant along the path of a particle.

Implicit differentiation results in the **dynamical equation** for the conservation law. This equation describes the dynamics or change, hence it is expressed in terms of the rate of change dy/dx.

The formal verification of (c) depends upon the **chain rule for 2 variables**

$$\frac{dF(x,y)}{dt} = \frac{\partial F}{\partial x}\frac{dx}{dt} + \frac{\partial F}{\partial y}\frac{dy}{dt}.$$

Technical assumptions which allow y to be found as a function of x in the equation $F(x,y)=0$ appear in the *implicit function theorem*, page 1041, where the critical assumption $F_y(x_0,y_0)\neq 0$ is made.

The chain rule is applied to the equation $F(x,y)=0$, setting $x=t$, $y=y(t)$, to give

$$F_x(t,y(t))(1) + F_y(t,y(t))\frac{dy(t)}{dt} = 0.$$

Substitution of $t=x$ and $y'=f(x,y)$ into this equation justifies Part (c) of the test.

Computer Algebra Methods

The ideas outlined above for checking an explicit or implicit equation can be implemented in most computer algebra systems (abbreviation `CAS`). It suffices to create the two `CAS` symbols LHS and RHS and then test for equality of LHS and RHS in all the relevant variables.

It sometimes transparent that LHS and RHS are equal, due to automatic `CAS` simplifications. There are instances where equality is completely opaque, because of insufficient `CAS` simplifications. To the rescue comes this idea: define `ZERO` to be the difference of LHS and RHS. The `CAS` symbol `ZERO` should reduce to zero, after simplifications are performed. See Examples A.21, A.22, page 1026 for details.

Realistically, engineers and scientists will migrate to CAS verifications, after intuition has been gained from many hand computations. Even in the simplest applications, something can go wrong, so experts advise: *verify the results by hand and by machine, verify it more than once, and check it from different viewpoints.*[1]

Example A.19 (Verify an Explicit Solution)

Verify the explicit solution $y = x - 1 + 2e^{-x}$ for $y' = x - y$, $y(0)=1$.

Solution: The *initial condition* $y(0)=1$ is verified as follows.

$y(0) = \left(x - 1 + 2e^{-x}\right)\big\vert_{x=0}$	Compute the left side of $y(0)=1$ where $y(x) = x - 1 + 2e^{-x}$.
$= 0 - 1 + 2e^0$	Evaluate.

[1] Picture a person walking in the rain, dripping wet, holding in one hand a closed umbrella.

$= 1.$	Therefore, the two sides of $y(0) = 1$ are equal.

The *differential equation* is verified in a slightly different way, by independent expansion of the left and right sides.

$\text{LHS} = y'$	The left side of $y' = x - y$ is y'.
$= (x - 1 + 2e^{-x})'$	Insert $y(x) = x - 1 + 2e^{-x}$.
$= 1 - 2e^{-x},$	Apply derivative rules.
$\text{RHS} = x - y$	The right side of $y' = x - y$ is $x - y$.
$= x - (x - 1 + 2e^{-x})$	Insert $y(x) = x - 1 + 2e^{-x}$.
$= 1 - 2e^{-x}.$	Simplified RHS.

Therefore, LHS=RHS.

Example A.20 (Verify an Implicit Solution)

Verify the implicit solution $3x^2 + y^2 = c$ for the equation $y' = -3x/y$.

Solution:

$f(x, y) = -3x/y$	The right side of $y' = -3x/y$ is called $f(x, y)$.
$F(x, y) = 3x^2 + y^2$	The level curve $F(x, y) = c$ duplicates the proposed solution $3x^2 + y^2 = c$.
$F_x(x, y) = 6x$	Partial derivative in x.
$F_y(x, y) = 2y$	Partial derivative in y.
$Z = F_x(x, y) + F_y(x, y)f(x, y)$	Test the differential equation. Expect Z to be zero.
$= 6x + 2y(-3x/y)$	Substitute partials and $f(x, y) = -3x/y$.
$= 0.$	Simplify.

Therefore, implicit differentiation of $F(x, y) = c$ reproduces the differential equation $y' = -3x/y$; see page 1025.

Example A.21 (Verify Explicit Solution by Computer)

Verify the explicit solution $y = e^{-x}$ for $y' = -y$, $y(0) = 1$ using a computer algebra system.

Solution: The illustration will be for `maple`.

`y:=x->exp(-x):`	The `maple` code for solution $y = e^{-x}$.
`LHS:=diff(y(x),x):`	The left side of $y' = -y$ is $y'(x)$.
`RHS:=-y(x):`	The right side of $y' = -y$ is $-y(x)$.
`ZERO:=LHS-RHS;`	The expression `ZERO` depends symbolically on x.
`Z:=y(0)-1;`	Write `Z` as the difference of the left and right sides of the equation $y(0) = 1$.

Evaluation of `ZERO` should give the symbolic answer 0, because LHS $=$ RHS is equivalent to LHS $-$ RHS $= 0$. Evaluation of the constant Z should give constant 0. This verifies the differential equation and initial condition by computer algebra methods. In unusual cases, it may be necessary to force simplifications or to interpret the answers. Simplification is forced by the `maple` command `simplify(ZERO)` while interpretation may be required to conclude that an expression, e.g., `sin(n*Pi)`, evaluates to zero.

Since `maple V 5.1`, there is a special function `odetest`, designed to do the above test. It is valuable because it eliminates errors made by re-typing formulas.

Example A.22 (Verify Implicit Solution by Computer)

Verify the implicit solution $x^2+y^2=c$ for the equation $y'=-x/y$ using a computer algebra system.

Solution: The illustration will be for `maple`.

`F:=(x,y)->x*x+y*y-c`	Write $x^2+y^2=c$ as $F=0$ where $F=x^2+y^2-c$.
`f:=(x,y)->-x/y`	The right side of $y'=-x/y$ is $f(x,y)$.
`Fx:=(x,y)->diff(F(x,y),x)`	Partial derivative in x.
`Fy:=(x,y)->diff(F(x,y),y)`	Partial derivative in y.
`ZERO:=Fx(x,y)+f(x,y)*Fy(x,y)`	Variable `ZERO` is the left side of $F_x+F_yy'=0$ with $y'=f(x,y)$.

Evaluation of `ZERO` should give the answer 0. This verifies the implicit solution of the differential equation by computer algebra methods.

In `maple V 5.1`, the function `odetest` will test implicit solutions.

Exercises A.3

Verify an Explicit Solution Apply the methods in Example A.19, page 1025, to verify the given solution of the initial value problem.

1. $I(t)=I_0e^{-2t}$,
$I'+2I=0$, $I(0)=I_0$.

2. $Q(t)=Q_0e^{-0.2t}$,
$Q'=-0.2Q$, $Q(0)=Q_0$.

3. $A(t)=100e^{kt}$,
$A'=kA$, $A(0)=100$.

4. $P(t)=1000e^{ht}$,
$P'=hP$, $P(0)=1000$.

5. $y(x)=-1+\sqrt(4+x^2-2x)$,
$y'=\dfrac{x-1}{y+1}$, $y(0)=1$.

6. $y(x)=-1+\sqrt{2+2e^x-2x}$,
$y'=\dfrac{e^x-1}{y+1}$, $y(0)=1$.

7. $y(x)=e^{x^2/2}$,
$y'=xy$, $y(0)=1$.

8. $y(x)=e^{x^3/3}$,
$y'=x^2y$, $y(0)=1$.

9. $y(x)=e^{1-\cos(x)}$,
$y'=\sin(x)y$, $y(0)=1$.

10. $y(x)=e^{\sin(x)}$,
$y'=\cos(x)y$, $y(0)=1$.

Verify an Implicit Solution Apply the methods in Example A.20, page 1026, to verify the given implicit solution of the differential equation. If an initial condition is given, then verify it also.

11. $xy^2+x^2y+xy=c$,
$y'=-\dfrac{y\,(y+2\,x+1)}{x\,(2\,y+x+1)}$.

12. $x^2y^2+x^3y+xy^2=c$,
$y'=-\dfrac{y\,(2\,xy+3\,x^2+y)}{x\,(2\,xy+x^2+2\,y)}$.

13. $x\sin y + \cos(xy) = c$,
$$y' = -\frac{-\sin(y) + \sin(xy)y}{x\,(-\cos(y) + \sin(xy))}.$$

14. $x^2\cos(y) + \sin(xy^2) = c$,
$$y' = \frac{2\,x\cos(y) + \cos(xy^2)y^2}{x\,(x\sin(y) - 2\,\cos(xy^2)y)}.$$

15. $x^2e^y + e^{x-y} = 1 + e$,
$$y' = -\frac{2\,xe^y + e^{x-y}}{x^2e^y - e^{x-y}},\ y(1) = 0.$$

16. $x^3e^{-y} + xe^{2x-y} = 1 + e^2$,
$$y' = \frac{3\,x^2 + e^{2\,x} + 2\,xe^{2\,x}}{x\,(x^2 + e^{2\,x})},$$
$y(1) = 0$.

Verify an Explicit Solution by Computer

Apply the methods in Example A.21, page 1026, to verify the given solution of the initial value problem.

17. $y(x) = \sqrt[3]{3x}$,
$y' = 1/y^2$, $y(1/3) = 1$.

18. $y(x) = \sqrt[4]{4x}$,
$y' = 1/y^3$, $y(1/4) = 1$.

19. $y(x) = e^{-x^2/2}$,
$y' = -xy$, $y(0) = 1$.

20. $y(x) = \pi e^{-x^3/3}$,
$y' = -x^2y$, $y(0) = \pi$.

21. $y(x) = xe^{\cos(x)-1}$,
$y' = (1/x - \sin(x))y$,
$y(2\pi) = 2\pi$.

22. $y(x) = \tan x + e^{\sin(x)}$,
$y' = \sec^2 x - \sin x + y\cos(x)$,
$y(0) = 1$.

Verify Implicit Solution by Computer

Apply the methods in Example A.22, page 1027, to verify the given implicit solution of the differential equation. If an initial condition is given, then verify it also.

23. $xy = 2$, $y' = -y/x$, $y(2) = 1$.

24. $x^2y = 2$, $y' = -2y/x$, $y(1) = 2$.

25. $xe^y + ye^x = c$, $y' = -\dfrac{e^y + ye^x}{xe^y + e^x}$.

26. $xe^{-y} + ye^{-x} = c$,
$$y' = \frac{e^{-y} - y^2e^{-x}}{xe^{-y} - 2\,ye^{-x}}.$$

27. $x\sin y + \cos(xy) = c$,
$$y' = \frac{\sin(y) - \sin(xy)y}{x\,(\sin(xy) - \cos(y))}.$$

28. $x^2\cos(y) + \sin(xy^2) = c$,
$$y' = \frac{2\,x\cos(y) + \cos(xy^2)y^2}{x\,(-x\sin(y) + 2\,\cos(xy^2)y)}.$$

A.4 Numerical and Graphical Answers

Given here are tests for numeric tables and graphics, as applied to the initial value problem

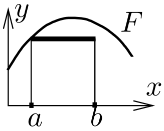
(1) $$\begin{aligned} y' &= f(x,y), \\ y(x_0) &= y_0. \end{aligned}$$

The numerical tests are based upon numerical integration methods from calculus. The ideas lead to the numerical methods of Euler, Heun and Runge-Kutta, which are studied in the text.

Numerical Integration Approximations

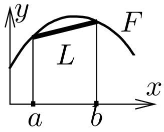
Reproduced here for future reference are calculus topics: the **rectangular rule** , the **trapezoidal rule** and **Simpson's rule** for the numerical approximation of an integral $\int_a^b F(x)dx$. The approximations are valid for $b-a$ small. Larger intervals must be subdivided, then the rule applies to the small subdivisions.

Rectangular Rule. The approximation uses Euler's idea of replacing the integrand by a constant. The value of the integral is approximately the area of a rectangle of width $b-a$ and height $F(a)$.



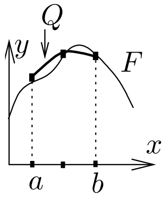
(2) $$\int_a^b F(x)dx \approx (b-a)F(a).$$

Trapezoidal Rule. The rule replaces the integrand $F(x)$ by a linear function $L(x)$ which connects the planar points $(a, F(a))$, $(b, F(b))$. The value of the integral is approximately the area under the curve L, which is the area of a trapezoid.



(3) $$\int_a^b F(x)dx \approx \frac{b-a}{2}\left(F(a)+F(b)\right).$$

Simpson's Rule. The rule replaces the integrand $F(x)$ by a quadratic polynomial $Q(x)$ which connects the planar points $(a, F(a))$, $((a+b)/2, F((a+b)/2))$, $(b, F(b))$. The value of the integral is approximately the area under the quadratic curve Q.



(4) $$\int_a^b F(x)dx \approx \frac{b-a}{6}\left(F(a)+4F((a+b)/2)+F(b)\right).$$

Simpson's Polynomial Rule. If $Q(x)$ is a linear, quadratic or cubic polynomial, then (proof on page 1036)

$$(5) \qquad \int_a^b Q(x)dx = \frac{b-a}{6}\left(Q(a) + 4Q((a+b)/2) + Q(b)\right).$$

Integrals of linear, quadratic and cubic polynomials can be evaluated *exactly* using Simpson's polynomial rule (5); see Example A.26, page 1035.

Remarks on Simpson's Rule. The right side of (4) is exactly the integral of $Q(x)$, which is evaluated by equation (5). The appearance of F instead of Q on the right in equation (4) is due to the relations $Q(a) = F(a)$, $Q((a+b)/2) = F((a+b)/2)$, $Q(b) = F(b)$, which arise from the requirement that Q connect three points along curve F.

The quadratic interpolation polynomial $Q(x)$ is determined uniquely from the three data points; see page 1037 for a formula for Q and a derivation. It is interesting that Simpson's rule depends only upon the uniqueness and not upon the actual formula for Q!

Graphic and Numeric Table Test

Studied here is a general problem:

> Find a test which verifies a given graphic or numeric table, given only the xy-pairs and $y' = f(x,y)$.

The test should work with a hand calculator, a spreadsheet or a computer algebra system. Important to the test is the ability to *spot-check* the graphic or table, testing just one or two data items.

To be presented here are the *Euler*, *trapezoidal* and *Simpson tests*. They detect errors in graphics by pixel criteria; see page 1032 for details. All tests have limitations and flaws. If the data items are far apart, then the approximation is poor and the test fails. Use is limited to *detection of gross errors*.

Equivalent Integral Equation. Fundamental to understanding the *tests* is the **equivalent integral equation**

$$(6) \qquad y(x) = y(x_0) + \int_{x_0}^x f(t, y(t))dt$$

for the first order initial value problem $y' = f(x,y)$, $y(x_0) = y_0$. Equation (6) is justified on page 1036.

Raw Data. Graphics produced in computer algebra systems or in computer spreadsheets require raw plot data, either implicitly or explicitly supplied. It will be assumed that this data is available as a table of xy-values, or equivalently, as a list of pairs

$$(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n).$$

It is necessary in the tests to evaluate $f(x, y)$ at the points of this list. No other evaluations of f are used, for the simple tests.

A linear connection ("connect-the-dots") of the data points is used by many computer programs; many points are required for a smooth result. Typical detail is shown in Figure 13.

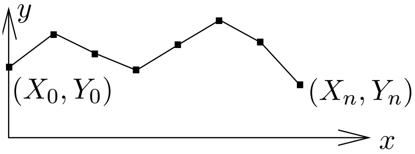
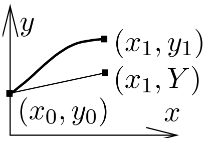


Figure 13. Linear connection of raw data points (X_0, Y_0), (X_1, Y_1), …, (X_n, Y_n) in a computer graphic.



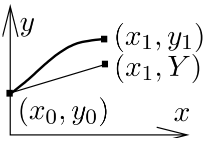
Euler's Test. The test applies to one pair of consecutive points from the raw plot data list. **Euler's test** is related to **Euler's numerical method**, which is the oldest and simplest numerical method for first order differential equations.

The test is named after Leonhard Euler[2] (1707-1783), Swiss physicist and mathematician. The test is justified on page 1037.

Step 1. Let (x_0, y_0) and (x_1, y_1) denote consecutive pairs from the raw plot data list (X_0, Y_0), (X_1, Y_1), …, (X_n, Y_n).

Step 2. Compute $h = x_1 - x_0$ and $Y = y_0 + hf(x_0, y_0)$.

Step 3. Test equality of y_1 and Y.



Trapezoidal Test. The tests applies to a consecutive pair of points from the raw plot data list. The **trapezoidal test** is related to the **modified Euler numerical method** , or **Heun's method**. The justification appears on page 1037.

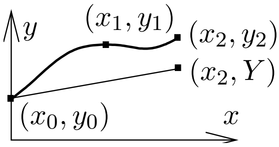
Step 1. Let (x_0, y_0) and (x_1, y_1) denote consecutive pairs from the raw plot data list (X_0, Y_0), (X_1, Y_1), …, (X_n, Y_n).

Step 2. Compute $h = x_1 - x_0$ and $Y = y_0 + \dfrac{h}{2}(f(x_0, y_0) + f(x_1, y_1))$.

Step 3. Test equality of y_1 and Y.

[2] His name is pronounced *Oiler*, and not *Yuler*.

Simpson's Test. The test is applied to three consecutive pairs from the raw plot data list. Assume uniformly-spaced X-data. The **Simpson test** is related to the **Runge-Kutta numerical method** for first order differential equations. Justification is on page 1037.



Step 1. Let (x_0, y_0), (x_1, y_1) and (x_2, y_2) denote three consecutive pairs from the raw plot data list. It is assumed that $x_1 = (x_0 + x_2)/2$.

Step 2. Let $Y = y_0 + \dfrac{x_2 - x_0}{6}(f(x_0, y_0) + 4f(x_1, y_1) + f(x_2, y_2))$.

Step 3. Test equality of y_2 and Y.

Pass and Fail. A given test can **pass** or **fail** according to how the resulting approximation is judged. A graph *passes* the test if the ideal data point (x, y) and the approximate data point (x, Y) land on the same pixel, that is, the dots cannot be distinguished in the graphic. Arithmetically, the test is an inequality

$$\frac{|y - Y|}{|d - c|} < \frac{1}{M},$$

where M is the number of y-pixels in the graphic on $c \le y \le d$. Otherwise, the graph *fails*.

Exercises in this text use the **standard graphic** , a $3\frac{1}{4}$-inch square graphic at 300 dots per inch, which is about 1000×1000 pixels. The same graphic displayed on a video monitor uses considerably fewer pixels.

There are two standard ways to measure the approximations:

Absolute Error. The **absolute error** is $E = |y - Y|$. The standard graphic of 1000 pixels on $c \le y \le d$ will *pass* the test if $E < \frac{d-c}{1000}$.

Relative Error. The **relative error** is $E = |y - Y|/|y|$. Since $Y = (1 \pm E)y$, it measures the percentage error.

Mostly, it is used for y-ranges $c \le y \le d$ where $c > 0$ or $d < 0$ (division by zero is problematic). The standard graphic of 1000 pixels on $0 < c \le y \le d$ will *pass* the test if $E < \frac{d-c}{1000d}$.

To distinguish the two measurements, apply the definitions to $y = 1000$ and $Y = 1001$: the absolute error is 1 and the relative error is 1/1000.

Uniformly-Spaced and Adaptive Data. In computer workbenches like `matlab` or `scilab`, the x-values will be uniformly spaced. In other systems, uniform spacing can be arranged, but the default may be **non-uniform data** or **adaptive data** , e.g., `maple`. Graphics systems normally document how to print out the plot data used for the graphic, even if the plot was done by implicit or automatic means. To get uniformly spaced data in `maple`, some preparation is required, as the following illustration shows. Uniformly spaced data is required in the Simpson test, page 1032.

```
y:=x->20*exp(-3*x):a:=0.0: b:=1.0:
# Adaptive plot saved in variable P
P:=plot(y(x),x=a..b);
# Uniform x-data plot saved in Q. Maple V 5.1
Q:=plot(y(x),x=a..b,adaptive=false,
        sample=[seq(i*h,i=0..(b-a)/h)]);
```

Example A.23 (Spot Check)
A graphic for the differential equation $y' = x+y$ has window $0 \le x \le 0.5$, $1 \le y \le 2$ and uses 1000×1000 pixels. Two adjacent plot data entries are $(0.180, 1.21443)$ and $(0.195, 1.23562)$. Spot-check these entries with Euler's test.

Solution: The plot data passes Euler's test, page 1031, because the target value 1.23562 is close to the Euler approximation 1.2353464, with less than one pixel difference in the plot. The steps of the justification appear below.

$x_0 = 0.180$, $y_0 = 1.21443$	The first data point $(0.180, 1.21443)$.
$x_1 = 0.195$, $y_1 = 1.23562$	The second data point $(0.195, 1.23562)$.
$h = x_1 - x_0 = 0.015$	Define the step size.
$Y = y_0 + h(x_0 + y_0)$	Apply Euler's test, page 1031.
$= 1.21443$ $+ 0.015(0.18 + 1.21443)$	Substitute x_0, y_0, h.
$= 1.2353464.$	Expected to be close to $y_1 = 1.23562$.

The *absolute error* is $E = 0.0002736$, which is less than the cutoff value of $E^* = (d - c)/1000 = 0.001$. The data passes Euler's test.

Example A.24 (Trapezoidal Test)
A graphic for the differential equation $y' = x+y$ has window $0 \le x \le 0.5$, $1 \le y \le 2$ and uses 1000×1000 pixels. Find the worst absolute error $|y_1 - Y|$ made according to the trapezoidal test for the associated plot data below and report pass or fail.

$$(0.180, 1.21443), \ (0.195, 1.23562), \ (0.210, 1.25736),$$
$$(0.225, 1.27965), \ (0.240, 1.30250), \ (0.255, 1.32592).$$

Solution: The cutoff value for the absolute error is $(d - c)/1000 = 0.001$. It will be justified below that the worst absolute error according to the Trapezoidal test is 0.0000057. In short, the data passes the test.

The first pair of points in the plot data passes the Trapezoidal test, page 1031, because the target value $y_1 = 1.23562$ is close to the test's value $Y = 1.2356179$, with absolute error $|y_1 - Y| = 0.0000021$. The steps of the justification appear below.

$x_0 = 0.180,\ y_0 = 1.21443$	Initial point $(0.180, 1.21443)$.
$x_1 = 0.195,\ y_1 = 1.23562$	Next point $(0.195, 1.23562)$.
$h = 0.015$	The value $h = x_1 - x_0$ should be small.
$f(x, y) = x + y$	Right side of the differential equation.
$Y = y_0 + \frac{h}{2}\left(f(x_0, y_0) + f(x_1, y_1)\right)$	Trapezoidal test, page 1031.
$= y_0 + \frac{h}{2}\left(x_0 + y_0 + x_1 + y_1\right)$	Expand functional expressions.
$= 1.21443 + \frac{0.015}{2}(2.82505)$	Expand expressions.
$= 1.2356179.$	Calculator result. The absolute error $E = \lvert y_1 - Y\rvert$ is 0.0000021.

This *process* can be carried out on the other four pairs of points, in a similar way, to find the five absolute errors

$$0.0000021,\ 0.0000052,\ 0.000000075,\ 0.0000036,\ 0.0000057.$$

The largest error is 0.0000057.

Details of a `maple` implementation appear below. The errors made with its ten-digit exact arithmetic may differ from those of a calculator.

```
# Execute the first group once.
f:=(x,y)->x+y:
L:=[[.180, 1.21443],[.195, 1.23562],
    [.210, 1.25736],[.225, 1.27965],
    [.240, 1.30250],[.255, 1.32592]]:
 n:=1:
# Execute the second group 5 times.
 x0:=L[n][1]:y0:=L[n][2]:
 x1:=L[n+1][1]:y1:=L[n+1][2]:
 Y:=y0+(x1-x0)*0.5*(f(x0,y0)+f(x1,y1)):
 n:=n+1: ABSerror:=abs(y1-Y);
```

New to the `maple` code is the *list* `L` of pairs of points. The syntax `L[n]` refers to item n of the 6 items in the list, a pair. Syntax `L[n][1]` means the first entry of that pair.

Example A.25 (Simpson Test)

A graphic for the differential equation $y' = x+y$ has window $0 \le x \le 0.5$, $1 \le y \le 2$ and uses 1000×1000 pixels. Given the data set below, compute the Simpson test prediction for each triple of data points. Report the four absolute errors and judge pass or fail.

$$(0.180, 1.21443),\ (0.195, 1.23562),\ (0.210, 1.25736),$$
$$(0.225, 1.27965),\ (0.240, 1.30250),\ (0.255, 1.32592).$$

Solution: The cutoff value for the absolute error is $(d-c)/1000 = 0.001$. It will be justified below that the four absolute errors according to the Simpson test are 0.0000087, 0.0000065, 0.0000023 and 0.0000079. The data passes Simpson's test.

The first three pairs of points in the plot data pass the Simpson test, page 1032, because the target value $y_1 = 1.25736$ is close to the test's value $Y = 1.257351350$, with absolute error $|y_1 - Y| = 0.0000087$. The steps of the justification are below.

$x_0 = 0.180,\ y_0 = 1.21443$	Initial point $(0.180, 1.21443)$.
$x_1 = 0.195,\ y_1 = 1.23562$	Second point $(0.195, 1.23562)$.
$x_2 = 0.210,\ y_2 = 1.25736$	Third point $(0.210, 1.23562)$.
$f(x,y) = x + y$	The right side of the differential equation.
$Y = y_0 + \frac{x_2 - x_0}{6}\,(f(x_0,y_0)$ $+4f(x_1,y_1) + f(x_2,y_2))$	Simpson test, page 1032.
$= y_0 + 0.005\,(x_0 + y_0$ $+4(x_1+y_1) + x_2 + y_2)$	Expand functional expressions.
$= 1.21443 + 0.005\,(8.58427)$	Substitute constants.
$= 1.2573513.$	Absolute error $E = \lvert y_2 - Y\rvert = 0.0000087$.

This *process* can be carried out in a similar way on the other triples of points:

Second: $(0.195, 1.23562)$, $(0.210, 1.25736)$, $(0.225, 1.27965)$,
Third: $(0.210, 1.25736)$, $(0.225, 1.27965)$, $(0.240, 1.30250)$,
Fourth: $(0.225, 1.27965)$, $(0.240, 1.30250)$, $(0.255, 1.32592)$.

The absolute errors for these last three cases are 0.0000065, 0.0000023 and 0.0000079.

Details of a `maple` implementation appear below.

```
# Execute the first group once.
f:=(x,y)->x+y:
L:=[[.180, 1.21443],[.195, 1.23562],
    [.210, 1.25736],[.225, 1.27965],
    [.240, 1.30250],[.255, 1.32592]]:
 n:=1:
# Execute the second group 4 times.
 x0:=L[n][1]:y0:=L[n][2]:
 x1:=L[n+1][1]:y1:=L[n+1][2]:
 x2:=L[n+2][1]:y2:=L[n+2][2]:
 Y:=y0+(x2-x0)*(f(x0,y0)+
    4*f(x1,y1)+f(x2,y2))/6:
 n:=n+1: ABSerror:=abs(y2-Y);
```

Example A.26 (Polynomial Quadrature)

Apply Simpson's polynomial rule (5) to verify $\int_1^2 (x^3 - 16x^2 + 4)dx = -355/12$.

Solution: The application proceeds as follows:

$I = \int_1^2 Q(x)dx$	Evaluate integral I using $Q(x) = x^3 - 16x^2 + 4$.

$= \dfrac{2-1}{6}\left(Q(1) + 4Q(3/2) + Q(2)\right)$	Apply Simpson's polynomial rule (5).
$= \dfrac{1}{6}\left(-11 + 4(-229/8) - 52\right)$	Use $Q(x) = x^3 - 16x^2 + 4$.
$= -\dfrac{355}{12}.$	Equality verified.

Integral Equation Justification. Let $f(x, y)$ be continuous for $a < x < b$, $-\infty < y < \infty$. Assume (x_0, y_0) is in the domain. It will be justified that the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ is equivalent to the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt.$$

The case $x \geq x_0$ will be considered, the other case $x \leq x_0$ being similar. *Equivalence* means a solution of the initial value problem is a solution of the integral equation, and conversely.

The integral equation is obtained from the initial value problem as follows: details.

$y'(t) = f(t, y(t))$	The given equation with x replaced by t.
$\int_{x_0}^{x} y'(t)dt = \int_{x_0}^{x} f(t, y(t))dt$	Integrate both sides on $x_0 \leq t \leq x$. It is assumed that y, y', f are continuous, which insures both integrals are defined.
$y(x) - y(x_0) = \int_{x_0}^{x} f(t, y(t))dt$	Apply the fundamental theorem of calculus, page 1008, part (a).

Conversely, if the integral equation is assumed, then $y(x)$ is differentiable by the fundamental theorem of calculus, page 1008, part (b). Differentiate across both sides of the integral equation to obtain $y' = f(x, y)$. Finally, substitute $x = x_0$ into the integral equation to obtain the initial condition $y(x_0) = y_0$.

Simpson's Polynomial Rule Proof. Let $Q(x)$ be a linear, quadratic or cubic polynomial. It will be verified that

$$\int_a^b Q(x)dx = \frac{b-a}{6}\left(Q(a) + 4Q((a+b)/2) + Q(b)\right). \tag{7}$$

If the formula holds for polynomial Q and c is a constant, then the formula also holds for the polynomial cQ. Similarly, if the formula holds for polynomials Q_1 and Q_2, then it also holds for $Q_1 + Q_2$. Consequently, it suffices to show that the formula is true for the special polynomials 1, x, x^2 and x^3, because then it holds for all combinations $Q(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$.

Only the special case $Q(x) = x^3$ will be treated here. The other cases are left to the exercises. The details:

$\text{RHS} = \dfrac{b-a}{6}\left(Q(a) + 4Q\left(\dfrac{a+b}{2}\right) + Q(b)\right)$	Evaluate the right side of equation (7).
$= \dfrac{b-a}{6}\left(a^3 + \dfrac{1}{2}(a+b)^3 + b^3\right)$	Substitute $Q(x) = x^3$.
$= \dfrac{b-a}{6}\dfrac{3}{2}\left(a^3 + a^2b + ab^2 + b^3\right)$	Expand $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ and simplify.

$= \frac{3}{12}\left(b^4 - a^4\right)$	Multiply and simplify.
$\text{LHS} = \int_a^b Q(x)dx$	Evaluate the left hand side (LHS) of equation (7).
$= \int_a^b x^3 dx$	Substitute $Q(x) = x^3$.
$= (b^4 - a^4)/4$	Evaluate.
$= \text{RHS}.$	Compare with the RHS.

Euler Test Proof: To justify Euler's test, page 1031, apply the equivalent integral equation (6) and the rectangular rule (2) with $F(t) = f(t, y(t))$, $a = x_0$ and $b = x_0 + h = x_1$. This gives a first approximation

$$y(x_0 + h) \approx y(x_0) + hF(x_0). \tag{8}$$

Then apply approximation $y(x_0) \approx y_0$ to the right side of (8) to give the approximation $Y = y_0 + hf(x_0, y_0)$.

Trapezoidal Test Proof: To justify the trapezoidal test, page 1031, begin with the equivalent integral equation (6) and approximate the integral using the trapezoidal rule (3), with $F(t) = f(t, y(t))$, $a = x_0$ and $b = x_0 + h = x_1$. This gives a first approximation

$$y(x_0 + h) \approx y(x_0) + \frac{h}{2}(F(x_0) + F(x_1)) \tag{9}$$

Then apply approximations $y(x_0) \approx y_0$ and $y(x_1) \approx y_1$ to the right side of (9) to give the approximation $Y = y_0 + \frac{h}{2}(f(x_0, y_0) + f(x_1, y_1))$.

Simpson Test Proof: To justify the Simpson test, page 1032, begin with the equivalent integral equation (6) and approximate the integral using Simpson's rule (4), with $F(t) = f(t, y(t))$, $a = x_0$ and $b = x_2$. This gives a first approximation

$$y(x_0 + h) \approx y(x_0) + \frac{x_2 - x_0}{6}(F(x_0) + 4F(x_1) + F(x_2)) \tag{10}$$

Then apply approximations $y(x_0) \approx y_0$, $y(x_1) \approx y_1$ and $y(x_2) \approx y_2$ to the right side of (10) to give the approximation

$$Y = y_0 + \frac{x_2 - x_0}{6}(f(x_0, y_0) + 4f(x_1, y_1) + f(x_2, y_2)).$$

Quadratic Interpolation Proof: Given $a < b$ and the three data points (a, Y_0), $((a + b)/2, Y_1))$, $(b, Y_2))$, it will be verified that the quadratic curve $Q(X)$ which connects the points is given by

$$\begin{aligned} Q(X) = Y_0 &+ (4Y_1 - Y_2 - 3Y_0)\frac{X - a}{b - a} \\ &+ (2Y_2 + 2Y_0 - 4Y_1)\frac{(X - a)^2}{(b - a)^2}. \end{aligned}$$

The term *quadratic* is meant loosely: it can be a constant or linear function as well. The solution is presented as two lemmas.[3] The first lemma contains the essential ideas. The second simply translates the variables.

[3] What's a lemma? It's a helper theorem, used to dissect long proofs into short pieces.

Lemma A.1 Given y_1 and y_2, define $A = y_2 - y_1$, $B = 2y_1 - y_2$. Then the quadratic $y = x(Ax + B)$ fits the data items $(0, 0)$, $(1, y_1)$, $(2, 2y_2)$.

Lemma A.2 Given Y_0, Y_1 and Y_2, define $y_1 = Y_1 - Y_0$, $y_2 = \frac{1}{2}(Y_2 - Y_0)$, $A = y_2 - y_1$, $B = 2y_1 - y_2$ and $x = 2(X - a)/(b - a)$. Then quadratic $Y(X) = Y_0 + x(Ax + B)$ fits the data items (a, Y_0), $((a + b)/2, Y_1)$, (b, Y_2).

To verify the first lemma, the formula $y = x(Ax + B)$ is tested to go through the given data points $(0, 0)$, $(1, y_1)$ and $(2, 2y_2)$. For example, the last pair is tested by the steps

$y(2) = 2(2A + B)$	Apply $y = x(Ax + B)$ with $x = 2$.
$= 4y_2 - 4y_1 + 4y_1 - 2y_2$	Use $A = y_2 - y_1$ and $B = 2y_1 - y_2$.
$= 2y_2$.	Therefore, the quadratic fits data item $(2, 2y_2)$.

The other two data items are tested similarly, details omitted here.

To verify the second lemma, observe that it is just a change of variables in the first lemma, $Y = Y_0 + y$. The data fit is checked as follows:

$Y(b) = Y_0 + y(2)$	Apply formulas $Y(X) = Y_0 + y(x)$, $y(x) = x(Ax+B)$ with $X = b$ and $x = 2$.
$= Y_0 + 2y_2$	Apply data fit $y(2) = 2y_2$.
$= Y_2$.	The quadratic fits the data item (b, Y_2).

The other two items are checked similarly, details omitted here. This completes the proof of the two lemmas. The formula for Q is obtained from the second lemma as $Q = Y_0 + Bx + Ax^2$ with substitutions for A, B and x performed to obtain the given equation for Q in terms of Y_0, Y_1, Y_2, a, b and X.

Exercises A.4

Euler Test: Spot Check Apply the methods of Example A.23, page 1033, to compute for the given differential equation the absolute error made by Euler's test for the given data. Report *pass* or *fail* for each exercise. Assume absolute error cutoff value $(d-c)/1000 = 0.001$.

1. $y' = 2y + \sin(x)$,
$(0.1, 0.005346)$,
$(0.2, 0.022884)$,
$(0.3, 0.055148)$.

2. $y' = -y + \cos(x)$,
$(0.1, 0.095000)$,
$(0.2, 0.180003)$,
$(0.3, 0.255019)$.

3. $y' = y(1-y) + 5$,
$(0.400, 1.877093)$,
$(0.405, 1.893746)$,
$(0.410, 1.910168)$.

4. $y' = y(2-y) + 10$,
$(0.400, 3.547216)$,
$(0.405, 3.569489)$,
$(0.410, 3.591196)$.

5. $y' = 1 + y^2$,
$(0.100, 0.100335)$,
$(0.105, 0.105388)$,
$(0.110, 0.110446)$.

6. $y' = 4 + 4y^2$,
$(0.100, 0.422793)$,
$(0.105, 0.446573)$,
$(0.110, 0.470781)$.

Trapezoidal Test Apply the methods of Example A.24, page 1033, to compute for the given differential equation the relative error $E = |y_1 - Y|/|y_1|$ made by the Trapezoidal test for the given data. Report for each exercise *pass* or *fail* and the three error values. Assume the given relative error cutoff value E^*.

7. $y' = 2y + \sin(x)$, $E^* = 0.001$,
$(0.1, 0.005346)$, $(0.2, 0.022884)$,
$(0.3, 0.055148)$, $(0.4, 0.105129)$.

8. $y' = -y + \cos(x)$, $E^* = 0.00009$,
$(0.1, 0.095000)$, $(0.2, 0.180003)$,
$(0.3, 0.255019)$, $(0.4, 0.320080)$.

9. $y' = y(1-y) + 5$, $E^* = 0.00024$,
$(0.100, 0.516828)$,
$(0.125, 0.647873)$,
$(0.150, 0.777953)$,
$(0.175, 0.621714)$.

10. $y' = y(2-y) + 10$, $E^* = 0.0013$,
$(0.100, 1.067919)$,
$(0.125, 1.341712)$,
$(0.150, 1.610877)$,
$(0.175, 1.871962)$.

11. $y' = \dfrac{1-x}{1+y}$, $E^* = 0.0004$,
$(0.100, 0.090871)$,
$(0.125, 0.111024)$,
$(0.150, 0.130265)$,
$(0.175, 0.148641)$.

12. $y' = \dfrac{1+x}{1-y}$, $E^* = 0.00047$,
$(0.100, 0.111181)$,
$(0.125, 0.143043)$,
$(0.150, 0.176896)$,
$(0.175, 0.212996)$.

Simpson Test Apply the ideas in Example A.25, page 1034, to compute for the given differential equation the relative error $E = |y_2 - Y|/|y_2|$ made by the Simpson test for the given data. Report for each exercise *pass* or *fail* and the three error values. Assume the given relative error cutoff value E^*.

13. $y' = 2y + \sin(x)$, $E^* = 0.0008$,
$(0.2, 0.022884)$, $(0.3, 0.055148)$,
$(0.4, 0.105129)$.

14. $y' = -y + \cos(x)$, $E^* = 0.00044$,
$(0.2, 0.180003)$, $(0.3, 0.255019)$,
$(0.4, 0.320080)$.

15. $y' = y(1-y) + 5$, $E^* = 0.000451$,
$(0.2, 1.031950)$, $(0.3, 1.495883)$,
$(0.4, 1.877093)$.

16. $y' = y(2-y)+10$, $E^* = 0.0004$, $(0.2, 2.121932)$, $(0.3, 2.970036)$, $(0.4, 3.547216)$.

17. $y' = \dfrac{1-x}{1+y}$, $E^* = 0.0004$, $(0.2, 0.166190)$, $(0.3, 0.228821)$, $(0.4, 0.280625)$.

18. $y' = \dfrac{1+x}{1-y}$, $E^* = 0.00068$, $(0.2, 0.251669)$, $(0.3, 0.443224)$, $(0.4, 0.800000)$.

Simpson's Rule The following exercises use formulas and techniques found in the proof on page 1036 and in Example A.26, page 1035.

19. Verify with Simpson's rule (5) for cubic polynomials the equality $\int_1^2 (x^3+16x^2+4)dx = 541/12$.

20. Verify with Simpson's rule (5) for cubic polynomials the equality $\int_1^2 (x^3+x+14)dx = 77/4$.

21. Let $f(x)$ satisfy $f(0) = 1$, $f(1/2) = 6/5$, $f(1) = 3/4$. Apply Simpson's rule with one division to verify that $\int_0^1 f(x)dx \approx 131/120$.

22. Let $f(x)$ satisfy $f(0) = -1$, $f(1/2) = 1$, $f(1) = 2$. Apply Simpson's rule with one division to verify that $\int_0^1 f(x)dx \approx 5/6$.

23. Verify Simpson's equality (5), assuming $Q(x) = 1$ and $Q(x) = x$.

24. Verify Simpson's equality (5), assuming $Q(x) = x^2$.

Quadratic Interpolation The following exercises use formulas and techniques from the proof on page 1037.

25. Verify directly that the quadratic polynomial $y = x(7-4x)$ goes through the points $(0,0)$, $(1,3)$, $(2,-2)$.

26. Verify directly that the quadratic polynomial $y = x(8-5x)$ goes through the points $(0,0)$, $(1,3)$, $(2,-4)$.

27. Compute the quadratic interpolation polynomial $Q(x)$ which goes through the points $(0,1)$, $(0.5,1.2)$, $(1,0.75)$.

28. Compute the quadratic interpolation polynomial $Q(x)$ which goes through the points $(0,-1)$, $(0.5,1)$, $(1,2)$.

29. Verify the remaining cases in Lemma A.1, page 1038.

30. Verify the remaining cases in Lemma A.2, page 1038.

A.5 Implicit Functions

The subject of **implicit function theory** treats the problem of solving an equation $F(x, y) = 0$ for y in terms of x. In differential equations, it is the theoretical basis for extracting an explicit solution $y(x)$ from an implicit solution $F(x, y) = 0$.

Theorem A.2 (Implicit Function Theorem)
Let $F(x, y)$, $F_x(x, y)$, $F_y(x, y)$ be defined and continuous in an open region D in the plane. Assume (x_0, y_0) is the center of a disk contained entirely in D and $F_y(x_0, y_0) \neq 0$. Then there is a number $H > 0$ and a function $y = y(x)$ such that

[1] $y(x_0) = y_0$,

[2] y is continuous on $|x - x_0| < H$,

[3] $(x, y(x))$ is in D for $|x - x_0| < H$,

[4] $F(x, y(x)) = 0$ for $|x - x_0| < H$.

Further, if another function $y = Y(x)$ satisfies [1]–[4] on $|x - x_0| < H$, then $y(x) = Y(x)$ for $|x - x_0| < H$.

The proof of Theorem A.2 appears in various references, for example, see Taylor-Mann [Taylor-M] and Marsden-Tromba [Marsden]. Results of this type are theoretical, that is, devoid of a method for finding the function $y(x)$.

Practical Numerical Methods

Item **[4]** in Theorem A.2 together with the chain rule $\frac{d}{dt}F(x(t), y(t)) = F_x x'(t) + F_y y'(t)$ implies that $y(x)$ satisfies the initial value problem

$$y' = -\frac{F_x(x, y)}{F_y(x, y)}, \quad y(x_0) = y_0. \tag{1}$$

Problem (1) is the basis for practical numerical methods which are used in applications to calculate and graph the implicit solution $y(x)$ of the equation $F(x, y) = 0$. See Example A.27, page 1042.

Computer Algebra Methods

Computer algebra systems `maple` and `mathematica` have facilities for solving an equation $F(x, y) = 0$ for y in terms of x. Limited support exists for making graphics directly from the implicit equation $F(x, y) = 0$. See Example A.28, page 1042.

Work-alike systems such as `matlab`, `octave` and `scilab` can be applied to solve implicit equations, although the work involved is always more tedious. One idea of merit is to model the implicit equation $F(x, y) = 0$ as several initial value problems, then apply differential equation numerical solution methods to graph the solutions. See Example A.29, page 1043.

Example A.27 (Modeling an Implicit Function Problem)

Model the implicit equation $x^2 + 4y^4 = 4$ at $x = 0$, $y = 1$ as an initial value problem for a function $y(x)$ defined near $x = 0$.

Solution: Let $F(x,y) = x^2 + 4y^4$. Then $x^2 + 4y^4 = 4$ can be written as $F(x,y) = 4$. We verify $F(0,1) = 4$. The chain rule $(d/dt)F(x(t),y(t)) = F_x x'(t) + F_y y'(t)$ with $x = t$ and $y = y(t)$ gives from $F(x,y) = 4$ the equation $\frac{dy}{dt} = -F_x(t,y(t))/F_y(t,y(t))$. Compute $F_x = 2x$ and $F_y = 16y^3$. The initial value problem is

$$\frac{dy}{dt} = -\frac{t}{8y^3}, \quad y(0) = 1.$$

Example A.28 (Solving $\mathbf{F(x,y) = 0}$ Symbolically)

Solve symbolically for y as a function of x in the implicit equation $x^2 + 4y^4 = 4$ at $x = 0$, $y = 1$ both by hand and by computer.

Solution: College algebra methods apply to solve $x^2 + 4y^4 = 4$ for y in terms of x, giving $y(x) = \sqrt[4]{1 - x^2/4}$. The graph is defined on $-2 \le x \le 2$; see Figure 14. The college algebra details:

$4y^4 = 4 - x^2$	Start with $x^2 + 4y^4 = 4$ and isolate y on the left.
$\lvert y\rvert = \sqrt[4]{1 - x^2/4}$	Divide by 4 and take the fourth root of both sides.
$y = \sqrt[4]{1 - x^2/4}$	Replace $\lvert y\rvert$ by $\pm y$ and resolve the sign with $y = 1$ at $x = 0$.

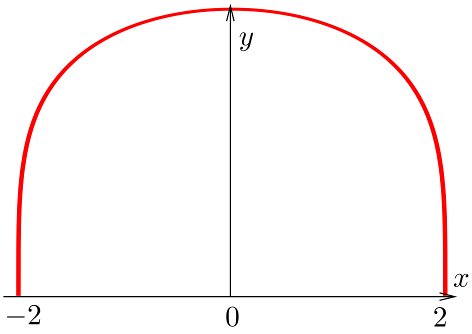


Figure 14. Implicit solution $y(x)$ of $x^2 + 4y^4 = 4$ at $x = 0$, $y = 1$.

The computer algebra system `maple` partially solves the problem with the command

```
solve(x^2+4*y^4=4,y));
```

Reported are four answers:

$$\begin{array}{ll} \frac{1}{2}\sqrt[4]{-4\,x^2+16}, & \frac{1}{2}\sqrt{-1}\sqrt[4]{-4\,x^2+16}, \\ -\frac{1}{2}\sqrt[4]{-4\,x^2+16}, & -\frac{1}{2}\sqrt{-1}\sqrt[4]{-4\,x^2+16}. \end{array}$$

Only one satisfies $y = 1$ at $x = 0$, namely the first. It is typical in computer algebra systems to spend time sorting out the system's answer.

Remark on Algebraic Complexity. The more complicated implicit equation $x^2 + 4y^4 + xy = 4$ does not have a useful or simple college algebra solution. To understand the complications, execute the following maple code, which displays several pages of answers involving cube roots of sixth degree polynomials. None of the answers are useful, it being easier to employ the ideas of Example A.27, page 1042.

```
allvalues([solve(x^2+4*y^4+x*y=4,y)]);
```

Example A.29 (Solving $F(x, y) = 0$ Numerically)
Solve numerically by computer for y as a function of x in the implicit equation $x^2 + 4y^4 = 4$ at $x = 0$, $y = 1$. Plot $y(x)$ on $0 \le x \le 2$.

Solution: It was shown in Example A.27, page 1042, that the problem is equivalent to the differential equation problem

$$\frac{dy}{dx} = -\frac{x}{8y^3}, \quad y(0) = 1.$$

The plot on $0 \le x \le 2$ will look like the right half of Figure 14. The maple code:

```
with(DEtools):
de:=diff(y(x),x)=-x/(8*y(x)^3):
DEplot(de,y(x),x=0..2,[[y(0)=1]],arrows=NONE);
```

A more simplistic approach, which is also capable of direct computation of values of $y(x)$, is to use the maple function dsolve.

```
# Maple V 5.1
de:=diff(y(x),x)=-x/(8*y(x)^3):ic:=y(0)=1:
p:=dsolve({de,ic},y(x),numeric);
Y:=x->rhs(p(x)[2]);
plot(Y,0..2);
```

Exercises A.5

Modeling an Implicit Function Problem

Apply the ideas in Example A.27, page 1042 to model the given implicit equation as an initial value problem for a function $y(x)$ defined near $x = 0$.

1. $x^2 + xy^4 + y = 1$,
$x = 0$, $y = 1$.

2. $x + xy^4 + y = 1$,
$x = 0$, $y = 1$.

3. $x + y^2 \ln(x + 1) + y = 2$,
$x = 0$, $y = 2$.

4. $e^x + y^2 \ln(x + 1) + y = 1$,
$x = 0$, $y = 2$.

5. $\sin x + y^3 \cos x + y^2 = 2$,
$x = 0$, $y = 1$.

6. $\tan x + y^2 \sec x + y^3 = 2$,
$x = 0$, $y = 1$.

7. $e^x + y^2x^2 + xy + 2y = 3$,
$x = 0$, $y = 1$.

8. $e^{-x} + -y^2x^2 + xy + 2y = 3$,
$x = 0$, $y = 1$.

Solve $\mathbf{F(x, y) = 0}$ Symbolically

Solve symbolically for y as a function of x in the given implicit equation both by hand and by computer. Apply the methods of Example A.28, page 1042.

9. $x^2 + 5y^4 = 5$,
$x = 0$, $y = 1$.

10. $x^2 + 5y^2 = 5$,
$x = 0$, $y = 1$.

11. $x^2 + y^2 + 2y = 3$,
$x = 0$, $y = 1$.

12. $x^2 + 4y^2 - 2y = 2$,
$x = 0$, $y = 1$.

13. $\sin x + y^4 = 1$,
$x = 0$, $y = 1$.

14. $\sin x + y^4 + 2y^2 = 3$,
$x = 0$, $y = 1$.

15. $-\sin x + \cos y = 1$,
$x = 0$, $y = 0$.

16. $\sin x + \cos y = 1$,
$x = 0$, $y = 0$.

Solve $F(x, y) = 0$ Numerically

Solve numerically by computer for y as a function of x in the given implicit equation. Plot $y(x)$ on an interesting interval. See Example A.29, page 1043 for methods.

17. $x^2 + x + 4 + \cos y = 5$,
$x = 0$, $y = 0$.

18. $x^2 + x + 6 - \cos(y) = 5$,
$x = 0$, $y = 0$.

19. $x^2 + y^3 + 2y = 3$,
$x = 0$, $y = 1$.

20. $x^2 + 4y^3 - 2y = 2$,
$x = 0$, $y = 1$.

21. $\sin x + y^4 + y = 2$,
$x = 0$, $y = 1$.

22. $\sin x + y^4 + 2y = 3$,
$x = 0$, $y = 1$.

23. $-\sin x + y + \cos y = 1$,
$x = 0$, $y = 0$.

24. $\sin x - y + \cos y = 1$,
$x = 0$, $y = 0$.

Index

Bibliography

[Abram-St] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York, Dover Publications, 1965. 654

[Bateman] H. Bateman, *The solution of a system of differential equations occurring in the theory of radioactive transformations.*
Proceedings Cambridge Philosophical Society 15 (1910), No. pt V, pp. 423–427. Archived:
`https://archive.org/details/cbarchive_1/22715_solutionofasystemofdifferentia1843` 592, 595

[Birkhoff] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, 4th Ed., New York, MacMillan, 1977. 721

[BergMcG] P. W. Berg and J. L. McGregor, *Elementary Partial Differential Equations*, 1st Ed., Holden–Day, 1966.

[BirkRota] G. Birkhoff and G. C. Rota, *Ordinary Differential Equations*, 3rd Ed., New York, John Wiley and Sons, 1978. 61, 965, 971

[Borrelli] R. Borrelli and C. Coleman, *Differential Equations*, 2nd ed., New York, John Wiley and Sons, 2004.

[Braun1986] M. Braun, *Differential Equations and Their Applications*, 3rd Ed., New York, Springer-Verlag, 1986. 9, 20

[BCD1983] M. Braun, C. S. Coleman and D. A. Drew, Editors, *Differential Equation Models*, New York, Springer-Verlag, 1983.

[Bret] O. Bretcher, *Linear Algebra with Applications*, Second Edition, New Jersey, Prentice-Hall, 2001.

[BurFair] R. L. Burden and J. D. Faires, *Numerical Analysis*, Seventh Edition, Pacific Grove, Brooks-Cole Publishing Co., 2001. 245, 250, 251

[Codd-L] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, New York, McGraw–Hill, 1955.

[ChurB1990] R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, 5th Ed., New York, McGraw–Hill, 1990.

[ChurB1978] R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd Ed., New York, McGraw–Hill, 1978.

[Cheney-K] W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*, 2nd Ed., Monterey, Brooks/Cole Publishing Co., 1985. 250, 251, 255

[Chur1972] R. V. Churchill, *Operational Mathematics*, 3rd Ed., New York, McGraw–Hill, 1972.

[Cushing] J. M. Cushing, *Differential Equations: An Applied Approach*, New Jersey, Pearson Prentice Hall, 2004. 821, 823

[D-S] H. F. Davis and A. D. Snider, *Introduction to Vector Analysis*, 6th Ed., Dubuque, William C. Brown Publishers, 1991.

[DenH] J. P. Den Hartog, *Mechanical Vibrations*, 4th Ed., New York, Dover Publications, 1985.

[Enright] W. H. Enright, *The Relative Efficiency of Alternative Defect Control Schemes for High Order Continuous Runge-Kutta Formulas*, Department of Computer Science, University of Toronto, Technical Report 252/91, June, 1991. See `dverk78`. 255

[EP] C. H. Edwards and D. E. Penney, *Differential Equations and Linear Algebra*, New Jersey, Prentice-Hall, 2001. 379, 587

[EP2] C. H. Edwards and D. E. Penney, *Differential Equations and Linear Algebra*, Second Edition, New Jersey, Prentice-Hall, 2005. 134, 587

[EPbvp] C. H. Edwards and D. E. Penney, *Differential Equations and Boundary Value Problems*, Second Edition, New Jersey, Prentice-Hall, 2000.

[Erdelyi] A. Erdelyi et al, *Tables of Integral Transforms*, Volumes I and II, New York, McGraw–Hill, 1954.

[FMM] G. E. Forsythe, M. A. Malcolm and C. B. Moler, *Computer Methods for Mathematical Computations*, New Jersey, Prentice-Hall, 1977. 255

[Friedman] A. Friedman, *Advanced Calculus*, 2007 Dover Edition republication of original: New York, Holt, Reinhart and Winston, 1971.

[Garab1964] P. R. Garabedian, *Partial Differential Equations*, New York, John Wiley and Sons, 1964.

[Gross-D] S. I. Grossman and W. R. Derrick, *Advanced Engineering Mathematics*, New York, Harper and Row, 1988.

[Gear] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, New Jersey, Prentice-Hall, 1971. 255

[Gupta] R. C. Gupta, *On linear differential equation with constant coefficients: a recursive alternative to the method of undetermined coefficients*, Int. J. Math. Edu. Sci. Technol. 27 (1996), 757-760. 577, 587

[Giord1991] F. R. Giordano and M. D. Weir, *Differential Equations: A Modeling Approach*, Massachussets, Addison-Wesley, 1991.

[Henrici] P. Henrici, *Elements of Numerical Analysis*, New York, John Wiley and Sons, 1965.

[Jackson] D. Jackson, *Fourier Series and Orthogonal Polynomials*, Washington, D.C., Mathematical Association of America, 1941.

[Keener] J. P. Keener, *Principles of Applied Mathematics*, Massachussets, Addison–Wesley, 1988.

[KKOP] Kreider, Kuller, Ostberg and Perkins, *An Introduction to Linear Analysis*, Massachussets, Addison–Wesley, 1966 587, 983

[Kreyszig] E. Kreyszig, *Advanced Engineering Mathematics*, 7th Ed., New York, John Wiley and Sons, 1993. 587

[Laham] M. F. Laham, et al, *Fish Harvesting Management Strategies Using Logistic Growth Model*, Sains Malaysiana 41(2)(2012), 171–177 Archived: M.F. Laham 2012 286, 287

[Lerch] M. Lerch, *Sur un point de la théorie des fonctions génératrices d'Abel*, Acta Mathematica 27 (1903), 339–351. 596

[Love1989] E. R. Love, *Particular solutions of constant coefficient linear differential equations*, IMA Bulletin 25 (1989), 165-166. 587

[May] R.M. May, editor, *Theoretical Ecology: Principles and Applications*, W.B. Saunders, 1976

[MayBCHL] R.M. May, J.R. Beddington, C.W. Clark, S.J. Holt and R.M. Lewis, Management of Multispecies Fisheries, Science 205 (July 1979), pp 256-277

[Marsden] J. E. Marsden and A. J. Tromba, *Vector Calculus*, Fourth Edition, San Francisco, W. H. Freeman and Company 1996. 1041

[McLach] N. W. McLachlan, *Bessel Functions for Engineers*, 2nd Ed., Oxford, Clarendon Press, 1961.

[Noble] B. Noble, *Applied Linear Algebra*, 3rd Ed., New Jersey, Prentice Hall, 1988.

[Press] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, London, Cambridge University Press, 1986.

[ODEP] A. C. Hindmarsh, *Odepack, a Systemized Collection of ODE Solvers*, in *Scientific Computing* by R. S. Stepleman et al. (eds.), Amsterdam, North-Holland 1983. 255

[Rice1972] D. Barton, I. M. Willers and R. V. M. Zahar, In *Mathematical Software*, J. R. RICE, Ed., New York, Academic Press, 1972. 255

[Rudin] W. A. Rudin, *Principles of Mathematical Analysis*, New York, McGraw Hill, 1976, ISBN10: 007054235X. 648, 651

[Schein1996] E. R. Scheinerman, *Introduction to Dynamical Systems*, New Jersey, Prentice Hall, 1996.

[Strauss] W. A. Strauss, *Introduction to Partial Differential Equations*, New York, John Wiley and Sons, 2008.

[Strang] W. G. Strang, *Linear Algebra and Its Applications*, New York, Academic Press, 1980. 369

[Taylor-M] Taylor and Mann, *Advanced Calculus*, New York, John Wiley and Sons, 1983. 1041

[Varberg] D. Varberg and E. J. Purcell, *Calculus*, 7th Ed., New Jersey, Prentice Hall, 1997.

[Watson] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd Ed., London, Cambridge University Press, 1944. 982, 983

[Weis] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*, 2nd Ed., New York, Chapman and Hall/CRC, 2002. 596

[Widd1975] D. V. Widder, *The Heat Equation*, New York, Academic Press, 1975.

[Widd1941] D. V. Widder, *The Laplace Transform*, New Jersey, Princeton University Press, 1941. 596

[Zill-C] Dennis G. Zill and Michael R. Cullen, *Advanced Engineering Mathematics*,
Boston, PWS-Kent Publishing Co.,1993. 21, 150

PDF Sources

Text, Solutions and Corrections

Author: Grant B. Gustafson, University of Utah, Salt Lake City 84112.

Paperback Textbook: There are 12 chapters on differential equations and linear algebra, book format 7 x 10 inches, 1077 pages. Copies of the textbook are available in two volumes at **Amazon** Kindle Direct Publishing for Amazon's cost of printing and shipping. No author profit. Volume I chapters 1-7, ISBN 9798705491124, 661 pages. Volume II chapters 8-12, ISBN 9798711123651, 479 pages. Both paperbacks have extra pages of backmatter: background topics Chapter A, the whole book index and the bibliography.

Textbook PDF with Solution Manual: Packaged as one PDF (13 MB) with hyperlink navigation to displayed equations and theorems. The header in an exercise set has a blue hyperlink ↗ to the same section in the solutions. The header of the exercise section within a solution Appendix has a red hyperlink ↗ to the textbook exercises. Solutions are organized by chapter, e.g., Appendix 5 for Chapter 5. Odd-numbered exercises have a solution. A few even-numbered exercises have hints and answers. Computer code can be mouse-copied directly from the PDF. Free to use or download, no restrictions for educational use.

Sources at Utah:

https://math.utah.edu/g̃ustafso/indexUtahBookGG.html

Sources for a Local Folder No Internet: The same PDF can be downloaded to a tablet, computer or phone to be viewed locally. After download, no internet is required. Best for computer or tablet using a PDF viewer (Adobe Reader, Evince) or web browser with PDF support (Chrome, FireFox). Smart phones can be used in landscape mode.

Sources at GitHub and GitLab Projects: Utah sources are duplicated at

`https://github.com/ggustaf/github.io` and mirror
`https://gitlab.com/ggustaf/answers`.

Communication: To contribute a solution or correction, ask a question or request an answer, click the link below, then create a GitHub issue and post. Contributions and corrections are credited, privacy respected.

`https://github.com/ggustaf/github.io/issues`