MATH 5620/6860
INTRO TO NUMERICAL ANALYSIS

Chapter 1. Preliminaries

Def LIMIT.

Let $f: \mathbb{R} \to \mathbb{R}$ be a function. The limit of f at x_0 is written as (if it exists)

$$\lim_{x \to c} f(x) = L$$

And means that for any positive ε there is a δ positive such that

$$0 < |x - x_0| < \delta \implies |f(x) - L| < \varepsilon.$$

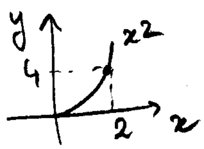
We shall write this in short using quantifiers:

$$\forall \varepsilon > 0 \ \exists \delta > 0 \quad 0 < |x - x_0| < \delta \Rightarrow |f(x) - L| < \varepsilon.$$

Example. $\lim_{x \to 2} x^2 = 4$

$|x-2| < \delta$

$$|x^2 - 4| = |x-2|\,|\underbrace{x+2}_{x-2+4}| < \delta(\delta + 4) = \varepsilon$$

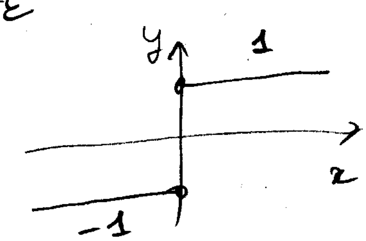


So for a given ε we can find δ s.t. limit def is satisfied.

$$\delta = -2 + \sqrt{4 + \varepsilon}$$

Example

$$f(x) = \frac{|x|}{x} = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$



Clearly limit does not exist at $x = 0$. Why?

Let $\varepsilon = 1$ and suppose $\exists \delta > 0$ and L s.t.

$$|x - 0| < \delta \implies \left|\frac{|x|}{x} - L\right| < \varepsilon = 1$$

If we let $x_1 = \frac{\delta}{2}$ then $|x_1| < \delta \Rightarrow |1 - L| < 1$

$x_2 = -\frac{\delta}{2}$ — $|x_2| < \delta \Rightarrow |-1 - L| < 1$

Which leads to a contradiction since the absolute value ineq imply:

$$0 < L < 2 \quad \text{and} \quad -2 < L < 0.$$

so limit does not exist.

- When f is defined only on $X \subset \mathbb{R}$ then the limit def becomes:

$$\forall \varepsilon > 0 \quad \exists \delta > 0, 0 < |x - x_0| < \delta \text{ and } x \in X \Rightarrow |f(x) - L| < \varepsilon$$

Def (Continuity) A function f is said to be continuous at x_0 if

$$\lim_{x \to x_0} f(x) = f(x_0)$$

A function is said to be continuous on some set X if it is cont. for all $x_0 \in X$.

Def (limit of sequences) Let $\{x_n\}_{n=1}^{\infty}$ be an infinite seq of real or complex numbers. The sequence has a limit x (or converges to x) as $n \to \infty$ if:

$$\forall \varepsilon > 0 \quad \exists N_0 \text{ s.t. } \forall n \geq N_0 \quad |x_n - x| < \varepsilon$$

Theorem (continuity with sequences)

f is continuous at x_0 $\Longleftrightarrow$ For any sequence $\{x_n\}_{n=1}^{\infty}$ converging to x_0

$$\lim_{n \to \infty} f(x_n) = f(x_0)$$

<u>Def</u> Let f be a function defined on an open interval containing x_0. The function f is differentiable at x_0 if the limit:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists. $f'(x_0)$ is called the <u>derivative</u> of f at x_0.

A function <u>f is differentiable on X</u> if it is diff'ble at all $x \in X$.

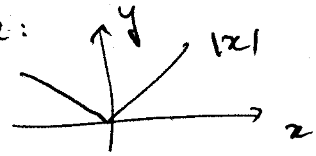
Clearly if f is differentiable at x_0, f is also continuous at x_0:

$$\lim_{x \to x_0} f(x) - f(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \, x - x_0$$

$$= f'(x_0) \lim_{x \to x_0} x - x_0 = 0$$

Of course the converse is not true:

$$f(x) = |x|$$



is continuous at $x = 0$ but not differentiable

We shall use the following notation for continuous, diff'ble fun:

$C^0(X) = C(X) =$ continuous functions on X

$C'(X) =$ functions w/ 1 continuous derivative

$\vdots$

$C^n(X) =$ ———— n ————

$\vdots$

$C^\infty(X) =$ ———— all derivatives continuous.

$$C^\infty(X) \subset \cdots \subset C^2(X) \subset C^1(X) \subset C(X)$$

Examples of $C^\infty(\mathbb{R})$ functions are $f(x)=e^x$, polynomials, cos, sin ...

<u>Theorem</u> (Taylor theorem with Lagrange remainder)

If $f \in C^n[a,b]$ and if $f^{(n+1)}$ exists on (a,b) then for any $x, x_0 \in [a,b]$ we have:

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + E_n(x)$$

where the error term is:

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-x_0)^{n+1}$$

for some ξ between x and x_0.

The case when $x_0 = 0$ is called <u>Mac Laurin series</u>

Some examples of Mac Laurin series

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}$$

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

valid for $x \in \mathbb{R}$

$$\frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k$$

valid for $|x|<1$

$$\ln(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} x^k$$

valid for $|x|<1$

In practical terms Taylor theorem gives a (local) <u>polynomial approximation</u> to a function, and also gives an expression for the error of approx. ($E_n(x)$).

<u>Example</u>:

$$\ln(1+x) = \underbrace{\sum_{k=1}^{n} \frac{(-1)^{k-1}}{k} x^k}_{\substack{\text{poly of degree } n \text{ approx.} \\ \ln(1+x) \text{ for } x \text{ small}}} + \frac{(-1)^n}{(n+1)} \frac{x^{n+1}}{(1+\xi)^{n+1}}$$

$$\frac{d^n}{dx^n}\left[\ln(1+x)\right] = \frac{(-1)^{n-1}(n-1)!}{(1+x)^n}$$

because we are doing expansion around $x=0$.

$$E_n(x) = \frac{1}{(n+1)!} \frac{(-1)^n n!}{(1+\xi)^{n+1}} (x-0)^{n+1}$$

How many terms in the Taylor series do we need to approximate $\ln(2)$ with 10^{-8} accuracy?

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots + (-1)^{n-1}\frac{1}{n} + E_n(1)$$

and we have:

$$|E_n(1)| = \frac{1}{n+1} \frac{1}{|1+\xi|^{n+1}} \le \frac{1}{n+1}$$

$0 < \xi < 1$

Thus we need:

$$|E_n(1)| \le \frac{1}{n+1} \le 10^{-8} \Rightarrow n+1 \ge 10^8$$

→ 100 million terms! Of course Taylor approx is better closer to the point where we did expansion ($x = \frac{1}{2}$ only 22 terms needed)

There are better ways of approx $\ln 2$!

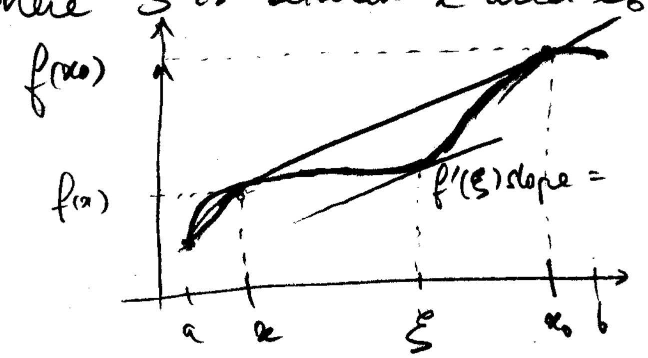
"All" you need to know is Taylor's theorem. Many theorems from calculus can be seen as a special case of some form of Taylor's theorem.

Theorem (Mean value theorem) ($n = 0$ Taylor's theorem)

If $f \in C[a,b]$ and f' exists in open interval (a,b), then for $x, x_0 \in [a,b]$:

$$f(x) = f(x_0) + f'(\xi)(x - x_0)$$

where ξ is between x and x_0.



$\dfrac{f(x) - f(x_0)}{x - x_0}$

A special case of MVT is Rolle's theorem:

Theorem (Rolle's theorem)

If $f \in C[a,b]$ and f' exists in open interval (a,b), and $f(a) = f(b)$ then $\exists\, \xi \in (a,b)$ s.t. $f'(\xi) = 0$.

(you probably did the following: Rolle's $\Rightarrow$ MVT $\Rightarrow$ Taylor)

Here is another version of Taylor's theorem that will be useful to us and that is easy to prove.

<u>Theorem. (Taylor's theorem with integral remainder)</u>

If $f \in C^{n+1}[a,b]$ then for any $x, x_0 \in [a,b]$

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + R_n(x)$$

where
$$R_n(x) = \frac{1}{n!} \int_{x_0}^{x} \underbrace{f^{(n+1)}(t)}_{\uparrow} \underbrace{(x-t)^n}_{\downarrow} dt$$

<u>Proof</u>: by successive integration by parts of the residual:

$$R_n = \frac{1}{n!} f^{(n)}(t)\,(x-t)^n \Big|_{t=x_0}^{x} + \frac{n}{n!} \int_{x_0}^{x} f^{(n)}(t)\,(x-t)^{n-1} dt$$

$$= \frac{-1}{n!} f^{(n)}(t)\,(x-x_0)^n + R_{n-1} \quad \leftarrow \text{repeat IBP with this residual}$$

$$= -\sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + R_0$$

But $R_0 = \int_{x_0}^{x} f'(t)\, dt = f(x) - f(x_0)$

which gives result. $f(x) = f(x_0) + \sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k$

$+ R_n(x)$.

<u>Question</u>: how do you show Taylor's theorem with Lagrange remainder from Taylor's theorem with integral reminder? Simply using MVT for integrals:

$\exists\, \xi \in [x, x_0]$ s.t.

$$\frac{1}{n!} \int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n dt = \frac{f^{(n+1)}(\xi)}{n!} \int_{x_0}^{x} (x-t)^n dt = \frac{f^{(n+1)}(\xi)(x-x_0)^{n+1}}{(n+1)!}$$

Here is yet another way of presenting Taylor's theorem:

Theorem (Taylor's theorem with $h = x - x_0$)

If $f \in C^{n+1}[a,b]$ and $x, x+h \in [a,b]$ then:

$$f(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + E_n(h)$$

where $E_n(h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$ for some ξ between x and $x+h$.

Order of convergence We saw from computer lab that even if we have

$$\lim_{n\to\infty} x_n = x$$

the convergence "speed" can be very slow. How do we quantify this?

- **Linear convergence**: $\exists\, c < 1$ and N s.t. $\forall n > N$:

$$|x_{n+1} - x| \leq c\,|x_n - x|$$

- **super linear convergence** There is $\{\varepsilon_n\}$ such that $\varepsilon_n \to 0$ and

$$|x_{n+1} - x| \leq \varepsilon_n |x_n - x|$$

- **quadratic convergence** $\exists\, C > 0$ s.t.

$$|x_{n+1} - x| \leq C\,|x_n - x|^2$$

- **order α** $\exists\, C > 0$ s.t.

$$|x_{n+1} - x| \leq C\,|x_n - x|^{\alpha}$$

Of course it is easy to show that:

order $\alpha \geq 2$ conv $\Rightarrow$ quadratic conv. $\Rightarrow$ superlinear conv. $\Rightarrow$ linear conv. $\Rightarrow$ conv.

In practical terms, linear convergence means convergence is at least as good as that of a geometric series. Quadratic convergence is quite good: the accurate # of digits in approx doubles at each iteration. (example: Newton's method). order $\alpha > 2$ convergence is quite rare (we will see only one example in this class)

Big O and little o notation: Handy notation to compare convergence rates of sequences. Let x_n and α_n be two sequences.

Def We say $x_n = O(\alpha_n)$ (read: x_n is big oh of α_n) if
$\exists\ C > 0,\ n_0 \in \mathbb{N}$ s.t. $n \geq n_0 \Rightarrow |x_n| \leq C|\alpha_n|$

In the common case where $\alpha_n \to 0$ and $x_n \to 0$, saying $x_n = O(\alpha_n)$ means $x_n \to 0$ at least as fast as $\alpha_n \to 0$.

Def We say $x_n = o(\alpha_n)$ (read: x_n is little oh of α_n) if

$\exists\ \varepsilon_n \to 0$ s.t. $|x_n| \leq \varepsilon_n |\alpha_n|$

in the particular case where $\alpha_n \neq 0$, this means $\lim_{n\to\infty} \frac{|x_n|}{|\alpha_n|} = 0$

i.e. that x_n is very small (negligible) w.r.t α_n.

Example:

$$\frac{n+1}{n^2} = O\left(\frac{1}{n}\right)$$

$$\frac{1}{n \ln n} = o\left(\frac{1}{n}\right)$$

$$\frac{5}{n} + 2^{-n} = O\left(\frac{1}{n}\right)$$

$$2^{-n} = o\left(\frac{1}{n^3}\right)$$

To get a feeling for convergence rates let us go back to the example from computer lab:

$$\ln(1+1) - \sum_{k=1}^{n-1} \frac{(-1)^{k-1}}{k} = \mathcal{O}\left(\frac{1}{n}\right) \quad \left(\text{since } |E_n(1)| \leq \frac{1}{n+1}\right)$$

$$e^x - \sum_{k=0}^{n-1} \frac{1}{k!} x^k = \frac{1}{n!} e^{\xi} x^n = \mathcal{O}\left(\frac{1}{n!}\right)$$

↑ if we assume that $|x| < 1$
we have $|x|^n < 1$
and $e^{\xi} < e$

which converges faster?

The same big $\mathcal{O}$, little o notation is used for functions e.g.:

$$\cos x = 1 - \frac{x^2}{2} + \mathcal{O}(x^4) \quad \text{as } x \to 0$$

$$= 1 - \frac{x^2}{2} + o(x^2) \quad \text{as } x \to 0.$$

Here is what it means exactly:

Def We say $f(x) = \mathcal{O}(g(x))$ as $x \to x_*$ if:
$\exists C > 0, \exists \delta > 0$ s.t. $|x - x_*| < \delta \Rightarrow |f(x)| \leq C|g(x)|$

When $x_* = \infty$ we take "neighborhoods" of ~~infinity~~ infinity:
$\exists C > 0, \exists r > 0$ s.t. $x \geq r \Rightarrow |f(x)| \leq C|g(x)|$

Def We say $f(x) = o(g(x))$ as $x \to x_*$ if:
$\exists$ function $h(x)$ with $h(x) \to 0$, $\exists \delta > 0$
s.t. $|x - x_*| < \delta \Rightarrow |f(x)| \leq h(x)|g(x)|$

(+ similar def when $x_* \to \infty$)

Note: When using big O / little o notation for functions it is important to include the point of convergence otherwise statement can be misinterpreted. Take for example:

$$\frac{1}{x^2} = o\left(\frac{1}{x}\right) \quad \text{as } x \to \infty$$

but

$$\frac{1}{x} = o\left(\frac{1}{x^2}\right) \quad \text{as } x \to 0.$$

I already used the following mean value theorem for integrals to show how Taylor's theorem with integral remainder implies Taylor's theorem with Lagrange remainder.

Theorem (Mean Value Theorem for Integrals)

Let u, v be continuous real valued functions on $[a,b]$ and suppose $v \geq 0$. Then:

$$\exists \xi \in [a,b]. \text{ s.t.}$$

$$\int_a^b u(x)\, v(x)\, dx = u(\xi) \int_a^b v(x)\, dx$$

proof: Since $u(x)$ is continuous on $[a,b]$ we have:

$$\alpha \leq u(x) \leq \beta$$

$v(x) \geq 0 \Rightarrow \quad \alpha v(x) \leq v(x) u(x) \leq \beta v(x)$

Taking integrals $\quad \alpha I \leq \int_a^b v(x) u(x) dx \leq \beta I$, where $I = \int_a^b v(x)\, dx$.

If $I = 0$, there is nothing to prove since $v(x) \equiv 0$.

If $I \neq 0 \qquad \alpha \leq I^{-1} \int_a^b v(x) u(x) dx \leq \beta$.

By the intermediate value theorem for continuous functions, $\exists \xi \in [a,b]$ s.t.

$$u(\xi) = I^{-1} \int_a^b v(x)\, u(x)\, dx$$ ⇝ this proves result.

Floating point arithmetic:

For a detailed discussion on this topic I recommend:

Overton, Numerical computing with IEEE Floating point arithmetic, SIAM, 2001.

Floating point is based on the exponential (or scientific) notation.

$$x = \pm S \times 10^E \quad \text{where } 1 \leq S < 10 \text{ and } E \in \mathbb{Z}$$

(signed integer)

S = significand
E = exponent. (we shall only consider $x \neq 0$)

In computers it is more natural to use base 2 (since operations are binary on a computer):

$$x = \pm S \times 2^E, \quad \text{where } 1 \leq S < 2, \text{ and } E \in \mathbb{Z}$$

S can be expanded in base 2:

$$S = (b_0 . b_1 b_2 b_3 \ldots)_2 = \sum_{k=0}^{\infty} b_k 2^{-k}$$

e.g. $$\frac{11}{2} = (1.011)_2 \times 2^2 = \left(1 + \frac{1}{4} + \frac{1}{8}\right) \times 4 = \frac{11}{8} \times 4$$

Here $S = (1.\underbrace{b_1 b_2 b_3 \ldots}_{\text{fractional part}})_2$ → only one we need to keep.

Numbers are roughly stored as follows in a computer:

0	EXP. BITS	FRACTIONAL PART

SIGN BIT

The number of bits we assign to fractional part and exponents are important to know precision at which we work.

The precision of a floating point system is the # of bits used to represent significand (counting the hidden bit from normalization)

The most common floating point systems you will encounter are:

	Single precision	double precision
storage/number	32 bit = 4 bytes	64 bit = 8 bytes
precision	24 bits	53 bits
exponent	8 bits	11 bits
Matlab	N/A	by default all variables are double precision
C	Float	double
Fortran	REAL, REAL*4	DOUBLE PRECISION, REAL*8

Floating point numbers are a compromise since we cannot represent any real number with them. However we can get an idea of the error in representing a real number with floating point:

If we have precision p:

$$x = \pm (1.b_1 b_2 \ldots b_{p-1})_2 \times 2^E$$

The smallest x larger than 1 is:

$$(1.\underbrace{00..0}_{p-2}1)_2 \times 2^E = 1 + 2^{-(p-1)}$$

Let $\varepsilon = 2^{-(p-1)}$ = gap between 1 and next number

= <u>machine epsilon</u> (eps command in matlab)