Math 6610 : Analysis of Numerical Methods I

Chee Han Tan

Last modified : August 18, 2017

Contents

Abstract: These notes are largely based on **Math 6610: Analysis of Numerical Methods I** course, taught by Yekaterina Epshteyn in Fall 2016, at the University of Utah. Additional examples or remarks or results from other sources are added as I see fit, mainly to facilitate my understanding. These notes are by no means accurate or applicable, and any mistakes here are of course my own. Please report any typographical errors or mathematical fallacy to me by email tan@math.utah.edu

Chapter 1

Introduction

We review some basic facts about linear algebra, in particular on viewing matrix-vector multiplication as linear combination of column matrices; this plays an important role in understanding key ideas behind many algorithms of numerical linear algebra. We review orthogonality, where many of the best algorithms are based upon. Finally, we discuss about vector norms and matrix norms, as these provide a way of measuring approximations and convergence of numerical algorithm.

1.1 Linear Algebra

Let $A \in \mathbb{C}^{m\times n}$ be an $m \times n$ matrix. The map $x \mapsto Ax$ is **linear**, *i.e.* the following holds for any $x, y \in \mathbb{C}^n$ and scalars $\alpha, \beta \in \mathbb{C}$:

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay.$$

Conversely, any linear map from $\mathbb{C}^n$ to $\mathbb{C}^m$ can be expressed as multiplication by an $m \times n$ matrix. The **matrix-vector product** $b = Ax \in \mathbb{C}^m$ is defined as

$$b_i = \sum_{j=1}^{n} a_{ij}x_j \quad \text{for every } i = 1, \ldots, m.$$

It is not too difficult to see that matrix-vector product can also be view as linear combination of columns $\{a_1, \ldots, a_n j\}$ of A, *i.e.*

$$b = Ax = \sum_{j=1}^{n} x_j a_j. \tag{1.1.1}$$

This easily generalises to matrix-matrix product $B = AC$, in which each column of B is a linear combination of the columns of A. More precisely, if $A \in C^{m\times l}$ and $C \in \mathbb{C}^{l\times n}$, then $B \in \mathbb{C}^{m\times n}$ with

$$b_{ij} = \sum_{k=1}^{l} a_{ik}c_{kj} \quad \text{for each } i = 1, \ldots, m,\ j = 1, \ldots, n,$$

or equivalently,

$$b_k = \sum_{j=1}^{l} c_{jk}a_j \quad \text{for each } k = 1, \ldots, n.$$

Example 1.1.1. The **outer product** is the product of a column vector $u \in \mathbb{C}^m$ and a row vector $v^* \in \mathbb{C}^n$, which gives a **rank-one-matrix** $A = uv^* \in \mathbb{C}^{m\times n}$. Symbolically,

$$A = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} = \begin{bmatrix} v_1 u & v_2 u & \dots & v_n u \end{bmatrix}.$$

Definition 1.1.2. Given a matrix $A \in \mathbb{C}^{m\times n}$,

(a) The **nullspace** $\mathcal{N}(A)$ of A is the set of vectors $x \in \mathbb{C}^n$ such that $Ax = \mathbf{0}$.

(b) The **range** $\mathcal{R}(A)$ of A is the set of vectors $y \in \mathbb{C}^m$ such that $y = Ax$ for some $x \in \mathbb{C}^n$. It is clear from (1.1.1) that $\mathcal{R}(A)$ is the vector space spanned by columns of A:

$$\mathcal{R}(A) = \operatorname{span}\{a_1, \dots, a_n\}.$$

Consequently, $\mathcal{R}(A)$ is also called the **column space** of A.

(c) The **column rank** of A is the dimension of its column space. The **row rank** of A is the dimension of its row space.

It can be shown that the column rank is always equal to the row rank of a matrix. Thus, the rank of a matrix is well-defined. A matrix $A \in \mathbb{C}^{m\times n}$ of **full rank** is one that has the maximal possible rank $\min\{m, n\}$. This means that a matrix of full rank with $m \geq n$ must have n linearly independent columns.

Theorem 1.1.3. *A matrix $A \in \mathbb{C}^{m\times n}$ with $m \geq n$ has full rank if and only if it maps no two distinct vectors to the same vector.*

Proof. Suppose A is of full rank, then its columns $\{a_1, \dots, a_n\}$ form a linearly independent set of vectors in $\mathbb{C}^n$. Suppose $Ax = Ay$, we need to show that $x = y$ but this is true since

$$A(x - y) = \mathbf{0} \implies \sum_{j=1}^{n} (x_j - y_j) a_j = \mathbf{0} \implies x_j - y_j = 0 \quad \text{for each } j = 1, \dots, n.$$

Conversely, suppose A maps no two distinct vectors to the same vector. To show that A is of full rank, it suffices to prove that its columns $\{a_1, \dots, a_n\}$ are linearly independent in $\mathbb{C}^n$. Suppose

$$\sum_{j=1}^{n} x_j a_j = \mathbf{0}.$$

This is equivalent to $Ax = \mathbf{0}$ with $x = (x_1, \dots, x_n)^* \in \mathbb{C}^n$, and we see that x must be the zero vector. Otherwise there exists a nonzero vector $y \in \mathbb{C}^n$ such that $Ay = \mathbf{0} = A(\mathbf{0})$ and this contradicts the assumption.

■

Theorem 1.1.4. *For $A \in \mathbb{C}^{m\times m}$, the following are equivalent:*

(a) A has an inverse $A^{-1} \in \mathbb{C}^{m\times m}$ satisfying $AA^{-1} = A^{-1}A = I_m$.

(b) $\text{rank}(A) = m$.

(c) $\mathcal{R}(A) = \mathbb{C}^m$.

(d) $\mathcal{N}(A) = \{\mathbf{0}\}$.

(e) 0 is not an eigenvalue of A.

(f) 0 is not a singular value of A.

(g) $\det(A) \neq 0$.

When writing the product $x = A^{-1}b$, we should understand x as the unique vector that satisfies the equation $Ax = b$. This means that x is the vector of coefficients of the unique linear expansion of b in the basis of columns of A. Multiplication by A^{-1} is a **change of basis** operation. More precisely, if we view b as coefficients of the expansion of b in $\{e_1, \ldots, e_m\}$, then multiplication of A^{-1} results in coefficients of the expansion of b in $\{a_1, \ldots, a_m\}$.

1.2 Orthogonal Vectors and Matrices

Given a matrix $A \in \mathbb{C}^{m\times n}$, we denote its **Hermitian conjugate** or **adjoint** by A^*. For example,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \implies A^* = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & \bar{a}_{31} \\ \bar{a}_{12} & \bar{a}_{22} & \bar{a}_{32} \end{bmatrix}.$$

A is said to be **Hermitian** if $A = A^*$. Note that a Hermitian matrix must be square by definition. Among the nice properties about finite-dimensional vector space is the notion of orthogonality. In a plane $\mathbb{R}^2$, two vectors are orthogonal if they make an angle of 90°; this can be extended into higher-dimensional Euclidean space by introducing the notion of inner product.

Definition 1.2.1.

(a) The **inner product** of two column vectors $x, y \in \mathbb{C}^m$ is defined as

$$(x, y) = x^*y = \sum_{j=1}^{m} \bar{x}_j y_j.$$

(b) The **Euclidean length** of $x \in \mathbb{C}^n$ is defined as

$$\|x\|_2 = \sqrt{x^*x} = \left(\sum_{j=1}^{m} |x_j|^2 \right)^{\frac{1}{2}}.$$

(c) The cosine of the angle α between x and y can be expressed in terms of the inner product

$$\cos(\alpha) = \frac{x^*y}{\|x\|_2\|y\|_2}.$$

Remark 1.2.2. Over $\mathbb{C}$, the inner product is **sesquilinear**, *i.e.* $x \mapsto (x, z)$ is linear and $y \mapsto (z, y)$ is conjugate linear. Over $\mathbb{R}$, the inner product is **bilinear**.

Definition 1.2.3. A set of nonzero vectors S is said to be **orthogonal** if its elements are pairwise orthogonal, that is,

$$x, y \in S,\ x \neq y \implies (x, y) = x^*y = 0.$$

S is said to be **orthonormal** if S is orthogonal and $\|x\|_2 = 1$ for every $x \in S$.

Theorem 1.2.4. *The vectors in an orthogonal set S are linearly independent. Consequently, if an orthogonal set $S \subset \mathbb{C}^m$ contains m vectors, then it is a basis for $\mathbb{C}^m$.*

Proof. Suppose, by contradiction, that the set of orthogonal vectors S is not linearly independent. This means that at least one of the vectors $v_k \in S$ can be written as a non-trivial linear combination of the remaining vectors in S, *i.e.*

$$v_k = \sum_{j \neq k} \alpha_j v_j.$$

Taking the inner product of v_k against v_k and using orthogonality of the set S gives

$$(v_k, v_k) = \sum_{j \neq k} (\alpha_j v_j, v_k) = 0,$$

which contradicts the assumption that all vectors in S are nonzero.

■

An important consequence of inner product is that it can be used to decompose arbitrary vectors into orthogonal components. More precisely, suppose $\{q_1, q_2, \ldots, q_n\}$ is an orthonormal set, and let v be an arbitrary vector. We decompose v into $(n+1)$ orthogonal components as

$$v = r + \sum_{j=1}^{n} (q_j, v) q_j = r + \sum_{j=1}^{n} (q_j q_j^*) v. \tag{1.2.1}$$

We see that r is the part of v orthogonal to $\{q_1, q_2, \ldots, q_n\}$ and for every $j = 1, 2, \ldots, n$, $(q_j, v)q_j$ is the part of v in the direction of q_j.

If $\{q_j\}$ is a basis for $\mathbb{C}^m$, then n must be equal to m and r must be the zero vector, so v is completely decomposed into m orthogonal components in the direction of the q_j. In (1.2.1), we see that we have two different expressions. In the first case, we view v as a sum of coefficients q_j^*v times vectors q_j. In the second case, we view v as a sum of orthogonal projections of v onto the various directions q_j. The jth projection operation is achieved by the very special rank-one matrix $q_j q_j^*$.

A square matrix $Q \in \mathbb{C}^{m\times m}$ is **unitary** (or **orthogonal** in the real case) if $Q^* = Q^{-1}$, that is, $Q^*Q = QQ^* = I_m$. In terms of the columns of Q, we have the relation $q_i^* q_j = \delta_{ij}$. This means that columns of a unitary matrix Q form an orthonormal basis for $\mathbb{C}^m$. In the real case, multiplication by an orthogonal matrix Q corresponds to a rigid rotation if $\det(Q) = 1$ or reflection if $\det(Q) = -1$.

Lemma 1.2.5. *The inner product is invariant under unitary transformation, i.e. for any unitary matrix $Q \in \mathbb{C}^{m\times m}$, $(Qx, Qy) = (x, y)$ for any $x, y \in \mathbb{C}^m$. Such invariance means that angles between vectors and their lengths are preserved under unitary transformation.*

Proof. We simply expand (Qx, Qy) and obtain

$$(Qx, Qy) = (Qx)^* Qy = x^* Q^* Qy = x^* y = (x, y).$$

In particular, we have that

$$\|Qx\|_2 = \|x\|_2 \quad \text{for any } x \in \mathbb{C}^m.$$

■

Remark 1.2.6. Note that the lemma is still true for any matrices with orthornormal columns.

1.3 Norms and Inequalities

We already see from Section 1.2 on how to quantify the length of a vector using the inner product, called the Euclidean length, which is a generalisation of distance in a plane. In practice, it is useful to consider other notions of length in a vector space, which give rise to the following:

Definition 1.3.1. A **norm** is a function $\|\cdot\| : \mathbb{C}^n \longrightarrow \mathbb{R}$ satisfying the following properties for all vectors $x, y \in \mathbb{C}^n$ and scalars $\alpha \in \mathbb{C}$:

(N1) $\|x\| \geq 0$ and $\|x\| = 0 \implies x = \mathbf{0}$.

(N2) $\|\alpha x\| = |\alpha| \|x\|$.

(N3) $\|x + y\| \leq \|x\| + \|y\|$.

Below are a few important examples of vector norms in $\mathbb{C}^n$:

$$\|x\|_1 = \sum_{j=1}^{n} |x_j| \qquad (l^1 \text{ norm})$$

$$\|x\|_2 = \left(\sum_{j=1}^{n} |x_j|^2 \right)^{\frac{1}{2}} \qquad (l^2 \text{ norm})$$

$$\|x\|_\infty = \max_{1\leq j\leq n} |x_j| \qquad (l^\infty \text{ norm})$$

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{\frac{1}{p}}, \quad p \geq 1. \qquad (l^p \text{ norm})$$

For any nonsingular matrix $W \in \mathbb{C}^{n\times n}$, we can define the **weighted p-norms**, given by

$$\|x\|_W = \|Wx\|_p = \left(\sum_{i=1}^{n} \left|\sum_{j=1}^{n} w_{ij}x_j\right|^p\right)^{\frac{1}{p}}.$$

1.3.1 Matrix Norms

We can easily generalise vector norms to matrix norms acting on the vector space of all matrices $A \in \mathbb{C}^{m\times n}$. There is a special norm that is sometimes more useful than the general matrix norms, which is defined by viewing matrix as a linear operator from $\mathbb{C}^n$ to $\mathbb{C}^m$:

Definition 1.3.2. Given $A \in \mathbb{C}^{m\times n}$, the **induced matrix norm** $\|A\|$ is defined as

$$\|A\| = \sup_{x\in\mathbb{C}^n, x\neq \mathbf{0}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\|.$$

Equivalently, $\|A\|$ is the smallest number $C \geq 0$ such that the inequality

$$\|Ax\| \leq C\|x\| \quad \text{holds for all } x \in \mathbb{C}^n.$$

Geometrically, $\|A\|$ is the maximum factor by which A can "stretch" a vector x.

Example 1.3.3. Let $D \in \mathbb{C}^{m\times m}$ be a diagonal matrix

$$D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix}.$$

To find the $\|D\|_2$ geometrically, observe that the image of the 2-norm unit sphere under D is an m-dimensional ellipse whose semiaxis lengths are given by the numbers $|d_j|$. The unit vectors amplified most by D are those that are mapped to the longest semiaxis of the ellipse, of length $\max\{|d_j|\}$. Thus, we have that

$$\|D\|_2 = \max_{1\leq j\leq m} |d_j|.$$

This result for the 2-norm generalises to any $p \geq 1$: if D is diagonal, then

$$\|D\|_p = \max_{1\leq j\leq m} |d_j|.$$

We can prove this algebraically. First,

$$\|Dx\|_p^p = \sum_{j=1}^{m} |x_j d_j|^p \leq \max_{1\leq j\leq m} |d_j|^p \sum_{j=1}^{m} |x_j|^p = \left(\max_{1\leq j\leq m} |d_j|^p\right) \|x\|_p^p.$$

Taking the pth root of each side, and then the supremum over all $x \in \mathbb{C}^m$ with $\|x\|_p = 1$ yields the upper bound

$$\|D\|_p \leq \max_{1\leq j\leq m} |d_j|.$$

To obtain $\|D\|_p \geq \max_{1\leq j\leq m} |d_j|$, we choose the standard basis vector $x = e_k$, where k is such that $|d_k|$ is the largest diagonal entry. Note that $\|e_k\|_p = 1$ and

$$\|D\|_p \leq \frac{\|De_k\|_p}{\|e_k\|_p} = \|De_k\|_p = \|d_k e_k\|_p = |d_k| = \max_{1\leq j\leq m} |d_j|.$$

Lemma 1.3.4. *For any $A \in \mathbb{C}^{m\times n}$, the induced matrix 1-norm and ∞-norm are equal to the "maximum column sum" and "maximum row sum" of A respectively, i.e.*

$$\|A\|_1 = \max_{1\leq j\leq n} \|a_j\|_1.$$
$$\|A\|_\infty = \max_{1\leq i\leq m} \|a_i^*\|_1.$$

Proof. Let $\{a_1, \ldots, a_n\}$ be columns of A. Viewing Ax as linear combinations of $\{a_1, \ldots, a_n\}$ gives

$$\|Ax\|_1 = \left\| \sum_{j=1}^n a_j x_j \right\|_1 \leq \sum_{j=1}^n |x_j| \|a_j\|_1 \leq \max_{1\leq j\leq n} \|a_j\|_1 \sum_{j=1}^n |x_j|$$
$$= \left(\max_{1\leq j\leq n} \|a_j\|_1 \right) \|x\|_1.$$

Taking supremum over all $x \in \mathbb{C}^n$ with $\|x\|_1 = 1$, we have that

$$\|A\|_1 \leq \max_{1\leq j\leq n} \|a_j\|_1.$$

To obtain $\|A\|_1 \geq \max_{1\leq j\leq n} \|a_j\|_1$, we choose the standard basis vector $x = e_k$, where k is such that $\|a_k\|_1$ is maximum. Note that $\|e_k\|_1 = 1$ and

$$\|A\|_1 \geq \frac{\|Ae_k\|_1}{\|e_k\|_1} = \|Ae_k\|_1 = \|a_k\|_1 = \max_{1\leq j\leq n} \|a_j\|_1.$$

For the induced ∞-norm, we first write $Ax = b = (b_1, b_2, \ldots, b_m)^* \in \mathbb{C}^m$. Using the definition of a matrix-vector product, we have that for any $i = 1, \ldots, m$

$$|b_i| = \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \sum_{j=1}^n |a_{ij}||x_j| \leq \max_{1\leq j\leq n} |x_j| \left(\sum_{j=1}^n |a_{ij}| \right)$$
$$= \|x\|_\infty \left(\sum_{j=1}^n |a_{ij}| \right)$$

Taking supremum over all $i = 1, \ldots, m$, we obtain

$$\max_{1\leq i\leq m} |b_i| = \|b\|_\infty = \|Ax\|_\infty \leq \left(\max_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}| \right) \|x\|_\infty.$$

Taking supremum over all $x \in \mathbb{C}^n$ of norm 1, we have

$$\|A\|_\infty \leq \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

To obtain $\|A\|_\infty \geq \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$, choose $x = (1, \ldots, 1)^* \in \mathbb{C}^n$. Note that $\|x\|_\infty = 1$ and

$$\|A\|_\infty \geq \frac{\|Ax\|_\infty}{\|x\|_\infty} = \|Ax\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

■

1.3.2 Cauchy-Schwarz and Holder Inequalities

Inner products can be bounded in terms of p-norms using **Hölder's inequality** and **Cauchy-Schwarz inequality**, the latter being the special case of Hölder's inequality. Note that these two inequalities are tight in the sense that these inequalities become equalities for certain choices of vectors.

Theorem 1.3.5 (Young's inequality)**.** *Let $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. For any two nonnegative real numbers a, b, we have*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{Young}$$

Proof. Observe that the inequality is trivial if either a or b are zero, so suppose both a and b are any positive real numbers. Choose any $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, the constraint on p and q suggests a possible convexity argument. Indeed, using the fact that exponential function is a convex function, we have that

$$\begin{aligned} ab = \exp(\ln(ab)) &= \exp(\ln(a) + \ln(b)) \\ &= \exp\left(\frac{p}{p}\ln(a) + \frac{q}{q}\ln(b)\right) \\ &\leq \frac{1}{p}\exp(p\ln(a)) + \frac{1}{q}\exp(q\ln(b)) \\ &= \frac{a^p}{p} + \frac{b^q}{q}. \end{aligned}$$

Since $p, q > 1$ were arbitrary numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$, this proves the Young's inequality. ■

Theorem 1.3.6 (Hölder's inequality)**.** *Let $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. For any $u, v \in \mathbb{C}^n$, we have*

$$|u^*v| = \left|\sum_{j=1}^{n} u_j^* v_j\right| \leq \|u\|_p \|v\|_q. \tag{Hölder}$$

Proof. Observe that the inequality is trivial if either u or v are the zero vector, so suppose $u, v \neq \mathbf{0}$. Choose any $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Young's inequality (Young) yields

$$|a_j^* b_j| = |a_j||b_j| \leq \frac{|a_j|^p}{p} + \frac{|b_j|^q}{q}.$$

Summing over $j = 1, \ldots, n$, we get

$$\sum_{j=1}^{n} |a_j^* b_j| \leq \frac{1}{p}\left(\sum_{j=1}^{n} |a_j|^p\right) + \frac{1}{q}\left(\sum_{j=1}^{n} |b_j|^q\right).$$

In particular, for any $a = (a_1, \ldots, a_n)^*, b = (b_1, \ldots, b_n)^* \in \mathbb{C}^n$ satisfying

$$\|a\|_p^p = \sum_{j=1}^{n} |a_j|^p = 1 = \sum_{j=1}^{n} |b_j|^q = \|b\|_q^q, \tag{1.3.1}$$

we have

$$\sum_{j=1}^{n} |a_j^* b_j| \leq \frac{1}{p} + \frac{1}{q} = 1. \tag{1.3.2}$$

Now, for any nonzero $u = (u_1, \ldots, u_n)^*, v = (v_1, \ldots, v_n)^* \in \mathbb{C}^n$, define vectors $a = (\tilde{a}_1, \ldots, \tilde{a}_n)^*, b = (\tilde{b}_1, \ldots, \tilde{b}_n)^*$ such that

$$\tilde{a}_j = \frac{u_j}{\|u\|_p}, \quad \tilde{b}_j = \frac{v_j}{\|v\|_q} \quad \text{for all } j = 1, \ldots, n.$$

By construction, both a and b satisfy (1.3.1) and substituting $\tilde{a}_j, \tilde{b}_j$ into (1.3.2) yields

$$\frac{1}{\|u\|_p \|v\|_q} \sum_{j=1}^{n} |u_j^* v_j| \leq 1 \implies \left|\sum_{j=1}^{n} u_j^* v_j\right| \leq \sum_{j=1}^{n} |u_j v_j| \leq \|u\|_p \|v\|_q.$$

Since u, v were arbitrary nonzero vectors in $\mathbb{C}^n$, this proves the Hölder's inequality. ∎

Example 1.3.7. Consider a matrix A containing a single row, i.e. $A = a^*$, where $a \neq \mathbf{0}$ is a fixed column vector in $\mathbb{C}^n$. For any $x \in \mathbb{C}^n$, Cauchy-Schwarz inequality yields

$$\|Ax\|_2 = |a^* x| \leq \|a\|_2 \|x\|_2.$$

Taking supremum over all $x \in \mathbb{C}^n$ of norm 1, we get $\|A\|_2 \leq \|a\|_2$. To obtain $\|A\|_2 \geq \|a\|_2$, choose the particular $x = a$. Then

$$\|A\|_2 \geq \frac{\|Aa\|_2}{\|a\|_2} = \frac{\|a\|_2^2}{\|a\|_2} = \|a\|_2.$$

Example 1.3.8. Consider the rank-one outer product $A = uv^*$, where $u \in \mathbb{C}^m$ and $v \in \mathbb{C}^n$. For any $x \in \mathbb{C}^n$, Cauchy-Schwarz inequality yields

$$\|Ax\|_2 = \|uv^*x\|_2 = |v^*x|\|u\|_2 \leq \|u\|_2\|v\|_2\|x\|_2.$$

Taking supremum over all $x \in \mathbb{C}^n$ of norm 1, we get $\|A\|_2 \leq \|u\|_2\|v\|_2$. To obtain $\|A\|_2 \geq \|u\|_2\|v\|_2$, choose the particular $x = v$. Then

$$\|A\|_2 \geq \frac{\|Av\|_2}{\|v\|_2} = \frac{\|uv^*v\|_2}{\|v\|_2} = \frac{\|u\|_2\|v\|_2^2}{\|v\|_2} = \|u\|_2\|v\|_2.$$

Lemma 1.3.9. *Let $A \in \mathbb{C}^{m\times l}, B \in \mathbb{C}^{l\times n}$. The induced matrix norm of AB satisfies the inequality*

$$\|AB\| \leq \|A\|\|B\|$$

Consequently, the induced matrix norm of A satisfies

$$\|A^n\| \leq \|A\|^n \quad \textit{for any } n \geq 1.$$

Proof. For any $x \in \mathbb{C}^n$,

$$\|ABx\| \leq \|A\|\|Bx\| \leq \|A\|\|B\|\|x\|.$$

Taking supremum over all $x \in \mathbb{C}^n$ of norm 1 gives the desired result. ■

This does not hold for matrix norms in general. Choose

$$\|A\| = \max_{1\leq i,j\leq m} |a_{ij}| \quad \text{and} \quad A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Then $\|AB\| = 2$ but $\|A\|\|B\| = 1$. An important matrix norm which is not induced by any vector norm is the **Hilbert-Schmidt** or **Frobenius norm**, defined by

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}.$$

If $\{a_1, \ldots, a_n\}$ are the columns of A, we have

$$\|A\|_F = \left(\sum_{j=1}^{n} \|a_j\|_2^2\right)^{\frac{1}{2}}.$$

An equivalent definition of $\|\cdot\|_F$ is in terms of trace

$$\|A\|_F = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\mathrm{tr}(AA^*)}.$$

Viewing the matrix $A \in \mathbb{C}^{m\times n}$ as a vector in $\mathbb{C}^{mn}$, the Frobenius norm can be seen as the usual l^2 norm. Replacing l^2 norm with l^p norm gives rise to the **Schatten *p*-norm**.

Lemma 1.3.10. *For any $A \in \mathbb{C}^{m\times l}, B \in \mathbb{C}^{l\times n}$, the Frobenius norm of AB satisfies*

$$\|AB\|_F \leq \|A\|_F \|B\|_F.$$

Proof. Let $C = AB = (c_{ij})$, where the entries of C is given by $c_{ij} = a_i^* b_j$ with a_i^*, b_j the ith-row and jth-column of the matrix A and B respectively. Cauchy-Schwarz inequality gives

$$|c_{ij}| \leq \|a_i\|_2 \|b_j\|_2.$$

Squaring both sides and summing over all i, j, we obtain

$$\begin{aligned}\|AB\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} |c_{ij}|^2 &\leq \sum_{i=1}^{m}\sum_{j=1}^{n} (\|a_i\|_2 \|b_j\|_2)^2 \\ &= \left(\sum_{i=1}^{m} \|a_i\|_2^2\right)\left(\sum_{j=1}^{n} \|b_j\|_2^2\right) \\ &= \|A\|_F^2 \|B\|_F^2.\end{aligned}$$

■

Theorem 1.3.11. *For any $A \in \mathbb{C}^{m\times n}$ and unitary matrix $Q \in \mathbb{C}^{m\times m}, V \in \mathbb{C}^{n\times n}$, we have*

$$\|QA\|_2 = \|A\|_2 = \|AV\|_2, \quad \|QA\|_F = \|A\|_F.$$

Proof. Using the trace definition of $\|\cdot\|_F$,

$$\|QA\|_F^2 = \mathrm{tr}[(QA)^*(QA)] = \mathrm{tr}[A^*Q^*QA] = \mathrm{tr}(A^*A) = \|A\|_F^2.$$

Since the 2-norm is invariant under unitary transformation,

$$\|QA\|_2 = \sup_{x\in\mathbb{C}^n, \|x\|_2=1} \|QAx\|_2 = \sup_{x\in\mathbb{C}^n, \|x\|_2=1} \|Ax\|_2 = \|A\|_2.$$

For any $x \in \mathbb{C}^n$, let $y = Vx \in \mathbb{C}^n$. Then $x = V^*y$ and $\|x\|_2 = \|V^*y\|_2 = \|y\|_2$ since unitary transformation preserves $\|\cdot\|_2$. Consequently,

$$\|AV\|_2 = \sup_{x\in\mathbb{C}^n, \|x\|_2=1} \|AVx\|_2 = \sup_{y\in\mathbb{C}^n, \|y\|_2=1} \|Ay\|_2 = \|A\|_2.$$

■

1.4 Problems

1. Show that if a matrix A is both triangular and unitary, then it is diagonal.

Solution: The statement is trivial if $A \in \mathbb{C}^{m\times m}$ is both upper and lower triangular, so suppose A is upper-triangular. This implies that $A^* = A^{-1}$ is lower-triangular. The result then follows if we show that A^{-1} is also upper-triangular. Since $A^{-1}A = I_{m\times m}$, we have that

$$\begin{bmatrix} b_1 & \dots & b_m \end{bmatrix} \begin{bmatrix} a_1 & \dots & a_m \end{bmatrix} = \begin{bmatrix} e_1 & \dots & e_m \end{bmatrix},$$

where a_j, b_j are columns of A and A^{-1} respectively and e_j are the standard basis vectors in $\mathbb{C}^m$. Interpreting e_j as the linear combination of the columns b_j together with the assumption that A is upper-triangular, we obtain the relation

$$e_j = \sum_{i=1}^{m} a_{ij} b_i = \sum_{i=1}^{j} a_{ij} b_i \quad \text{for any } j = 1, \dots, m.$$

More precisely, we have

$$\begin{aligned} e_1 &= a_{11} b_1 \\ e_2 &= a_{12} b_1 + a_{22} b_2 \\ \vdots &= \quad \vdots \quad \vdots \\ e_m &= a_{1m} b_1 + a_{2m} b_2 + \dots + a_{mm} b_m. \end{aligned}$$

This implies that $b_{ij} = 0$ for all $i > j$, $j = 1, \dots, m$, *i.e.* A^{-1} is upper-triangular.

2. Let $A \in \mathbb{C}^{m\times m}$ be Hermitian. An eigenvector of A is a nonzero vector $x \in \mathbb{C}^m$ such that $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$, the corresponding eigenvalue.

   (a) Prove that all eigenvalues of A are real.

Solution: Let λ be any eigenvalue of a Hermitian matrix $A \in \mathbb{C}^{m\times m}$, with corresponding eigenvector $x \in \mathbb{C}^m$. Since $Ax = \lambda x$, we have

$$\begin{aligned} (\lambda x)^* x &= (Ax)^* x \\ \bar{\lambda}(x^* x) &= x^* A^* x \\ &= x^* A x \qquad \Big[A \text{ is Hermitian.}\Big] \\ &= \lambda x^* x. \end{aligned}$$

Since $x \neq 0$, $x^* x = \|x\|_2^2 \neq 0$ and we must have $\lambda = \bar{\lambda}$, *i.e.* λ is real. Since λ was arbitrary eigenvalue of A, the result follows.

(b) Prove that if x and y are eigenvectors corresponding to distinct eigenvalues, then x and y are orthogonal.

Solution: Suppose x and y are eigenvectors of a Hermitian matrix A corresponding to distinct eigenvalues λ and μ respectively, i.e.

$$Ax = \lambda x \quad \text{and} \quad Ay = \mu y.$$

Using the result that eigenvalues of a Hermitian matrix are real,

$$\lambda x^* y = (\lambda x)^* y = (Ax)^* y = x^* A^* y = x^* A y = \mu x^* y.$$

Consequently, $x^* y = 0$ since $\lambda \neq \mu$. Since λ, μ were arbitrary distinct eigenvalues of A, the result follows.

3. What can be said about the eigenvalues of a unitary matrix?

Solution: Choosea any eigenvalue λ of a unitary matrix Q with corresponding eigenvector $x \neq 0$. Since the 2-norm is invariant under unitary transformations,

$$|\lambda| \|x\|_2 = \|\lambda x\|_2 = \|Qx\|_2 = \|x\|_2 \implies |\lambda| = 1.$$

Hence, the eigenvalues of a unitary matrix must lie on the unit circle in $\mathbb{C}$.

4. If u and v are m-vectors, the matrix $A = I + uv^*$ is known as a *rank-one perturbation of the identity*. Show that if A is nonsingular, then its inverse has the form $A^{-1} = I + \alpha uv^*$ for some scalar α, and give an expression for α. For what u and v is A singular? If it is singular, what is $\mathcal{N}(A)$?

Solution: The result is trivial if either u or v is the zero vector, so suppose $u, v \neq \mathbf{0}$. Suppose A is nonsingular, with its inverse $A^{-1} = I + \alpha uv^*$, then

$$\begin{aligned} I = AA^{-1} &= (I + uv^*)(I + \alpha uv^*) \\ &= I + \alpha uv^* + uv^* + \alpha uv^* uv^* \\ &= I + uv^*(1 + \alpha + \alpha v^* u). \end{aligned}$$

Since $uv^* \neq \mathbf{0}_m$, we must have

$$1 + \alpha + \alpha v^* u = 0 \implies \alpha(1 + v^* u) = -1 \implies \alpha = -\frac{1}{1 + v^* u}.$$

Note that division by $1 + v^* u$ is allowed here, since $1 + v^* u \neq 0$ if A is nonsingular, as we shall prove now. Suppose A is singular, there exists a nonzero $x \in \mathbb{C}^m$ such that $Ax = \mathbf{0}$. In particular, we have

$$Ax = (I + uv^*)x = \mathbf{0} \implies uv^* x = -x. \tag{1.4.1}$$

For any nonzero scalars $\beta \in \mathbb{C}$, let $x = \beta u$. Substituting this into (1.4.1) yields

$$uv^*(\beta u) = -\beta u \implies \cancel{\beta}(v^* u)u = -\cancel{\beta} u \implies (v^* u)u = -u.$$

Hence, we see that if $v^*u = -1$, then A is singular and $\mathcal{N}(A) = \text{span}(u)$.

5. Let $\|\cdot\|$ denote any norm on $\mathbb{C}^m$ and also the induced matrix norm on $\mathbb{C}^{m\times m}$. Show that $\rho(A) \leq \|A\|$, where $\rho(A)$ is the *spectral radius* of A, i.e., the largest absolute value $|\lambda|$ of an eigenvalue λ of A.

Solution: Choose any eigenvalue λ of a matrix $A \in \mathbb{C}^{m\times m}$, with corresponding eigenvector $x \in \mathbb{C}^m$. Since $Ax = \lambda x$, we have

$$|\lambda|\|x\| = \|\lambda x\| = \|Ax\| \leq \|A\|\|x\|,$$

where we use the assumption that $\|A\|$ is an induced matrix norm for the inequality. Dividing each side of the inequality by $\|x\| \neq 0$ yields $|\lambda| \leq \|A\|$. The desired inequality follows from taking the supremum over all eigenvalues of A.

6. (a) Let $N(x) := \|\cdot\|$ be any vector norm on $\mathbb{C}^n$ (or $\mathbb{R}^n$). Show that $N(x)$ is a continuous function of the components $x_1, x_2, \ldots, x_n$ of x.

Solution: Consider the canonical basis $\{e_1, \ldots, e_n\}$ for $\mathbb{C}^n$. Then

$$x - y = \sum_{j=1}^{n} (x_j - y_j)e_j,$$

and

$$\Big|\|x\| - \|y\|\Big| \leq \|x - y\| \leq \sum_{j=1}^{n} |x_j - y_j|\|e_j\| \leq \|x - y\|_\infty \left(\sum_{j=1}^{n} \|e_j\|\right).$$

Continuity follows by taking $\|x - y\|_\infty \longrightarrow 0$.

(b) Prove that if $W \in \mathbb{C}^{m\times m}$ is an arbitrary nonsingular matrix, and $\|\cdot\|$ is any norm on $\mathbb{C}^m$, then $\|x\|_W = \|Wx\|$ is a norm on $\mathbb{C}^m$.

Solution: Let $W \in \mathbb{C}^{m\times m}$ be an arbitrary nonsingular matrix, and w_j the jth column of W. Let $\|\cdot\|$ be any norm on $\mathbb{C}^m$, and $x = (x_1, \ldots, x_m)^*, y = (y_1, \ldots, y_m)^* \in \mathbb{C}^m$. It is clear that $\|x\|_W = \|Wx\| \geq 0$. Suppose $\|x\|_W = 0$. Then

$$0 = \|x\|_W = \|Wx\| \implies Wx = \mathbf{0}.$$

Viewing Wx as a linear combination of columns of W, we obtain

$$\mathbf{0} = Wx = \sum_{j=1}^{m} x_j w_j.$$

Since W is nonsingular, its columns $\{w_1, w_2, \ldots, w_m\}$ form a linearly independent set of vectors in $\mathbb{C}^m$ and this implies that $x_j = 0$ for all $j = 1, \ldots, m$. Hence, $\|x\|_W = 0 \implies x = \mathbf{0}$. For any $\alpha \in \mathbb{C}$ we have that

$$\|\alpha x\|_W = \|W(\alpha x)\| = \|\alpha(Wx)\| = |\alpha|\|Wx\| = |\alpha|\|x\|_W.$$

Finally, the triangle inequality of $\|\cdot\|$ gives

$$\|x+y\|_W = \|Wx+Wy\| \leq \|Wx\| + \|Wy\| = \|x\|_W + \|y\|_W.$$

Hence, we conclude that $\|\cdot\|_W$ is a norm on $\mathbb{C}^m$.

7. (a) Explain why $\|I\| = 1$ for every induced matrix norm.

Solution: It follows directly from the definition of an induced matrix norm. More precisely,

$$\|I\| = \sup_{x\in\mathbb{C}^m, x\neq\mathbf{0}} \frac{\|Ix\|}{\|x\|} = \sup_{x\in\mathbb{C}^m, x\neq\mathbf{0}} \frac{\|x\|}{\|x\|} = 1.$$

(b) What is $\|I_{n\times n}\|_F$?

Solution: From the definition of Frobenius norm,

$$\|I_{n\times n}\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}} = \left(\sum_{j=1}^{n} |1|^2\right)^{\frac{1}{2}} = \sqrt{n}.$$

(c) Show that Frobenius norm is not induced by any vector norm.

Solution: If $\|\cdot\|_F$ were induced by any vector norm, then $\|I\|_F$ must equal to 1 from 7(a). But 7(b) shows that for every $n > 1$, $\|I\|_F = \sqrt{n} \neq 1$. Consequently, Frobenius norm is not induced by any vector norm on $\mathbb{C}^n$ for $n > 1$.

Chapter 2

Matrix Decomposition and Least Squares Problems

Matrix decomposition has been of fundamental importance in modern sciences. In the context of numerical linear algebra, matrix decomposition serves the purpose of rephrasing through a series of easier subproblems a task that may be relatively difficult to solve in its original form, for instance solving linear systems. In the context of applied statistics, matrix decomposition offers a way of obtaining some form of low-rank approximation to some large "data" matrix containing numerical observations; this is crucial in understanding the structure of the matrix, in particular exploring and identifying the relationship within data. In this chapter, we will study the *singular value decomposition (SVD)* and *QR factorisation*, and demonstrate how to solve linear least squares problems using these decompositions.

2.1 The Singular Value Decomposition

Throughout this section, we will assume that $A \in \mathbb{C}^{m\times n}, m \geq n$ has full rank for simplicity. The central theme of this section is that SVD is just another formulation of the following geometric fact in terms of linear algebra:

The image of the unit sphere under linear transformations is a hyperellipse.

It involves three geometrical transformations: rotation, reflection and scaling. Given $A \in \mathbb{C}^{m\times n}$, $m \geq n$, there exists a decomposition, called **reduced singular value decomposition**, or **reduced SVD** of the form $A = \widehat{U}\widehat{\Sigma}V^*$, where

$$\widehat{U} = \left[\begin{array}{c|c|c|c} & & & \\ u_1 & u_2 & \dots & u_n \\ & & & \end{array}\right] \in \mathbb{C}^{m\times n}.$$

$$\widehat{\Sigma} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

$$V = \left[\begin{array}{c|c|c|c} & & & \\ v_1 & v_2 & \dots & v_n \\ & & & \end{array}\right] \in \mathbb{C}^{n\times n}.$$

(a) $\{u_1, u_2, \dots, u_n\}$ and $\{v_1, v_2, \dots, v_n\}$ are the **left and right singular vectors** of A; columns of $\widehat{U}$ are orthonormal, V is unitary and $\widehat{U}^*\widehat{U} = V^*V = I_n$;

(b) $\{\sigma_j\}_{j=1}^n$ are **singular values** of A, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. These are the lengths of the n *principal semiaxes* of the hyperellipse in the case of real matrices A.

(c) These singular vectors and singular values satisfy the relation

$$Av_j = \sigma_j u_j, \quad j = 1, \dots, n. \tag{2.1.1}$$

Example 2.1.1. Consider any matrix $A \in \mathbb{C}^{2\times 2}$. It is clear that $H = A^*A \in \mathbb{C}^{2\times 2}$ is Hermitian. Moreover, for any $x \in \mathbb{C}^2$ we have

$$x^*Hx = x^*(A^*A)x = (Ax)^*(Ax) = \|Ax\|_2^2 \geq 0.$$

Consequently, H has nonnegative eigenvalues λ_1, λ_2 and $H = VDV^* = V\Sigma^2V^*$, where V is unitary and

$$D = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix},$$

such that $\sigma_1 \geq \sigma_2 \geq 0$ and $\sigma_1^2 = \lambda_1, \sigma_2^2 = \lambda_2$. Assume $\sigma_1 \geq \sigma_2 > 0$, we claim that $U = AV\Sigma^{-1}$. Indeed,

$$\begin{aligned} U^*U = (\Sigma^{-1}V^*A^*)(AV\Sigma^{-1}) &= \Sigma^{-1}V^*HV\Sigma^{-1} \\ &= \Sigma^{-1}\Sigma^2\Sigma^{-1} = I_2. \end{aligned}$$

Hence, $AV = U\Sigma \implies A = U\Sigma V^*$.

2.1.1 Geometric Intepretation

[**Include diagrams**] Let S be the unit circle in $\mathbb{R}^n$, then any $x \in S$ can be written as

$$x = V\hat{x} = \begin{bmatrix}v_1 & v_2\end{bmatrix}\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}.$$

Observe that $\|\hat{x}\|_2 = 1$ since $\cos^2\theta + \sin^2\theta = 1$. It follows from $AV = U\Sigma$ that

$$Ax = A(v_1\cos\theta + v_2\sin\theta) = \sigma_1 u_1\cos\theta + \sigma_2 u_2\sin\theta.$$

In terms of (v_1, v_2) coordinates, the vector $(\cos\theta, \sin\theta)$ gets mapped onto $(z_1, z_2) = (\sigma_1\cos\theta, \sigma_2\sin\theta)$ in (u_1, u_2) coordinates. Moreover,

$$\left(\frac{z_1}{\sigma_1}\right)^2 + \left(\frac{z_2}{\sigma_2}\right)^2 = \cos^2\theta + \sin^2\theta = 1,$$

i.e. S is being transformed to ellipse. We claim that $\|A\|_2 = \sigma_1$. On one hand, we obtain using orthonormality of $\{u_1, u_2\}$

$$\begin{aligned}\|Ax\|_2^2 &= (\sigma_1 u_1 \cos\theta + \sigma_2 u_2 \sin\theta)^*(\sigma_1 u_1 \cos\theta + \sigma_2 u_2 \sin\theta) \\ &= \sigma_1^2 \cos^2\theta u_1^* u_1 + \sigma_2^2 \sin^2\theta u_2^* u_2 \\ &\leq \sigma_1^2 \cos^2\theta + \sigma_1^2 \sin^2\theta = \sigma_1^2.\end{aligned}$$

On the other hand, choosing $x = v_1$ gives $\|Av_1\|_2^2 = \|\sigma_1 u_1\|_2^2 = \sigma_1^2$.

We see that the image of unit circle under A is an ellipse in the 2-dimensional subspace of $\mathbb{R}^m$ defined by $\text{span}\{u_1, u_2\}$. If $A \in \mathbb{R}^{m\times n}$ is of full rank with $m \geq n$, then the image of the unit sphere in $\mathbb{R}^n$ under A is a hyperellipsoid in $\mathbb{R}^m$.

2.1.2 Full SVD

From the *reduced* SVD of A, columns of $\widehat{U}$ form an orthonormal set in $\mathbb{C}^m$, but $\widehat{U}$ is not unitary. By adjoining an additional $m-n$ orthonormal columns, $\widehat{U}$ can be extended to a unitary matrix $U \in \mathbb{C}^{m\times m}$. Consequently, we must concatenated $\widehat{\Sigma}$ together with an additional $m-n$ rows of zero vector so that the product remains unchanged upon replacing $\widehat{U}$ by U. This process yields the **full SVD** of $A = U\Sigma V^*$, where

$$U = \left[\begin{array}{c:c|c|c} & & & \\ \widehat{U} & u_{n+1} & \cdots & u_m \\ & & & \end{array}\right] \in \mathbb{C}^{m\times m}.$$

$$\Sigma = \left[\begin{array}{c} \widehat{\Sigma} \\ \hdashline \mathbf{0}^* \\ \hline \vdots \\ \hline \mathbf{0}^* \end{array}\right] \in \mathbb{R}^{m\times n}$$

$$V = \left[\begin{array}{c|c|c|c} & & & \\ v_1 & v_2 & \cdots & v_n \\ & & & \end{array}\right] \in \mathbb{C}^{n\times n}.$$

Note that in full SVD form, Σ has the same size as A, and U, V are unitary matrices.

Theorem 2.1.2. *Every matrix $A \in \mathbb{C}^{m\times n}$ can be factored as $A = U\Sigma V^*$, where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary and $\Sigma \in \mathbb{R}^{m\times n}$ is a rectangular matrix whose only nonzero entries are nonnegative entries on its diagonal.*

Proof. The statement is trivial if A is the zero matrix, so assume $A \neq \mathbf{0}$. Let $\sigma_1 = \|A\|_2 > 0$, there exists $v_1 \in \mathbb{C}^n$ such that $\|v_1\|_2 = 1$ and $\|Av_1\|_2 = \|A\|_2 = \sigma_1$. Such v_1 exists since the induced matrix norm is by definition a minimisation problem of a continuous functional (in this case the norm) over a compact nonempty subset of $\mathbb{C}^n$. Define $u_1 = Av_1/\sigma_1 \in \mathbb{C}^m$, clearly $u_1 \neq \mathbf{0}$ and $\|u_1\|_2 = 1$ by construction.

Consider any extension of u_1, v_1 to an orthonormal basis $\{u_1, u_2, \ldots, u_m\}, \{v_1, v_2, \ldots, v_n\}$ of $\mathbb{C}^m, \mathbb{C}^n$ respectively. Construct the following unitary matrices

$$\widehat{U}_1 = \begin{bmatrix} u_2 & \ldots & u_m \end{bmatrix} \in \mathbb{C}^{m\times(m-1)}, \quad \widehat{V}_1 = \begin{bmatrix} v_2 & \ldots & v_n \end{bmatrix} \in \mathbb{C}^{n\times(n-1)}.$$

and define two unitary matrices

$$U_1 = \begin{bmatrix} u_1 & \widehat{U}_1 \end{bmatrix} \in \mathbb{C}^{m\times m}, \quad V_1 = \begin{bmatrix} v_1 & \widehat{V}_1 \end{bmatrix} \in \mathbb{C}^{n\times n}.$$

We then have:

$$\begin{aligned} A_1 \coloneqq U_1^* A V_1 = \begin{bmatrix} u_1^* \\ \widehat{U}_1^* \end{bmatrix} A \begin{bmatrix} v_1 & \widehat{V}_1 \end{bmatrix} &= \begin{bmatrix} u_1^* A v_1 & u_1^* A \widehat{V}_1 \\ \widehat{U}_1^* A v_1 & \widehat{U}_1^* A \widehat{V}_1 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 u_1^* u_1 & w^* \\ \sigma_1 \widehat{U}_1^* u_1 & \widehat{A} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & w^* \\ \mathbf{0} & \widehat{A} \end{bmatrix}, \end{aligned}$$

where $w \in \mathbb{C}^{(n-1)}$ and $\widehat{A} \in \mathbb{C}^{(m-1)\times(n-1)}$. We claim that $w = \mathbf{0}$. The first thing is to observe that

$$\begin{aligned} \left\| A_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2 &= \left\| \begin{bmatrix} \sigma_1^2 + w^* w \\ \widehat{A} w \end{bmatrix} \right\|_2^2 \\ &= (\sigma_1^2 + w^* w)^2 + \|\widehat{A} w\|_2^2 \\ &\geq (\sigma_1^2 + w^* w)^2 \\ &= (\sigma_1^2 + w^* w) \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2 \end{aligned}$$

Since $\|A_1\|_2 = \|U_1^* A V_1\|_2 = \|A\|_2 = \sigma_1$, we have

$$0 \leq (\sigma_1^2 + w^* w) \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2 \leq \left\| A_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2 \leq \sigma_1^2 \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2,$$

which implies that

$$0 \leq \sigma_1^2 + w^* w \leq \sigma_1^2,$$

i.e. $w^* w = 0 \implies w = \mathbf{0}$.

We now proceed by induction. The result is trivial if $m = 1$ or $n = 1$. Suppose $\widehat{A}$ has an SVD

$$U_2^* \widehat{A} V_2 = \Sigma_2 \in \mathbb{R}^{(m-1)\times(n-1)},$$

where $U_2 \in \mathbb{C}^{(m-1)\times(m-1)}, V_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ are unitary. Observe that

$$\begin{aligned} U_1^* A V_1 = \begin{bmatrix} \sigma_1 & \mathbf{0}^* \\ \mathbf{0} & \widehat{A} \end{bmatrix} &= \begin{bmatrix} \sigma_1 & \mathbf{0}^* \\ \mathbf{0} & U_2 \Sigma_2 V_2^* \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & \mathbf{0}^* \\ \mathbf{0} & \Sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & V_2^* \end{bmatrix}. \end{aligned}$$

Consequently, the unitary matrices U, V are naturally defined as

$$U = U_1 \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U_2 \end{bmatrix} = \begin{bmatrix} u_1 & \widehat{U}_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ \widehat{U}_1 U_2 \end{bmatrix} \in \mathbb{C}^{m \times m}$$
$$V = V_1 \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & V_2 \end{bmatrix} = \begin{bmatrix} v_1 & \widehat{V}_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & V_2 \end{bmatrix} = \begin{bmatrix} v_1 & \widehat{V}_1 V_2 \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

Since product of unitary matrices are unitary, we only need to show that the vector u_1 is orthogonal to each column $u_2, \ldots, u_m$ of the matrix $\widehat{U}_1 U_2$, but this must be true since $\{u_1, u_2, \ldots, u_m\}$ is an orthonormal basis by construction. A similar argument shows that V is also unitary. ■

Remark 2.1.3. In the case $m \leq n$, we simply consider the SVD of its conjugate tranpose A^*. If A is singular with rank $r < \min\{m, n\}$, the full SVD is still appropriate. What changes is that not n; but only r of the left singular vectors u_j are determined by the geometry of the hyperellipse. To construct the unitary matrix U and V, we introduce an additional $(m - r)$ and $(n - r)$ arbitrary orthonormal columns respectively.

It is well known that a nondefective square matrix can be expressed as a diagonal matrix Λ of eigenvalues, if the range and domain are represented in a basis of eigenvectors. SVD gene ralises this fact to any matrix $A \in \mathbb{C}^{m \times n}$, in that SVD allows us to say that A reduces to diagonal matrix Σ when the range is expressed in the basis of columns of U and the domain is expressed in the basis of columns of V. More precisely, any $b \in \mathbb{C}^m$ can be expanded in the basis of columns $\{u_1, \ldots, u_m\}$ of U and any $x \in \mathbb{C}^n$ can be expanded in the basis of columns $\{v_1, \ldots, v_n\}$ of V. The coordinate vectors for these expansions are

$$b = Ub' \iff b' = U^* b \quad \text{and} \quad x = Vx' \iff x' = V^* x.$$

Hence,

$$b = Ax \iff U^* b = U^* Ax = U^* U \Sigma V^* x = \Sigma V^* x \iff b' = \Sigma x'.$$

There are fundamental differences between the SVD and the eigenvalue decomposition.

(a) SVD uses two different bases (the sets of left and right singular vectors), whereas the eigenvalue decomposition uses just one (the eigenvectors).

(b) SVD uses orthonormal bases, whereas the eigenvalue decomposition uses a basis that generally is not orthogonal.

(c) Not all matrices (even square ones) have an eigenvalue decomposition, but all matrices (even rectangular ones) have a SVD.

(d) In practice, eigenvalues tend to be relevant to problems involving the behaviour of iterated forms of A, such as matrix powers A^n or matrix exponentials e^{tA}, whereas singular vectors tend to be relevant to problems involving the behaviour of A itself, or its inverse.

2.1.3 Matrix Properties via SVD

For the following discussion, we assume that $A \in \mathbb{C}^{m\times n}$ and denote $p = \min\{m, n\}$ and $r \leq p$ the number of nonzero singular values of A.

Theorem 2.1.4. *The rank of A is r, the number of nonzero singular values. Moreover,*

$$\mathcal{R}(A) = span\{u_1, \ldots, u_r\} \quad and \quad \mathcal{N}(A) = span\{v_{r+1}, \ldots, v_n\}.$$

Proof. Since U, V are unitary, they have full rank. Thus, $\text{rank}(A) = \text{rank}(\Sigma) =$ numbers of its nonzero entries. For any $x \in \mathbb{C}^n$, we have $Ax = U\Sigma V^* x = U\Sigma y$, where $y \in \mathbb{C}^n$ is arbitrary. The $\mathcal{R}(A)$ is then deduced from the fact that $\mathcal{R}(\Sigma) = \text{span}\{e_1, \ldots, e_r\}$. To find the nullspace of A, expanding $Az = 0$ yields

$$Az = U\Sigma V^* z = \mathbf{0} \implies \Sigma V^* z = \mathbf{0} \quad \text{since } U \text{ is of full rank,}$$

from which we deduce that $\mathcal{N}(A)$ is the span of $\{v_{r+1}, \ldots, v_n\}$ since $\mathcal{N}(\Sigma) = \text{span}\{e_{r+1}, \ldots, e_n\}$. ■

Theorem 2.1.5. $\|A\|_2 = \sigma_1$ *and* $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_r^2}$.

Proof. Since $\|\cdot\|_2$ and $\|\cdot\|_F$ are both invariant under unitary transformation, we have that

$$\|A\|_2 = \|U\Sigma V^*\|_2 = \|\Sigma\|_2 = \max_{1\leq j\leq p} |\sigma_j| = \sigma_1,$$

and

$$\|A\|_F = \|U\Sigma V^*\|_F = \|\Sigma\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_r^2}.$$

■

Theorem 2.1.6. *The nonzero singular values of A are the square roots of the nonzero eigenvalues of A^*A or AA^*. (These matrices have the same nonzero eigenvalues.)*

Proof. Observe that $A^*A \in \mathbb{C}^{n\times n}$ is similar to $\Sigma^*\Sigma$ since

$$A^*A = (U\Sigma V^*)^*(U\Sigma V^*) = V\Sigma^T U^* U\Sigma V^* = V(\Sigma^T\Sigma)V^*,$$

and hence has the same n eigenvalues. $\Sigma^*\Sigma$ is a diagonal matrix with p eigenvalues $\sigma_1^2, \ldots, \sigma_p^2$ and $n-p$ additional zero eigenvalues if $n > p$. A similar calculation applies to the m eigenvalues of AA^*.

■

Theorem 2.1.7. *If $A = A^*$, then the singular values of A are the absolute values of the eigenvalues of A.*

Proof. Since A is Hermitian, it has an eigendecomposition of the form $A = Q\Lambda Q^*$ for some unitary matrix Q and real diagonal matrix Λ consisting of eigenvalues λ_j of A. We rewrite it as

$$A = Q\Lambda Q^* = Q|\Lambda|\text{sign}(\Lambda)Q^*,$$

where $|\Lambda|$ and $\text{sign}(\Lambda)$ denote the diagonal matrices whose entries are $|\lambda_j|$ and $\text{sign}(\lambda_j)$ respectively. Since $\text{sign}(\Lambda)Q^*$ is unitary whenever Q is unitary, the expression above is an SVD of A, with the singular values equal to the diagonal entries of $|\Lambda|$. These can be put into nonincreasing order by inserting suitable permutation matrices as factors in Q and $\text{sign}(\Lambda)Q^*$ if required. ■

Theorem 2.1.8. *For $A \in \mathbb{C}^{m\times m}$, $|\det(A)| = \prod_{i=1}^{m} \sigma_i$.*

Proof. Using the fact that unitary matrices have determinant ± 1, we obtain

$$|\det(A)| = |\det(U)||\det(\Sigma)||\det(V^*)| = |\det(\Sigma)| = \prod_{j=1}^{n} \sigma_j.$$

■

2.1.4 Low-Rank Approximations

Low-rank approximation has been applied in a wide variety of areas such as dimension reduction, signal processing, classification and clustering. The basic problem is as follows: Given a data matrix A, we want to identify the "best" way of approximating A with matrices having rank less than ν for some ν . This constrained optimisation problem can be solved analytically using SVD. Essentially, the idea is to consider the **outer-product representation** of A, given by

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^*, \tag{2.1.2}$$

which can be deduced from the SVD of A by writing Σ as a sum of r matrices

$$\Sigma_j = \text{diag}(0, \ldots, 0, \sigma_j, 0, \ldots, 0).$$

There are many ways to decompose A into rank-one matrices, but (2.1.2) has a deeper property: its vth partial sum captures as much of the energy of A as possible, in the sense of the 2-norm of the Frobenius norm.

Theorem 2.1.9. *For any ν with $0 \le \nu \le r$, define*

$$A_\nu = \sum_{j=1}^{\nu} \sigma_j u_j v_j^*;$$

if $\nu = p = \min\{m, n\}$, define $\sigma_{\nu+1} = 0$. Then

$$\|A - A_\nu\|_2 = \inf_{\substack{B\in\mathbb{C}^{m\times n} \\ \text{rank}(B)\le\nu}} \|A - B\|_2 = \sigma_{\nu+1}.$$

$$\|A - A_\nu\|_F = \inf_{\substack{B\in\mathbb{C}^{m\times n} \\ \text{rank}(B)\le\nu}} \|A - B\|_F = \sqrt{\sigma_{\nu+1}^2 + \ldots + \sigma_r^2}.$$

Proof. Suppose there is a matrix B with $\text{rank}(B) \le \nu$ such that

$$\|A - B\|_2 < \|A - A_\nu\|_2 = \sigma_{\nu+1}.$$

There is an $(n - \nu)$-dimensional subspace $W \subset \mathbb{C}^n$ such that $B(W) = \mathbf{0}$. For any $w \in W$ we have

$$\|Aw\|_2 = \|(A - B)w\|_2 \le \|A - B\|_2\|w\|_2 < \sigma_{\nu+1}\|w\|_2.$$

On the other hand, for any $z \in \text{span}\{v_1, \ldots, v_{\nu+1}\} \coloneqq Z \subset \mathbb{C}^n$ we have

$$\begin{aligned}
\|Az\|_2^2 = \left\| A\left(\sum_{j=1}^{\nu+1} \alpha_j v_j\right)\right\|_2^2 &= \left\|\sum_{j=1}^{\nu} \alpha_j \sigma_j u_j\right\|_2^2 && \Big[\text{Since } Av_j = \sigma_j u_j.\Big] \\
&= \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \ldots + \alpha_{\nu+1}^2\sigma_{\nu+1}^2 && \Big[\text{From Pythagorean theorem.}\Big] \\
&\geq \sigma_{\nu+1}^2[\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_{\nu+1}^2] \\
&= \sigma_{\nu+1}^2\|z\|_2^2,
\end{aligned}$$

Since W, Z are subspaces of $\mathbb{C}^n$,

$$\dim(W + Z) = \dim(W) + \dim(Z) - \dim(W \cap Z),$$

and since the sum of the dimensions of W and Z exceeds n, there must be a nonzero vector in $W \cap Z$ and we arrive at a contradiction.

■

The MATLAB command for computing the reduced and full SVD is [U,S,V] = svd(A,0) and [U,S,V] = svd(A) respectively.

2.2 Projectors

Projection is an important concept in designing algorithms for certain linear algebra problems. Geometrically, projection is a generalisation of *graphical projection*. In functional analysis, a **projection** P is a bounded linear operator such that $P^2 = P$; in finite-dimensional vector space, P is a square matrix in $\mathbb{C}^{n\times n}$ and it is said to be **idempotent**. Observe that if $y \in \mathcal{R}(P)$, then $y = Px$ for some $x \in \mathbb{C}^n$ and

$$Py = PPx = P^2x = Px = y.$$

What if $y \neq Py$? For any particular $y \in \mathbb{C}^n$, consider the vector y to Py, $Py - y$. Applying the projector to $Py - y$ gives

$$P(Py - y) = P^2y - Py = Py - Py = \mathbf{0}.$$

i.e. $Py - y \in \mathcal{N}(P)$. Geometrically, this means that P projects onto $\mathcal{R}(P)$ along $\mathcal{N}(P)$.

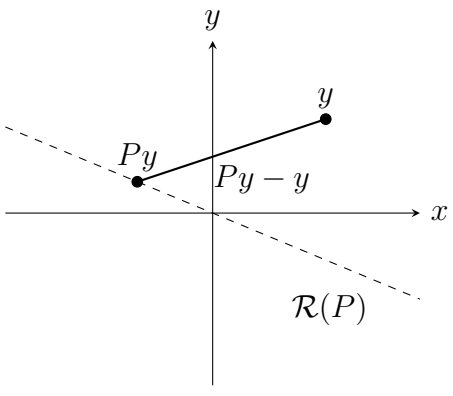

Figure 2.1: An oblique (non-orthogonal) projection.

Example 2.2.1. Consider the matrix $P = uv^*$, where $u, v \in \mathbb{C}^n$ such that $v^*u = 1$. P is a projector since $P^2 = uv^*uv^* = uv^* = P$. Note that $\mathcal{N}(P) = \{x \in \mathbb{C}^n : v^*x = 0\}$ and $\mathcal{R}(P) = \text{span}\{u\}$. We verify that $Px - x \in \mathcal{N}(P)$ for any $x \in \mathbb{C}^n$.

$$v^*(Px - x) = v^*(uv^*x) - v^*x = v^*x(v^*u - 1) = 0.$$

2.2.1 Complementary Projectors

Theorem 2.2.2. *Let $P \in \mathbb{C}^{n\times n}$ be a projector and consider the matrix $Q = I - P$.*

*(a) Q is also a projector, and we called Q the **complementary projector** to P.*

(b) $PQ = P(I - P) = \mathbf{0}$.

(c) $\mathcal{N}(P) = \mathcal{R}(Q)$ and $\mathcal{R}(P) = \mathcal{N}(Q)$.

Proof. Expanding Q^2 gives

$$Q^2 = (I - P)^2 = I^2 - 2P + P^2 = I - 2P + P = I - P = Q.$$

For the second result, $P(I - P) = P - P^2 = \mathbf{0}$. Suppose $x \in \mathcal{N}(P)$, then

$$Px = \mathbf{0} \implies Qx = x - Px = x \in \mathcal{R}(Q) \implies \mathcal{N}(P) \subset \mathcal{R}(Q).$$

Suppose $y \in \mathcal{R}(Q)$,

$$y = Qy = y - Py \implies Py = \mathbf{0} \implies y \in \mathcal{N}(P) \implies \mathcal{R}(Q) \subset \mathcal{N}(P).$$

Combining these two set inequalities show the first equation in (c). The second equation in (c) now follows from applying the previous result to $I - P$:

$$\mathcal{N}(Q) = \mathcal{N}(I - P) = \mathcal{R}(I - (I - P)) = \mathcal{R}(P).$$

■

Theorem 2.2.2 actually shows that a projector decomposes $\mathbb{C}^n$ into subspaces $\mathcal{R}(P)$ and $\mathcal{N}(P)$ such that $\mathbb{C}^n = \mathcal{R}(P) \oplus \mathcal{N}(P)$. Such a pair are said to be **complementary subspaces**. Indeed, suppose $x = Px + z$, then

$$z = x - Px = Q(x) \in \mathcal{R}(Q) = \mathcal{N}(P), \quad i.e.\ \mathbb{C}^n = \mathcal{R}(P) + \mathcal{N}(P).$$

To see that $\mathcal{R}(P) \cap \mathcal{N}(P) = \{\mathbf{0}\}$, note that any $v \in \mathcal{R}(P) \cap \mathcal{N}(P)$ satisfies

$$v = v - Pv = (I - P)v = \mathbf{0},$$

since $\mathcal{R}(P) = \mathcal{N}(I - P)$. Conversely, for any pair of complementary subspaces S_1, S_2 of $\mathbb{C}^n$, there exists a projector $P \in \mathbb{C}^{n\times n}$ such that $S_1 = \mathcal{R}(P)$ and $S_2 = \mathcal{N}(P)$. We then say that P is a projector onto S_1 along S_2.

2.2.2 Orthogonal Projectors

In general, $S_1 = \mathcal{R}(P)$ and $S_2 = \mathcal{N}(P)$ might not be orthogonal, *i.e.* there exists $x_1 = Px \in \mathcal{R}(P), x_2 = (I - P)y \in \mathcal{N}(P)$ such that

$$x_1^* x_2 = (Px)^*(I - P)y = x^* P^*(I - P)y \neq 0.$$

With this in mind, a projector $P \in \mathbb{C}^{n\times n}$ is an **orthogonal projector** if $P^*(I - P) = \mathbf{0}$; otherwise it is an **oblique** projector. Geometrically, orthogonal projectors P projects any given vector x orthogonally onto $\mathcal{R}(P)$ along $\mathcal{N}(P)$, *i.e.* $\mathcal{R}(P) \perp \mathcal{N}(P)$. Orthogonal projectors are not to be confused with orthogonal matrices! Surprisingly, orthogonal projectors have a rather simple characterisation, which is the result of the next theorem.

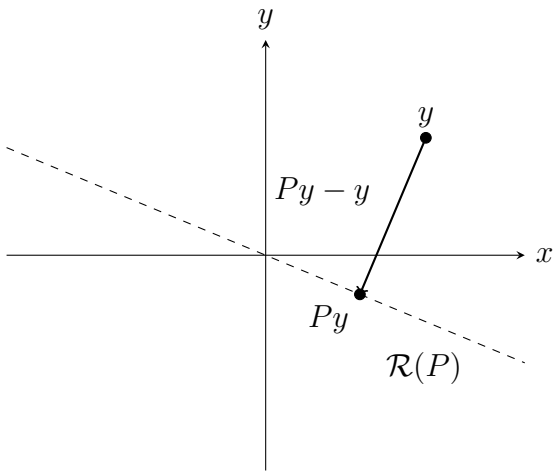

Figure 2.2: An orthogonal projection.

Theorem 2.2.3. *A projector $P \in \mathbb{C}^{n\times n}$ is orthogonal $\iff$ P is Hermitian, that is, $P = P^*$.*

Proof. If $P = P^*$, then

$$P^*(I - P) = P(I - P) = P - P^2 = \mathbf{0},$$

and it follows from the algebraic definition that P is orthogonal. Conversely, suppose P is an orthogonal projector, then

$$P^*(I - P) = \mathbf{0}, \quad \text{or} \quad P^* = P^*P.$$

Consider the minimal rank SVD of $P = U_r \Sigma_r V_r^*$, where $r \leq n$ is the rank of P, $U_r^* U_r = I_r = V_r^* V_r$ and Σ_r is nonsingular. Substituting the SVD of P into $P^* = P^*P$ yields

$$V_r \Sigma_r U_r^* = P^* = P^*P = (V_r \Sigma_r U_r^*)(U_r \Sigma_r V_r^*) = V_r \Sigma_r^2 V_r^*,$$

and left multiplying both sides by $\Sigma_r^{-1} V_r^*$ gives $U_r^* = \Sigma_r V_r^*$. Hence,

$$P = U_r(\Sigma_r V_r^*) = U_r U_r^* \quad \text{and} \quad P^* = (U_r U_r^*)^* = U_r U_r^* = P.$$

■

We demonstrate how to construct a full SVD of an orthogonal projector P. Let $\{q_1, \ldots, q_r\}$ be a basis for $\mathcal{R}(P)$ and $\{q_{r+1}, \ldots, q_n\}$ a basis for $\mathcal{N}(P)$. Define a unitary matrix Q with columns $\{q_1, \ldots, q_n\}$. Since

$$Pq_j = \begin{cases} q_j & \text{if } j = 1, \ldots, r \\ \mathbf{0} & \text{if } j = r+1, \ldots, n. \end{cases}$$

we obtain

$$\begin{aligned} PQ &= \begin{bmatrix} q_1 & \ldots & q_r & \mathbf{0} & \ldots & \mathbf{0} \end{bmatrix} \\ \Longrightarrow \quad Q^*PQ &= \begin{bmatrix} q_1^* \\ \vdots \\ q_n^* \end{bmatrix} \begin{bmatrix} q_1 & \ldots & q_r & \mathbf{0} & \ldots & \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} I_r & \\ & \mathbf{0}_{n-r} \end{bmatrix} = \Sigma. \end{aligned}$$

Consequently, the singular values of orthogonal projectors consists of 1's and 0's. Because some singular values are zero, it is advantageous to drop the columns $\{q_{r+1}, \ldots, q_n\}$ of Q which leads to

$$P = \widehat{Q}\widehat{Q}^*, \quad \text{where} \quad \widehat{Q} = \begin{bmatrix} q_1 & \ldots & q_r \end{bmatrix} \in \mathbb{C}^{n \times r}.$$

Remark 2.2.4. Orthogonal projectors doesn't necessarily have the form $\widehat{Q}\widehat{Q}^*$. We will show in Section 2.2.4 that $P = A(A^*A)^{-1}A^*$ is an orthogonal projection onto $\mathcal{R}(A)$ for any $A \in \mathbb{C}^{m \times n}$.

2.2.3 Projection with an Orthonormal Basis

Any natural extension of the discussion above is that in fact any matrix $\widehat{Q}$ with orthonormal columns can generate an orthogonal projector. For $r \leq n$, let $\{q_1, \ldots, q_r\}$ be any set of r orthonormal vectors in $\mathbb{C}^n$ and $\widehat{Q}$ the corresponding $n \times r$ matrix. We decompose any $v \in \mathbb{C}^n$ into $(r+1)$ orthogonal components

$$v = w + \sum_{j=1}^{r} (q_j^* v) q_j = w + \sum_{j=1}^{r} (q_j q_j^*) v.$$

More precisely, $v \in \mathbb{C}^n$ is decomposed into components in $\mathcal{R}(\widehat{Q})$ plus component in $\mathcal{R}(\widehat{Q})^\perp$. It follows that the map

$$v \mapsto \sum_{j=1}^{r} (q_j q_j^*) v,$$

represents an orthogonal projection onto $\mathcal{R}(\widehat{Q})$, *i.e.* the matrix $P = \widehat{Q}\widehat{Q}^*$ is an orthogonal projector onto $\mathcal{R}(Q)$, regardless of how $\{q_1, \ldots, q_r\}$ was obtained. Note that its complement projector $I - \widehat{Q}\widehat{Q}^*$ is an orthogonal projector onto $\mathcal{R}(\widehat{Q})^\perp$.

In the case of $r = 1$, we have the rank-one orthogonal projector that isolates the component in a single direction. More precisely, for any given $q \in \mathbb{C}^n$, the matrix

$$P_q = \frac{qq^*}{q^*q},$$

projects any vector $v \in \mathbb{C}^n$ onto $\text{span}\{q\}$. Its complement

$$P_{\perp q} = I - \frac{qq^*}{q^*q},$$

is the rank $n-1$ orthogonal projector onto $\mathbb{C}^n \setminus \text{span}\{q\}$.

2.2.4 Projection with an Arbitrary Basis

One can also define an orthogonal projection onto a subspace of $\mathbb{C}^n$ with an arbitrary basis, not necessarily orthogonal. This avoids the need to transform a given set of basis into orthonormal basis. Assume this subspace of $\mathbb{C}^n$ is spanned by a set of linearly independent vectors $\{a_1, \ldots, a_r\}$, with $r \leq n$. Define

$$A \coloneqq \begin{bmatrix} a_1 & \ldots & a_r \end{bmatrix} \in \mathbb{C}^{n \times r},$$

Geometrically, projecting any $v \in \mathbb{C}^n$ orthogonally onto $y = Ax \in \mathcal{R}(A)$ is equivalent to requiring $y - v \perp \mathcal{R}(A) = \mathcal{N}(A^*)^\perp$. This means that

$$a_j^*(y - v) = 0 \quad \text{for all } j = 1, \ldots, r,$$

or

$$A^*(Ax - v) = \mathbf{0} \implies A^*Ax = A^*v.$$

Since A is of full rank, A^*A is also of full rank and x is uniquely given by

$$x = (A^*A)^{-1}A^*v \implies Pv = y = Ax = A(A^*A)^{-1}A^*v \implies P = A(A^*A)^{-1}A^*.$$

Note that this is a generalisation of the rank-one orthogonal projector. If A has orthonormal columns, then we recover $P = AA^*$ as before.

2.3 QR Factorisation

We now study the second matrix factorisation in the course: *QR factorisation*. Assume for now that $A \in \mathbb{C}^{m \times n}, m \geq n$ is of full rank, but we will see later that this is not necessary. The idea of QR factorisation is to construct a sequence of orthonormal vectors $\{q_1, q_2, \ldots\}$ that spans the nested succesive spaces $\text{span}\{a_1, a_2, \ldots\}$, *i.e.*

$$\text{span}\{q_1, \ldots, q_j\} = \text{span}\{a_1, \ldots, a_j\} \quad \text{for } j = 1, \ldots, n.$$

In order for $\text{span}\{a_1, \ldots, a_j\}$ to be successive spaces, the vector a_j must be linear combination of the vectors $\{q_1, \ldots, q_j\}$. Writing this out

$$a_1 = r_{11}q_1 \tag{2.3.1a}$$
$$a_2 = r_{12}q_1 + r_{22}q_2 \tag{2.3.1b}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \tag{2.3.1c}$$
$$a_n = r_{1n}q_1 + r_{2n}q_2 + \ldots + r_{nn}q_n. \tag{2.3.1d}$$

In matrix form,

$$A = \left[\begin{array}{c|c|c|c} & & & \\ a_1 & a_2 & \dots & a_n \\ & & & \end{array}\right] = \left[\begin{array}{c|c|c|c} & & & \\ q_1 & q_2 & \dots & q_n \\ & & & \end{array}\right] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{in} \\ & r_{22} & \ddots & \vdots \\ & & \ddots & r_{(n-1)n} \\ & & & r_{nn} \end{bmatrix} = \hat{Q}\hat{R},$$

where $\widehat{Q} \in \mathbb{C}^{m\times n}$ has orthonormal columns and $\widehat{R} \in \mathbb{C}^{n\times n}$ is upper-triangular. Such a factorisation is called a **reduced *QR* factorisation** of A.

One can define a **full QR factorisation** in a similar fashion as how we define a full SVD, by adding $m-n$ orthonormal columns (in an arbitrary fashion) to $\widehat{Q}$ so that it becomes a unitary matrix $Q \in \mathbb{C}^{m\times m}$; in doing so, $m-n$ rows of zeros needs to be added to $\widehat{R}$ and it becomes an upper-triangular matrix $R \in \mathbb{C}^{m\times n}$. In the full QR factorisation, the columns $\{q_{n+1}, \dots, q_m\}$ are orthogonal to $\mathcal{R}(A)$ by construction, and they constitute an orthonormal basis for $\mathcal{R}(A)^\perp = \mathcal{N}(A^*)$ if A is of full rank n.

2.3.1 Gram-Schmidt Orthogonalisation

Equation (2.3.1) suggests the following method for computing the reduced QR factorisation. Given $a_1, a_2, \dots$, construct the vectors $q_1, q_2, \dots$ and entries r_{ij} by a process of successive orthogonalisation. This idea is known as the *Gram-Schmidt orthogonalisation*.

More precisely, at the jth step, we want to find a unit vector $q_j \in \operatorname{span}\{a_1, \dots, a_j\}$ such that $q_j \perp \{q_1, \dots, q_{j-1}\}$. This is done by projecting the vector a_j onto each component $\{q_1, \dots, q_{j-1}\}$. We then obtain

$$a_j = v_j + (q_1^* a_j)q_1 + \dots + (q_{j-1}^* a_j)q_{j-1}. \tag{2.3.2}$$

By construction, $v_j \perp \{q_1, \dots, q_{j-1}\}$ and $v_j \neq \mathbf{0}$, since otherwise a_j is a nontrivial linear combination of $\{a_1, \dots, a_{j-1}\}$, contradicting the assumption that A is of full rank. The orthonormal vectors are given by

$$q_j = \frac{1}{r_{jj}}\left(a_j - \sum_{i=1}^{j-1} r_{ij}q_i\right), \quad j = 1, \dots, n,$$

where the coefficients r_{ij} for each $i = 1, \dots, n$ are

$$r_{ij} = \begin{cases} q_i^* a_j & \text{if } i \neq j, \\ \pm \left\| a_j - \sum_{i=1}^{j-1} r_{ij}q_i \right\|_2 & \text{if } i = j. \end{cases}$$

The sign of r_{jj} is not determined and if desired we may choose $r_{jj} > 0$ so that $\hat{R}$ has positive diagonal entries. Gram-Schmidt iteration is numerically unstable due to rounding errors on a computer. To emphasise the instability, we refer to this algorithm as the **classical Gram-Schmidt iteration**.

Theorem 2.3.1. *Every matrix $A \in \mathbb{C}^{m\times n}, m \geq n$ has a full QR factorisation, hence also a reduced QR factorisation.*

Proof. The case where A has full rank follows easily from the Gram-Schmidt orthogonalisation, so suppose A does not have full rank. At one or more steps j, it will happen that $v_j = \mathbf{0}$; at this point, simply pick q_j arbitrarily to be any unit vector orthogonal to $\{q_1, \ldots, q_{j-1}\}$, and then continue the Gram-Schmidt orthogonalisation process. Previous step gives us a reduced QR factorisation of A. One can construct a full QR factorisation by introducing arbitrary $m - n$ orthonormal vectors in the same style as in Gram-Schmidt process.

■

Suppose $A = \widehat{Q}\widehat{R}$ is a reduced QR factorisation of A, then multiplying the ith column of $\hat{Q}$ by z and the ith row of $\hat{R}$ by z^{-1}, where $z \in \mathbb{C}$ such that $|z| = 1$ gives us another reduced QR factorisation of A. The next theorem asserts that this is the only way to obtain a unique reduced QR factorisation if A is of full rank.

Theorem 2.3.2. *Every matrix $A \in \mathbb{C}^{m\times n}, m \geq n$ of full rank has a unique reduced QR factorisation $A = \widehat{Q}\widehat{R}$, with $r_{jj} > 0$ for each $j = 1, \ldots, n$.*

Proof. The Gram-Schmidt orthogonalisation determines r_{ij} and q_j fully, except for the sign of r_{jj}, but this is now fixed by the condition $r_{jj} > 0$.

■

Algorithm 2.1: Classical Gram-Schmidt (unstable)

for $j = 1$ **to** n
  $v_j = a_j$
  for $i = 1$ **to** $j - 1$
    $r_{ij} = q_i^* a_j$
    $v_j = v_j - r_{ij} q_i$
  end
  $r_{jj} = \|v_j\|_2$
  $q_j = v_j / r_{jj}$
end

Suppose we want to solve $Ax = b$ for x, where $A \in \mathbb{C}^{m\times m}$ is nonsingular. If $A = QR$ is a (full) QR factorisation, then we can write

$$QRx = b \quad \text{or} \quad Rx = Q^* b.$$

The linear system $Rx = Q^*b$ can be solved easily using backward substitution since R is upper-triangular. This suggests the following method for solving $Ax = b$:

1. Compute a QR factorisation $A = QR$.
2. Compute $y = Q^*b$.
3. Solve $Rx = y$ for $x \in \mathbb{C}^m$.

2.3.2 Modified Gram-Schmidt Algorithm

At each jth step, the classical Gram-Schmidt iteration computes a single orthogonal projection of rank $m-(j-1)$ onto the space orthogonal to $\{q_1,\dots,q_{j-1}\}$, given by

$$v_j = P_j a_j, \quad j = 1,\dots,n.$$

In contrast, the modified Gram-Schmidt iteration computes the same result by a sequence of $(j-1)$ projections of rank $(m-1)$. Let $P_{\perp q} = I - qq^*$ be the rank $(m-1)$ orthogonal projector onto the space orthogonal to the nonzero vector $q \in \mathbb{C}^m$. It can be shown that

$$P_j = P_{\perp q_{j-1}} P_{\perp q_{j-2}} \dots P_{\perp q_1} \quad \text{for each } j = 1,\dots,n \text{ with } P_1 = I.$$

The operations are equivalent, but we decompose the projection to obtain numerical stability. The modified Gram-Schmidt algorithm computes v_j as follows (in order):

$$\begin{aligned}
v_j^{(1)} &= P_1 a_j &&= a_j \\
v_j^{(2)} &= P_{\perp q_1} v_j^{(1)} &&= (I - q_1 q_1^*) v_j^{(1)} \\
&\vdots &&\vdots \\
v_j = v_j^{(j)} &= P_{\perp q_{j-1}} v_j^{(j-1)} &&= (I - q_{j-1} q_{j-1}^*) v_j^{(j-1)}.
\end{aligned}$$

Algorithm 2.2: Modified Gram-Schmidt

for $j = 1$ **to** n
  $v_j = a_j$
  for $i = 1$ **to** $j-1$
    $r_{ij} = q_i^* v_j$ (Step by step projection)
    $v_j = v_j - r_{ij} q_i$
  end
  $r_{jj} = \|v_j\|_2$
  $a_j = v_j / r_{jj}$
end

Algorithm 2.3: (Efficient) Modified Gram-Schmidt

for $i = 1$ **to** n
  $v_i = a_i$
end
for $i = 1$ **to** n
  $r_{ii} = \|v_i\|_2$
  $q_i = v_i / r_{ii}$
  for $j = i+1$ **to** n
    $r_{ij} = q_i^* v_j$ (Compute P_{q_i} as soon as q_i is found
    $v_j = v_j - r_{ij} q_i$ and then apply to all $v_{i+1},\dots,v_n$)
  end
end

Consider three vectors

$$a_1 = \begin{bmatrix} 1 \\ \varepsilon \\ 0 \\ 0 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 1 \\ 0 \\ \varepsilon \\ 0 \end{bmatrix}, \quad a_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \varepsilon \end{bmatrix},$$

and make the approximation $\varepsilon^2 \approx 0$ for $\varepsilon \ll 1$ that accounts for rounding error. Applying the classical Gram-Schmidt gives

$$\begin{cases} v_1 = a_1 \\ r_{11} = \|a_1\|_2 = \sqrt{1+\varepsilon^2} \approx 1 \\ q_1 = \dfrac{v_1}{r_{11}} \quad \approx (1, \varepsilon, 0, 0)^T \end{cases}$$

$$\begin{cases} v_2 = a_2 \\ r_{12} = q_1^T a_2 \quad\quad = 1 \\ v_2 = v_2 - r_{12} q_1 = (0, -\varepsilon, \varepsilon, 0)^T \\ r_{22} = \|v_2\|_2 \quad\quad = \sqrt{2}\varepsilon \\ q_2 = \dfrac{v_2}{r_{22}} \quad\quad = \dfrac{1}{\sqrt{2}}(0, -1, 1, 0)^T \end{cases}$$

$$\begin{cases} v_3 = a_3 \\ r_{13} = q_1^T a_3 \quad\quad = 1 \\ v_3 = v_3 - r_{13} q_1 = (0, -\varepsilon, 0, \varepsilon)^T \\ r_{23} = q_2^T a_3 \quad\quad = 0 \\ v_3 = v_3 - r_{23} q_2 = (0, -\varepsilon, 0, \varepsilon)^T \\ r_{33} = \|v_3\|_2 \quad\quad = \sqrt{2}\varepsilon \\ q_3 = \dfrac{v_3}{r_{33}} \quad\quad = \dfrac{1}{\sqrt{2}}(0, -1, 0, 1)^T. \end{cases}$$

However, $q_2^T q_3 = 1/2 \neq 0$. We see that small perturbation results in instability, in the sense that we lose orthogonality due to round off errors. On the other hand, to apply the modified Gram-Schmidt, it is not difficult to see that q_1, q_2 remains unchanged and q_3 is obtained as

$$\begin{cases} v_3 = a_3 \\ r_{13} = q_1^T v_3 \quad\quad = 1 \\ v_3 = v_3 - r_{13} q_1 = (0, -\varepsilon, 0, \varepsilon)^T \\ r_{23} = q_2^T v_3 \quad\quad = \dfrac{\varepsilon}{\sqrt{2}} \\ v_3 = v_3 - r_{23} q_2 = \left(0, -\dfrac{\varepsilon}{2}, -\dfrac{\varepsilon}{2}, \varepsilon\right)^T \\ r_{33} = \|v_3\|_2 \quad\quad = \dfrac{\sqrt{6}\varepsilon}{2} \\ q_3 = \dfrac{v_3}{r_{33}} \quad\quad = \dfrac{1}{\sqrt{6}}(0, -1, -1, 2)^T. \end{cases}$$

We recover $q_2^T q_3 = 0$ in this case.

2.3.3 Operation Count

To assess the cost of both the Gram-Schmidt algorithms, we count the number of **floating point operations** called flops. Each addition, subtraction, multiplication, division or square root counts as one floop. We make no distinction between real and complex arithmetic, and no consideration of memory access or other aspects. From Algorithm 2.3, we see that:

$$\begin{aligned}
\#\text{ of addition} &= \sum_{i=1}^{n}\left(m-1+\left(\sum_{j=i+1}^{n} m-1\right)\right)\\
&= n(m-1)+\sum_{i=1}^{n}(m-1)(n-i)\\
&= \frac{1}{2}n(n+1)(m-1)\\
\#\text{ of subtraction} &= \sum_{i=1}^{n}\sum_{j=i+1}^{n} m = \sum_{i=1}^{n} m(n-i) = \frac{1}{2}mn(n-1)\\
\#\text{ of multiplication} &= \sum_{i=1}^{n}\left(m+\sum_{j=i+1}^{n} 2m\right)\\
&= mn+\sum_{i=1}^{n} 2m(n-i)\\
&= mn^2\\
\#\text{ of division} &= \sum_{i=1}^{n} m = mn\\
\#\text{ of square root} &= \sum_{i=1}^{n} 1 = n.
\end{aligned}$$

Hence, the number of flops is

$$\begin{aligned}
&\frac{1}{2}n(n+1)(m-1)+\frac{1}{2}mn(n-1)+mn^2+mn+n\\
&= 2mn^2-\frac{1}{2}n^2+mn+\frac{1}{2}n\\
&\sim 2mn^2,
\end{aligned}$$

where " $\sim$" means that

$$\lim_{m,n\to\infty}\frac{\text{number of flops}}{2mn^2}=1.$$

When m and n are large, this can also be obtained by considering only the dominating operations which occurs in the innermost loop of Algorithm 2.3

$$\begin{aligned}
r_{ij} &= q_i^* v_j && \Big[m \text{ multiplications and } m-1 \text{ additions.}\Big]\\
v_j &= v_j - r_{ij}q_i && \Big[m \text{ multiplications and } m-1 \text{ subtractions.}\Big]
\end{aligned}$$

Thus, the number of flops is asymptotic to

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n}(4m-1)\sim\sum_{i=1}^{n}(i)4m\sim 2mn^2.$$

2.4 Least Squares Problems

Consider a linear system of m equations having n unknowns, with $m > n$. In matrix formulation, we want to solve for $x \in \mathbb{C}^n$, the matrix equation $Ax = b$, where $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. In general, such a problem has no solution unless $b \in \mathcal{R}(A)$, since $b \in \mathbb{C}^m$ and $\mathcal{R}(A)$ is of dimension at most $n < m$. We say that a rectangular system of equations with $m > n$ is **overdetermined**.

Since the residual vector $r = Ax - b \in \mathbb{C}^m$ cannot be made to be zero for certain $b \in \mathbb{C}^m$, minimising it seems like a reasonable thing to do and measuring the "size" of r involves choosing a norm. For the 2-norm, the problem takes the following form:

$$\begin{aligned} &\text{Given } A \in \mathbb{C}^{m\times n},\ m \geq n,\ b \in \mathbb{C}^m, \\ &\text{find } x \in \mathbb{C}^n \text{ such that } \|Ax - b\|_2 \text{ is minimised.} \end{aligned} \tag{2.4.1}$$

This is called the general (linear) **least squares problem**. The 2-norm is chosen due to certain geometric and statistical reasons, but the more important reason is it leads to simple algorithms since the derivative of a quadratic function, which must be set to zero for minimisation, is linear. Geometrically, (2.4.1) means that we want to find a vector $x \in \mathbb{C}^n$ such that the vector $Ax \in \mathbb{C}^m$ is the closest point in $\mathcal{R}(A)$ to $b \in \mathbb{C}^m$.

Example 2.4.1. For a curve fitting problem, given a set of data $(y_1, b_1), \ldots, (y_m, b_m)$, we want to find a polynomial $p(y)$ such that $p(y_j) = b_j$ for every $j = 1, \ldots, m$. If the points $\{x_1, \ldots, x_m\} \in \mathbb{C}$ are distinct, it can be shown that there exists a unique **polynomial interpolant** to these data, which is a polynomial of degree at most $m - 1$. However, the fit is often bad, in the sense that they tend to get worse rather than better if more data are utilised. Even the fit is good, the interpolation process may be sensitive to perturbations of the data. One way to avoid such complications is to choose a nonuniform set of interpolation points, but in applications this will not always be possible.

Surprisingly, one can do better by reducing the degree of the polynomial. For some $n < m$, consider a degree $n - 1$ polynomial of the form

$$p(y) = x_0 + x_1 y + \ldots + x_{n-1} y^{n-1}.$$

In matrix form, the problem $Ax = b$ has the form

$$Ax = \begin{bmatrix} 1 & y_1 & \ldots & y_1^{n-1} \\ 1 & y_2 & \ldots & y_2^{n-1} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & y_m & \ldots & y_m^{n-1} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = b.$$

Such a polynomial is a least squares fit to the data if it minimises the residual vector in the 2-norm, that is,

$$\min_{\text{polynomials of degree } n-1} \left(\sum_{i=1}^{m} |p(y_i) - b_i|^2 \right)^{1/2}.$$

2.4.1 Existence and Uniqueness

Geometrically, a vector $x \in \mathbb{C}^n$ that minimises the residual $r = Ax - b$ in the 2-norm satisfies $Ax = Pb$, where $P \in \mathbb{C}^{m\times m}$ is the orthogonal projector onto $\mathcal{R}(A)$. In other words, the residual

$r = Ax - b$ must be orthogonal to $\mathcal{R}(A)$.

Theorem 2.4.2. *Let $A \in \mathbb{C}^{m\times n}, m \geq n$ and $b \in \mathbb{C}^m$ be given. A vector $x \in \mathbb{C}^n$ minimises the residual norm $\|r\|_2 = \|b - Ax\|_2$, thereby solving the least squares problem (2.4.1), if and only if $r \perp \mathcal{R}(A)$, that is, $A^*r = \mathbf{0}$, or equivalently,*

$$A^*Ax = A^*b, \tag{2.4.2}$$

or again equivalently,

$$Pb = Ax,$$

*where $P \in \mathbb{C}^{m\times m}$ is the orthogonal projector onto $\mathcal{R}(A)$. The $n \times n$ system of equations (2.4.2), known as the **normal equations**, is nonsingular if and only if A has full rank. Consequently, the solution $x \in \mathbb{C}^n$ is unique if and only if A has full rank.*

Proof. The equivalence of $A^*r = \mathbf{0}$ and (2.4.2) follows from the definition of r. The equivalence of $A^*r = \mathbf{0}$ and $Pb = Ax$ follows from the properties of orthogonal projectors, see Subsection 2.2.4. To prove that $y = Pb$ is the unique point in $\mathcal{R}(A)$ that minimises $\|b - y\|_2$, suppose z is another point in $\mathcal{R}(A)$. Since $z - y \perp b - y$, the Pythagorean theorem gives

$$\|b - z\|_2^2 = \|b - y\|_2^2 + \|y - z\|_2^2 > \|b - y\|_2^2.$$

Finally, suppose $A^*A \in \mathbb{C}^{n\times n}$ is nonsingular. For any $x \in \mathbb{C}^n$ satisfying $Ax = \mathbf{0}$, we have

$$Ax = \mathbf{0} \implies (A^*A)x = A^*\mathbf{0} = \mathbf{0} \implies x = \mathbf{0},$$

and so A has full rank. Conversely, suppose $A \in \mathbb{C}^{m\times n}$ is nonsingular and $A^*Ax = \mathbf{0}$ for some $x \in \mathbb{C}^n$. Then

$$x^*A^*Ax = x^*\mathbf{0} = 0 \implies (Ax)^*Ax = \|Ax\|_2^2 = 0 \implies Ax = \mathbf{0}.$$

Since $m \geq n$ by assumption, the rank of A is $\min\{m, n\} = n$ and the nullity of A is $n - n = 0$. Hence, the nullspace of A is trivial and $Ax = \mathbf{0} \implies x = \mathbf{0}$, which implies that A^*A is nonsingular.

■

If A is of full rank, it follows from Theorem 2.4.2 that the unique solution to the least squares problem is given by

$$x = (A^*A)^{-1}A^*b.$$

where the matrix $A^+ = (A^*A)^{-1}A^* \in \mathbb{C}^{n\times m}$ is called the **pseudoinverse** of A. The full-rank linear least squares problem (2.4.1) can then be solved by computing one of both vectors

$$x = A^+b, \quad y = Pb,$$

where P is the orthogonal projector onto $\mathcal{R}(A)$.

2.4.2 Normal Equations

The classical way to solve (2.4.1) is to solve the normal equations (2.4.2). If $A \in \mathbb{C}^{m\times n}$ has full rank with $m \geq n$, then A^*A is a square Hermitian positive-definite matrix. Indeed, for any nonzero $x \in \mathbb{C}^n$ we have

$$x^*(A^*A)x = (Ax)^*Ax = \|Ax\|_2^2 > 0.$$

The standard method of solving such a system is by **Cholesky factorisation**, which constructs a factorisation $A^*A = R^*R$, where $R \in \mathbb{C}^{n\times n}$ is upper-triangular. Consequently, $(A^*A)x = A^*b$ becomes $R^*Rx = A^*b$.

Algorithm 2.4: Least Squares via Normal Equations

1. Form the matrix A^*A and the vector A^*b.
2. Compute the Cholesky factorisation $A^*A = R^*R$.
3. Solve the lower-triangular system $R^*w = A^*b$ for $w \in \mathbb{C}^n$, using forward substitution.
4. Solve the upper-triangular system $Rx = w$ for $x \in \mathbb{C}^n$, using backward substitution.

The steps that dominate the work for this computation are the first two. Exploiting the symmetry of the problem, the computation of A^*A and the Cholesky factorisation require only mn^2 flops and $n^3/3$ flops respectively. Thus the total operation count is $\sim mn^2 + n^3/3$ flops.

2.4.3 QR Factorisation

Given a reduced QR factorisation $A = \widehat{Q}\widehat{R}$, the orthogonal projector $P \in \mathbb{C}^{m\times m}$ onto $\mathcal{R}(A)$ can be written as $P = \widehat{Q}\widehat{Q}^*$. Since $Pb \in \mathcal{R}(A)$, the system $Ax = Pb$ has an exact solution and

$$\widehat{Q}\widehat{R}x = \widehat{Q}\widehat{Q}^*b \implies \widehat{R}x = \widehat{Q}^*b.$$

Algorithm 2.5: Least Squares via QR Factorisation

1. Compute the reduced QR factorisation $A = \widehat{Q}\widehat{R}$.
2. Form the vector $\widehat{Q}^*b \in \mathbb{C}^n$.
3. Solve the upper-triangular system $\widehat{R}x = \widehat{Q}^*b$ for $x \in \mathbb{C}^n$, using backward substitution.

Note that the same reduction can also be derived from the normal equations (2.4.2).

$$A^*Ax = A^*b \implies (\widehat{R}^*\widehat{Q}^*)(\widehat{Q}\widehat{R})x = \widehat{R}^*\widehat{Q}^*b \implies \widehat{R}x = \widehat{Q}^*b.$$

The operation count for this computation is dominated by the cost of the QR factorisation, which is $\sim 2mn^2 - 2n^3/3$ flops if Householder reflections are used.

2.4.4 SVD

Given a reduced SVD $A = \widehat{U}\widehat{\Sigma}V^*$, it follows from Theorem 2.1.4 that the orthogonal projector $P \in \mathbb{C}^{m\times m}$ onto $\mathcal{R}(A)$ can be written as $P = \widehat{U}\widehat{U}^*$. The system $Ax = Pb$ reduces to

$$\widehat{U}\widehat{\Sigma}V^*x = \widehat{U}\widehat{U}^*b \implies \widehat{\Sigma}V^*x = \widehat{U}^*b.$$

Algorithm 2.6: Least Squares via SVD

1. Compute the reduced SVD $A = \widehat{U}\widehat{\Sigma}V^*$.
2. Form the vector $\widehat{U}^*b \in \mathbb{C}^n$.
3. Solve the diagonal system $\widehat{\Sigma}w = \widehat{U}^*b$ for $w \in \mathbb{C}^n$.
4. Set $x = Vw \in \mathbb{C}^n$.

Note that the same reduction can also be derived from the normal equations (2.4.2).

$$(V\widehat{\Sigma}\widehat{U}^*)(\widehat{U}\widehat{\Sigma}V^*)x = V\widehat{\Sigma}\widehat{U}^*b \implies \widehat{\Sigma}V^*x = \widehat{U}^*b.$$

The operation count for this computation is dominated by the cost of the SVD. For $m \gg n$ this cost is approximately the same as for QR factorisation, but for $m \approx n$ the SVD is more expensive. A typical estimate is $\sim 2mn^2 + 11n^3$ flops.

Algorithm 2.4 may be the best if we only care about the computational speed. However, solving the normal equations is not always numerically stable and so Algorithm 2.5 is the "modern standard" method for least squares problem. However if A is close to rank-deficient, it turns out that Algorithm 2.5 has less-than-ideal stability properties and Algorithm 2.6 is chosen instead.

2.5 Problems

1. Two matrices $A, B \in \mathbb{C}^{m\times m}$ are *unitary equivalent* if $A = QBQ^*$ for some unitary $Q \in \mathbb{C}^{m\times m}$. Is it true or false that A and B are unitarily equivalent if and only if they have the same singular values?

   Solution: Observe that for a square matrix, the reduced SVD and full SVD has the same structure. **The "only if" statement is true.** Suppose $A = QBQ^*$ for some unitary matrix $Q \in \mathbb{C}^{m\times m}$ and let $B = U_B\Sigma_B V_B^*$ be the SVD of B. Then

   $$A = QBQ^* = (QU_B)\Sigma_B(V_B^*Q^*),$$

   is a SVD of A since product of unitary matrices are unitary. Consequently, the singular values of A must be the same as B. **The "if" statement is false.** Consider the following two matrices

   $$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Since A is diagonal with positive entries, it has a SVD of the form

$$A = I_2 I_2 I_2 = AAA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

with singular value $\sigma_1^A = \sigma_2^A = 1$. Since $BB^* = I_2$, B is unitary and it has a SVD of the form

$$B = BI_2 I_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

with singular values $\sigma_1^B = \sigma_2^B = 1$. Suppose A and B are unitary equivalent, then

$$A = Q^* A Q = Q^*(QBQ^*Q) = B,$$

and we arrive at a contradiction.

2. Using the SVD, prove that any matrix in $C^{m\times n}$ is the limit of a sequence of matrices of full rank. In other words, prove that the set of full-rank matrices is a dense subset of $\mathbb{C}^{m\times n}$. Use the 2-norm for your proof. (The norm doesn't matter, since all norms on a finite-dimensional space are equivalent.)

Solution: We may assume WLOG that $m \geq n$. We want to show that for any matrix $A \in \mathbb{C}^{m\times n}$, there exists a sequence of full rank matrices $(A_k) \in \mathbb{C}^{m\times n}$ such that

$$\|A_k - A\|_2 \longrightarrow 0 \quad \text{as } k \longrightarrow \infty.$$

The result is trivial if A has full rank, since we may choose $A_k = A$ for each $k \geq 1$, so suppose A is rank-deficient. Let $r < \min\{m, n\} = n$ be the rank of A, which is also the number of nonzero singular values of A. Consider the reduced SVD $A = \widehat{U}\widehat{\Sigma}V^*$, where $V \in \mathbb{C}^{n\times n}$ is unitary, $U \in \mathbb{C}^{m\times n}$ has orthonormal columns and

$$\widehat{\Sigma} = \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

The fact that the 2-norm is invariant under unitary transformation suggests perturbing $\widehat{\Sigma}$ in such a way that it has full rank. More precisely, consider $A_k = \widehat{U}\widehat{\Sigma}_k V^*$, where

$$\widehat{\Sigma}_k = \widehat{\Sigma} + \frac{1}{k} I_n.$$

A_k has full rank by construction since it has n nonzero singular values and

$$\|A_k - A\|_2 = \|\widehat{U}(\widehat{\Sigma}_k - \widehat{\Sigma})V^*\|_2 = \frac{1}{k}\|I_n\|_2 = \frac{1}{k} \longrightarrow 0 \quad \text{as } k \longrightarrow \infty.$$

Since $A \in \mathbb{C}^{m\times n}$ was arbitrary, this shows that the set of full-rank matrices is a dense subset of $\mathbb{C}^{m\times n}$.

3. Consider the matrix

$$A = \begin{bmatrix} -2 & 11 \\ -10 & 5 \end{bmatrix}.$$

(a) Determine, on paper, a real SVD of A in the form $A = U\Sigma V^T$. The SVD is not unique, so find the one that has the minimal number of minus signs in U and V.

Solution: Since A is nonsingular, Theorem 2.1.6 says that the singular values of A are square roots of the eigenvalues of A^TA. Computing A^TA gives

$$A^TA = \begin{bmatrix} -2 & -10 \\ 11 & 5 \end{bmatrix} \begin{bmatrix} -2 & 11 \\ -10 & 5 \end{bmatrix} = \begin{bmatrix} 104 & -72 \\ -72 & 146 \end{bmatrix},$$

with characteristic equation

$$\lambda^2 - \mathrm{Tr}(A^TA)\lambda + \det(A^TA) = \lambda^2 - 250\lambda + 10000 = 0.$$

Solving this using quadratic formula gives the eigenvalues

$$\lambda = \frac{250 \pm \sqrt{250^2 - 4(10000)}}{2} = 125 \pm 75.$$

Thus, $\lambda_1 = 200 \implies \sigma_1 = \sqrt{\lambda_1} = 10\sqrt{2}$ and $\lambda_2 = 50 \implies \sigma_2 = \sqrt{\lambda_2} = 5\sqrt{2}$.

Denote $U = [u_1|u_2] \in \mathbb{R}^{2\times 2}$ and $V = [v_1|v_2] \in \mathbb{R}^{2\times 2}$, where u_1, u_2 and v_1, v_2 are column vectors of U and V respectively in the SVD of $A = U\Sigma V^T$. Observe that v_1, v_2 are normalised eigenvectors of A^TA corresponding to eigenvalues λ_1, λ_2 respectively since $A^TA = V\Sigma^2V^T$. It can be shown that

$$V = [v_1|v_2] = \begin{bmatrix} -3/5 & 4/5 \\ 4/5 & 3/5 \end{bmatrix}.$$

To find u_1, u_2, we use the relation $Av_j = \sigma_j u_j$

$$Av_1 = \begin{bmatrix} 10 \\ 10 \end{bmatrix} = 10\sqrt{2} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \sigma_1 u_1.$$
$$Av_2 = \begin{bmatrix} 5 \\ -5 \end{bmatrix} = 5\sqrt{2} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \sigma_2 u_2.$$

Hence, a real SVD of A with minimal number of minus signs in U and V is

$$A = U\Sigma V^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 10\sqrt{2} & 0 \\ 0 & 5\sqrt{2} \end{bmatrix} \begin{bmatrix} -3/5 & 4/5 \\ 4/5 & 3/5 \end{bmatrix}.$$

(b) List the singular values, left singular vectors, and right singular vectors of A. Draw a careful, labeled picture of the unit ball in $\mathbb{R}^2$ and its image under A, together with the singular vectors, with the coordinates of their vertices marked.

Solution: The singular values of A are $\sigma_1 = 10\sqrt{2}, \sigma_2 = 5\sqrt{2}$. The left singular vectors and right singular vectors of A are

$$v_1 = \begin{bmatrix} -3/5 \\ 4/5 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}, \quad u_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}.$$

(c) What are the $1-$, $2-$, ∞-, and Frobenius norms of A?

Solution: Let $A = (a_{ij})_{i,j=1,2}$. Then

$$\begin{aligned} \|A\|_1 &= \max_{j=1,2}\{|a_{1j}| + |a_{2j}|\} = \max\{12, 16\} = 16 \\ \|A\|_2 &= \sigma_1 = 10\sqrt{2} \\ \|A\|_\infty &= \max_{i=1,2}\{|a_{i1}| + |a_{i2}|\} = \max\{13, 15\} = 15 \\ \|A\|_F &= \sqrt{\operatorname{tr}(A^T A)} = \sqrt{250} = 5\sqrt{10}. \end{aligned}$$

(d) Find A^{-1} not directly, but via the SVD.

Solution: Using SVD,

$$A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T = \begin{bmatrix} 1/20 & -11/100 \\ 1/10 & -1/50 \end{bmatrix}.$$

(e) Find the eigenvalues λ_1, λ_2 of A.

Solution: Solving the characteristic equation

$$\lambda^2 - \operatorname{Tr}(A)\lambda + \det(A) = \lambda^2 - 3\lambda + 100 = 0,$$

yields

$$\lambda = \frac{3 \pm \sqrt{9 - 4(100)}}{2} = \frac{3}{2} \pm \frac{\sqrt{391}i}{2}.$$

(f) Verify that $\det(A) = \lambda_1\lambda_2$ and $|\det(A)| = \sigma_1\sigma_2$.

Solution:

$$\lambda_1\lambda_2 = \left(\frac{3}{2} + \frac{\sqrt{391}i}{2}\right)\left(\frac{3}{2} - \frac{\sqrt{391}i}{2}\right) = \frac{9}{4} + \frac{391}{4} = \frac{400}{4} = 100 = \det(A).$$

$$\sigma_1\sigma_2 = (10\sqrt{2})(5\sqrt{2}) = 50(2) = 100 = |\det(A)|.$$

(g) What is the area of the ellipsoid onto which A maps the unit ball of $\mathbb{R}^2$?

Solution: The ellipse onto which A maps the unit ball of $\mathbb{R}^2$ has major radius $a = \sigma_1$ and minor radius $b = \sigma_2$. Thus, its area is $\pi ab = \pi\sigma_1\sigma_2 = 100\pi$.

4. Let $P \in \mathbb{C}^{m\times m}$ be a nonzero projector. Show that $\|P\|_2 \geq 1$, with equality if and only if P is an orthogonal projector.

Solution: Since any projector $P \in \mathbb{C}^{m\times m}$ satisfies $P^2 = P$,

$$\|Px\|_2 = \|P^2x\|_2 \leq \|P\|_2^2\|x\|_2.$$

Taking the supremum over all $x \in \mathbb{C}^m$ with $\|x\|_2 = 1$ gives

$$\|P\|_2 \leq \|P\|_2^2 \implies \|P\|_2 \geq 1 \quad \text{since } \|P\|_2 \neq 0.$$

Suppose P is an orthogonal projector with its SVD $P = U\Sigma V^*$, where its singular values are 1's and 0's. It follows from Theorem 2.1.4 that $\|P\|_2 = \sigma_1 = 1$. Conversely, suppose P is not an orthogonal projector. By definition, this means that

$$\mathcal{R}(P) \not\perp \mathcal{N}(P) = \mathcal{R}(I - P).$$

5. Let $A = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

(a) Determine the reduced QR factorisation of A.

Solution: Denote by a_j the jth column of A, $j = 1, 2, 3$. Following the Gram-Schmidt iteration notation from Section 2.3,

$$\begin{aligned}
r_{11} = \|a_1\|_2 = \sqrt{3} &\implies q_1 = \frac{a_1}{r_{11}} = \frac{1}{\sqrt{3}}(1,1,1,0)^*. \\
r_{12} = q_1^* a_2 &= \sqrt{3}. \\
v_2 = a_2 - r_{12}q_1 &= a_2 - \sqrt{3}q_1 = (-1,1,0,1)^*. \\
r_{22} = \|v_2\|_2 = \sqrt{3} &\implies q_2 = \frac{v_2}{r_{22}} = \frac{1}{\sqrt{3}}(-1,1,0,1)^*. \\
r_{13} = q_1^* a_3 = -\sqrt{3}, r_{23} &= q_2^* a_3 = \sqrt{3}. \\
v_3 = a_3 - r_{13}q_1 - r_{23}q_2 &= a_3 + \sqrt{3}q_1 - \sqrt{3}q_2 = (1,1,-2,0)^*. \\
r_{33} = \|v_3\|_2 = \sqrt{6} &\implies q_3 = \frac{v_3}{r_{33}} = \frac{1}{\sqrt{6}}(1,1,-2,0)^*.
\end{aligned}$$

Hence, $A = \hat{Q}\hat{R}$, where

$$\hat{Q} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & -1 & 1/\sqrt{2} \\ 1 & 1 & 1/\sqrt{2} \\ 1 & 0 & -\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix}, \hat{R} = \sqrt{3} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & \sqrt{2} \end{bmatrix}.$$

(b) Use the QR factors from part (a) to determine the least square solution to $Ax = b$.

Solution: We follow Algorithm 11.2., page 83. Computing $\hat{Q}^*b$ yields

$$\hat{Q}^*b = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ 1/\sqrt{2} & 1/\sqrt{2} & -\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}.$$

Thus, $\hat{R}x = \hat{Q}^*b$ becomes

$$\sqrt{3} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/3 \\ 0 \end{bmatrix}.$$

Performing back substitution gives

$$\begin{aligned} x_3 &= 0. \\ x_2 &= \frac{1}{3} - x_3 = \frac{1}{3}. \\ x_1 &= 1 - x_2 + x_3 = 1 - \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

Hence, $x = (x_1, x_2, x_3)^* = (2/3, 1/3, 0)^*$.

6. Let A be an $m \times n$ matrix ($m \geq n$), and let $A = \hat{Q}\hat{R}$ be a reduced QR factorisation.

   (a) Show that A has rank n if and only if all the diagonal entries of $\hat{R}$ are nonzero.

Solution: Let a_j be the jth column of the matrix $A \in \mathbb{C}^{m\times n}, m \geq n$. Observe that to prove the claim above, it suffices to show that the set of vectors $\{a_1, \ldots, a_n\}$ is linearly independent in $\mathbb{C}^m$ if and only if all the diagonal entries of $\hat{R}$ are nonzero. Indeed, $\operatorname{rank}(A) \leq \min\{m, n\} = n$, so $\operatorname{rank}(A) = n$ if and only if A is of full rank.

Suppose the set of vectors $\{a_1, \ldots, a_n\}$ is linearly independent in $\mathbb{C}^m$. Recall that for every $j = 1, \ldots, n$, a_j can be expressed as a linear combination of $\{q_1, \ldots, q_j\}$. More precisely,

$$a_j = r_{1j}q_1 + r_{2j}q_2 + \ldots + r_{jj}q_j = \sum_{i=1}^{j} r_{ij}q_j, \quad j = 1, \ldots, n. \tag{2.5.1}$$

Suppose by contradiction that there exists an $j_0 \in \{1, \ldots, n\}$ such that $r_{j_0 j_0} = 0$, (2.5.1) implies that $a_{j_0} \in \operatorname{span}\{a_1, \ldots, a_{j_0-1}\}$, which contradicts the linear independency of $\{a_1, \ldots, a_n\}$. Hence all the diagonal entries of $\hat{R}$ must be nonzero.

Conversely, suppose all the diagonal entries of $\hat{R}$ are nonzero. Suppose the

following equation holds

$$\beta_1 a_1 + \ldots + \beta_n a_n = \mathbf{0}. \tag{2.5.2}$$

Substituting (2.5.1) into (2.5.2) yields

$$\gamma_1 q_1 + \ldots + \gamma_n q_n = \mathbf{0},$$

where

$$\gamma_j = \sum_{k=j}^{n} \beta_k r_{jk}, \quad j = 1, \ldots, n. \tag{2.5.3}$$

Since $\{q_1, \ldots, q_n\}$ is an orthonormal set of vectors in $\mathbb{C}^m$, it is linearly independent and so we must have $\gamma_1 = \ldots = \gamma_n = 0$. We claim that this and (2.5.3) implies $\beta_1 = \ldots = \beta_n = 0$. First,

$$\gamma_n = \beta_n r_{nn} = 0 \implies \beta_n = 0 \text{ since } r_{nn} \neq 0.$$

Next,

$$\gamma_{n-1} = \beta_{n-1} r_{(n-1)(n-1)} + \cancel{\beta_n} r_{(n-1)n} = \beta_{n-1} r_{(n-1)(n-1)} = 0$$
$$\implies \beta_{n-1} = 0 \text{ since } r_{(n-1)(n-1)} \neq 0.$$

Carrying the exact computation inductively from $j = n-2$ to $j = 1$, together witih $r_{jj} \neq 0$ proves the claim. Hence, (2.5.2) has only trivial solution $\beta_1 = \ldots = \beta_n = 0$, which by definition means that the set of vectors $\{a_1, \ldots, a_n\}$ is linearly independent in $\mathbb{C}^m$.

(b) Suppose $\hat{R}$ has k nonzero diagonal entries for some k with $0 \leq k < n$. What does this imply about the rank of A? Exactly k? At least k? At most k? Give a precise answer, and prove it.

Solution: Suppose $\hat{R}$ has k nonzero diagonal entries for some k with $0 \leq k < n$, i.e. $\hat{R}$ has at least one zero diagonal entry. Let a_j be the jth column of A, and $A_j \in \mathbb{C}^{m \times j}$ be the matrix defined by $A_j = [a_1 | a_2 | \ldots | a_j]$.

- First,
$$\text{rank}(A_1) = \begin{cases} 1 & \text{if } r_{11} \neq 0, \\ 0 & \text{if } r_{11} = 0. \end{cases}$$
- For $j = 2, \ldots, n$, regardless of the value of r_{jj}, either
$$a_j \notin \text{span}\{a_1, \ldots, a_{j-1}\} \implies \text{rank}(A_j) = \text{rank}(A_{j-1}) + 1. \tag{2.5.4}$$
or
$$a_j \in \text{span}\{a_1, \ldots, a_{j-1}\} \implies \text{rank}(A_j) = \text{rank}(A_{j-1}). \tag{2.5.5}$$
This means that the rank of A cannot be at most k.
- For any $j = 2, \ldots, n$, if $r_{jj} \neq 0$, then (2.5.1) implies that (2.5.4) must hold. However, if $r_{jj} = 0$, then either (2.5.4) or (2.5.5) holds. We illustrate this

two cases by looking at 3×3 matrix $\hat{R}$, similar idea applies to "higher dimensional" $\hat{R}$ too.

- One example where (2.5.4) holds is the case where $r_{11} = r_{22} = r_{33} = 0$ but $r_{12} = r_{13} = r_{23} \neq 0$. In this case,

$$a_1 = \mathbf{0}, a_2 = r_{12}q_1, a_3 = r_{13}q_1 + r_{23}q_2$$

  and it is clear that $a_3 \notin \text{span}\{a_1, a_2\}$.
- One example where (2.5.5) holds is the case where $r_{11} = r_{22} = r_{33} = r_{23} = 0$ but $r_{12} = r_{13} \neq 0$. In this case,

$$a_1 = \mathbf{0}, a_2 = r_{12}q_1, a_3 = r_{13}q_1$$

  and it is clear that $a_3 \in \text{span}\{a_1, a_2\}$.

Summarising everything, we conclude that the rank of A is **at least** k.

7. Let A be an $m \times m$ matrix, and let a_j be its jth column. Give an algebraic proof of *Hadamard's inequality*:

$$|\det A| \leq \prod_{j=1}^{m} \|a_j\|_2.$$

Also give a geometric interpretation of this result, making use of the fact that the determinant equals the volume of a parallelepiped.

Solution: The inequality is trivial if A is a singular matrix, so suppose not. Consider the QR factorisation $A = \hat{Q}\hat{R}$. Since $\hat{Q} \in \mathbb{C}^{m\times m}$ is unitary, $\det(\hat{Q}) = \pm 1$; since $\hat{R} \in \mathbb{C}^{m\times m}$ is upper triangular, $\det(\hat{R})$ equals the product of its diagonal entries. Using these two facts and product property of determinant,

$$\begin{aligned} |\det(A)| = |\det(\hat{Q}\hat{R})| &= |\det(\hat{Q})||\det(\hat{R})| \\ &= |\det(\hat{R})| \\ &= \prod_{j=1}^{m} |r_{jj}| = \prod_{j=1}^{m} \|v_j\|_2, \end{aligned}$$

where $v_j = a_j - \sum_{i=1}^{j-1} (q_i^* a_j) q_i$, with the convention that $q_0 = \mathbf{0} \in \mathbb{C}^m$. For any $j = 1, \ldots, m$, since $\{v_j, q_1, \ldots, q_{j-1}\}$ are mutually orthogonal, **Pythagorean theorem** gives

$$\begin{aligned} \|a_j\|_2^2 &= \left\| v_j + \sum_{i=1}^{j-1} (q_i^* a_j) q_i \right\|_2^2 \\ &= \|v_j\|_2^2 + \sum_{i=1}^{j-1} \|(q_i^* a_j) q_i\|_2^2 \\ &\geq \|v_j\|_2^2 \end{aligned}$$

where we crucially use the fact that $\|\cdot\|_2 \geq 0$. Hence,

$$|\det(A)| \leq \prod_{j=1}^{m} \|v_j\|_2 \leq \prod_{j=1}^{m} \|a_j\|_2.$$

Since $|\det(A)|$ is the volume of the parallelepiped with sides given by the vector $\{a_1, a_2, \ldots, a_m\}$, the Hadamard's inequality asserts that this is bounded above by the volume of a rectangular parallelepiped with sides of length $\|a_1\|_2, \|a_2\|_2, \ldots, \|a_m\|_2$.

8. Consider the inner product space of real-valued continuous functions defined on $[-1, 1]$, where the inner product is defined by

$$f \cdot g = \int_{-1}^{1} f(x)g(x)\,dx.$$

Let M be the subspace that is spanned by the three linearly independent polynomial $p_0 = 1, p_1 = x, p_2 = x^2$.

(a) Use the Gram-Schmidt process to determine an orthonormal set of polynomials (Legendre polynomials) q_0, q_1, q_2 that spans M.

Solution: Following Gram-Schmidt iteration notation from lectures,

$$\begin{aligned}
q_0 &= \frac{p_0}{(p_0 \cdot p_0)^{1/2}} = \frac{1}{\sqrt{2}}. \\
r_{12} &= q_0 \cdot p_1 = \int_{-1}^{1} \frac{x}{\sqrt{2}}\,dx = 0. \\
\implies v_1 &= p_1 - \cancel{r_{12}} q_0 = x. \\
(r_{22})^2 &= v_1 \cdot v_1 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}. \\
\implies q_1 &= \frac{v_1}{r_{22}} = \sqrt{\frac{3}{2}}x. \\
r_{13} &= q_0 \cdot p_2 = \int_{-1}^{1} \frac{x^2}{\sqrt{2}}\,dx = \frac{\sqrt{2}}{3}. \\
r_{23} &= q_1 \cdot p_2 = \int_{-1}^{1} \sqrt{\frac{3}{2}}x^3\,dx = 0. \\
\implies v_2 &= p_2 - r_{13}q_0 - \cancel{r_{23}} q_1 = x^2 - \frac{1}{3}. \\
(r_{33})^2 &= v_2 \cdot v_2 = \int_{-1}^{1} \left(x^2 - \frac{1}{3}\right)^2 dx = \frac{8}{45}. \\
\implies q_2 &= \frac{v_2}{r_{33}} = \sqrt{\frac{45}{8}}\left(x^2 - \frac{1}{3}\right) = \sqrt{\frac{5}{8}}(3x^2 - 1).
\end{aligned}$$

Hence, $q_0(x) = \frac{1}{\sqrt{2}}, q_1(x) = \sqrt{\frac{3}{2}}x, q_2(x) = \sqrt{\frac{5}{8}}(3x^2 - 1)$.

(b) Check that q_n satisfies $(1 - x^2)y'' - 2xy' + n(n + 1)y = 0$ for $n = 0, 1, 2$.

Solution: It is clear that q_0 satisfies the given ODE for $n = 0$ since $q_0' = q_0'' = 0$ and $n(n+1)|_{n=0} = 0$. Because differentiation is a linear operation, it suffices to show that v_1, v_2 (from part (a)) satisfies the given ODE for $n = 1, 2$ respectively. For n=1,

$$(1-x^2)v_1'' - 2xv_1' + 1(1+1)v_1 = (1-x^2)(0) - 2x(1) + 2(x) = 0.$$

For $n = 2$,

$$\begin{aligned}(1-x^2)v_2'' - 2xv_2' + 2(2+1)v_2 &= (1-x^2)(2) - 2x(2x) + 6\left(x^2 - \frac{1}{3}\right) \\ &= 2 - 2x^2 - 4x^2 + 6x^2 - 2 = 0.\end{aligned}$$

9. Let $A \in \mathbb{R}^{m\times n}$ with $m < n$ and of full rank. Then min $\|Ax - b\|_2$ is called an Underdetermined Least-Squares Problem. Show that the solution is an $n - m$ dimensional set. Show how to compute the unique mininum norm solution using QR decomposition and SVD approach.

Solution: Let $A \in \mathbb{R}^{m\times n}$ with $m < n$ and of full rank. Since $m < n$, $Ax = b$ is an underdetermined system and $\|Ax - b\|_2$ attains its mininum 0 in this case, where the solution set, S is given by

$$S = \{x_p - z \in \mathbb{R}^n : z \in \mathcal{N}(A)\},$$

where x_p is the particular solution to $Ax = b$ and $\mathcal{N}(A)$ denotes the null space of A. Note that S is not a vector subspace of $\mathbb{R}^n$ (unless $b = \mathbf{0} \in \mathbb{R}^m$). Invoking the **Rank-Nullity theorem** gives

$$\dim(\mathcal{N}(A)) = n - \text{rank}(A) = n - m.$$

i.e. the solution set S is an $n - m$ dimensional set.

Now that we know solutions to an Underdetermined Least-Squares problem must belong to S , we seek the mininum norm solution. More precisely, we look for $x_0 = x_p - z_0 \in S$ that solves the following minimisation problem:

$$\min_{x\in S} \|x\|_2 = \min_{z\in\mathcal{N}(A)} \|x_p - z\|_2. \tag{2.5.6}$$

Since $\mathcal{N}(A)$ is a closed subspace of $\mathbb{R}^n$, (2.5.6) has a unique solution z_0 satisfying $x_0 = x_p - z_0 \in \mathcal{N}(A)^\perp$, where $\mathcal{N}(A)^\perp$ denotes the orthogonal complement of $\mathcal{N}(A)$. Geometrically, z_0 is precisely the orthogonal projection of x_p onto $\mathcal{N}(A)$. We will not prove it here, but one can show that $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$, the range of A^T. Since $x_0 \in \mathcal{N}(A)^\perp = \mathcal{R}(A^T)$, there exists an $v_0 \in \mathbb{R}^m$ such that $A^T v_0 = x_0$. Substituting this into $Ax_0 = b$ thus gives

$$AA^T v_0 = b \implies v_0 = (AA^T)^{-1}b \implies x_0 = A^T(AA^T)^{-1}b, \tag{2.5.7}$$

where $(AA^T)^{-1}$ exists since A has full rank implies A^TA (and also AA^T) is nonsingular.

- Suppose $A^T \in \mathbb{R}^{n\times m}$ has a reduced QR factorisation $A^T = \hat{Q}\hat{R}$. Then

$$(AA^T)^{-1} = (\hat{R}^T\hat{Q}^T\hat{Q}\hat{R})^{-1} = (\hat{R}^T\hat{R})^{-1} = (\hat{R})^{-1}(\hat{R}^T)^{-1}.$$

  Substituting this into (2.5.7) yields

$$x_0 = A^T(AA^T)^{-1}b = \hat{Q}\hat{R}(\hat{R})^{-1}(\hat{R}^T)^{-1}b = \hat{Q}(\hat{R}^T)^{-1}b.$$

- Suppose $A^T \in \mathbb{R}^{n\times m}$ has a reduced SVD $A^T = \hat{U}\hat{\Sigma}V$. Then

$$(AA^T)^{-1} = (V^T\hat{\Sigma}^T\hat{U}^T\hat{U}\hat{\Sigma}V)^{-1} = (V^T\hat{\Sigma}^2V)^{-1} = V^T(\hat{\Sigma}^2)^{-1}V.$$

  where $V^{-1} = V^T$ since $V \in \mathbb{R}^{m\times m}$ is unitary. Substituting this into (2.5.7) yields

$$x_0 = A^T(AA^T)^{-1}b = \hat{U}\hat{\Sigma}VV^T(\hat{\Sigma}^2)^{-1}Vb = \hat{U}\hat{\Sigma}(\hat{\Sigma}^2)^{-1}Vb = \hat{U}(\hat{\Sigma})^{-1}Vb.$$

Here, the assumption that A is full rank is crucial, in that it ensures the existence of $(\hat{R}^T)^{-1}$ and $(\hat{\Sigma})^{-1}$. Indeed, Q1(b)(i) says that $\hat{R}$ has all nonzero diagonal entries, which implies that $\hat{R}$ (and also $\hat{R}^T$) is nonsingular since $\hat{R}$ is upper-triangular; Theorem 5.1, page 33, tells us that all singular values of A, which are the diagonal entires of $\hat{\Sigma}$, are nonzero, which implies that $\hat{\Sigma}$ is nonsingular since $\hat{\Sigma}$ is diagonal.

Chapter 3

Conditioning and Stability

3.1 Conditioning and Condition Numbers

One can view a problem as a function $f\colon X \longrightarrow Y$ from a normed vector space X of data to a normed vector space Y of solutions. A **well-conditioned problem** is one with the property that all small perturbations of x lead to only small changes in $f(x)$; an **ill-conditioned problem** is one with the property that some small perturbations of x lead to a large change in $f(x)$.

Definition 3.1.1. Let δx be a small perturbation of x, and $\delta f = f(x + \delta x) - f(x)$. The **absolute condition number** $\hat{\kappa} = \hat{\kappa}(x)$ of the problem f at x is defined as

$$\hat{\kappa} = \hat{\kappa}(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f\|}{\|\delta x\|}.$$

- It can be interpreted as a supremum over all infinitesimal perturbations δx, thus it can be written as
$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}.$$

- If $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is differentiable, then there exists an element $J(x) \in \mathbb{R}^{m \times n}$, called the Jacobian, such that
$$f(x + \delta x) - f(x) = J(x)\delta x + o(\|\delta x\|).$$
In the limit $\|\delta x\| \longrightarrow 0$, the above simplifies to $\delta f = J(x)\delta x$ and the absolute condition number then becomes
$$\hat{\kappa}(x) = \|J(x)\|,$$

Definition 3.1.2. The **relative condition number** $\kappa = \kappa(x)$ of the problem f at x is defined as

$$\kappa = \kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right),$$

or assuming $\delta x, \delta f$ are infinitesimal,

$$\kappa = \kappa(x) = \sup_{\delta x} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right).$$

- If $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is differentiable, then it can be expressed in terms of the Jacobian:

$$\kappa = \sup_{\|\delta x\|} \frac{\|J(x)\|}{\|f(x)\|/\|x\|}.$$

- A problem is *well-conditioned* if κ is *small* (*e.g.* 1, 10, 100) and *ill-conditioned* if κ is *large* (*e.g.* $10^6, 10^{16}$).

Example 3.1.3. Consider $f(x) = \alpha x, x \in \mathbb{C}$. Then $J(x) = f'(x) = \alpha$ and

$$\hat{\kappa} = \|J(x)\| = |\alpha| \quad \text{but} \quad \kappa = \frac{\|J(x)\|}{\|f(x)\|/\|x\|} = \frac{|\alpha|}{|\alpha x|/|x|} = 1.$$

Thus, this problem is well-conditioned.

Example 3.1.4. Consider $f(x) = \sqrt{x}, x > 0$. Then $J(x) = f'(x) = 1/(2\sqrt{x})$ and

$$\hat{\kappa} = \|J(x)\| = \frac{1}{2\sqrt{x}} \quad \text{but} \quad \kappa = \frac{\|J(x)\|}{\|f(x)\|/\|x\|} = \left(\frac{1}{2\sqrt{x}} \middle/ \frac{\sqrt{x}}{x}\right) = \frac{1}{2}.$$

Thus, this problem is well-conditioned.

Example 3.1.5. Consider $f(x) = x_1 - x_2, x = (x_1, x_2)^* \in (\mathbb{C}^2, \|\cdot\|_\infty)$. Then $J(x) = (1, -1)$ and

$$\hat{\kappa} = \|J(x)\|_\infty = 2 \quad \text{but} \quad \kappa = \frac{\|J(x)\|_\infty}{\|f(x)\|_1/\|x\|_\infty} = \frac{2}{|x_1 - x_2|/\max\{|x_1|, |x_2|\}}.$$

The absolute condition number blows up if $|x_1 - x_2| \approx 0$. Thus, this problem is severely ill-conditioned when $x_1 \approx x_2$, an issue which $\hat{\kappa}$ would not reveal.

Condition of Matrix-Vector Multiplication

Given $A \in \mathbb{C}^{m\times n}$, consider $f(x) = Ax, x \in \mathbb{C}^n$. If x has some perturbation δx, then for arbitrary vector norm we have

$$\kappa = \sup_{\delta x} \left(\frac{\|J\|}{\|f(x)\|/\|x\|}\right) = \frac{\|A\|\|x\|}{\|Ax\|}.$$

- If A is square and non-singular, then

$$\frac{\|x\|}{\|Ax\|} = \frac{\|A^{-1}Ax\|}{\|Ax\|} \leq \|A^{-1}\| \implies \kappa \leq \|A\|\|A^{-1}\|.$$

- For $\|\cdot\|_2$, this bound is actually attained since $\|A\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = 1/\sigma_m$, where $\sigma_m > 0$ since A is non-singular. Indeed, choosing x to be the mth right singular vector of A yields

$$\frac{\|x\|_2}{\|Ax\|_2} = \frac{\|v_m\|_2}{\|Av_m\|_2} = \frac{\|v_m\|_2}{\sigma_m\|u_m\|_2} = \frac{1}{\sigma_m}.$$

Theorem 3.1.6. *Let $A \in \mathbb{C}^{m\times m}$ be non-singular and consider the equation $Ax = b$.*

(a) Consider $f(x) = Ax = b$. The problem of computing b, given x, has condition number

$$\kappa(x) = \frac{\|A\|\|x\|}{\|Ax\|} = \frac{\|A\|\|x\|}{\|b\|} \leq \|A\|\|A^{-1}\|$$

with respect to perturbations of x. If $\|\cdot\| = \|\cdot\|_2$, then equality holds if x is a multiple of a mth right singular vector v_m of A corresponding to the minimal singular value σ_m.

(b) Consider $f(b) = A^{-1}b = x$. The problem of computing x, given b, has condition number

$$\kappa(b) = \frac{\|A^{-1}\|\|b\|}{\|A^{-1}b\|} = \frac{\|A^{-1}\|\|b\|}{\|x\|} \leq \|A^{-1}\|\|A\|$$

with respect to perturbations of b. If $\|\cdot\| = \|\cdot\|_2$, then equality holds if b is a multiple of a 1st left singular vector u_1 of A corresponding to the maximal singular value σ_1.

Condition of a System of Equations

Theorem 3.1.7. *Consider the problem $f(A) = A^{-1}b = x$, where $A \in \mathbb{C}^{m\times m}$ is non-singular. The problem of computing x, given A, has condition number*

$$\kappa(A) \leq \|A^{-1}\|A\|$$

with respect to perturbations of A.

- Consider the problem $f(A) = A^{-1}b = x$, where now A has some perturbation δA instead of b. Then

$$(A + \delta A)(x + \delta x) = b \implies \delta Ax + A\delta x \approx 0 \implies \delta x \approx -A^{-1}\delta Ax$$

and

$$\kappa(A) = \sup_{\delta A}\left(\frac{\|\delta x\|}{\|x\|}\Big/\frac{\|\delta A\|}{\|A\|}\right) \leq \sup_{\delta A}\left(\frac{\|A^{-1}\|\|\delta A\|\|x\|}{\|x\|}\right)\left(\frac{\|A\|}{\|\delta A\|}\right) = \|A^{-1}\|\|A\|.$$

- Equality holds whenever δA is such that

$$\|A^{-1}\delta Ax\| = \|A^{-1}\|\|\delta A\|\|x\|.$$

It can be shown that such perturbations δA exists for any given $A \in \mathbb{C}^{m\times m}, b \in \mathbb{C}^m$ and any chosen norm $\|\cdot\|$.

The product $\|A\|\|A^{-1}\|$ appears so often that we decided to call it the **condition number** of A (relative to the norm $\|\cdot\|$), denoted by $\kappa(A)$. A is said to be *well-conditioned* if $\kappa(A)$ is *small* and *ill-conditioned* if $\kappa(A)$ is large. In the case where A is singular, we write $\kappa(A) = \infty$.

For a rectangular matrix $A \in \mathbb{C}^{m\times n}$ of full rank, $m \geq n$, the condition number is defined in terms of the pseudoinverse, *i.e.*

$$\kappa(A) = \|A\|\|A^+\| = \|A\|\|(A^*A)^{-1}A^*\|, \quad A^+ \in \mathbb{C}^{n\times m}.$$

3.2 Floating Point Arithmetic

Computer uses binary system to represent real numbers. Some examples are

$$(1101.11)_2 = 2^3 + 2^2 + 2^0 + \frac{1}{2} + \frac{1}{4} = (13.75)_{10}.$$
$$(\underbrace{11\ldots11}_{n})_2 = 2^{n-1} + 2^{n-2} + \ldots + 2^1 + 2^0 = (2^n - 1)_{10}.$$

How exactly does one goes from decimal (base 10) to binary (base 2)?

- Suppose $x \in \mathbb{Z}_+$ in decimal, we divide by 2 and denote the remainder by a_0; this process continues until we reach 0 and

  $$x = (a_n a_{n-1} \ldots a_1 a_0)_2 = a_n \cdot 2^n + a_{n-1} \cdot 2^{n-1} + \ldots + a_1 \cdot 2^1 + a_0 \cdot 2^0.$$

  Let $x = 17$, then

  $$\begin{aligned}
  \frac{17}{2} &= 8 \ \text{ remainder } \ 1 \implies a_0 = 1.\\
  \frac{8}{2} &= 4 \ \text{ remainder } \ 0 \implies a_1 = 0.\\
  \frac{4}{2} &= 2 \ \text{ remainder } \ 0 \implies a_2 = 0.\\
  \frac{2}{2} &= 1 \ \text{ remainder } \ 0 \implies a_3 = 0.\\
  \frac{1}{2} &= 0 \ \text{ remainder } \ 1 \implies a_4 = 1.
  \end{aligned}$$

  Thus, $x = (10001)_2 = 17$.

- Suppose x has decimal digits now. We can write x in binary as follows:

  $$x = (0.a_1 a_2 a_3 \ldots)_2 = a_1 \cdot 2^{-1} + a_2 \cdot 2^{-2} + a_3 \cdot 2^{-3} + \ldots.$$

  where

  $$\begin{aligned}
  x_1 &= \text{frac}(2x) \quad \text{and} \quad a_1 = \text{Int}(2x).\\
  x_2 &= \text{frac}(2x_1) \quad \text{and} \quad a_2 = \text{Int}(2x_1).\\
  x_3 &= \text{frac}(2x_2) \quad \text{and} \quad a_3 = \text{Int}(2x_2).\\
  &\vdots \qquad\qquad \vdots \qquad\qquad \vdots
  \end{aligned}$$

  Take $x = 0.75$, then

  $$\begin{aligned}
  2x = 1.5 &\implies x_1 = \text{frac}(2x) = 0.5 \quad \text{and} \quad a_1 = \text{Int}(2x) = 1.\\
  2x_1 = 1.0 &\implies x_1 = \text{frac}(2x_1) = 0.0 \quad \text{and} \quad a_2 = \text{Int}(2x_1) = 1.
  \end{aligned}$$

  Thus, $x = (0.11)_2 = 0.75$.

For any nonzero decimal numbers x, express it in the following form:

$$x = \sigma \bar{x} \beta^e, \quad \text{where} \quad \sigma = \text{sign}(x) = \pm 1,$$

$$\begin{aligned}\beta &= \text{chosen base},\\ e &= \text{exponent},\\ \bar{x} &= \text{mantissa of } x, \text{ and } \quad (0.1)_\beta \le \bar{x} < 1.\end{aligned}$$

Observe that $(0.1)_{10} = 0.1$ for decimal while $(0.1)_2 = 0.5$ for binary. For example,

$$(12.462)_10 = 1 \cdot (0.12462) \cdot 10^2.$$
$$(1101.10111)_2 = 1 \cdot (0.110110111) \cdot 2^4.$$

There exists two types of *floating-point format*:

$$\text{Single-precision (32 bits)} : \sigma \quad \underbrace{\text{exponent}}_{8 \text{ bits}} \quad \underbrace{\text{mantissa}}_{23 \text{ bits}}$$
$$\text{Double-precision (64 bits)} : \sigma \quad \underbrace{\text{exponent}}_{11 \text{ bits}} \quad \underbrace{\text{mantissa}}_{52 \text{ bits}}$$

The exponent is stored as is if it is within the given range, otherwise the number is **overflow** if e is too large or **underflow** if e is too small.

1. An example of an overflow operation is $\sqrt{x^2+y^2}$ when x is large. To avoid this, we rewrite it as
$$\sqrt{x^2+y^2} = \begin{cases} |x| \left\{1 + \left(\dfrac{y}{x}\right)^2\right\}^{1/2} & \text{if } x > y, \\ |y| \left\{1 + \left(\dfrac{x}{y}\right)^2\right\}^{1/2} & \text{if } x < y. \end{cases}$$

2. An example of an underflow operation is $\sqrt{x+1} - \sqrt{x}$. Observe that the quantity is approximately 0 if x is large. To avoid this, we rationalise the function
$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

Definition 3.2.1. Suppose x has a representation of the form

$$x = \sigma \cdot (0.a_1 a_2 \ldots a_n a_{n+1} \ldots) \cdot 2^e,$$

but its floating point representation $\text{fl}(x)$ can only include n digits for the mantissa. There are two ways to truncate x when stored:

(a) **Chopping**, which amounts to truncating remaining digits after a_n,

(b) **Rounding**, based on the digit a_{n+1}:

$$\text{fl}(x) = \begin{cases} \sigma \cdot (0.a_1 a_2 \ldots a_{n-1} a_n) \cdot 2^e & \text{if } a_{n+1} = 0, \\ \sigma \cdot (0.a_1 a_2 \ldots a_{n-1} 1) \cdot 2^e & \text{if } a_{n+1} = 1. \end{cases}$$

- One can view the floating point representation $\text{fl}(x)$ as a perturbation of x, *i.e.* there exists an $\varepsilon = \varepsilon(x)$ such that

$$\text{fl}(x) = x(1+\varepsilon) \quad \text{or} \quad \frac{\text{fl}(x) - x}{x} = \varepsilon.$$

It can be shown that ε has certain range depending on the truncation method:

$$\text{Chopping} \quad : \quad -2^{-n+1} \leq \varepsilon \leq 0. \tag{3.2.1}$$

$$\text{Rounding} \quad : \quad -2^{-n} \quad \leq \varepsilon \leq 2^{-n}. \tag{3.2.2}$$

- Suppose chopping is used. Assuming $\sigma = 1$, we have that

$$\begin{aligned} 0 \leq x - \text{fl}(x) &= (0.\underbrace{0\ldots0}_{n} a_{n+1}a_{n+2}\ldots)_2 \cdot 2^e \\ &\leq (0.\underbrace{0\ldots0}_{n} 11\ldots)_2 \cdot 2^e \\ &= \left\{ \left(\frac{1}{2}\right)^{n+1} + \left(\frac{1}{2}\right)^{n+2} + \ldots \right\} \cdot 2^e \\ &= \left(\frac{1}{2}\right)^{n+1} \left(1 + \frac{1}{2} + \ldots\right) \cdot 2^e \\ &= \left(\frac{1}{2}\right)^{n+1} 2 \cdot 2^e = 2^{e+1-n-1} = 2^{-n+e}. \end{aligned}$$

  Thus,

$$0 \leq \frac{x - \text{fl}(x)}{x} \leq \frac{2^{-n+e}}{(0.a_1a_2\ldots)_2 \cdot 2^e} = \frac{2^{-n}}{(0.a_1a_2\ldots)_2} \leq \frac{2^{-n}}{2^{-1}} = 2^{-n+1}.$$

- Suppose rounding is used. A similar calculation as above shows that

$$\begin{aligned} 0 \leq |x - \text{fl}(x)| &\leq (0.\underbrace{0\ldots0}_{n-1} a_n a_{n+1}\ldots)_2 \cdot 2^e \\ &\leq \left(\frac{1}{2}\right)^{n} 2 \cdot 2^e = 2^{-n+e+1}. \end{aligned}$$

  Thus,

$$0 \leq \left|\frac{x - \text{fl}(x)}{x}\right| \leq \frac{2^{-n+e+1}}{(0.a_1a_2\ldots)_2 \cdot 2^e} \leq \frac{2^{-n+1}}{2^{-1}} = 2^{-n}.$$

- The worst possible error for chopping is twice as large as when rounding is used. It can be seen from (3.2.1), (3.2.2) that $x - \text{fl}(x)$ has the same sign as x for chopping but possibly different sign for rounding. This means that there might be cancellation of error if rounding is used!

Definition 3.2.2. The **machine epsilon**, denoted by $\varepsilon_{\text{machine}}$ is the difference between 1 and the next larger floating point number. In a relative sense, the machine epsilon is as large as the gaps between floating point number get. For a double-precision computer, $\varepsilon_{\text{machine}} = 2^{-52} \approx \mathcal{O}(10^{-16})$.

Axioms of Floating Point Arithmetic

1. For all $x \in \mathbb{R}$, there exists a floating point $\text{fl}(x)$ such that

$$\frac{|\text{fl}(x) - x|}{|x|} \leq \varepsilon_{\text{machine}}.$$

Equivalently, for all $x \in \mathbb{R}$, there exists an ε with $|\varepsilon| \leq \varepsilon_{\text{machine}}$ such that $\text{fl}(x) = x(1+\varepsilon)$. That is, the difference between a real number and its (closest) floating point approximation is always smaller than $\varepsilon_{\text{machine}}$ in relative terms.

2. Basic floating point operations consists of $\oplus, \ominus, \otimes, \oslash$. Denote the floating point operation by $\circledast$. For any floating points x, y, there exists an ε with $|\varepsilon| \leq \varepsilon_{\text{machine}}$ such that

$$x \circledast y = \text{fl}(x * y) = (x * y)(1+\varepsilon).$$

That is, every operation of floating point arithmetic is exact up to a relative error of size at most $\varepsilon_{\text{machine}}$.

Common sources of error include mathematical modelling of a physical problem, uncertainty in physical data, machine errors and truncation errors.

3.3 Stability

Definition 3.3.1. An algorithm $\tilde{f}$ for a problem f is **accurate** if for each $x \in X$,

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

In other words, there exists a constant $C > 0$ such that for all sufficiently small $\varepsilon_{\text{machine}}$ we have that

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq C\varepsilon_{\text{machine}}.$$

- In practice, C can be large. For ill-conditioned problems, the definition of accuracy can be too restrictive.

Definition 3.3.2.

1. An algorithm $\tilde{f}$ for a problem f is **stable** if for each $x \in X$,

$$\frac{\|\tilde{f}(x) - f(\tilde{x})\|}{\|f(\tilde{x})\|} = \mathcal{O}(\varepsilon_{\text{machine}})$$

for some $\tilde{x}$ with

$$\frac{\|\tilde{x} - x\|}{\|x\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

In words, a stable algorithm gives nearly the right answer to nearly the right question.

2. An algorithm $\tilde{f}$ for a problem f is **backward stable** if for each $x \in X$,

$$\tilde{f}(x) = f(\tilde{x}) \quad \text{for some } \tilde{x} \text{ with } \quad \frac{\|\tilde{x} - x\|}{\|x\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

In words, a backward stable algorithm gives exactly the right answer to nearly the right question.

Theorem 3.3.3. *For problems f and algorithms $\tilde{f}$ defined on finite-dimensional spaces X and Y, the properties of accuracy, stability and backward stability all hold or fail to hold independently of the choice of norms in X and Y.*

3.4 More on Stability

Theorem 3.4.1. *The four floating point operations $\oplus, \ominus, \otimes, \oslash$ are all backward stable.*

Proof. We will only prove this in the case of a subtraction. Consider the subtraction $f(x_1, x_2) = x_1 - x_2$, with floating point

$$\tilde{f}(x_1, x_2) = \text{fl}(x_1) \ominus \text{fl}(x_2).$$

From the first axiom of floating point arithmetic, there exists $\varepsilon_1, \varepsilon_2, \varepsilon_3$ with $|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \leq \varepsilon_{\text{machine}}$ such that

$$\begin{aligned} &\text{fl}(x_1) = x_1(1+\varepsilon_1), \quad \text{fl}(x_2) = x_2(1+\varepsilon_2), \\ &\text{fl}(x_1) \ominus \text{fl}(x_2) = (\text{fl}(x_1) - \text{fl}(x_2))(1+\varepsilon_3). \end{aligned}$$

Thus,

$$\begin{aligned} \tilde{f}(x_1, x_2) = \text{fl}(x_1) \ominus \text{fl}(x_2) &= [x_1(1+\varepsilon_1) - x_2(1+\varepsilon_2)](1+\varepsilon_3) \\ &= x_1(1+\varepsilon_4) - x_2(1+\varepsilon_5) \\ &= \tilde{x}_1 - \tilde{x}_2 = f(\tilde{x}_1, \tilde{x}_2) \end{aligned}$$

for some $|\varepsilon_4|, |\varepsilon_5| \leq 2\varepsilon_{\text{machine}} + \mathcal{O}(\varepsilon_{\text{machine}}^2)$. Backward stability follows directly since

$$\frac{|\tilde{x}_1 - x_1|}{|x_1|} = \mathcal{O}(\varepsilon_{\text{machine}}), \quad \frac{|\tilde{x}_2 - x_2|}{|x_2|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

■

Accuracy of a Backward Stable Algorithm

Theorem 3.4.2. *If a backward stable algorithm is used to solve a problem $f\colon X \longrightarrow Y$ with condition number κ, then the relative error satisfies the following estimates:*

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\kappa(x)\varepsilon_{machine}).$$

Proof. Since f is backward stable, $\tilde{f}(x) = f(\tilde{x})$ for some $\tilde{x} \in X$ satisfying

$$\frac{\|\tilde{x} - x\|}{\|x\|} = \mathcal{O}(\varepsilon_{\text{machine}}.$$

Definition of $\kappa(x)$ yields:

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \Big[\kappa(x) + o(1)\Big] \frac{\|\tilde{x} - x\|}{\|x\|},$$

where $o(1)$ denotes a quantity that converges to 0 as $\varepsilon_{\text{machine}} \longrightarrow 0$. The desired inequality follows from combining these bounds. ■

3.5 Stability of Back Substitution

Lower and upper triangular systems arise in QR factorisation, Gaussian elimination and Cholesky factorisation. These systems are easily solved by a process of successive substitution, called **forward substitution** if the system is lower-triangular and **back substitution** if the system is upper-triangular.

1. Given a non-singular, lower-triangular matrix $L \in \mathbb{R}^{m\times m}$, the solution to $Lx = b$ is given by

$$\begin{aligned} x_1 &= \frac{b_1}{l_{11}}, \\ x_i &= \frac{1}{l_{ii}}\left(b_1 - \sum_{j=1}^{i-1} l_{ij}x_j\right), \quad i = 2, \ldots, m. \end{aligned}$$

2. Given a non-singular, upper-triangular matrix $U \in \mathbb{R}^{m\times m}$, the solution to $Ux = b$ is given by

$$\begin{aligned} x_m &= \frac{b_m}{u_{mm}}, \\ x_i &= \frac{1}{u_{ii}}\left(b_i - \sum_{j=i+1}^{m} u_{ij}x_j\right), i = 1, \ldots, m-1. \end{aligned}$$

3. The operational count for both forward and backward substitution is $\sim m^2$ flops, since

$$\begin{aligned} \text{addition and substraction} &\sim \frac{m(m-1)}{2} \text{ flops.} \\ \text{multiplication and division} &\sim \frac{m(m+1)}{2} \text{ flops.} \end{aligned}$$

Theorem 3.5.1. *The backward substitution algorithm applied to $Ux = b$ is backward stable. The computed solution $\tilde{x} \in \mathbb{R}^m$ satisfies $(U + \delta U)\tilde{x} = b$, where the upper-triangular matrix $\delta U \in \mathbb{R}^{m\times m}$ satisfies*

$$\frac{\|\delta U\|}{\|U\|} = \mathcal{O}(\varepsilon_{machine}),$$

or for all i, j,

$$\frac{|\delta u_{ij}|}{|u_{ij}|} \leq m\varepsilon_{machine} + \mathcal{O}(\varepsilon^2_{machine}).$$

- What about its accuracy? With $\kappa(A) = \|A^{-1}\|\|A\|$,

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|} = \kappa(U)\frac{\|\delta U\|}{\|U\|} = \mathcal{O}(\kappa(U)\varepsilon_{\text{machine}}).$$

3.6 Problems

1. Assume that the matrix norm $\|\cdot\|$ satisfies the submultiplicative property $\|AB\| \leq \|A\|\|B\|$. Show that if $\|X\| < 1$, then $I - X$ is invertible, $(I - X)^{-1} = \sum_{j=0}^{\infty} X^j$ and $\|(I - X)^{-1}\| \leq 1/(1 - \|X\|)$.

Solution: This is a classical result about Neumann series, which is the infinite series $\sum_{j=0}^{\infty} X^j$. Assuming $(I - X)$ is invertible, with its inverse $(I - X)^{-1}$ given by the Neumann series, using the submultiplicative property and triangle inequality for norms we have that

$$\|(I - X)^{-1}\| = \left\| \sum_{j=0}^{\infty} X^j \right\| \leq \sum_{j=0}^{\infty} \|X\|^j = \frac{1}{1 - \|X\|} \tag{3.6.1}$$

where the second infinite series, which is a geometric series, converges since $\|X\| < 1$ by assumption. This proves the desired inequality and moreover it shows that the Neumann series $\sum_{j=0}^{\infty} X^j$ converges absolutely in the matrix norm, and thus converges in the matrix norm too. To conclude the proof, we need to show that $I - X$ is in fact invertible, with its inverse given by the Neumann series. A direct computation shows that

$$(I - X)\left(\sum_{j=0}^{n} X^j\right) = (I - X)(I + X + \ldots + X^n) = I - X^{n+1}. \tag{3.6.2}$$

Since the geometric series in (3.6.1) converges, $\|X\|^n \longrightarrow 0$ as $n \longrightarrow \infty$. It follows that $\|X^n - \mathbf{0}\| = \|X^n\| \leq \|X\|^n \longrightarrow 0$ as $n \longrightarrow \infty$. Taking limit of (3.6.2) as $n \longrightarrow \infty$ yields

$$\lim_{n\to\infty} (I - X)\left(\sum_{j=0}^{n} X^n\right) = (I - X)\left(\sum_{j=0}^{\infty} X^j\right) = I,$$

A symmetric argument also shows that $\left(\sum_{j=0}^{\infty} X^j\right)(I - X) = I$. Hence, we have that $(I - X)^{-1} = \sum_{j=0}^{\infty} X^j$, *i.e.* $I - X$ is invertible.

2. Let $A \in \mathbb{R}^{n\times n}$ be nonsingular matrix. Assume that $b \neq \mathbf{0}$, x satisfies $Ax = b$, and $\tilde{x}$ is an approximate solution to this linear system. Denote $e := x - \tilde{x}$ the error vector and $r := b - A\tilde{x}$ the residual vector. Show the following inequalities and explain their importance.

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}.$$

Solution: This is a relatively straightforward bound, using the assumption that A is non-singular so that A^{-1} exists, and the fact that we are working with an induced matrix norm, *i.e.* for any $x \in \mathbb{R}^n$, $\|Ax\| \leq \|A\|\|x\|$ and $\|A^{-1}x\| \leq \|A^{-1}\|\|x\|$. Since $x = A^{-1}b$, we have

$$\|x\| = \|A^{-1}b\| \leq \|A^{-1}\|\|b\|. \tag{3.6.3}$$

On the other hand, since $b = AA^{-1}b$, we have

$$\|r\| = \|b - A\tilde{x}\| = \|A(A^{-1}b - \tilde{x})\| = \|A(x - \tilde{x})\| = \|Ae\| \leq \|A\|\|e\|. \tag{3.6.4}$$

Combining (3.6.3), (3.6.4) and rearranging yields the left inequality. Next, since $Ax = b$, we have

$$\|b\| = \|Ax\| \leq \|A\|\|x\|. \tag{3.6.5}$$

On the other hand, since $Ax - b + b - A\tilde{x} = r$, we have

$$\|e\| = \|A^{-1}Ae\| = \|A^{-1}(Ae - b + b)\| = \|A^{-1}r\| \leq \|A^{-1}\|\|r\|. \tag{3.6.6}$$

Combining (3.6.5), (3.6.6) and rearranging yields the right inequality.

We know that $\kappa(A) = \|A\|\|A^{-1}\|$ is by definition the condition number of the matrix A; moreover $\kappa(A) \geq \|AA^{-1}\| = 1$. The terms $\|e\|/\|x\|, \|r\|/\|b\|$ can be interpreted as the relative solution error and the relative residual eror respectively. Thus, the right inequality

$$\frac{\|x - \tilde{x}\|}{\|x\|} = \frac{\|e\|}{\|x\|} \leq \kappa(A)\frac{\|r\|}{\|b\|} = \kappa(A)\frac{\|(b + r) - b\|}{\|b\|}$$

tells us that the ratio between the relative solution error and the relative residual error is controlled by the condition number of A. In other words, suppose $x_1 \in \mathbb{R}^n$ is such that $Ax_1 = b_1$ and suppose we perturb b_1 by some $\varepsilon > 0$. Then the corresponding solution can only differ from x_1 at most $\kappa(A)\varepsilon/\|b\|$ in relative terms.

This estimate also shows that if $\kappa(A)$ is not large then the residual r gives a good representation of the error e. However, if $\kappa(A)$ is large then the residual r is not a good estimate of the error e.

3. Suppose that $A \in \mathbb{R}^{n\times n}$ is non-singular, and consider the two problems:

$$Ax = b \qquad \text{and } (A + \delta A)\tilde{x} = b + \delta b,$$

where we assume that $\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1$, so that $(A + \delta A)$ is nonsingular (Why?). Show that

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right),$$

where $\kappa(A)$ is the condition number of the matrix.

Solution: Observe that since A is non-singular, we can rewrite $A+\delta A$ as

$$A+\delta A = A(I+A^{-1}\delta A) = A\Big[I-(-A^{-1}\delta A)\Big].$$

Since $\|-A^{-1}\delta A\| = \|A^{-1}\delta A\| < 1$, Problem 1 together with the assumption that A is also invertible shows that $A+\delta A$ is invertible, *i.e* $A+\delta A$ is non-singular. Since

$$A\tilde{x}-Ax = (A\tilde{x}-b)+(b-Ax) = A\tilde{x}-b = \delta b-\delta A\tilde{x},$$

we have that

$$\frac{\|\tilde{x}-x\|}{\|x\|} = \frac{\|A^{-1}(A\tilde{x}-Ax)\|}{\|x\|} = \frac{\|A^{-1}(\delta b-\delta A\tilde{x})\|}{\|x\|} \tag{3.6.7a}$$

$$\leq \frac{\|A^{-1}\|\delta b\|}{\|x\|} + \frac{\|A^{-1}\|\|\delta A\|\|\tilde{x}\|}{\|x\|}. \tag{3.6.7b}$$

Using $\kappa(A) = \|A^{-1}\|\|A\|$ and $\|b\| \leq \|A\|\|x\|$ yield the bound

$$\frac{\|A^{-1}\|\|\delta b\|}{\|x\|} = \frac{\kappa(A)\|\delta b\|}{\|A\|\|x\|} \leq \kappa(A)\frac{\|\delta b\|}{\|b\|}. \tag{3.6.8a}$$

$$\frac{\|A^{-1}\|\|\delta A\|\|\tilde{x}\|}{\|x\|} = \kappa(A)\left(\frac{\|\delta A\|}{\|A\|}\right)\left(\frac{\|\tilde{x}\|}{\|x\|}\right). \tag{3.6.8b}$$

Denote the following quantity

$$C = \frac{\|\tilde{x}-x\|}{\|x\|}, \quad D = \frac{\|\delta b\|}{\|b\|}, \quad E = \frac{\|\delta A\|}{\|A\|}. \tag{3.6.9}$$

Substituting (3.6.8) into (3.6.7) and using triangle inequality yield

$$\begin{aligned} C \leq \kappa(A)\left(D+E\frac{\|\tilde{x}\|}{\|x\|}\right) &\leq \kappa(A)\left(D+E\left(\frac{\|\tilde{x}-x\|+\|x\|}{\|x\|}\right)\right) \\ &= \kappa(A)\Big[D+E(C+1)\Big] \\ \implies \Big[1-\kappa(A)E\Big]C &\leq \kappa(A)[D+E] \\ \implies C &\leq \frac{\kappa(A)}{1-\kappa(A)E)}[D+E]. \end{aligned}$$

The desired inequality follows from substituting (3.6.9) into the above inequality.

4. Show that for Gaussian elimination with partial pivoting (permutation by rows) applied to a matrix $A \in \mathbb{R}^{n\times n}$, the growth factor $\rho = \dfrac{\max_{ij}|u_{ij}|}{\max_{ij}|a_{ij}|}$ satisfies the estimate $\rho \leq 2^{n-1}$.

Solution:

5. Show that if all the principal minors of a matrix $A \in \mathbb{R}^{n\times n}$ are nonzero, then there exists diagonal matrix D, unit lower triangular matrix L and unit upper triangular matrix U,

such that $A = LDU$. Is this factorisation unique? What happens if A is symmetric matrix?

Solution: Since all the principal minors of a matrix $A \in \mathbb{R}^{n\times n}$ are nonzero, there exists a unique LU decomposition such that L has unit diagonal entries. Note that U, an upper-triangular matrix itself, must have non-zero diagonal entries since all the principal minors of A including $\det(A)$ itself is non-zero. However, these diagonal entries might not be unit; fortunately this can be achieved with left-multiplying U by a diagonal matrix D with entries $d_{ii} = u_{ii}, i = 1, \ldots, n$. The consequence of doing this however is that we need to scale each rows u_i of U by u_{ii}, $i = 1, \ldots, n$, which of course is valid since $u_{ii} \neq 0$ for every $i = 1, \ldots, n$. More precisely,

$$
\begin{aligned}
U &= \begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ & u_{22} & u_{23} & \ldots & u_{2n} \\ & & u_{33} & \ldots & u_{3n} \\ & & & \ddots & \vdots \\ & & & & u_{nn} \end{bmatrix} \\
&= \begin{bmatrix} u_{11} & & & & \\ & u_{22} & & & \\ & & u_{33} & & \\ & & & \ddots & \\ & & & & u_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12}/u_{11} & u_{13}/u_{11} & \ldots & u_{1n}/u_{11} \\ & 1 & u_{23}/u_{22} & \ldots & u_{2n}/u_{22} \\ & & 1 & \ldots & u_{3n}/u_{33} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} = D\tilde{U}
\end{aligned}
$$

Since such LU decomposition and the way we factored out pivots of U are both unique, we conclude that there exists a unique LDU factorisation of A with all the desired properties for L, D, U.

If A is symmetric, then its LDU decomposition of the required form might not exist, and might not be unique even if it does exists. Consider $A = (a_{ij}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ which is symmetric and suppose we want to decompose A into the form

$$
\begin{aligned}
A &= \begin{bmatrix} 1 & & \\ a & 1 & \\ b & c & 1 \end{bmatrix} \begin{bmatrix} A & & \\ & B & \\ & & C \end{bmatrix} \begin{bmatrix} 1 & d & e \\ & 1 & f \\ & & 1 \end{bmatrix} \\
&= \begin{bmatrix} A & & \\ aA & B & \\ bA & cB & C \end{bmatrix} \begin{bmatrix} 1 & d & e \\ & 1 & f \\ & & 1 \end{bmatrix} \\
&= \begin{bmatrix} A & Ad & Ae \\ aA & aAd + B & aAe + Bf \\ bA & bAd + cB & bAe + cBf + C \end{bmatrix}.
\end{aligned}
$$

Comparing the first two diagonal entries a_{11}, a_{22} gives $A = B = 0$, but then

$$aAe + Bf = 0 \neq a_{23} = 1.$$

We see that an LDU decomposition of the required form for this particular matrix A does not exist. Next, consider $B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ which is symmetric. Observe that for any $\alpha, \beta \in \mathbb{R}$, B has infinite LDU decomposition as follows

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix}.$$

The issue seems to be that A is singular. If A is symmetric positive-definite (SPD) however, then all principal minors of A are nonzero and A has a unique LDU decomposition of the required form. It turns out that the LDU decomposition of a SPD matrix has a simpler form. Indeed, since $A = A^T$,

$$LDU = A = A^T = U^T D^T L^T = U^T D L^T.$$

Uniqueness of such decomposition implies $U^T = L$, *i.e.*

$$A = LDU = U^T DU = LDL^T.$$

Chapter 4

Systems of Equations

4.1 Gaussian Elimination

The goal is to solve $Ax = b$ by transforming the system into an upper-triangular one by applying simple linear transformations on the left. Consider a non-singular square matrix $A = A^{(1)} = (a_{ij}) \in \mathbb{R}^{n\times n}, b = b^{(1)} \in \mathbb{R}^n$

1. Assume $a_{11} \neq 0$. Introducing the **multipliers**

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, \ldots, n.$$

Subtracting multiples of first row from rows $2, \ldots n$ yields

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}, \qquad i, j = 2, \ldots, n. \\ b_i^{(2)} &= b_i^{(1)} - m_{i1}b_1^{(1)}, \qquad i = 2, \ldots, n. \end{aligned}$$

and

$$A^{(2)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}.$$

2. Assuming $a_{22}^{(2)} \neq 0$. Introducing the **multipliers**

$$m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}, \quad i = 3, \ldots, n.$$

Subtracting multiples of second row of $A^{(2)}$ from rows $3, \ldots, n$ yields

$$\begin{aligned} a_{ij}^{(3)} &= a_{ij}^{(2)} - m_{i2}a_{2j}^{(2)}, \qquad i, j = 3, \ldots, n. \\ b_i^{(3)} &= b_i^{(2)} - m_{i2}b_2^{(2)}, \qquad i = 3, \ldots, n. \end{aligned}$$

and

$$A^{(3)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \ldots & a_{nn}^{(3)} \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_n^{(3)} \end{bmatrix}.$$

3. Under the assumption that $a_{ii}^{(i)} \neq 0, i = 1, \ldots, k-1$, we will have $A^{(k)}x = b^{(k)}$, $k = 2, \ldots, n$, where

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & \ldots & \ldots & \ldots & a_{(1n)}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & \ldots & \ldots & a_{2n}^{(2)} \\ & & \ddots & \ddots & \vdots & \vdots & \vdots \\ & & & a_{kk}^{(k)} & a_{k(k+1)}^{(k)} & \ldots & a_{kn}^{(k)} \\ & & & \vdots & \vdots & \vdots & \vdots \\ & & & a_{nk}^{(k)} & \ldots & \ldots & a_{nn}^{(k)} \end{bmatrix}, \quad b^{(k)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix},$$

where

$$\begin{aligned} m_{i(k-1)} &= \frac{a_{i(k-1)}^{(k-1)}}{a_{(k-1)(k-1)}^{(k-1)}}, & i &= k, \ldots, n. \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{i(k-1)} a_{(k-1)j}^{(k-1)}, & i, j &= k, \ldots, n. \\ b_i^{(k)} &= b_i^{(k-1)} - m_{i(k-1)} b_{k-1}^{(k-1)}, & i &= k, \ldots, n. \end{aligned}$$

Finally, $A^{(n)} = U$ is an upper-triangular matrix. By setting $L = M = (m_{ij})$, where m_ij is as in above for $j < i$, we obtain the LU decomposition.

4. A sufficient condition for the pivots $a_{kk}^{(k)} \neq 0, k = 1, \ldots, n-1$ is that the matrix A be symmetric positive-definite. In basic form without pivoting, the Gaussian Elimination Method (GEM) can be excuted in general on

   - strictly diagonally dominant (SDD) matrices,
   - symmetric positive-definite (SPD) matrices.

Example 4.1.1. Consider the Hilbert matrix

$$A^{(1)} = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}, \qquad b = \begin{bmatrix} 11/6 \\ 13/12 \\ 47/60 \end{bmatrix}.$$

This is a classic example of an ill-conditioned matrix. It motivates the LU decomposition.

Example 4.1.2. If one needs to solve $Ax = b$ for different b's, factor $A = LU$ and store L, U for multiple uses. Consider $A = A^{(1)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$. Then

$$M_1 = \begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix} \implies A^{(2)} = M_1 A^{(1)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix}.$$

$$M_2 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -3 & 1 & \\ & -4 & & 1 \end{bmatrix} \Longrightarrow A^{(3)} = M_2 A^{(2)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix}.$$

$$M_3 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -1 & 1 \end{bmatrix} \Longrightarrow A^{(4)} = M_3 A^{(3)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix}.$$

Thus, $U = A^{(4)} = M_3 M_2 M_1 A^{(1)} \implies A = LU = (M_3 M_2 M_1)^{-1} U$ and

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix} = LU.$$

General Idea

Define $m_k = [0, \ldots, 0, m_{k+1k}, \ldots, m_{nk}]^T \in \mathbb{R}^n$ and consider the ***k*th Gaussian transformation matrix** M_k defined by

$$M_k = I_n - m_k e_k^T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{k+1k} & 1 & \\ & & \vdots & & \ddots \\ & & -m_{nk} & & 1 \end{bmatrix}.$$

where e_k is the kth canonical basis vector in $\mathbb{R}^n$. Component-wise, we have

$$(M_k)_{ip} = \delta_{ip} - (m_k e_k^T)_{ip} = \delta_{ip} - m_{ik}\delta_{kp}, \quad 1 \le i, p \le n.$$

To obtain $A^{(k+1)}$ from $A^{(k)}$,

$$\begin{aligned} a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} &= a_{ij}^{(k)} - m_{ik}\delta_{kk} a_{kj}^{(k)} \\ &= \sum_{p=1}^{n} (\delta_{ip} - m_{ik}\delta_{kp}) a_{pj}^{(k)} \\ &= \sum_{p=1}^{n} (M_k)_{ip} a_{pj}^{(k)} = (M_k A^{(k)})_{ij}. \end{aligned}$$

Hence, $A^{(k)} \longrightarrow A^{(k+1)}$ is given by $A^{(k+1)} = M_k A^{(k)}$.

- The inverse of M_k is $M_k^{-1} = I_n + m_k e_k^T$. Indeed, $e_k^T m_k = 0$ since m_k has nonzero entries starting from $k+1$, $k = 1, \ldots, n-1$. Thus,

$$M_k M_k^{-1} = (I_n - m_k e_k^T)(I_n + m_k e_k^T) = I_n - m_k e_k^T m_k e_k^T = I_n.$$

- Choose $L = (M_{n-1} \dots M_1)^{-1} = M_1^{-1} \dots M_{n-1}^{-1}$. A similar reason shows that $e_k^T m_{k+1} = 0$ for all $k = 1, \dots, n-2$. Thus,

$$M_k^{-1} M_{k+1}^{-1} = (I_n + m_k e_k^T)(I_n + m_{k+1} e_{k+1}^T) = I_n + m_k e_k^T + m_{k+1} e_{k+1}^T.$$

$$\implies \prod_{j=1}^{n-1} M_j^{-1} = I_n + \sum_{j=1}^{n-1} m_j e_k^T = \begin{bmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ m_{n1} & m_{n2} & \dots & m_{nn-1} & 1 \end{bmatrix}.$$

We now can solve the system $Ax = b$ using GEM.

1. Compute the LU factorisation of A using GEM. This requires

$$\frac{2(n-1)n(n+1)}{3} + n(n-1) \sim \frac{2}{3} n^3 \text{ flops.}$$

2. Solve for $y \in \mathbb{R}^n$ the lower-triangular system $Ly = b$, using forward substitution. This requires $\sim n^2$ flops.

3. Solve for $x \in \mathbb{R}^n$ the upper-triangular system $Ux = y$, using backward substitution. This requires $\sim n^2$ flops.

Theorem 4.1.3. *Let $A \in \mathbb{R}^{n \times n}$. The LU factorisation of A with $l_{ii} = 1, i = 1, \dots, n$ exists and is unique if and only if the ith order leading principal submatrix A_i of A, $i = 1, \dots, n-1$ are non-singular.*

- If $a_{11} = 0$, the LU decomposition as how we defined above does not exist.

- The ith order leading principal submatrix A_i is constructed by the first i rows and i columns of A. Its determinant is called leading dominating minors.

- If A_i is singular, then LU factorisation (with $l_{ii} = 1$) may not exist, or will not be unique. We demonstrate this with the following examples:

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 2-\beta \end{bmatrix}.$$

Proof. We begin by proving the "if" direction. By using induction, we want to show that if $\det(A_i) \neq 0, i = 1, \dots, n-1$, then the LU factorisation of A_i (as defined above) exists and is unique. The case $i = 1$ is trivial since $a_{11} \neq 0$. Suppose the case $(i-1)$ is true, there exists a unique LU decomposition of A_{i-1} such that

$$A_{i-1} = L^{(i-1)} U^{(i-1)}, \quad \text{with} \quad l_{kk}^{(i-1)} = 1, k = 1, \dots, i-1.$$

We look for a factorisation of the form

$$\begin{bmatrix} A_{i-1} & c \\ d^T & a_{ii} \end{bmatrix} = A_i = L^{(i)} U^{(i)} = \begin{bmatrix} L^{(i-1)} & \mathbf{0} \\ l^T & 1 \end{bmatrix} \begin{bmatrix} U^{(i-1)} & u \\ \mathbf{0}^T & u_{ii} \end{bmatrix}.$$

where $\mathbf{0}, l, u, c, d \in \mathbb{R}^{i-1}$. Note that $l_{ii}^{(i)} \neq 0$ since $\det(A_i) \neq 0$. Comparing terms in the factorisation yields

$$L^{(i-1)}u = c, \quad l^T U^{(i-1)} = d^T, \quad l^T u + u_{ii} = a_{ii}. \tag{4.1.1}$$

Since $\det(A_{i-1}) \neq 0$ by induction assumption, we also have $\det(L^{(i-1)}), \det(U^{(i-1)}) \neq 0$. Thus, there exists a unique u, l, u_{ii} solving (4.1.1).

Conversely, assume there exists a unique LU factorisation $A = LU$, with $l_{ii} = 1, i = 1, \ldots, n$. There are two separate cases to consider:

1. A is non-singular. Recall that for every $i = 1, \ldots, n$, A_i has an LU factorisation of the form
$$A_i = L^{(i)}U^{(i)} = \begin{bmatrix} L^{(i-1)} & \mathbf{0} \\ l^T & 1 \end{bmatrix} \begin{bmatrix} U^{(i-1)} & u \\ \mathbf{0}^T & u_{ii} \end{bmatrix}.$$
Thus, $\det(A_i) = \det(L^{(i)}) \det(U^{(i)}) = u_{11}u_{22} \ldots u_{ii}, i = 1, \ldots, n$. In particular, $\det(A_n) = u_{11} \ldots u_{nn}$; but since A is non-singular, $u_{ii} \neq 0$ for all $i = 1, \ldots, n$. Hence, we must have $\det(A_i) \neq 0$ for every $i = 1, \ldots, n$.

2. A is singular. Analysis above shows that U must have at least one zero entry on the main diagonal. Let u_{kk} be the first zero entry of U on the main diagonal. LU factorisation process then breaks down at $(k+1)$th step, because then l^T will not be unique due to U^k being singular (refer to (4.1.1)). In other words, if $u_{kk} = 0$ for some $k \leq n-1$, then we loose existence and uniqueness of LU factorisation at $(k+1)$th step. Hence, in order to have a unique LU factorisation of A, we must have $u_{jj} \neq 0$ for every $j = 1, \ldots, n-1$ and $u_{nn} = 0$.

■

We provide a simple algorithm for the Gaussian elimination method without pivoting. This pseudocode is not optimal, in the sense that both matrices U and M can be stored in the same array as A.

$$\begin{aligned}
&U = A, L = I. \\
&\text{for } k = 1 \text{ to } n-1 \\
&\qquad \text{for } j = k+1 \text{ to } n \\
&\qquad\qquad m_{jk} = \frac{u_{jk}}{u_{kk}} \\
&\qquad\qquad u_{j,k:n} = u_{j,k:n} - m_{jk}u_{k,k:n}
\end{aligned}$$

4.2 Pivoting

We begin by exploring the cruel fact that Gaussian elimination method without pivoting is neither stable nor backward stable, mainly due to sensitivity of rounding errors. Fortunately, this instability can be rectified by permutating the order of the rows of the matrix in a certain way! This operation is called **pivoting**.

Motivation for Pivoting

Example 4.2.1. Consider the following 2×2 linear system $Ax = b$ given by

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The obvious solution is to interchange rows. Now suppose we perturb a_{11} by some small number $\varepsilon > 0$ so that

$$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Performing GEM yields

$$\begin{bmatrix} \varepsilon & 1 \\ 0 & 1-1/\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2-1/\varepsilon \end{bmatrix} \implies x_2 = \frac{2-1/\varepsilon}{1-1/\varepsilon} \approx 1, \quad x_1 = \frac{1-x_2}{\varepsilon} \approx 0.$$

However, the actual solution is given by $x_2 = \dfrac{1-2\varepsilon}{1-\varepsilon} \approx 1, x_1 = \dfrac{1}{1-\varepsilon} \approx 1 \neq 0$. If we interchange rows, we have

$$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \implies \begin{bmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1-2\varepsilon \end{bmatrix}.$$

The solution is given by $x_2 = \dfrac{1-2\varepsilon}{1-\varepsilon} \approx 1, x_1 = 2 - x_2 \approx 1$.

Example 4.2.2. Consider the following 2×2 linear system $Ax = b$ given by

$$\begin{bmatrix} 1 & 1/\varepsilon \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\varepsilon \\ 2 \end{bmatrix}.$$

Performing GEM yields

$$\begin{bmatrix} 1 & 1/\varepsilon \\ 0 & 1-1/\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\varepsilon \\ 2-1/\varepsilon \end{bmatrix} \implies x_2 = \frac{2-1/\varepsilon}{1-1/\varepsilon} \approx 1, \quad x_1 = \frac{1}{\varepsilon} - \frac{1}{\varepsilon} x_2 \approx 0.$$

However, the actual solution is given by $x_1 = \dfrac{1}{1-\varepsilon} \approx 1 \neq 0$.

Main Idea

We demonstrate the main idea with a simple example. Consider $A = A^{(1)} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix}$.

$$\tilde{A}^{(1)} = P_1 A^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix}.$$

$$A^{(2)} = M_1 \tilde{A}^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -7 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{bmatrix}.$$

$$\tilde{A}^{(2)} = P_2 A^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & 1 \end{bmatrix}.$$

$$A^{(3)} = M_2 \tilde{A}^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, $M_2 P_2 M_1 P_1 A^{(1)} = U$. Define $P = P_1 P_2$ and $M = P_2 P_2 M_1 P_1$. M doesn't look lower-triangular at all, but $L = M^{-1} P$ happens to be lower-triangular and we have $PA = M^{-1} PU$.

4.3 Stability of Gaussian Elimination

4.4 Cholesky Factorisation

Chapter 5

Iterative Methods For Linear Systems

5.1 Consistent Iterative Methods and Convergence

The main idea of **iterative method** is to generate a sequence of vectors $(x^{(k)})_{k=1}^{\infty}$ with the property

$$x = \lim_{k\to\infty} x^{(k)},$$

where x is the true solution to $Ax = b$. In practice, we impose the following stopping criteria:

$$\|x^{(n)} - x\| < \varepsilon \quad \text{or} \quad \|x^{(n+1)} - x^{(n)}\| < \varepsilon,$$

where $\varepsilon > 0$ is some fixed tolerance. We could also look at the residual norm and demand

$$\|r^{(k)}\| = \|b - Ax^{(k)}\| < \varepsilon.$$

Let $e^{(k)} = x^{(k)} - x$ be the **error vector** at the kth step of the iteration process. We have the following relation:

$$x = \lim_{k\to\infty} x^{(k)} \iff \lim_{k\to\infty} e^{(k)} = \mathbf{0}.$$

Definition 5.1.1. Given some initial guess $x^{(0)} \in \mathbb{R}^n$, consider iterative methods of the form

$$x^{(k+1)} = Bx^{(k)} + f, \quad k \geq 0, \quad \text{where } B = n \times n \text{ iteration matrix}, \tag{5.1.1a}$$

$$f = \text{some } n\text{-vector obtained from } b. \tag{5.1.1b}$$

An iterative method of the form (5.1.1) is said to be **consistent** with the linear system $Ax = b$ if f and B are such that $x = Bx + f$.

Example 5.1.2. Observe that consistency of (5.1.1) does not imply its convergence. Consider the linear system $2Ix = b$. It is clear that the iterative method defined below is consistent:

$$x^{(k+1)} = -x^{(k)} + b.$$

However, this method is not convergent for every choice of initial guess $x^{(0)}$. Indeed, choosing $x^{(0)}$ gives

$$x^{(2k)} = \mathbf{0}, \quad x^{(2k+1)} = b \quad , k \geq 0.$$

On the other hand, the proposed iterative method converges to the true solution if $x^{(0)} = b/2$.

Let $e^{(k)} = x - x^{(k)}$. Subtracting the consistency equation $x = Bx + f$ from the iterative method (5.1.1) yields the recurrence relation for the error equation

$$e^{(k+1)} = Be^{(k)} = B^2e^{(k-1)} = \ldots\ldots = B^{(k+1)}e^{(0)} \quad \text{for each } k \geq 0.$$

In order for $e^{(k)} \longrightarrow 0$ as $k \longrightarrow \infty$ for any choices of $e^{(0)}$, we require $B^k \longrightarrow \mathbf{0}$ as $k \longrightarrow \infty$ and not surprisingly, this depends on the the magnitude of the largest eigenvalue of B.

Definition 5.1.3. Let $A \in \mathbb{C}^{n\times n}$ be a square matrix. A nonzero vector $x \in \mathbb{C}^n$ is an **eigenvector** of A, and $\lambda \in \mathbb{C}$ is its corresponding **eigenvalue** if $Ax = \lambda x$. The set of all eigenvalues of A is the spectrum of A, denoted by $\sigma(A)$.

Theorem 5.1.4. *Given any square matrix $A \in \mathbb{C}^{n\times n}$,*

$$\lim_{m\to\infty} A^m = \mathbf{0} \iff \rho(A) < 1,$$

where $\rho(A)$ is the spectral radius of A defined by

$$\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda|.$$

Proof. The result is trivial if $A = \mathbf{0}$, so suppose not. Suppose $\lim_{m\to\infty} A^m = \mathbf{0}$. Choose any $\lambda \in \sigma(A)$ with corresponding eigenvector $x \neq \mathbf{0}$. Since $A^m x = \lambda^m x$,

$$\lim_{m\to\infty} \lambda^m x = \lim_{m\to\infty} A^m x$$
$$\left(\lim_{m\to\infty} \lambda^m\right) x = \left(\lim_{m\to\infty} A^m\right) x = \mathbf{0}.$$

Since $x \neq \mathbf{0}$, it follows that

$$\lim_{m\to\infty} \lambda^m = 0 \implies |\lambda| < 1,$$

and this proves the only if statement since $\lambda \in \sigma(A)$ was arbitrary.

Conversely, suppose $\rho(A) < 1$. By continuity of norm and the fact that any norms are equivalent in finite-dimensional vector space, it suffices to prove that

$$\lim_{m\to\infty} \|A^m\| = 0.$$

Consider the Schur decomposition of A given by $A = QTQ^*$ (see Theorem 6.1.11) where Q is unitary and $T = D + U$ is upper-triangular, with D the diagonal matrix with eigenvalues of A on its diagonal and U the nilpotent matrix, *i.e.* there exists an $N > 0$ such that $U^m = \mathbf{0}$ for $m \geq N$. Since the 2-norm is invariant under unitary transformation, for m much larger than N we have that

$$\begin{aligned}\|A^m\|_2 = \|(D+U)^m\|_2 &\leq \sum_{k=0}^{m} \left(\frac{m!}{k!(m-k)!}\right) \|D\|_2^{m-k}\|U\|_2^k \\ &= \sum_{k=0}^{N-1} \left(\frac{m(m-1)\ldots(m-k+1)}{k!}\right) \|D\|_2^{m-k}\|U\|_2^k\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{k=0}^{N-1} \left(\frac{m^k}{k!}\right) \|D\|_2^m \left(\frac{\|U\|_2}{\|D\|_2}\right)^k \\
&\leq m^{N-1}\|D\|_2^m \left(\sum_{k=0}^{N-1} \left(\frac{\|U\|_2}{\|D\|_2}\right)^k\right) \\
&= Cm^{N-1}\rho(A)^m,
\end{aligned}$$

where C is independent of m. Let $a_m = m^{N-1}\rho(A)^m$. Since

$$\lim_{m\to\infty}\left(\frac{a_{m+1}}{a_m}\right) = \lim_{m\to\infty}\left(\frac{m+1}{m}\right)^{N-1}\rho(A) = \rho(A) < 1,$$

the sequence (a_m) converges to 0 as $m \to \infty$ by the **Ratio Test for sequences**. Consequently,

$$0 \leq \lim_{m\to\infty}\|A^m\| \leq C \lim_{m\to\infty}\left(m^{N-1}\rho(A)^m\right) = 0,$$

and the if statement follows.

■

Theorem 5.1.5. *Let (5.1.1) be a consistent iterative method. Then its iterates* $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ *converges to the solution of* $Ax = b$ *for any choice of initial guess* $x^{(0)}$ *if and only if* $\rho(B) < 1$.

Proof. The if statement follows from applying Theorem 5.1.4 to the error equation. To prove the only if statement, suppose $\rho(B) \geq 1$. There exists $\lambda \in \sigma(B)$ such that $|\lambda| \geq 1$ with corresponding eigenvector $x \neq 0$. Choosing $e^{(0)} = x$ yields

$$e^{(k)} = B^k e^{(0)} = \lambda^k e^{(0)} \not\longrightarrow \mathbf{0} \text{ as } k \longrightarrow \infty.$$

■

Remark 5.1.6. A sufficient but not necessary condition for convergence of consistent iterative method is $\|B\| < 1$ for any consistent matrix norm, since $\rho(B) \leq \|B\|$. The rate of convergence depends on how much less that 1 the spectral radius is. The smaller it is, the faster the convergence.

5.2 Linear Iterative Methods

A common approach to devise consistent iterative methods is based on an **additive splitting** of the matrix A. More precisely, writing $A = P - N$ where P is nonsingular, we obtain the consistent iterative method

$$Px^{(k+1)} = Nx^{(k)} + b \implies x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b = Bx^{(k)} + f.$$

where $B = P^{-1}N$ and $f = P^{-1}b$.

Example 5.2.1. Consider solving the following matrix equation

$$Ax = \begin{bmatrix} 7 & -6 \\ -8 & 9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix} = b,$$

Choosing $P = \text{diag}(A)$ and $N = P - A$, we propose the following method

$$x^{(k+1)} = \begin{bmatrix} 0 & 6/7 \\ 8/9 & 0 \end{bmatrix} x^{(k)} + \begin{bmatrix} 3/7 \\ -4/9 \end{bmatrix} = Bx^{(k)} + f.$$

Another way to deduce this is by rewriting the system of linear equations in the form

$$\begin{cases} x_1 = \dfrac{6}{7}x_2 + \dfrac{3}{7} \\ x_2 = \dfrac{8}{9}x_1 - \dfrac{4}{9}. \end{cases}$$

5.2.1 Jacobi Method

Jacobi method is applicable when $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant. By strictly diagonally dominant we mean that

$$|a_{ii}| > \sum_{j\neq i}^{n} |a_{ij}|.$$

Splitting $A = D + R$, where

$$D = \text{diag}(A) = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} \quad \text{and} \quad R = A - D = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}.$$

In matrix form, we have that

$$\begin{aligned} Dx &= -Rx + b \\ x^{(k+1)} &= \underbrace{-D^{-1}R}_{B} x^{(k)} + \underbrace{D^{-1}b}_{f}. \end{aligned}$$

Component-wise, the Jacobi iterative method has the form

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j\neq i}^{n} a_{ij} x_j^{(k)} \right), \quad i = 1, \dots, n. \tag{Jacobi}$$

This says that $x^{(k+1)}$ is found using $x^{(k)}$ only.

5.2.2 Gauss-Siedel Method

Gauss-Siedel method is applicable when $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant, and it is an improvement of the Jacobi method. More precisely, at the $(k+1)$th iteration, the readily computed values of $x_i^{(k+1)}$ are used to update the solution. This suggests splitting $A = D+L+U$, where

$$D = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}, \quad L = \begin{bmatrix} 0 & & & \\ a_{21} & 0 & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ & 0 & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & 0 \end{bmatrix}.$$

In matrix form, we have that

$$\begin{aligned}(D+L)x^{(k+1)} &= -Ux^{(k)} + b \\ x^{(k+1)} &= \underbrace{-(D+L)^{-1}U}_{B}\, x^{(k)} + \underbrace{(D+L)^{-1}b}_{f}.\end{aligned}$$

Component-wise, the Gauss-Siedel iterative method has the form

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1,\dots,n. \tag{Gauss-Siedel}$$

In the matrix form of Gauss Siedel, we claim that $(D+L)^{-1}$ exists. This is true because $(D+L)$ is strictly diagonally dominant by rows.

Theorem 5.2.2. *Strictly diagonally dominant (SDD) matrices are invertible.*

Proof. Suppose by contradiction that $A \in \mathbb{R}^{n\times n}$ is a strictly diagonally dominant matrix that is singular. There exists a nonzero $x \in \mathbb{R}^n$ such that $Ax = \mathbf{0}$. Let $J \in \{1,\dots,n\}$ be such that

$$|x_J| = \max_{j=1,\dots,n} |x_j|.$$

Expanding the Jth component of Ax yields

$$\begin{aligned}0 = (Ax)_J = \sum_{j=1}^{n} a_{Jj}x_j \implies a_{JJ} &= -\sum_{j\neq J}^{n} a_{Jj}\frac{x_j}{x_J} \\ \implies |a_{JJ}| \le \sum_{j\neq J}^{n} |a_{Jj}|\left|\frac{x_j}{x_J}\right| &\le \sum_{j\neq J}^{n} |a_{Jj}|,\end{aligned}$$

contradicting the assumption that A is strictly diagonally dominant.

■

Theorem 5.2.3. *If A is strictly diagonally dominant by rows, then the Jacobi and Gauss-Siedel methods are convergent.*

Proof. Choose any $\lambda \in \sigma(B)$ with corresponding eigenvector $x \neq \mathbf{0}$, where $B = -D^{-1}R$ is the iteration matrix of the Jacobi method. Rearranging $Bx = \lambda x$ yields

$$\begin{aligned}-D^{-1}Rx &= \lambda x \\ -Rx &= \lambda Dx \\ -\sum_{j\neq i}^{n} a_{ij}x_j &= \lambda a_{ii}x_i, \quad i = 1,\dots,n.\end{aligned}$$

WLOG, assume $\|x\|_\infty = 1$. Let i be the index such that

$$|x_j| \le |x_i| = 1 \quad \text{for all } j \neq i.$$

It follows that

$$|\lambda||a_{ii}| \le \sum_{j\neq i}^{n} |a_{ij}| \implies |\lambda| \le \sum_{j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

since A is SDD by rows from assumption. Since $\lambda \in \sigma(B)$ was arbitrary, this gives $\rho(B) < 1$ and the Jacobi method is convergent.

A similar argument with the iteration matrix in the Gauss-Siedel method $B = -(D+L)^{-1}U$ gives

$$\begin{aligned} -(D+L)^{-1}Ux = \lambda x &\implies -Ux &&= \lambda(D+L)x \\ &\implies \lambda Dx &&= -(\lambda L + U)x \\ &\implies \lambda a_{ii}x_i = -\lambda \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j, &&\quad i = 1, \dots, n. \end{aligned}$$

WLOG, assume $\|x\|_\infty = 1$. Let i be the index such that

$$|x_j| \leq |x_i| = 1 \quad \text{for all } j \neq i.$$

This together with the triangle inequality yields

$$\begin{aligned} |\lambda||a_{ii}| &\leq |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \\ \implies |\lambda| &\leq \frac{\displaystyle\sum_{j=i+1}^{n} |a_{ij}|}{\displaystyle |a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1, \end{aligned}$$

since A is SDD by rows from assumption. Since $\lambda \in \sigma(B)$ was arbitrary, this gives $\rho(B) < 1$ and the Gauss-Siedel method is convergent.

■

5.2.3 Successive Over Relaxation (SOR) Method

This is a variant of the Gauss-Siedel method that results in faster convergence by introducing a **relaxation parameter** $\omega \neq 0$. Splitting $A = D + L + U$ as in the Gauss-Siedel method, we can rewrite the linear system $Ax = b$ as follows

$$\begin{aligned} (D+L+U)x &= b \\ Dx &= b - Lx - Ux \\ x &= D^{-1}(b - Lx - Ux) \\ \omega x &= \omega D^{-1}(b - Lx - Ux) \\ x &= \omega D^{-1}(b - Lx - Ux) + (1-\omega)x. \end{aligned}$$

Component-wise, the SOR method has the form

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right) + (1-\omega)x_i^{(k)}, \quad i = 1, \dots, n. \tag{SOR}$$

In matrix form, we have that

$$\begin{aligned} x^{(k+1)} &= \omega D^{-1}(b - Lx^{(k+1)} - Ux^{(k)}) + (1-\omega)x^{(k)} \\ (D+\omega L)x^{(k+1)} &= \Big[(1-\omega)D - \omega U\Big]x^{(k)} + \omega b \\ x^{(k+1)} &= (D+\omega L)^{-1}\Big[(1-\omega)D - \omega U\Big]x^{(k)} + \omega(D+\omega L)^{-1}b \\ &= (D+\omega L)^{-1}\Big[D - \omega(D+U)\Big]x^{(k)} + \omega(D+\omega L)^{-1}b \\ &= B_\omega x^{(k)} + f_\omega. \end{aligned}$$

For $\omega = 1$, we recover the Gauss-Siedel method. For $\omega \in (0,1)$, the method is called **under-relaxation**; for $\omega > 1$, the method is called **over-relaxation**. Clearly there exists an optimal parameter ω_0 that produces the smallest spectral radius.

Theorem 5.2.4.

(a) If A is symmetric positive definite (SPD), then the SOR method is convergent if and only if $0 < \omega < 2$.

(b) If A is SDD by rows, then the SOR method is convergent if $0 < \omega \leq 1$.

(c) If $A, 2D - A$ are SPD, then the Jacobi method is convergent.

(d) If A is SPD, then the Gauss-Siedel method is convergent.

We can extrapolate the idea of a relaxation parameter to general consistent iterative methods (5.1.1). This results in a consistent iterative method for any $\gamma \neq 0$

$$x^{(k+1)} = \gamma(Bx^{(k)} + f) + (1-\gamma)x^{(k)},$$

where upon rearranging yields

$$x^{(k+1)} = \Big[\gamma B + (1-\gamma)\Big]x^{(k)} + \gamma f = B_\gamma x^{(k)} + f_\gamma.$$

From **Spectral Mapping Theorem**, it follows that if $\lambda \in \sigma(B)$, then $\gamma\lambda + (1-\gamma) \in \sigma(B_\gamma)$.

5.3 Iterative Optimisation Methods

In this section, we reformulate the linear system $Ax = b$ as a quadratic minimisation problem, in the case where $A \in \mathbb{R}^{n\times n}$ is symmetric positive-definite (SPD).

Lemma 5.3.1. *If $A \in \mathbb{R}^{n\times n}$ is SPD, solving $Ax = b$ is equivalent to minimising the quadratic form*

$$q(x) = \frac{1}{2}\langle x, Ax\rangle - \langle x, b\rangle.$$

Proof. Suppose x is a minimiser of $q(x)$, the first variation of $q(x)$ must equal to 0. More precisely, for any $v \in \mathbb{R}^n$ we must have

$$\lim_{\varepsilon \to 0} \frac{q(x+\varepsilon v) - q(x)}{\varepsilon} = 0.$$

Since A is symmetric, expanding $q(x+\varepsilon v)$ yields

$$\begin{aligned} q(x+\varepsilon v) &= \frac{1}{2}\langle x+\varepsilon v, A(x+\varepsilon v)\rangle - \langle x+\varepsilon v, b\rangle \\ &= \frac{1}{2}\langle x, Ax\rangle + \frac{\varepsilon}{2}\langle x, Av\rangle + \frac{\varepsilon}{2}\langle v, Ax\rangle + \frac{\varepsilon^2}{2}\langle v, Av\rangle - \langle x, b\rangle - \varepsilon\langle v, b\rangle \\ &= q(x) + \varepsilon\Big[\langle v, Ax\rangle - \langle v, b\rangle\Big] + \frac{\varepsilon^2}{2}\langle v, Av\rangle, \end{aligned}$$

which upon rearranging gives

$$\begin{aligned} 0 = \lim_{\varepsilon \to 0} \frac{q(x+\varepsilon v) - q(x)}{\varepsilon} &= \lim_{\varepsilon \to 0} \Big(\langle v, Ax\rangle - \langle v, b\rangle + \frac{\varepsilon}{2}\langle v, Av\rangle\Big) \\ &= \langle v, Ax\rangle - \langle v, b\rangle \\ &= \langle v, Ax - b\rangle \quad \text{for all } v \in \mathbb{R}^n. \end{aligned}$$

Choosing $v = Ax - b$, we have that $\|Ax-b\|_2^2 = 0 \implies Ax = b$. Observe that its mininum is given by

$$q(x) = \frac{1}{2}\langle A^{-1}b, b\rangle - \langle A^{-1}b, b\rangle = -\frac{1}{2}\langle A^{-1}b, b\rangle < 0.$$

Conversely, suppose $Ax = b$. For any $v \in \mathbb{R}^n$, we have

$$\begin{aligned} q(v) &= q(x+w) \\ &= \frac{1}{2}\langle x+w, A(x+w)\rangle - \langle x+w, b\rangle \\ &= \frac{1}{2}\langle x, Ax\rangle + \frac{1}{2}\langle x, Aw\rangle + \frac{1}{2}\langle w, Ax\rangle + \frac{1}{2}\langle w, Aw\rangle - \langle x, b\rangle - \langle w, b\rangle \\ &= q(x) + \Big(\langle w, Ax\rangle - \langle w, b\rangle\Big) + \frac{1}{2}\langle w, Aw\rangle \\ &= q(x) + \frac{1}{2}\langle w, Aw\rangle \geq q(x), \end{aligned}$$

where we use the assumption that A is positive-definite. Since $v \in \mathbb{R}^n$ was arbitrary, it follows that x is a minimiser of $q(x)$.

■

If A is SPD, then its minimiser is unique. Suppose there are two distinct minimisers $x, y \in \mathbb{R}^n$ of $q(x)$. They must satisfy $Ax = b = Ay$ or $A(x-y) = \mathbf{0}$, which implies that $x - y = \mathbf{0}$ since A is non-singular. In practice, $q(x)$ usually represents a significant quantity such as the energy of a system. In this case the solution to $Ax = b$ represents a state of minimal energy.

5.3.1 Steepest Descent/Gradient Descent Method

To compute the mininum of $E(x) = q(x)$, we propose an iterative method, called the **steepest descent method**, defined by

$$x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)} \tag{5.3.1a}$$

$$r^{(k)} = b - Ax^{(k)} \tag{5.3.1b}$$

where α_k will be chosen in such a way that $E(x^{(k+1)})$ is minimised.

Lemma 5.3.2. *Given $x^{(k)}$, $E(x^{(k+1)})$ is minimised if α_k is chosen to be*

$$\alpha_k = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k)}, Ar^{(k)} \rangle} = \frac{\|r^{(k)}\|_2^2}{\|r^{(k)}\|_A^2}, \quad k \geq 0.$$

where $\|\cdot\|_A$ is the energy norm.

Proof. Since A is symmetric,

$$\begin{aligned} E(x^{(k+1)}) &= E(x^{(k)} + \alpha_k r^{(k)}) \\ &= E(x^{(k)}) + \alpha_k \Big[\langle r^{(k)}, Ax^{(k)} \rangle - \langle r^{(k)}, b \rangle \Big] + \frac{\alpha_k^2}{2} \langle r^{(k)}, Ar^{(k)} \rangle \\ &= \frac{\alpha_k^2}{2} \langle r^{(k)}, Ar^{(k)} \rangle - \alpha_k \langle r^{(k)}, r^{(k)} \rangle + E(x^{(k)}). \end{aligned}$$

The last expression, denoted by $G(\alpha_k)$, is a quadratic equation in α_k. where $G(\cdot)$ is a quadratic equation in α_k. Since A is positive-definite, $\langle r^{(k)}, Ar^{(k)} \rangle > 0$ and there exists a unique minimum of $G(\alpha_k)$. This minimum must satisfies $G'(\alpha_k) = 0$ and solving this gives the desired expression for α_k.

■

Lemma 5.3.3. *For every $k \geq 0$, we have that $r^{(k)} \perp r^{(k+1)}$ with respect to $\langle \cdot, \cdot \rangle$.*

Proof. First, observe that substituting (5.3.1a) into (5.3.1b) yields

$$\begin{aligned} r^{(k+1)} = b - Ax^{(k+1)} &= b - A\Big[x^{(k)} + \alpha_k r^{(k)} \Big] \\ &= b - Ax^{(k)} - \alpha_k Ar^{(k)} \\ &= r^{(k)} - \alpha_k Ar^{(k)}. \end{aligned}$$

This together with the expression for α_k from Lemma 5.3.2 yields

$$\langle r^{(k+1)}, r^{(k)} \rangle = \langle r^{(k)}, r^{(k)} \rangle - \alpha_k \langle r^{(k)}, Ar^{(k)} \rangle = 0.$$

■

Remark 5.3.4. The residual vector given by $r^{(k+1)} = r^{(k)} - \alpha_k Ar^{(k)}$ is chosen to update the residual vector in the steepest descent method. It is more stable numerically compared to (5.3.1b) due to rounding error, since b can be very close to $Ax^{(k+1)}$ for large enough k.

Algorithm 5.1: Steepest Descent Method
Given an initial guess $x^{(0)} \in \mathbb{R}^n$, set $r^{(0)} = b - Ax^{(0)}$. For any $k = 0, 1, \ldots$, compute the following until desired tolerance.

$$\begin{aligned} \alpha_k &= \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k)}, Ar^{(k)} \rangle} \\ x^{(k+1)} &= x^{(k)} + \alpha_k r^{(k)} \\ r^{(k+1)} &= r^{(k)} - \alpha_k A r^{(k)}. \end{aligned}$$

Theorem 5.3.5. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. The steepest descent method is convergent for any initial condition $x^{(0)} \in \mathbb{R}^n$ and we have the following error estimate:*

$$\|e^{(k+1)}\|_A \leq \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\right) \|e^{(k)}\|_A,$$

where $e^{(k)} = x^{(k)} - x_{exact}$ and $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_1/\sigma_n$ is the condition number of A with respect to $\|\cdot\|_2$.

Although the steepest descent method is convergent, it does not imply that the error is monotonically decreasing. As such, the steepest descent method can be time consuming. It can happen that $r^{(k)}$ (steepest descent direction) oscillates. Indeed, Lemma 5.3.3 tells us that $r^{(k+2)}$ can almost be in the same direction as $\pm r^{(k)}$ with same magnitude.

5.3.2 Conjugate Gradient Method

The *conjugate gradient (CG) method* can be seen as an improvisation of the steepest descent method. It is defined by

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)},$$

where the **conjugate direction** $p^{(k)}$ is a linear combination of the steepest descent direction $r^{(k)} = b - Ax^{(k)}$ and previous change in position $x^{(k)} - x^{(k-1)}$, *i.e.*

$$\begin{aligned} p^{(k)} &= r^{(k)} + \gamma_k(x^{(k)} - x^{(k-1)}) \\ &= r^{(k)} + \gamma_k \alpha_{k-1} p^{(k-1)} \\ &= r^{(k)} + \beta_{k-1} p^{(k-1)}. \end{aligned}$$

$r^{(k)}$ takes another form

$$r^{(k)} = b - A\left(x^{(k-1)} + \alpha_{k-1}p^{(k-1)}\right) = r^{(k-1)} - \alpha_{k-1}Ap^{(k-1)}.$$

Thus, the conjugate gradient method takes the form

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \alpha_k p^{(k)} && \text{(5.3.2a)} \\ r^{(k+1)} &= r^{(k)} - \alpha_k A p^{(k)} && \text{(Residual)} \\ p^{(k+1)} &= r^{(k+1)} + \beta_k p^{(k)}. && \text{(Search)} \end{aligned}$$

where $\alpha_k, \beta_k, p^{(0)}$ are again chosen so that $E(x^{(k+1)})$ is minimised. We recover the steepest descent method if $\beta_k = 0$.

Lemma 5.3.6. *Given $x^{(k)}$, $E(x^{(k+1)})$ is minimised if $p^{(0)} = r^{(0)}$ and α_k, β_k are chosen to be*

$$\alpha_k = \frac{\langle r^{(k)}, r^{(k)}\rangle}{\langle p^{(k)}, Ap^{(k)}\rangle} = \frac{\|r^{(k)}\|_2^2}{\|p^{(k)}\|_A^2} \quad \textit{and} \quad \beta_k = \frac{\langle r^{(k+1)}, r^{(k+1)}\rangle}{\langle r^{(k)}, r^{(k)}\rangle} = \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}, \quad k \geq 0.$$

Proof. Since A is symmetric,

$$\begin{aligned} E(x^{(k+1)}) &= E(x^{(k)} + \alpha_k p^{(k)}) \\ &= E(x^{(k)}) + \alpha_k\Big[\langle p^{(k)}, Ax^{(k)}\rangle - \langle p^{(k)}, b\rangle\Big] + \frac{\alpha_k^2}{2}\langle p^{(k)}, Ap^{(k)}\rangle \\ &= \frac{\alpha_k^2}{2}\langle p^{(k)}, Ap^{(k)}\rangle - \alpha_k\langle p^{(k)}, r^{(k)}\rangle + E(x^{(k)}). \end{aligned}$$

A similar argument in Lemma 5.3.2 shows that in order to minimise $E(x^{(k+1)})$, α_k must satisfy $G'(\alpha_k)$ and solving this gives

$$\alpha_k = \frac{\langle p^{(k)}, r^{(k)}\rangle}{\langle p^{(k)}, Ap^{(k)}\rangle}, \quad k \geq 0. \tag{5.3.3}$$

We now rewrite $\langle p^{(k)}, r^{(k)}\rangle$ using (Search), (Residual) and (5.3.3) accordingly:

$$\begin{aligned} \langle p^{(k+1)}, r^{(k+1)}\rangle &= \langle r^{(k+1)}, r^{(k+1)}\rangle + \beta_k\langle p^{(k)}, r^{(k+1)}\rangle \\ &= \langle r^{(k+1)}, r^{(k+1)}\rangle + \beta_k\langle p^{(k)}, r^{(k)} - \alpha_k Ap^{(k)}\rangle \\ &= \langle r^{(k+1)}, r^{(k+1)}\rangle + \beta_k\Big[\langle p^{(k)}, r^{(k)}\rangle - \alpha_k\langle p^{(k)}, Ap^{(k)}\rangle\Big] \\ &= \langle r^{(k+1)}, r^{(k+1)}\rangle. \end{aligned}$$

Next, substituting α_k into $E(x^{(k+1)})$ yields

$$E(x^{(k+1)}) = E(x^{(k)}) - \frac{1}{2}\left(\frac{\langle r^{(k)}, r^{(k)}\rangle^2}{\langle p^{(k)}, Ap^{(k)}\rangle}\right). \tag{5.3.4}$$

For $k = 0$, $E(x^{(1)}) < E(x^{(0)})$ if we choose $p^{(0)} = r^{(0)}$, since A is positive-definite. To find β_k, we want to maximise the second term in (5.3.4), *i.e.* minimise $\langle p^{(k)}, Ap^{(k)}\rangle$. We write this expression in terms of β_k using (Search) and $A = A^T$ and get

$$\begin{aligned} \langle p^{(k)}, Ap^{(k)}\rangle &= \langle r^{(k)} + \beta_{k-1}p^{(k-1)}, A(r^{(k)} + \beta_{k-1}p^{(k-1)})\rangle \\ &= \langle r^{(k)}, Ar^{(k)}\rangle + 2\beta_{k-1}\langle r^{(k)}, Ap^{(k-1)}\rangle + \beta_{k-1}^2\langle p^{(k-1)}, Ap^{(k-1)}\rangle. \end{aligned}$$

Since the last expression is a quadratic equation in β_{k-1}, $E(x^{(k+1)})$ is minimised if β_{k-1} satisfies $H'(\beta_{k-1}) = 0$ and solving this gives

$$\beta_k = -\frac{\langle r^{(k+1)}, Ap^{(k)}\rangle}{\langle p^{(k)}, Ap^{(k)}\rangle}, \quad k \geq 1. \tag{5.3.5}$$

Observe that using (Search) gives an orthogonal relation for successive $p^{(k)}$ with respect to $\langle \cdot, A(\cdot)\rangle$:

$$\langle p^{(k+1)}, Ap^{(k)}\rangle = \langle r^{(k+1)}, Ap^{(k)}\rangle + \beta_k\langle p^{(k)}, Ap^{(k)}\rangle$$

$$= \langle r^{(k+1)}, Ap^{(k)} \rangle - \left(\frac{\langle r^{(k+1)}, Ap^{(k)} \rangle}{\cancel{\langle p^{(k)}, Ap^{(k)} \rangle}} \right) \cancel{\langle p^{(k)}, Ap^{(k)} \rangle} = 0,$$

which in turn gives

$$\langle p^{(k)}, Ap^{(k)} \rangle = \langle r^{(k)}, Ap^{(k)} \rangle + \beta_{k-1} \langle p^{(k-1)}, Ap^{(k)} \rangle = \langle r^{(k)}, Ap^{(k)} \rangle.$$

We also obtain an orthogonal relation for succesive $r^{(k)}$ with respect to $\langle \cdot, \cdot \rangle$, using (Residual) and $A = A^T$ to get

$$\begin{aligned} \langle r^{(k+1)}, r^{(k)} \rangle &= \langle r^{(k)}, r^{(k)} \rangle - \alpha_k \langle Ap^{(k)}, r^{(k)} \rangle \\ &= \langle r^{(k)}, r^{(k)} \rangle - \alpha_k \langle Ap^{(k)}, p^{(k)} \rangle = 0, \end{aligned}$$

which in turn gives

$$\begin{aligned} \langle r^{(k+1)}, Ap^{(k)} \rangle = \left\langle r^{(k+1)}, \frac{r^{(k)} - r^{(k+1)}}{\alpha_k} \right\rangle &= -\frac{1}{\alpha_k} \langle r^{(k+1)}, r^{(k+1)} \rangle \\ &= - \left(\frac{\langle p^{(k)}, Ap^{(k)} \rangle}{\langle r^{(k)}, r^{(k)} \rangle} \right) \langle r^{(k+1)}, r^{(k+1)} \rangle. \end{aligned}$$

Finally,

$$\beta_k = \left(\frac{\langle p^{(k)}, Ap^{(k)} \rangle}{\langle r^{(k)}, r^{(k)} \rangle} \right) \left(\frac{\langle r^{(k+1)}, r^{(k+1)} \rangle}{\langle p^{(k)}, Ap^{(k)} \rangle} \right) = \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}.$$

■

Lemma 5.3.7. *For the conjugate gradient method, the residuals and search directions satisfy the orthogonality:*

$$\langle r^{(j)}, r^{(k)} \rangle = \langle p^{(j)}, Ap^{(k)} \rangle = 0 \quad \textit{for all} \quad j \neq k.$$

Proof. We need to show the following statement for each $N \geq 1$:

$$\langle r^{(j)}, r^{(k)} \rangle = \langle p^{(j)}, Ap^{(k)} \rangle = 0 \quad \text{for all } 0 \leq k < j \leq N.$$

The following partial result was shown in the proof of Lemma 5.3.6:

$$\langle r^{(k+1)}, r^{(k)} \rangle = \langle p^{(k+1)}, Ap^{(k)} \rangle = 0 \quad \text{for all } k \geq 0.$$

The base case $N = 1$ follows from the partial result above:

$$\langle r^{(1)}, r^{(0)} \rangle = \langle p^{(1)}, Ap^{(0)} \rangle = 0.$$

Suppose

$$\langle r^{(j)}, r^{(k)} \rangle = \langle p^{(j)}, Ap^{(k)} \rangle = 0 \quad \text{for all } 0 \leq k < j \leq N.$$

We need to show that the same relation holds for all $0 \leq k < j \leq N+1$. This is true from the partial result if $j = N+1$ and $k = N$, so suppose $j = N+1$ and $k < N$. Then

$$\begin{aligned} \langle r^{(N+1)}, r^{(k)} \rangle &= \langle r^{(N)} - \alpha_N Ap^{(N)}, r^{(k)} \rangle && \big[\text{From (Residual).}\big] \\ &= -\alpha_N \langle Ap^{(N)}, r^{(k)} \rangle \\ &= -\alpha_N \langle Ap^{(N)}, p^{(k)} - \beta_k p^{(k-1)} \rangle && \big[\text{From (Search).}\big] \end{aligned}$$

$$
\begin{aligned}
&= 0. \\
\langle p^{(N+1)}, Ap^{(k)} \rangle &= \langle r^{(N+1)} + \beta_N p^{(N)}, Ap^{(k)} \rangle && \left[\text{From (Search).}\right] \\
&= \langle r^{(N+1)}, Ap^{(k)} \rangle \\
&= \left\langle r^{(N+1)}, \frac{r^{(k)} - r^{(k+1)}}{\alpha_k} \right\rangle && \left[\text{From (Residual).}\right] \\
&= 0,
\end{aligned}
$$

provided $\alpha_k \neq 0$, but $\alpha_k = 0$ means $r^{(k)} = 0$ and the method terminates.

■

Algorithm 2: Conjugate Gradient Method
Given an initial guess $x^{(0)} \in \mathbb{R}^n$, set $p^{(0)} = r^{(0)} = b - Ax^{(0)}$. For each $k = 0, 1, \ldots,$

$$
\begin{aligned}
\alpha_k &= \frac{\|r^{(k)}\|_2^2}{\langle p^{(k)}, Ap^{(k)} \rangle} \\
x^{(k+1)} &= x^{(k)} + \alpha_k p^{(k)} \\
r^{(k+1)} &= r^{(k)} - \alpha_k A p^{(k)} \\
\beta_k &= \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2} \\
p^{(k+1)} &= r^{(k+1)} + \beta_k p^{(k)}
\end{aligned}
$$

Theorem 5.3.8. *If $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, then the conjguate gradient method converges (pointwise) in at most n steps to the solution of $Ax = b$. Moreover, the error $e^{(k)} = x^{(k)} - x_{exact} \perp p^{(j)}$ for $j = 0, 1, \ldots, k-1$, $k < n$, and*

$$
\|e^{(k)}\|_A \leq \left(\frac{2C^k}{1 + C^{2k}} \right) \|e^{(0)}\|_A, \quad \textit{where} \quad C = \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}.
$$

Proof. We only prove the first part. Suppose

$$
\sum_{j=0}^{n-1} \delta_j r^{(j)} = \mathbf{0}. \tag{5.3.6}
$$

From Lemma 5.3.7, it follows that

$$
\delta_k \langle r^{(k)}, r^{(k)} \rangle = 0 \quad \text{for all } k = 0, 1, \ldots, n-1.
$$

Either $r^{(k)} = \mathbf{0}$ for some $k \leq n-1$ which means the iteration process stops at the kth step, or $\delta_k = 0$ for every $k = 0, 1, \ldots, n-1$, which means the set of residual vectors $\{r^{(0)}, r^{(1)}, \ldots, r^{(n-1)}\}$ form a basis of $\mathbb{R}^n$ and $r^{(n)} \equiv 0$. In both cases, we see that the conjugate gradient method converges in at most n steps.

■

5.4 Problems

1. Let A be a square matrix and let $\|\cdot\|$ be a consistent matrix norm (we say that $\|\cdot\|$ is compatible or consistent with a vector norm $\|\cdot\|$ if $\|Ax\| \leq \|A\|\|x\|$). Show that

$$\lim_{m\to\infty} \|A^m\|^{1/m} = \rho(A). \tag{5.4.1}$$

Solution: Since any consistent norms are equivalent in finite-dimensional vector space, it suffices to prove the statement in the case of $\|\cdot\|_2$. Consider the Schur decomposition of $A \in \mathbb{C}^{n\times n}$ given by $A = QTQ^*$ (see Theorem 6.1.11) where Q is unitary and $T = D + U$ is upper-triangular, with D the diagonal matrix with eigenvalues of A on its diagonal and U the nilpotent matrix, *i.e.* there exists an $N > 0$ such that $U^m = \mathbf{0}$ for $m \geq N$. Suppose $\rho(A) = 0$, then $D = \mathbf{0}$ and $T = U$. Since $\|\cdot\|_2$ is invariant under unitary transformation, for $m \geq N$ we have that

$$\|A^m\|_2 = \|(QTQ^*)^m\|_2 = \|QT^mQ^*\|_2 = \|QU^mQ^*\|_2 = \|Q\mathbf{0}Q^*\|_2 = 0,$$

and the equality (5.4.2) holds.

Choose any $\lambda \in \sigma(A)$, with corresponding eigenvector $x \neq \mathbf{0}$. Since $A^m x = \lambda^m x$, we have that

$$|\lambda^m|\|x\|_2 = \|\lambda^m x\|_2 = \|A^m x\|_2 \leq \|A^m\|_2\|x\|_2.$$

Since $x \neq \mathbf{0}$, dividing each side by $\|x\|_2$ gives

$$|\lambda|^m \leq \|A^m\|_2.$$

Taking the mth root of each side, and then the maximum over all $\lambda \in \sigma(A)$ yields

$$\max_{\lambda\in\sigma(A)} |\lambda| = \rho(A) \leq \|A^m\|_2^{1/m} \implies \rho(A) \leq \lim_{m\to\infty} \|A^m\|_2^{1/m}. \tag{5.4.2}$$

On the other hand, for m much larger than N we have that

$$\begin{aligned}
\|A^m\|_2 = \|(D+U)^m\|_2 &\leq \sum_{k=0}^{m} \left(\frac{m!}{k!(m-k)!}\right) \|D\|_2^{m-k}\|U\|_2^k \\
&= \sum_{k=0}^{N-1} \left(\frac{m(m-1)\dots(m-k+1)}{k!}\right) \|D\|_2^{m-k}\|U\|_2^k \\
&\leq \sum_{k=0}^{N-1} \left(\frac{m^k}{k!}\right) \|D\|_2^m \left(\frac{\|U\|_2}{\|D\|_2}\right)^k \\
&\leq m^{N-1}\|D\|_2^m \left(\sum_{k=0}^{N-1} \left(\frac{\|U\|_2}{\|D\|_2}\right)^k\right) \\
&= Cm^{N-1}\rho(A)^m,
\end{aligned}$$

where $C > 0$ is independent of m. Taking the mth root and then the limit as $m \to \infty$ yields

$$\lim_{m\to\infty} \|A^m\|_2^{1/m} \leq \lim_{m\to\infty} \left(C^{1/m} m^{(N-1)/m} \rho(A)\right) = \rho(A), \tag{5.4.3}$$

where we use the fact that $\lim_{m\to\infty} C^{1/m} = 1 = \lim_{m\to\infty} m^{1/m}$ for any nonnegative real number C. The result follows from combining (5.4.2) and (5.4.3).

2. Consider the 3×3 linear system of the form $A_j x = b_j$, where b_j is always taken in such a way that the solution of the system is the vector $x = (1, 1, 1)^T$, and the matrices A_j are

$$A_1 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & 1 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}.$$

Suggest strictly diagonally dominant by rows 3×3 matrix A_5. Implement Jacobi and Gauss-Siedel methods on A_1 to A_5. Explain theoretically your numerical observations.

Solution: We have that

$$b_1 = \begin{bmatrix} 7 \\ 13 \\ 2 \end{bmatrix}, \quad b_2 = \begin{bmatrix} -6 \\ -5 \\ 3 \end{bmatrix}, \quad b_3 = \begin{bmatrix} 6 \\ -7 \\ -14 \end{bmatrix}, \quad b_4 = \begin{bmatrix} 22 \\ 5 \\ -2 \end{bmatrix}.$$

We suggest the following strictly diagonally dominant by rows matrix $A_5 \in \mathbb{R}^{3\times 3}$ given by

$$A_5 = \begin{bmatrix} 15 & -7 & -7 \\ -1 & 8 & 2 \\ 3 & -6 & 11 \end{bmatrix}, \quad \text{with} \quad b_5 = \begin{bmatrix} 1 \\ 9 \\ 8 \end{bmatrix}.$$

Recall that for the Jacobi and Gauss-Siedel method, the corresponding iteration matrix B_J and B_{GS} is given by

$$B_J = I - D^{-1}A, \quad B_{GS} = -(D + L)^{-1}U,$$

where $A = L + D + U$ with D the diagonal, L the lower off diagonal and U the upper off diagonal. The number of iterations is $N = 200$, and we test the algorithm for three different random initial guesses

$$x_0^1 = \begin{bmatrix} 0.1214 \\ 0.1815 \\ 1.6112 \end{bmatrix}, \quad x_0^2 = \begin{bmatrix} 1.0940 \\ 1.7902 \\ 0.3737 \end{bmatrix}, \quad x_0^3 = \begin{bmatrix} 1.8443 \\ 1.3112 \\ 1.2673 \end{bmatrix}.$$

We choose to stop the iteration process if the Euclidean norm of the residual vector $\|b - Ax^{(k)}\|_2$ is less than the chosen tolerance $\epsilon = 10^{-12}$. We explain the numerical result using the spectral radius of the iteration matrix.

Matrix	$\rho(B_J)$	$\rho(B_{GS})$	Jacobi	Gauss-Siedel
A_1	1.1251	1.5833	Does not converges	Does not converge
A_2	0.8133	1.1111	Converge	Does not converge
A_3	0.4438	0.0185	Converge	Converge
A_4	0.6411	0.7746	Converge	Converge
A_5	0.5134	0.2851	Converge	Converge

Chapter 6

Eigenvalue Problems

Eigenvalues and eigenvectors of square matrices appear in the analysis of linear transformation and has a wide range of applications, such as facial recognition, image compression, spectral clustering, dimensionality reduction and ranking algorithm. These matrices may be sparse or dense and may have greatly varying order and structure. What is to be calculated affects the choice of method to be used, as well as the structure of the given matrix. We first discuss three matrix factorisations, where the eigenvalues are explicitly displayed. We then review three classical eigenvalue algorithms: power iteration, inverse iteration and Rayleigh quotient iteration.

6.1 Eigenvalue-Revealing Factorisation

Definition 6.1.1. Let $A \in \mathbb{C}^{m\times m}$ be a square matrix. A nonzero vector $x \in \mathbb{C}^m$ is an **eigenvector** of A, and $\lambda \in \mathbb{C}$ is its corresponding **eigenvalue** if $Ax = \lambda x$. The set of all eigenvalues of A is the **spectrum** of A, denoted by $\sigma(A)$.

6.1.1 Geometric and Algebraic Multiplicity

The set of eigenvectors corresponding to a single eigenvalue λ, together with the zero vector, forms a subspace of $\mathbb{C}^m$ known as an **eigenspace**, denoted by E_λ. Observe that E_λ is an *invariant subspace* of A, that is $AE_\lambda \subset E_\lambda$. The dimension of E_λ is known as the **geometric multiplicity** of λ. Equivalently, the geometric multiplicity is the dimension of $(\mathcal{N}(A - \lambda I))$, *i.e.* it is the maximum number of linearly independent eigenvectors with the same eigenvalue λ.

The **characteristic polynomial** of a square matrix $A \in \mathbb{C}^{m\times m}$ is the polynomial $p_A(z)$ of degree m defined by

$$p_A(z) = \det(zI - A).$$

From the definition of an eigenvalue,

$$\begin{aligned} \lambda \in \sigma(A) &\iff Ax = \lambda x \quad \text{for some } x \neq \mathbf{0}. \\ &\iff (\lambda I - A)x = 0 \quad \text{for some } x \neq \mathbf{0}. \\ &\iff \det(\lambda I - A) = p_A(\lambda) = 0. \end{aligned}$$

Consequently, eigenvalues of A are roots of the characteristic polynomial p_A and vice versa and we may write p_A as

$$p_A(z) = \prod_{j=1}^{m}(z-\lambda_j) = (z-\lambda_1)(z-\lambda_2)\ldots\ldots(z-\lambda_m),$$

where $\lambda_j \in \mathbb{C}$ are eigenvalues of A and they might be repeated. With this in mind, we define the **algebraic multiplicty** of $\lambda \in \sigma(A)$ as the multiplicity of λ as a root of $p_A(z)$; an eigenvalue is **simple** if its algebraic multiplicity is 1.

Theorem 6.1.2. *A matrix $A \in \mathbb{C}^{m\times m}$ has m eigenvalues, counted with algebraic multiplicity. In particular, A has m distinct eigenvalues if the roots of p_A are simple.*

Theorem 6.1.3. *Given a matrix $A \in \mathbb{C}^{m\times m}$, the following relation holds where eigenvalues are counted with algebraic multiplicity:*

$$\det(A) = \prod_{j=1}^{m}\lambda_j, \quad \operatorname{tr}(A) = \sum_{j=1}^{m}\lambda_j.$$

Proof. From the product property of the determinant,

$$\begin{aligned}\det(A) = (-1)^m\det(-A) &= (-1)^m p_A(0)\\ &= (-1)^m\left(\prod_{j=1}^{m}(z-\lambda_j)\right)\Big|_{z=0}\\ &= \prod_{j=1}^{m}\lambda_j.\end{aligned}$$

The second formula follows from equating the coefficient of z^{m-1} in $\det(zI-A)$ and $\prod_{j=1}^{m}(z-\lambda_j)$. ■

If $X \in \mathbb{C}^{m\times m}$ is nonsingular, then the map $A \mapsto X^{-1}AX$ is called a **similarity transformation** of A. We say that two matrices A and B are **similar** if there exists a nonsingular X such that $B = X^{-1}AX$.

Theorem 6.1.4. *If X is nonsingular, then A and $X^{-1}AX$ have the same characteristic polynomial, eigenvalues and algebraic and geometric multiplicities.*

Proof. Checking the characteristic polynomial of $X^{-1}AX$ yields

$$\begin{aligned}p_{X^{-1}AX}(z) &= \det(zI - X^{-1}AX)\\ &= \det(X^{-1}(zI-A)X)\\ &= \det(X^{-1})\det(zI-A)\det(X)\\ &= \det(zI-A) = p_A(z).\end{aligned}$$

Consequently, A and $B = X^{-1}AX$ have the same characteristic polynomial and also the eigenvalues and algebraic multiplicities. Finally, suppose E_λ is an eigenspace for A. For any $x \in E_\lambda$, we have that

$$Ax = \lambda x, \quad XBX^{-1}x = \lambda x, \quad BX^{-1}x = \lambda X^{-1}x,$$

i.e. $X^{-1}E_\lambda$ is an eigenspace for $B = X^{-1}AX$ (since X^{-1} is nonsingular), and conversely.

■

Theorem 6.1.5. *The algebraic multiplicity of an eigenvalue $\lambda \in \sigma(A)$ is at least as great as its geometric multiplicity.*

Proof. A proof can be found in [TBI97, p.185].

■

6.1.2 Eigenvalue Decomposition

Definition 6.1.6. An **eigenvalue decomposition** of a square matrix A is a factorisation $A = X\Lambda X^{-1}$, where X is nonsingular and Λ is diagonal. Equivalently, we have $AX = X\lambda$, or column-wise

$$Ax_j = \lambda_j x_j, \quad j = 1, \ldots, m.$$

This suggests that the jth column of X is an eigenvector of A, with its eigenvalue λ_j the jth diagonal entry of Λ.

Observe that the eigenvalue decomposition expresses a change of basis to "eigenvector coordinates" and this provides a way of reducing a coupled system to a system of scalar problems. For instance, suppose we want to solve $x' = Ax$, with $A \in \mathbb{R}^{m\times m}$ given. Suppose A is not diagonal but there exists a nonsingular X such that $A = X\Lambda X^{-1}$. Introducing a change of variable $y = X^{-1}x$, then

$$y' = X^{-1}x' = X^{-1}Ax = (X^{-1}AX)X^{-1}x = \Lambda y.$$

The system is now decoupled and it can be solved separately. The solutions are

$$y_j(t) = e^{\lambda_j t}y_j(0), \quad j = 1, \ldots, m,$$

which then gives

$$x(t) = XDX^{-1}x(0), \quad \text{where} \quad d_{jj} = e^{\lambda_j t},\ j = 1, \ldots, m.$$

Example 6.1.7. Consider the matrices

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

Both A and B have characteristic polynomial $(z-2)^3$, so there is a single eigenvalue $\lambda = 2$ of algebraic multiplicity 3. In the case of A, the eigenvalue has geometric multiplicty 3, which can be seen by choosing the standard basis of $\mathbb{R}^3$ as eigenvectors. In the case of B, the eigenvalue has geometric multiplicty 1, since the only eigenvectors are scalar multiples of $(1, 0, 0)^T$.

Definition 6.1.8. An eigenvalue whose algebraic multiplicity exceeds its geometric multiplicity is a **defective eigenvalue**. A matrix that has one or more defective eigenvalues is a **defective matrix**.

Theorem 6.1.9. *$A \in \mathbb{C}^{m\times m}$ is nondefective if and only if it has an eigenvalue decomposition $A = X\lambda X^{-1}$.*

Proof. Suppose A has an eigenvalue decomposition. Since A is similar to Λ, it follows from Theorem 6.1.4 that they share the same eigenvalues and the same multiplicities. Consequently, A is nondefective since diagonal matrices are nondefective. Conversely, suppose A is nondefective. Then A must have m linearly independent eigenvectors since one can show that eigenvectors with different eigenvalues must be linearly independent, and each eigenvalue can contribute as many linearly independent eigenvectors as its multiplicity. Defining X as the matrix whose columns are these m linearly independent eigenvectors, we see that $AX = X\Lambda$, or $A = X\Lambda X^{-1}$. ■

6.1.3 Unitary Diagonalisation

In some special cases, the eigenvectors can be chosen to be orthogonal. In this case, we say that the matrix A is **unitary diagonalisable** if there exists a unitary matrix Q and a diagonal matrix Λ such that $A = Q\Lambda Q^{-1}$. This factorisation is both an eigenvalue decomposition and a singular value decomposition, aside from the signs of entries of Λ. Surprisingly, unitary diagonalisable matrices have an elegant characterisation.

Theorem 6.1.10.

(a) A hermitian matrix is unitary diagonalisable, and its eigenvalues are real.

(b) A matrix A is unitary diagonalisable if and only if it is normal, that is $A^*A = AA^*$.

6.1.4 Schur Factorisation

Theorem 6.1.11. *Every square matrix $A \in \mathbb{C}^{m\times m}$ has a Schur factorisation, i.e. there exists a unitary matrix Q and an upper-triangular matrix T such that $A = QTQ^*$. Moreover, the eigenvalues of A necessarily appear on the diagonal of T since A and T are similar.*

Proof. The proof is similar to the existence of SVD. The case $m = 1$ is trivial, so suppose $m \geq 2$. Let q_1 be any eigenvector of A, with corresponding eigenvalue λ. WLOG, we may assume $\|q_1\|_2 = 1$. Consider any extension of q_1 to an orthonormal basis $\{q_1, \ldots, q_m\} \subset \mathbb{C}^m$ and construct the unitary matrix

$$Q_1 = \begin{bmatrix} q_1 & \widehat{Q}_1 \end{bmatrix} \in \mathbb{C}^{m\times m}, \quad \widehat{Q}_1 = \begin{bmatrix} q_2 & \cdots & q_m \end{bmatrix}.$$

We have that

$$T_1 \coloneqq Q_1^* A Q_1 = \begin{bmatrix} q_1^* \\ \widehat{Q}_1^* \end{bmatrix} A \begin{bmatrix} q_1 & \widehat{Q}_1 \end{bmatrix} = \begin{bmatrix} q_1^*Aq_1 & q_1^*A\widehat{Q}_1 \\ \widehat{Q}_1^*Aq_1 & \widehat{Q}_1^*A\widehat{Q}_1 \end{bmatrix} = \begin{bmatrix} \lambda & b^* \\ \mathbf{0} & \widehat{A} \end{bmatrix}.$$

By induction hypothesis, $\widehat{A}$ has a Schur factorisation $\widehat{A} = Q_2 T_2 Q_2^*$. Then

$$\begin{aligned} Q_1^* A Q_1 &= \begin{bmatrix} \lambda & b^* \\ 0 & Q_2 T_2 Q_2^* \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q_2 \end{bmatrix} \begin{bmatrix} \lambda & b^* Q_2 \\ \mathbf{0} & T_2 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2^* \end{bmatrix}, \end{aligned}$$

and

$$A = Q_1 \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q_2 \end{bmatrix} \begin{bmatrix} \lambda & b^* Q_2 \\ \mathbf{0} & T_2 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q_2^* \end{bmatrix} Q_1^* = QTQ^*.$$

To finish the proof, we need to show that Q is unitary, but this must be true since

$$Q = Q_1 \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q_2 \end{bmatrix} = \begin{bmatrix} q_1 & \widehat{Q}_1 Q_2^* \end{bmatrix},$$

and $\{q_1, \ldots, q_m\}$ is orthogonal by construction.

■

Remark 6.1.12. Among all three factorisations, the Schur decomposition exists for any matrix and it tends to be numerically stable since unitary transformations are involved. If A is normal, T will be diagonal and in particular, if A is hermitian, then we can take advantage of this symmetry to reduce the computational cost.

6.1.5 Localising Eigenvalues

Below we prove a result that locates and bounds the eigenvalues of a given matrix A. A crude bound is the following inequality regarding the spectral radius:

$$\rho(A) \coloneqq \max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\|,$$

for any consistent matrix norm.

Theorem 6.1.13 (Gershgorin Circle Theorem)**.** *The spectrum $\sigma(A)$ is contained in the union of the following m disks $D_i, i = 1, \ldots, m$ in $\mathbb{C}$, where*

$$D_i = \left\{ z \in \mathbb{C} \colon |z - a_{ii}| \leq \sum_{j \neq i}^{m} |a_{ij}| \right\}, \quad i = 1, \ldots, m.$$

Proof. Choose any eigenvector x of A, with corresponding eigenvalue λ. WLOG, we may assume that $\|x\|_\infty = 1$. Let i be the index in $\{1, \ldots, m\}$ such that $|x_i| = 1$. Then

$$\lambda x_i = (Ax)_i = \sum_{j=1}^{m} a_{ij} x_j,$$

and

$$|\lambda - a_{ii}| = |(\lambda - a_{ii}) x_i| = \left| \sum_{j \neq i}^{m} a_{ij} x_j \right| \leq \sum_{j \neq i}^{m} |a_{ij} x_j| \leq \sum_{j \neq i}^{m} |a_{ij}|.$$

■

Example 6.1.14. Consider the matrix $A = \begin{bmatrix} -1+i & 0 & 1/4 \\ 1/4 & 1 & 1/4 \\ 1 & 1 & 3 \end{bmatrix}$. Applying the Gershgorin Circle Theorem gives the following three disks in $\mathbb{C}$

$$\begin{aligned} |\lambda - (-1+i)| &\leq 0 + \frac{1}{4} = \frac{1}{4} \\ |\lambda - 1| &\leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\ |\lambda - 3| &\leq 1 + 1 = 2. \end{aligned}$$

Upon sketching these disks in $\mathbb{C}$, we see that $\frac{1}{2} \leq |\lambda| \leq 5$. [Draw the solution set in $\mathbb{C}$.]

6.2 Eigenvalue Algorithms

For the remaining section, we will assume that $A \in \mathbb{R}^{m \times m}$ is symmetric unless specified otherwise. In particular, this means that A has real eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$ and a complete set of orthogonal eigenvectors $\{q_1, \ldots, q_m\}$.

6.2.1 Shortcomings of Obvious Algorithms

The most obvious method would be to compute the roots of the characteristic polynomial $p_A(z)$. Unfortunately, polynomial-rootfinding is a severely ill-conditioned problem even when the underlying eigenvalue problem is well-conditioned. This is because roots of polynomial depend continuously on the coefficient of $p_A(z)$ and thus are extremely sensitive to errors, such as round-off error.

Actually, any polynomial rootfinding problem can be rephrased as an eigenvalue problem. Given a monic polynomial

$$p_m(z) = z^m + a_{m-1}z^{m-1} + \ldots + a_1 z + a_0,$$

We prove by induction that $p_m(z)$ is equal to $(-1)^m$ times the determinant of the $m \times m$ matrix

$$B_m = \begin{bmatrix} -z & & & & & -a_0 \\ 1 & -z & & & & -a_1 \\ & 1 & -z & & & -a_2 \\ & & 1 & \ddots & & \vdots \\ & & & \ddots & -z & -a_{m-2} \\ & & & & 1 & (-z - a_{m-1}) \end{bmatrix}.$$

The base case $m = 2$ is clear:

$$(-1)^2 \begin{vmatrix} -z & -a_0 \\ 1 & (-z - a_1) \end{vmatrix} = z(z + a_1) + a_0 = p_2(z).$$

Suppose the statement is true for $m = k - 1$, then

$$(-1)^k \det(B_k) = (-1)^k \Big[(-z)(-1)^{k-1}(z^{k-1} + a_{k-1}z^{k-2} + \ldots + a_2 z + a_1) + (-1)^{k+1}(-a_0) \Big]$$

$$= z(z^{k-1} + a_{k-1}z^{k-2} + \ldots + a_2 z + a_1) + a_0$$
$$= p_k(z).$$

It follows that roots of $p_m(z)$ are equal to the eigenvalues of the **companion matrix**

$$A = \begin{bmatrix} 0 & & & & & -a_0 \\ 1 & 0 & & & & -a_1 \\ & 1 & 0 & & & -a_2 \\ & & 1 & \ddots & & \vdots \\ & & & \ddots & 0 & -a_{m-2} \\ & & & & 1 & -a_{m-1} \end{bmatrix}.$$

Theorem 6.2.1. *For any $m \geq 5$, there exists a polynomial $p(z)$ of degree m with rational coefficients that has a real root r, with the property that r cannot be written using any expression involving rational numbers, addition, subtraction, multiplication, division and kth roots.*

This theorem says that no computer program can produce the exact roots of an arbitrary polynomial of degree ≥ 5 in a finite number of steps even in exact arithmetic, and it is because of this that **any eigenvalue solver must be iterative**.

6.2.2 Rayleigh Quotient

Given a symmetric matrix $A \in \mathbb{R}^{m\times m}$, the Rayleigh Quotient of a nonzero vector $x \in \mathbb{R}^m$ is the scalar

$$R(x) = \frac{x^T Ax}{x^T x}.$$

Choosing x to be an eigenvector of A gives $R(x) = \lambda$ the corresponding eigenvalue. A natural question arises: given a nonzero x, what scalar α behaves like an eigenvalue for x in the sense of minimising $\|Ax - \alpha x\|_2$? Since x is given, this is an $m \times 1$ least squares problem of the form:

"Find $\alpha \in \mathbb{R}$ such that $\|x\alpha - Ax\|_2$ is mimimised."

With $A = x$ and $b = Ax$, The normal equation is precisely the Rayleigh quotient

$$A^T A\alpha = A^T b, \quad (x^T x)\alpha = x^T Ax, \quad \alpha = \frac{x^T Ax}{x^T x} = R(x).$$

It is helpful to view $R(\cdot)$ as a function from $\mathbb{R}^m$ to $\mathbb{R}$. We investigate the local behavior of $R(x)$ when x is near the eigenvector. Computing the partial derivatives of $r(x)$ with respect to the coordinates x_j yields

$$\begin{aligned} \frac{\partial R(x)}{\partial x_j} &= \left(\frac{1}{x^T x}\right)\frac{\partial}{\partial x_j}(x^T Ax) - \left(\frac{x^T Ax}{(x^T x)^2}\right)\frac{\partial}{\partial x_j}(x^T x) \\ &= \frac{2(Ax)_j}{x^T x} - \frac{(x^T Ax)2x_j}{(x^T x)^2} \\ &= \frac{2}{x^T x}\Big((Ax)_j - R(x)x_j\Big) \end{aligned}$$

$$= \frac{2}{x^T x} \left(Ax - R(x)x\right)_j .$$

Consequently, the gradient of $R(x)$ is

$$\nabla R(x) = \frac{2}{x^T x} \left(Ax - R(x)x\right)^T .$$

We deduce the following properties of $R(x)$ from the formula of $\nabla R(x)$:

1. $R(x)$ is smooth except at $x = \mathbf{0}$.
2. $\nabla R(x) = 0$ at an eigenvector x of A. Conversely if $\nabla R(x) = 0$ with $x \neq \mathbf{0}$, then x is an eigenvector of A with corresponding eigenvalue $R(x)$.
3. Let q_J be an eigenvector of A. For any nonzero $x \in \mathbb{R}^m$ sufficiently close to q_J, a second order Taylor expansion yields the asymptotic relation

$$R(x) - R(q_J) = \mathcal{O}\left(\|x - q_J\|_2^2\right) \quad \text{if } x \text{ is close to } q_J. \tag{6.2.1}$$

Thus the Rayleigh quotient is a **quadratically accurate** estimate of an eigenvalue!

We give another proof of the asymptotic relation (6.2.1). We express x as a linear combination of the eigenvectors $\{q_1, \ldots, q_m\}$

$$x = \sum_{j=1}^m a_j q_j = \sum_{j=1}^m \langle x, q_j \rangle q_j, \quad \text{since } A \text{ is symmetric.}$$

Assuming $x \approx q_J$ and $\left|\frac{a_j}{a_J}\right| \leq \varepsilon$ for all $j \neq J$, it suffices to show that $R(x) - R(q_J) = \mathcal{O}(\varepsilon)^2$, since from Pythagorean theorem we have that

$$\|x - q_J\|_2^2 = \sum_{j \neq J}^m |a_j|^2 + |a_J - 1|^2 = |a_J|^2 \left(\sum_{j \neq J}^m \left|\frac{a_j}{a_J}\right|^2 + \left|\frac{a_J - 1}{a_J}\right|^2 \right) \approx C\varepsilon^2.$$

Substituting the expansion of x into the Rayleigh quotient yields

$$R(x) = \frac{\left\langle \sum_{j=1}^m a_j q_j, \sum_{j=1}^m \lambda_j a_j q_j \right\rangle}{\sum_{j=1}^m a_j^2} = \frac{\sum_{j=1}^m \lambda_j a_j^2}{\sum_{j=1}^m a_j^2},$$

which in turn gives

$$R(x) - R(q_J) = \frac{\sum_{j=1}^m \lambda_j a_j^2}{\sum_{j=1}^m a_j^2} - \lambda_J = = \frac{\sum_{j \neq J}^m a_j^2 (\lambda_j - \lambda_J)}{a_J + \sum_{j \neq J}^m a_j^2} = \frac{\sum_{j \neq J}^m \left(\frac{a_j}{a_J}\right)^2 (\lambda_j - \lambda_J)}{1 + \sum_{j \neq J}^m \left(\frac{a_j}{a_J}\right)^2} = \mathcal{O}(\varepsilon).$$

6.2.3 Power iteration

The *power iteration* is used to find the largest eigenvalue and its corresponding eigenvector of a given matrix. The algorithm is presented below, and we note that the asymptotic relation (6.2.1) can be used as a stopping criteria.

Algorithm 6.1: Power Iteration
Assume $v^{(0)}$ is a vector with $\|v^{(0)}\|_2 = 1$.
for $k = 1, 2, 3, \ldots\ldots$
$w = Av^{k-1}$
$v^{(k)} = \dfrac{w}{\|w\|_2}$
$\lambda^{(k)} = (v^{(k)})^T A v^{(k)}$.

Theorem 6.2.2. *Assume* $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_m| \geq 0$ *and* $q_1^T v^{(0)} \neq 0$. *Then the iterates of power iteration algorithm satisfy*

$$\|v^{(k)} - \pm q_1\|_2 = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right), \quad |\lambda^{(k)} - \lambda_1| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \quad \text{as } k \longrightarrow \infty.$$

The $\pm$ *sign means that at each step* k, *one or the other choice of sign is to be taken, and then the indicated bound holds.*

Proof. Write $v^{(0)}$ as a linear combination of the eigenvectors

$$v^{(0)} = a_1 q_1 + a_2 q_2 + \ldots + a_m q_m.$$

By definition of the power iteration,

$$\begin{aligned}
v^{(1)} &= C_1 A \left(\sum_{j=1}^m a_j q_j\right) &&= C_1 \left(\sum_{j=1}^m \lambda_j a_j q_j\right) \\
v^{(2)} &= C_2 A \left(\sum_{j=1}^m \lambda_j a_j q_j\right) &&= C_2 \left(\sum_{j=1}^m \lambda_j^2 a_j q_j\right) \\
&\vdots &&\vdots \\
v^{(k)} &= C_k A \left(\sum_{j=1}^m \lambda_j^{k-1} a_j q_j\right) &&= C_k \left(\sum_{j=1}^m \lambda_j^k a_j q_j\right),
\end{aligned}$$

where C_k are the normalisation constant. Factoring out λ_1^k gives

$$v^{(k)} = C_k \lambda_1^k \left(a_1 q_1 + \sum_{j=2}^m \left(\frac{\lambda_j}{\lambda_1}\right)^k a_j q_j\right) = \frac{\lambda_1^k \left(a_1 q_1 + \sum_{j=2}^m \left(\frac{\lambda_j}{\lambda_1}\right)^k a_j q_j\right)}{|\lambda_1|^k \left\| a_1 q_1 + \sum_{j=2}^m \left(\frac{\lambda_j}{\lambda_1}\right)^k a_j q_j \right\|_2}.$$

Provided $a_1 = q_1^T v^{(0)} \neq 0$, we see that

$$v^{(k)} \longrightarrow \frac{\lambda_1^k a_1 q_1}{|\lambda_1|^k |a_1| \|q_1\|_2} = \pm q_1 \quad \text{as } k \longrightarrow \infty,$$

depending on the sign of λ_1 and initial guess $v^{(0)}$ and the first equation follows. The second equation follows from the asymptotic relation (6.2.1) of the Rayleigh quotient.

■

If $\lambda_1 > 0$, then the signs of q_1 is controlled by the initial guess $v^{(0)}$ and so are all $+$ or all $-$. If $\lambda_1 < 0$, then the signs of q_1 alternate and $\|v^{(k)}\|_2^2 \longrightarrow \|q_1\|_2^2$ as $k \longrightarrow \infty$. One can show that the iterates of power iteration algorithm satisfy

$$\frac{\|v^{(k+1)} - \pm q_1\|_2}{\|v^{(k)} - \pm q_1\|_2} = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|\right) \quad \text{as } k \longrightarrow \infty.$$

Consequently, the rate of convergence for the power iteration is linear. Except for special matrices, the power iteration is very slow!

6.2.4 Inverse Iteration

There is a potential problem with the power iteration: what if $\lambda_2/\lambda_1 \approx 1$ which will result in very slow convergence? Building from the power iteration, one approach would be to transform A such that the new matrix, say B, has its largest eigenvalue λ_1^B much larger than its remaining eigenvalues. The **Spectral Mapping Theorem** tells us just how to find such B and as it turns out, the same exact idea applies to finding any eigenvalues of A!

Consider any $\lambda \in \sigma(A)$ with corresponding eigenvector x. For any $\mu \in \mathbb{R}$, $(A - \mu I)^{-1}$ also has the same eigenvector x, but with a different eigenvalue:

$$Ax = \lambda x, \quad (A - \mu I)x = (\lambda - \mu)x, \quad (A - \mu I)^{-1}x = (\lambda - \mu)^{-1}x.$$

The following relation is also true:

$$\lambda \in \sigma(A) \iff (\lambda - \mu)^{-1} \in \sigma((A - \mu I)^{-1}).$$

The upshot is if we choose μ sufficiently close to λ_J, then $(\lambda_J - \mu)^{-1}$ may be much larger than $(\lambda_j - \mu)^{-1}$ for all $j \neq J$. Consequently, applying the power iteration to the matrix $(A - \mu I)^{-1}$ gives a rapid convergence to q_J and this is precisely the idea of *inverse iteration*.

Algorithm 6.2: Inverse Iteration
Given $\mu \in \mathbb{R}$ and $v^{(0)}$ some initial guess such that $\|v^{(0)}\|_2 = 1$.
for $k = 1, 2, 3, \ldots\ldots$
Solve for w in the equation $(A - \mu I)w = v^{(k-1)}$
$v^{(k)} = \dfrac{w}{\|w\|_2}$
$\lambda^{(k)} = (v^{(k)})^T A v^{(k)}$.

Note that the first step of the algorithm involves solving a linear system at each iteration step and this raises an immediate question: what if $A-\mu I$ is so ill-conditioned that an accurate solution of the linear system is not possible? This however is not a problem at all and we shall not pursue this issue any further; interested reader may refer to Exercise 27.5 in [TBI97, p.210]. The following theorem is essentially a corollary of Theorem 6.2.2.

Theorem 6.2.3. *Suppose λ_J is the closest eigenvalue to μ and λ_m is the second closest, that is,*

$$|\mu - \lambda_J| < |\mu - \lambda_m| \leq |\mu - \lambda_j| \quad \text{for each } j \neq J.$$

Suppose $q_J^T v^{(0)} \neq 0$. Then the iterates of inverse iteration algorithm satisfy

$$\|v^{(k)} - \pm q_J\|_2 = \mathcal{O}\left(\left|\frac{\mu - \lambda_J}{\mu - \lambda_m}\right|^k\right), \quad |\lambda^{(k)} - \lambda_J| = \mathcal{O}\left(\left|\frac{\mu - \lambda_J}{\mu - \lambda_m}\right|^{2k}\right) \quad \text{as } k \longrightarrow \infty.$$

The $\pm$ sign means that at each step k, one or the other choice of sign is to be taken, and the inidicated bound holds.

In practice, the inverse iteration is used when a good approximation for the desired eigenvalue is known. Otherwise, the inverse iteration converges to the eigenvector of the matrix A corresponding to the closest eigenvalue to μ. As opposed to the power iteration, we can control the rate of linear convergence since this depends on μ.

6.2.5 Rayleigh Quotient Iteration

Even with a good choice of μ, the inverse iteration converges at best linearly. Extending the idea of inverse iteration, this rate of convergence can be improved as follows: at each new iteration step, μ is replaced with the Rayleigh quotient of the previous eigenvector approximation. This leads to the *Rayleigh Quotient Iteration*:

Algorithm 6.3: Rayleight Quotient Iteration
$v^{(0)}$ = some initial vector with $\|v^{(0)}\|_2 = 1$
$\lambda^{(0)} = (v^{(0)})^T A v^{(0)}$.
for $k = 1, 2, 3, \ldots\ldots$
  Solve for w in the equation $(A - \lambda^{(k)} I)w = v^{(k-1)}$
  $v^{(k)} = \dfrac{w}{\|w\|_2}$
  $\lambda^{(k)} = (v^{(k)})^T A v^{(k)}$.

Theorem 6.2.4. *Rayleigh quotient iteration converges to an eigenvalue/eigenvector pair for all except a set of measure zero of starting vectors $v^{(0)}$. When it converges, the convergence is ultimately cubic in the sense that if λ_J is an eigenvalue of A and $v^{(0)}$ is sufficiently close to the corresponding eigenvector q_J, then*

$$\|v^{(k+1)} - \pm q_J\|_2 = \mathcal{O}\left(\|v^{(k)} - \pm q_J\|_2^3\right), \quad |\lambda^{(k+1)} - \lambda_J| = \mathcal{O}\left(|\lambda^{(k)} - \lambda_J|^3\right) \quad \text{as } k \longrightarrow \infty.$$

The $\pm$ signs are not necessarily the same on the two sides of the first equation.

Remark 6.2.5. We have cubic convergence for the Rayleigh quotient iteration if A is symmetric, otherwise it only has quadratic convergence.

Bibliography

[TBI97] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. Vol. 50. Other Titles in Applied Mathematics. SIAM, 1997.