Numerical Analysis ?

← subject of writing numerical approx to problems
Idea: Numerics is the heart of studying something which is closing to/close but not the same as the desired quantity

- Algorithms
- Analysis
- Implementation

(approximation)

Idea: get as close as possible to the original problem, construct a good approx

And we want to make sure this approx works within certain errors

(errors)

(python)

We can only approximate Def well posed problems . i.e.

1) there exist a unique solution
2) it depends continuously on the data

Idea behind numerics and numerical approx is that we simplify in a controlled manner

This means that we make the following replacements

infinite	finite
differential	algebraic
non-linear	linear problems

Difficult → "easy"

Example:

Given a function "f", a point x_0, compute "$f'(x_0)$" (derivative of f in x_0)

$f \in C^1[a,b]$ continuous in $[a,b]$ and first derivative continuous in $[a,b]$ $x_0 \in (a,b)$

Is it well posed? yes, because exist a UNIQUE derivative of f at x_0
If f is not derivable ⇒ problem is not well posed ⇒ WE MUST CHECK DOMAIN!

→ use finite difference

$$f'(x_0) \approx \frac{f(x_0+h) - f(x_0)}{h}$$

with small h ⇕ (s.t. $x_0+h \in (a,b)$)

(!! I'm not assuming to know anything about f except the evaluation of its body)

anything you choose here will have an effect of what you can do and what you can't

Error: Typical ways to estimate errors is → Taylor expansion around x_0:

$$f(x_0+h) = f(x_0) + h f'(x_0) + \frac{h^2}{2} f''(x_0) + \dots$$

When you compute powers of 2 (2**i , i ...),
The accuracy of computer is perfect up to
$\sim 10^{-4}$, and in reality is still perfect unto
$15^{14} \cdot 10^{-15}$

In theory, as $h \to 0$, the finite approx should get close to the exact result (exact derivative)

In practice much else happens

$$f(x_0+h) \approx \sum_{i=0}^{n} \frac{f^{(i)}(x_0)\, h^i}{i!}$$

For $n \to \infty$, $\approx \to =$ provided that function is analytical, so that I can do it for n functions at each step

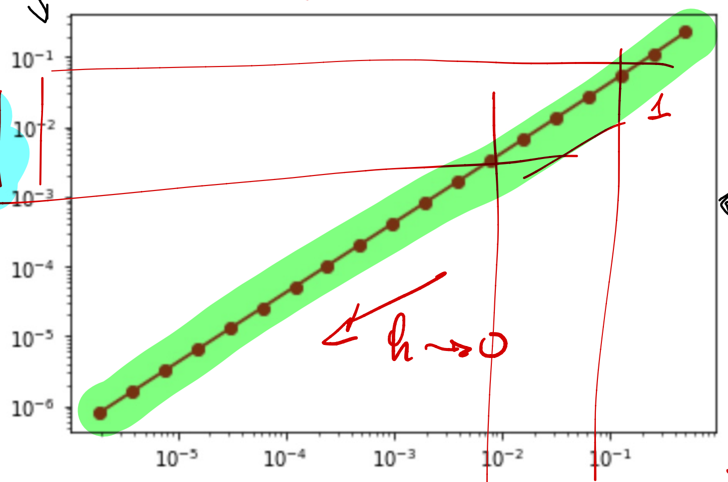
$$FD := \frac{f(x_0+h) - f(x_0)}{h} \simeq f'(x_0) + \frac{1}{2} f''(x_0)\, h + O(h^2)$$

exact derivative = + smth small that when $h \to 0$ means that it's way smaller compared to h

The error is computed as

Error: $|FD(f) - f'| \simeq \frac{f''(x_0)\, h}{2} + O(h^2)$

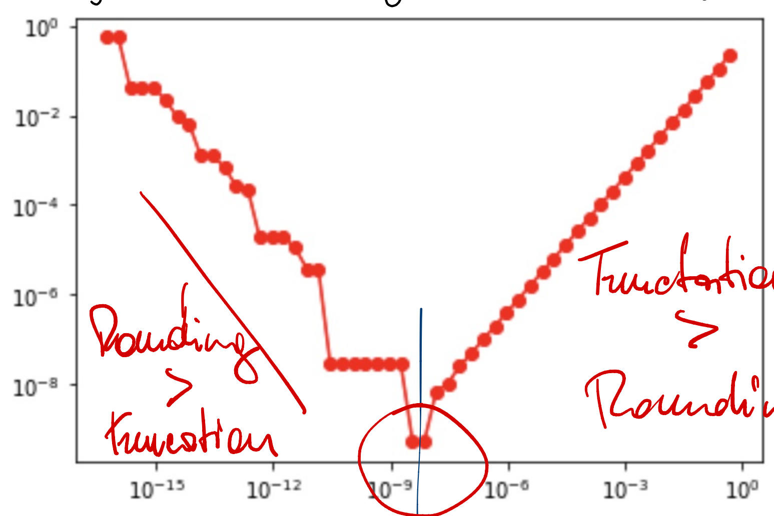
for h down to 2^{-20} ($\sim 10^{-6}$)



linearity in h (slope is = 1)

interesting things happens when h keeps getting smaller, but the algorithm above doesn't tell us that: it tells us that the error, being $\sim h$, will go to zero in a linear fashion as h goes down

h goes at lower values (10^{-15})



$\sim 10^{-15}$ $\quad 10^{-8}$

$\sqrt{\varepsilon} \sim$ optimal h

ε: precision of machine (Rounding precision)

Types of error are

- **Data error** → always present (Data) — If you have error in measurements of data, i.e. the x_0 above, and the function
- **Computational Error** (Algorithm) — depends on used algorithms — (gathering of the following)
  - **Truncation error** (Algorithm error in Exact Arithmetics) — not the one of PCs — ← error you make on the algorithm level by assuming that happens by multiplying exact of in the arts
  - **Rounding Error** (Finite Arithmetic) — the one of PC — Due to ≠ arithmetics of computers and numbers we deal with every day

diff between what you expect from algorithm by putting in exact, and what you get when using your algorithm (the approximation)

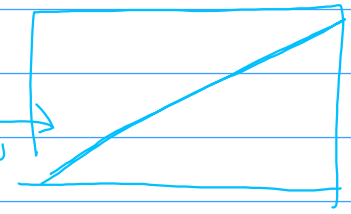
It doesn't take account of the fact that machines operate in a different arithmetics

arithmetics in pc is of power of 2

FINITE DIFF:

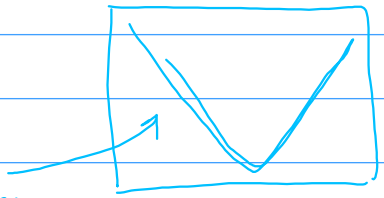
Infinite series - Truncated series

linear the graph above is TRUNCATION ERROR ONLY



because trunc. error dominates over rounding error

but, as h drops, rounding error starts becoming more relevant



here it dominates over the trunc error

The problem to solve is

Pb: Compute value of a function "F"

$$F : \mathbb{R} \longrightarrow \mathbb{R}$$

at a argument x

So we have

- x : true value of input
- $F(x)$: desired value of output ← what to compute
- $\hat{x}$: approximate input (may be "off" for rounding) — Error on input data; affected by rounding, data error, ...
- $\hat{F}(\hat{x})$: computed output (may be "off" for both rounding and truncation) — Error on outcome due to algorithm; due to approximation

$$\hat{F}(\hat{x}) - F(x) =$$

effect of rounding in input data on the algorithm → ROUNDING ERROR OR COMPUTATIONAL OUTPUT DUE TO ROUNDING ERROR IN INPUT (ROUNDING ERROR IN APPROXIMATED PROBLEM)

Truncation error is exact arithmetic

$$= \hat{F}(\hat{x}) - \hat{F}(x) + \hat{F}(x) - F(x)$$

(1) (2)

or

sensitivity of algorithm with respect to the original problem; consistency evaluated at # points

sensitivity of original problem with respect to perturbation in the data → ROUNDING ERROR IN EXACT PROBLEM

$$= \hat{F}(\hat{x}) - F(\hat{x}) + F(\hat{x}) - F(x)$$

(3) (4)

/!\ You only control the algorithm, and only know that rounding is there, but you can't remove it. You don't choose $h < 10^{-8}$!

COMPUTATIONAL ERROR (diff between algorithm and expected output)

(2) in exact arithmetic
(3) in rounding arithmetics

Rounding Precision (machine precision) : ε the largest number s.t.

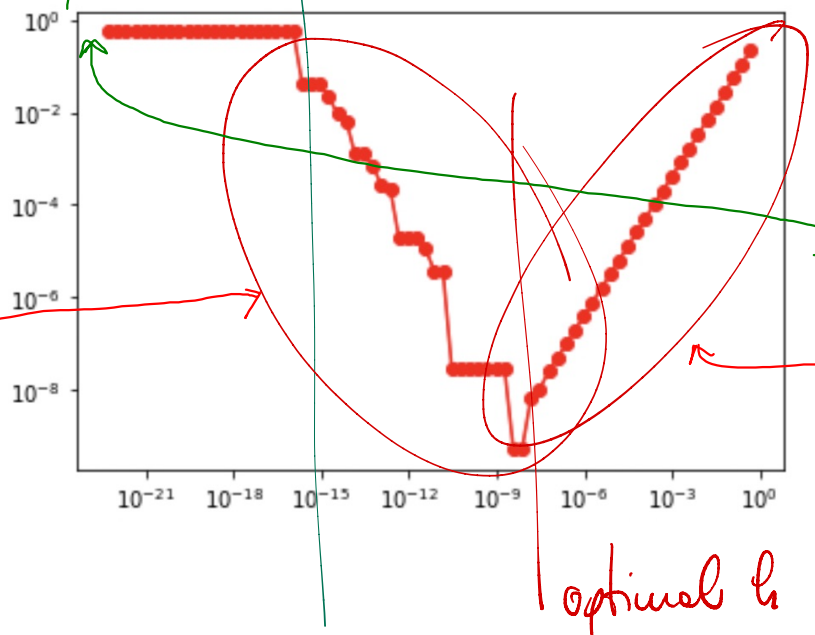
$$fl(1+\varepsilon) = fl(1)$$

= floating point representation of a number

gives optimal $h \sim \sqrt{\varepsilon}$
$h < \sqrt{\varepsilon}$ rounding error dominates

$fl(f(x+h)) = fl(f(x))$ in a computer $\rightarrow$ $f(x+h) - f(x) = 0$

false in exact arithmetic



in the example, the error here is exactly the value of the function at the point!

error here is order of 1!

Rounding error dominates

Truncation error dominates

optimal $h \simeq \sqrt{\varepsilon}$

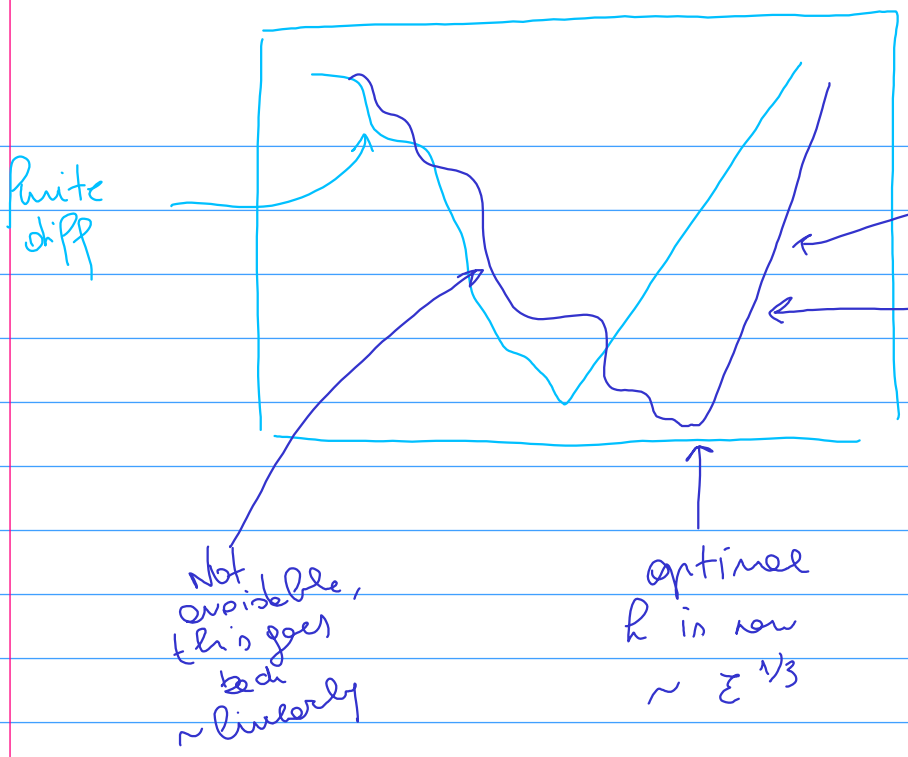
for 64 bits doubles (python) $\varepsilon \sim 10^{-15}$

$\sin(1) = 0.8414709848078965$

$$\frac{f(x+h) - f(x)}{h}$$ ← this h must be chosen according to property on f' and $\sqrt{\varepsilon}$, otherwise you'll get good results for some f' and bad for others.

All of this also depends on where we evaluate $f'(x_0)$ and truncation error at the same time

⚠ Don't use abs error as a measure of error, but use the RELATIVE ERROR


finite diff
central finite difference. BETTER, because here we have another power of h, which is power of 2 (slope two in 2!)
Not available, this goes to 0 linearly
optimal h is now ~ ε^{1/3}

Numerical Analysis - lec. 2

8.10.2020

problems can be written in a functional approach

$y = F(x)$ where $F: \mathcal{X} \longrightarrow \mathcal{Y}$

data $\rightarrow x \longrightarrow y = F(x)$ ← output/result

Functional Evaluation → F is a given function

EXAMPLES

1) $\mathcal{X}: \mathbb{R}$, $\mathcal{Y}: \mathbb{R}$ F: continuous func.

$$y = F(x)$$

2) **Sum of two numbers:**

$\mathcal{X}: \mathbb{R}^2$ ← pair of numbers, $\mathcal{Y}: \mathbb{R}$: F: sum of numbers.

$x = (a, b)$ $\quad F(x) = F((a,b)) = a+b$

3) **Computation of derivative**

$\mathcal{X}: C^1([a,b])$ $\quad \mathcal{Y}: \mathbb{R}$

F: evaluation of first derivative in $x_0 \in [a,b]$

$$y = F(x) := x'(x_0)$$

← function

Only well posed problems are characterized by 2 properties

1) $\forall x \in \mathcal{X}$, $\exists!\ y \in \mathcal{Y}$ s.t. $F(x) = y$

2) $\exists\ k$ (condition number) s.t. $\forall x, \hat{x} \in \mathcal{X}$

$$\| F(x) - F(\hat{x}) \|_{\mathcal{Y}} \leq k \| x - \hat{x} \|_{\mathcal{X}}$$

← basically F is continuous (bounded for every pair)

each of these are norms

We assume that $\mathbb{K}$, $\mathbb{V}$

Real Vector Spaces (Normed)

Def (Real vector spaces)

RVS: collection of objects for which it makes sense "V" to "sum" and "scale"

$$+ : V \times V \longrightarrow V$$

$$(u, v) \longrightarrow w = u + v$$

$$\cdot : V \times \mathbb{R} \longrightarrow V$$

such that

$$\forall\, u, v \in V, \quad \forall\, \alpha, \beta \in \mathbb{R}$$

$$\alpha u + \beta v = w \in V$$

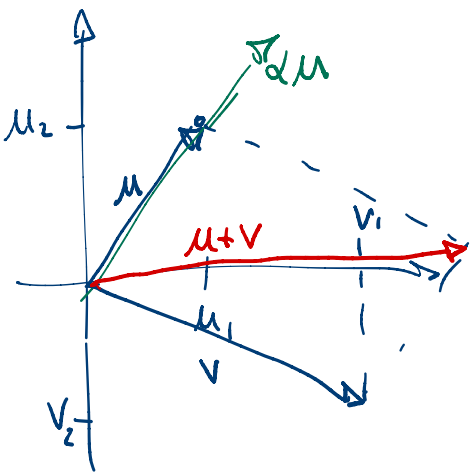
$\mathbb{R}^2$

$u = (u_1, u_2)$

$v = (v_1, v_2)$

$u + v = (u_1 + v_1, u_2 + v_2)$

$\alpha u = (\alpha u_1, \alpha u_2)$



$C^0([a,b])$

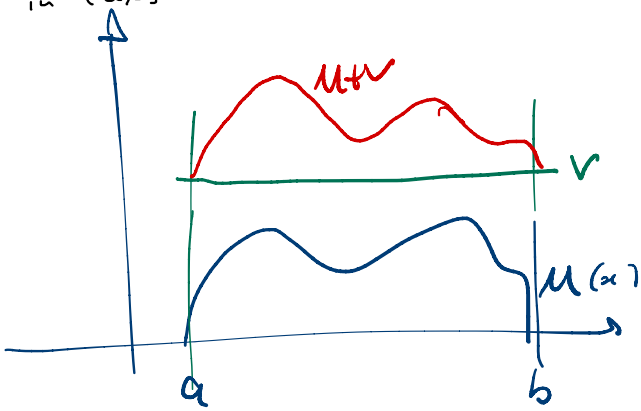
$u : [a,b] \longrightarrow \mathbb{R}$

$v : [a,b] \longrightarrow \mathbb{R}$

$w = u + v : [a,b] \longrightarrow \mathbb{R}$

because u, v are domained in $[a,b]$

$x \longrightarrow u(x) + v(x)$



Way to measure things

↓

introduce NORMS

Norms on Real Vector Spaces

a function $\|\cdot\| : V \longrightarrow \mathbb{R}_0^+$ ← positive or zero

1) $\|u\| \geq 0 \quad \forall u \in V$

2) $\|u+v\| \leq \|u\| + \|v\|$ Triangle inequality

3) $\|\alpha u\| = |\alpha| \|u\|$

optional 4) $\|u\| = 0 \Rightarrow u = 0$

without ④ it is a semi-norm

We will use

ℓ_p norms on $\mathbb{R}^n$

L_p norms on $\Omega \subset \mathbb{R}^d$

$\|\cdot\|_*$ operatorial norms induced by $\|\cdot\|_V$

Def:

$\mathbb{R}^n$: ℓ_p norm $\|u\|_p := \left(\sum_{i=1}^n |u_i|^p \right)^{\frac{1}{p}}$ ↘ $p \to \infty$

$\|u\|_\infty := \max_i |u_i|$ ℓ_∞ norm

↓ generalization to continuous domains

$\Omega \in \mathbb{R}^d$ L_p norms $\|u\|_p := \left(\int_\Omega |u|^p \right)^{\frac{1}{p}}$ ↘ $p \to \infty$

↙ L_∞ norm

$\|u\|_\infty := (\text{ess}) \sup_{x \in \Omega} |u(x)|$

$\|\cdot\|_*$ induced by vector space norm $\|\cdot\|_V$ on V and $\|\cdot\|_W$ on W

$A : V \longrightarrow W$ $\|A\|_* := \sup_{0 \neq x \in V} \frac{\|A(x)\|_W}{\|x\|_V}$

↑ used to measure how functionals / functions work / or what is their measure in a space

You def the norms of domain and codomain of A, and then you can def the operator norm

L_p norm of Matrices in $\mathbb{R}^{n\times m}$

(which is an obj of this kind)

$$A : \mathbb{R}^m \longrightarrow \mathbb{R}^n$$

$$A^{n\times m} \cdot x \qquad x \in \mathbb{R}^m$$

$$A \cdot x = y \in \mathbb{R}^n$$

$\|A\|_p$: L_p norm of the matrix → functional norm because I measure the norm of a function

$$\|A\|_p := \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|_p}{\|x\|_p} = \|A\|_*$$

$*$ operator id norm because A is a linear operator

induced by $\|\cdot\|_p$ on $\mathbb{R}^n$

$$A \in \mathbb{R}^{n\times m} \qquad A(x) = y \qquad \sum_{j=1}^{m} A_{ij} x_j = y_i$$

$$A \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$$

$\|\cdot\|_*$ norm on $\mathcal{L}(\mathbb{R}^m, \mathbb{R}^n) \equiv \mathbb{R}^{n\times m}$

Why we need them? Every quantity related to a well posed problem can be put in a way between two spaces and their norms.

well posed problems

1) $\forall x \in \underline{\underline{X}}$, $\exists!\ y \in \underline{\underline{Y}}$ s.t. $F(x) = y$

$\underline{\underline{X}}$: Normed Vector Space; $\underline{\underline{Y}}$: Normed Vector Space

2) $\exists\ K$ (condition number) s.t. $\forall x, \hat{x} \in \underline{\underline{X}}$

$$\|F(x) - F(\hat{x})\|_{\underline{\underline{Y}}} \leq (K_{Abs}) \|x - \hat{x}\|_{\underline{\underline{X}}}$$

← Here we use norms

3) Relative cond. number (only if $x \neq 0$ and $F(x) \neq 0$!)

K_{rel} s.t. $\forall x, \hat{x}$

↑ INDEPENDENT ON x

$$\frac{\|F(x)-F(\hat{x})\|_{\mathcal{Y}}}{\|F(x)\|_{\mathcal{Y}}} \leq \boxed{K_{rel}} \frac{\|x-\hat{x}\|_{\mathcal{X}}}{\|x\|_{\mathcal{X}}}$$

No denominator

Prop

A problem is well posed if $K_{rel/abs}$ is "small" ⟷ Relation must be CONTROLLABLE

↳ with respect to what? → relative: depends on problem handled

Example: sum of two numbers: (ℓ_1 norm)

$\mathcal{X}: \mathbb{R}^2 \qquad \mathcal{Y}: \mathbb{R}$

$x \in \mathbb{R}^2 \to \|x\|_1 := |x_1| + |x_2| \qquad \|y\|_1 = |y|$

We have that

$$\frac{\|F(x)-F(\hat{x})\|_1}{\|F(x)\|_1} = \frac{|x_1-\hat{x}_1+x_2-\hat{x}_2|}{|x_1+x_2|} \leftarrow \|\cdot\|_{\mathcal{Y}=\mathbb{R}}$$

$$\frac{\|x-\hat{x}\|_1}{\|x\|_1} = \frac{|x_1-\hat{x}_1|+|x_2-\hat{x}_2|}{|x_1|+|x_2|}$$

$\Delta y, \quad \Delta x \qquad \Delta x = x-\hat{x} \qquad \Delta y = F(x)-F(\hat{x})$

Now Find K_{abs}: $\exists K_{abs} \mid \forall x_1, x_2, \quad |\Delta y| \leq K \|\Delta x\|$?

$$|\Delta y| \leq K\|\Delta x\| \leq K\left(|\Delta x_1|+|\Delta x_2|\right)$$

Assume that $\Delta x \neq 0$

$$\frac{|\Delta y|}{|\Delta x|} \leq K$$

$$\frac{|\Delta x_1 + \Delta x_2|}{|\Delta x_1| + |\Delta x_2|} \leq \frac{|\Delta x_1| + |\Delta x_2|}{|\Delta x_1| + |\Delta x_2|} \leq 1$$

From pts of absolute K: this is perfect: you perturb data, and the perturbation of the result reduces proportionally the ptb of the inputs ($K_{abs} = 1$)

$K_{abs} = 1$ but ~~$\exists K < 1$ s.t.~~ Indeed:

$$\frac{|\Delta y|}{|\Delta x|} \cdot \frac{|x|}{|y|} \leq \frac{|x|}{|y|} \leq K_{rel}$$

$K_{abs} \leq 1$

because

$$\frac{|x|}{|y|} = \frac{|x_1| + |x_2|}{|x_1 + x_2|}$$

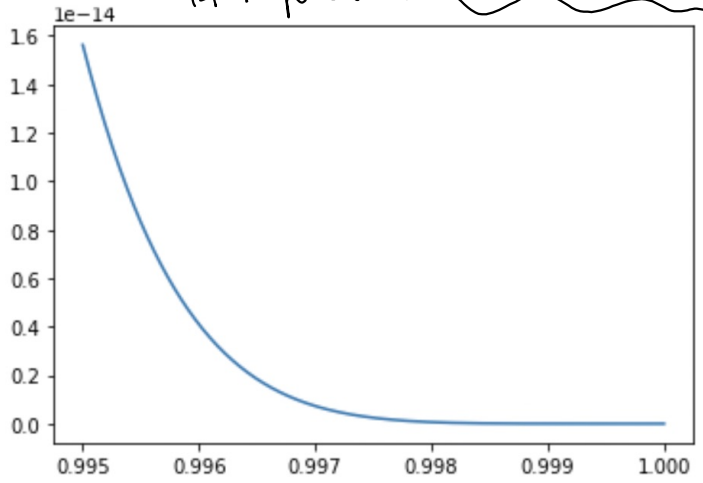
we cannot control this number!

can be as large as we want

=> If you try to sum 2 numbers with closing values and opposite signs, you're dealing with an ill posed problem, but the sum of 2 numbers is not ill posed by 6

Example: $(1-x)$ in $[1-\epsilon, 1]$ ϵ is small

If I plot




```
x = linspace(.995, 1,1029)
```

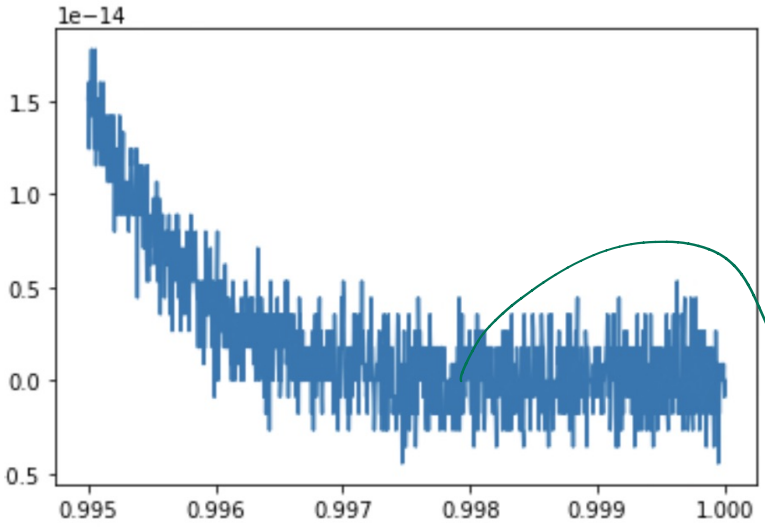
```
y = (x-1)**6
```

but if I expand the power I get a bad result

```
y = x**6 - 6*x**5 + 15*x**4 - 20*x**3 + 15*x**2 - 6*x + 1
```

↖ due to addition of very close numbers (in value) but with ≠ signs → we really that this is an ill posed problem

$x - \hat{x}$ = floating point errors (rounding errors)



Δy

Consider:

If $x_1 = -(x_2 + \varepsilon)$ with ε "small"

So what we obtain from procedure above

$$\frac{|x|}{|y|} = \frac{|x_2+\varepsilon| + |x_2|}{|-x_2 - \varepsilon + x_2|} \leq \frac{2|x_2|+|\varepsilon|}{|\varepsilon|}$$

($|x| \leq |x_2+\varepsilon| + |x_2|$; $x_1 = -x_2 - \varepsilon$)

if x_2 not close to zero then

$$\frac{|x|}{|y|} \leq \frac{2|x_2|}{|\varepsilon|} + 1$$

↖ if $|\varepsilon|$ is too small we get a problem

So, relating this to the true quantity

$$\frac{|\Delta y|}{|y|} \cdot \frac{|x|}{|\Delta x|} \leq K_{rel} \Rightarrow$$

which makes $K_{rel} = \infty$ to agree with its def (it has to be greater for every chosen ε)

$$\frac{|\Delta y|}{|\Delta x|} \frac{|x|}{|y|} \leq K_{rel}$$

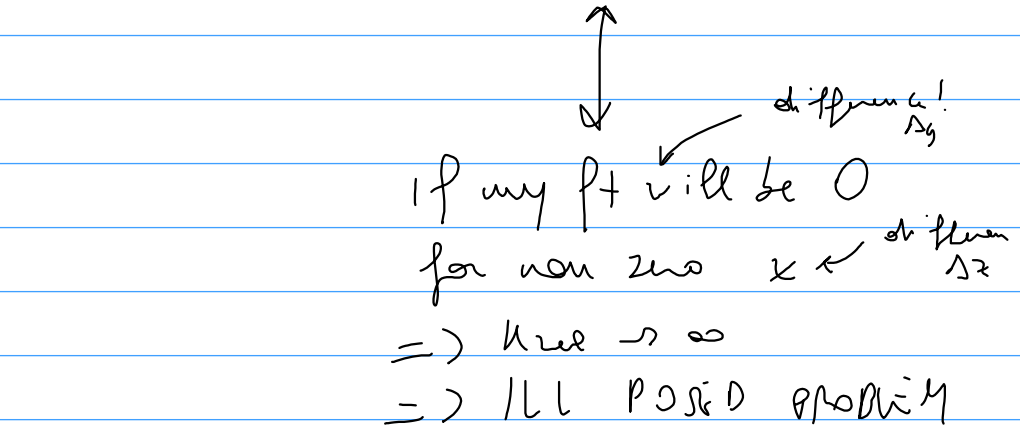
↖ in this case, it may also depend on x

Sum of 2 number is not a "RELATIVELY" well posed problem ($\leftrightarrow$ in relative terms $\leftrightarrow$ k rel)

But it is an "ABSOLUTELY" well posed problem

relative change in results caused by relative change in input, what you get is NOT controllable

THIS DEPENDS ON THE EVALUATION POINT



! Rounding errors are MAGNIFIED by big cancellation numbers

Numerical Analysis — Lec. 3 — LH — 13.10.2020

1) ∃! solution
2) pb. is stable ⟺ well posed ⟺ finite cond. Num.

X, Y Normed Vector spaces

$F: X \longrightarrow Y$ ← ∃! solution ⇒ F is an INVERTIBLE FUNCTION

We had

1) $K_{abs} \in \mathbb{R}^+ \cup \{+\infty\}$ ← positive and possibly infinite

s.t. $x, \hat{x} \in X$ LIPSCHITZ PROPERTY, $\|F(x) - F(\hat{x})\|_Y \leq K_{abs} \|x - \hat{x}\|_X$

used whenever is required to study stability of a problem and implies that there is a continuous dependency on data (how much you perturb the input ⟺ when you perturb the input data)

which is ⟺ to $K_{abs} \geq \dfrac{\|F(x) - F(\hat{x})\|_Y}{\|x - \hat{x}\|_X}$

You choose the smallest number greater than that fraction ⇒ $K_{abs} = \sup_{x, \hat{x} \in X,\ x - \hat{x} \neq 0} \dfrac{\|F(x) - F(\hat{x})\|_Y}{\|x - \hat{x}\|_X}$

If $K_{abs} \neq \infty$ ⇒ problem is stable (well conditioned)

1) $K_{rel} \in \mathbb{R}^+ \cup \{+\infty\}$

s.t. $x, \hat{x} \in X$

$$\frac{\|F(x) - F(\hat{x})\|_Y}{\|F(x)\|_Y} \leq K_{rel} \frac{\|x - \hat{x}\|_X}{\|x\|_X}$$

denominators make the relativity

Similarly as before, the smallest number greater than is

$$K_{rel} = \sup_{x, \hat{x} \in X,\ x - \hat{x} \neq 0,\ F(x) \neq 0} \frac{\|F(x) - F(\hat{x})\|_Y}{\|x - \hat{x}\|_X} \frac{\|x\|_X}{\|F(x)\|_Y}$$

I split the sup because I have INEQUALITIES and then I can bound the OVERALL definition and put K_{rel} equal to this

$= K_{abs} \sup_{x, \hat{x} \in X,\ x - \hat{x} \neq 0,\ F(x) = 0} \dfrac{\|F^{-1}(F(x)) - F^{-1}(F(\hat{x}))\|_X}{\|F(x)\|_Y}$ (remember F is invertible)

$= K_{abs} \|F^{-1}\|_*$

If F is not linear ⇒ K_{abs} is basically the Lipschitz constant of F^{-1}

use def of operatorial norm applied to F^{-1}: cond to be sure that F is invertible

$\delta x = x - \hat{x}$, $\delta y = y - \hat{y}$ (⟺ F is affine)

if operator is linear

Example: $Ax = y$ ($\Leftrightarrow y = F(x)$) $\quad y = F(x)$

Absolute stability: $A\hat{x} = \hat{y}$ ($\Leftrightarrow \hat{y} = F(\hat{x})$) → $A\,\delta x = \delta y$

$$\|Ax - A\hat{x}\| = \|A(x - \hat{x})\| = \|\delta y\|$$

I can collect the A thanks to the linearity of the problem

Expl. it property of operatorial norm of a matrix (follows from its definition)

$$\|\delta y\| \leq \|A\|_* \|\delta x\|$$

and we get Kabs

Finding Krel gives (by applying the def of Krel above)

$$\frac{\|\delta y\|}{\|y\|} \leq \|A\|_* \|A^{-1}\|_* \frac{\|\delta x\|}{\|x\|}$$

Condition number of A (this is a def: CONDITION NUMBER OF A MATRIX IS $\|matrix\| \|matrix^{-1}\|$)

it will depend on the norm you use

Prop

A problem is (Absolutely) Relatively well posed if

(Kabs) Krel is $< +\infty$.

What is an approximation?

Sequence of problems F_n $(n \in \mathbb{N})$ applied to a sequence of "approximated data" x_n (implied approximated sequence of domains $\mathbb{X}_n$)

s.t.

$$\lim_{n\to\infty} |F_n(x_n) - F(x)| = 0$$

← CONVERGENCE OF THE PROBLEM: this is really what we want to check

Note 1: "$\mathbb{X}_n \to \mathbb{X}$"

simple case: $\mathbb{X}_n \subset \mathbb{X}$

consistency ↦ Approximated problem converges to the exact problem that we have

2) Note 2: "$F_n \to F$"

this is the algorithm

Assume that $\mathbb{X}_n \equiv \mathbb{X}$ $\forall n$ → I can apply the algorithm to all pt of the "complicated space"

2) can be rewritten → $\lim_{n\to\infty} |F_n(x) - F(x)| \to 0$

I'm asking that the algo applied to exact input minus the exact problem (in modulus) goes to zero for increasing n

LAX-RICHTMEYR THEOREM:

For consistent pb $\left(\lim_{n\to\infty} |F_n(x) - F(x)| = 0\right)$

CONVERGENCE $\Longleftrightarrow$ STABILITY

err

input $\rightarrow x$ $\xrightarrow{F}$ y $\leftarrow$ output

Proof

if F is well posed $\Rightarrow$ sure exist $F^{-1}(y_n) = \hat{x}$ $\xrightarrow{F}$ y_n

can be an approx of x (a ntl or x itself) $\rightarrow x_n$ $\xrightarrow{F_n}$ y_n

evaluate it with algorithm

if you want to prove convergence you must show that

Convergence $\lim_{n\to\infty} |F_n(x_n) - F(x)| \rightarrow 0$

We have:

adding and subtracting

$$|F_n(x_n) - F(x_n) + F(x_n) - F(\hat{x}) + F(\hat{x}) - F(x)| = \dots \rightarrow 0$$

consistency $\rightarrow 0$

stability $\Leftarrow \leq K|x_n - \hat{x}_n|$

stability $\Downarrow$ $K|\hat{x} - x|$

All the errors falls into 3 categories

- **stability errors** (of continuous pb)
- " " (of discrete pb)
- **consistency error**

If we control this, convergence is implied by stability of the problem

If continuous prob is stable, and the discrete problem is stable and consistent $\Rightarrow$ we have convergence

Example: Interpolation (Polynomial interpolation)

for simplicity but it is gen topic

We consider the case of One dimension. $\mathcal{X} := C^0([a,b])$ ← continuous functions

and we study the following problem

Pb: Find a polynomial of order n (at most) that coincides with the input function at (given) $n+1$ interpolation points

! They are not inputs! They are part of the problem itself

CONSIDERATIONS

1. in $\mathcal{X}$ we use the L^∞ norm:

We want to understand what F is and what's the condition number of the problem

$$\|u\|_\infty := \max_{x\in[a,b]} |u(x)|$$

2. the space $\mathcal{Y}$ is $\mathbb{P}^n([a,b])$ (because the outcome is a polynomial of degree n): we use again $\|\cdot\|_\infty$ as a norm (same norm as norm chosen for $\mathcal{X}$, because polys are also continuous fcts)

$\mathcal{X}$ is FINITE DIM, while $\mathcal{Y}$ IS NOT!

Note 1: $\mathcal{Y}$ is finite dimensional (with dimension $n+1$), and we can write it as

$$\mathcal{Y} = \mathbb{P}^n([a,b]) = \text{span}\{v_i\}_{i=0}^{n}$$ (linear combination; same basis functions)

$$\Updownarrow$$

$$\forall p \in \mathcal{Y} = \mathbb{P}^n([a,b]) \quad \exists! \ \{p^i\}_{i=0}^{n} \in \mathbb{R}^{n+1} \text{ s.t.}$$ (p^i: coefficients)

$$p(x) = \sum_{i=0}^{n} p^i v_i(x)$$

linear independence of v_i, i.e. $\iff$ $p^i = 0 \quad i = 0, \dots, n \iff p(x) = 0 \ \forall x \in [a,b]$

Polynomial interpolation: (Given $n+1$ points $\{x_i\}_{i=0}^{n}$)

The algorithm takes a function (continuous) and returns a polynomial

$$F: \mathcal{X} = C^0([a,b]) \longrightarrow \mathcal{Y} = \mathbb{P}^n([a,b])$$

where u function $\longrightarrow$ p polynomial

$$p(x_i) = u(x_i) \quad i = 0, \dots, n$$

($n+1$ conditions; the given points)

$$\sum_{j=0}^{n} p^j v_j(x_i) = u(x_i)$$

(p^j: contravariant coefs)

(Can do this because I know that the space $\mathbb{P}^n$ is finite dimensional, meaning that I can uniquely identify every element of that space with a vector of coefficients with respect to a given basis)

$$\Downarrow$$

$$\sum_{j=0}^{n} V_{ij} p^j = u_i \iff V p = u$$

(u_i: covariant coefs)

must invent index summing to einstein notation

$$V_{ij} = v_j(x_i)$$

Vandermond Matrix (when $v_i(x) = \text{pow}(x, i)$)

what we doing: going from a vector of values of a pt in $C^0([a,b])$ to a vector of coefficient of a basis in $\mathbb{P}^n([a,b])$

Einstein notation: $V_{iJ}\, p^J \equiv \sum_{J=0}^{n} V_{iJ}\, p^J$

we set $\left(V_{iJ}\right)^{-1} = V^{iJ}$ so that $\boxed{V^{iJ} V_{Jk} = \delta^i_k = \begin{cases} 1 & \text{if } i=k \\ 0 & \text{if } i\neq k \end{cases}}$

So if needed

$$V_{iJ}\, p^J = u_i \quad \Updownarrow \quad p^J = V^{Ji}\, u_i \iff \underline{p} = \mathbb{V}^{-1}\, \underline{u}$$

If you want to solve the prob of interpolation $\Rightarrow$ solve a linear system $\Rightarrow$ condition number of problem is related to condition number of $\mathbb{V}$

we define the "Reduced problem": $\mathcal{X} \equiv \mathbb{R}^n \longrightarrow \mathcal{Y} \equiv \mathbb{R}^n$ ← LINEAR PROBLEM NOW!

where

$\mathcal{X}$:= set of values of u in x_i $\mathcal{Y}$:= coefficients of the polynomial

with norm $\|\underline{u}\|_2 := \left(\sum_{i=0}^{n} u_i^2\right)^{\frac{1}{2}}$ and same for $\mathcal{Y}$

The original problem can be thought in this problem with 2 additional functions

We want to find Condition number of F: $\underline{u} \longrightarrow \mathbb{V}\,\underline{u} = \underline{p}$ ← by similarity

that is input $\{u_i\}_{i=0}^n \longrightarrow$ output $\{p^J\}_{J=0}^n$

If I was suppose to perturb the input, starting from $\hat{u}$ I have $\hat{p} = \mathbb{V}^{-1}\hat{u}$

$$\|p-\hat{p}\|_2 \leq K_{abs}\, \|u-\hat{u}\|_2$$

We KNOW ALREADY THAT THE K_{abs} is given by the norm of the matrix (find out starting from $\|p-\hat{p}\|$ and using $\hat{p} = \mathbb{V}^{-1}\hat{u}$)

$$\|p-\hat{p}\|_2 \leq \|\mathbb{V}^{-1}\|\, \|u-\hat{u}\|_2$$

* As long as $\mathbb{V}$ is invertible, problem is always well posed, and $\|\mathbb{V}^{-1}\|_*$ is not infinite

The op norm is defined here as: $\|\mathbb{V}^{-1}\|_* := \left|\rho\left(\mathbb{V}^{-1}\right)\right| = \max_i \left|\lambda_i\left(\mathbb{V}^{-1}\right)\right| = \|\mathbb{V}^{-1}\|_2$

eigenvalues of $\mathbb{V}^{-1}$

op norm induced by the 2-norm

To make the problem better conditioned, we need to minimize $K_{abs} = \|\mathbb{V}^{-1}\|_2$ that is $\mathbb{V} \equiv \mathbb{I}$

then we have

$\mathbb{V} = I \Rightarrow V_J(x_i) = \delta_{Ji}$

which means we have a

Polynomial that is zero on every point x_i except on x_J where it is 1. We call this polynomial

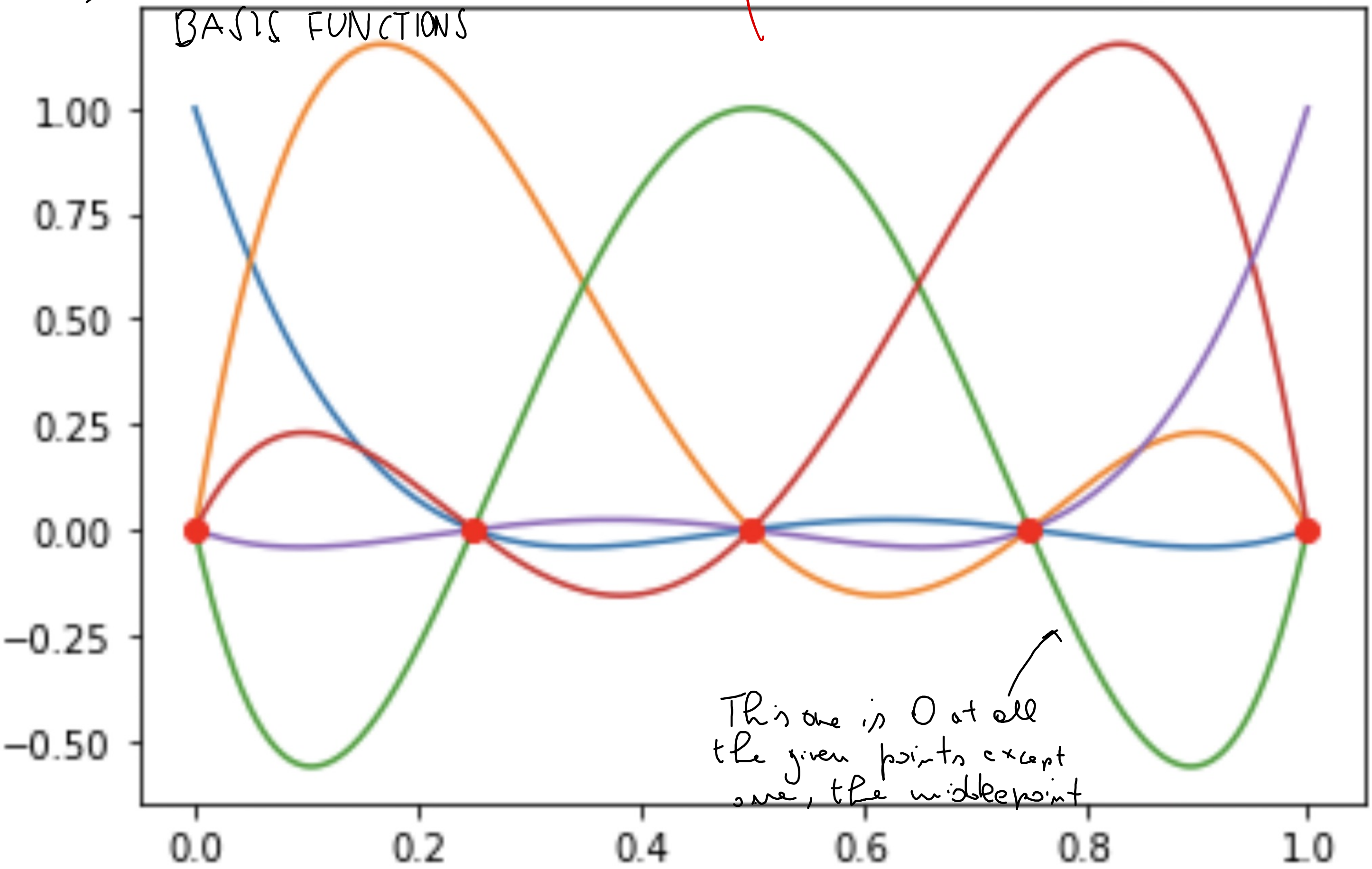
$$\ell_i := \prod_{\substack{i \neq J \\ J=0}}^{n} \frac{(x - x_J)}{(x_i - x_J)}$$

Lagrange Basis

polyn. of order n

$\Rightarrow \ell_i(x_J \neq x_i) = 0, \ell_i(x_i) = 1$

example with $n = 5$



This example shows you how to choose the basis function such that

$\mathbb{V} = I$

And you can use directly these basis functions to construct the polynomial interpolation

$$p(x) = \sum_{J=0}^{n} u(x_J)\, V_J(x)$$

where the coefficient are exactly

$$p^J = u(x_J)$$

because $\mathbb{V}^{-1} = \mathbb{I}$

⚠ If you have columns of $\mathbb{V}$ that are very similar to one another (i.e. because basis fct get closer and closer to one another) $\mathbb{V}$ will still be invertible but the condition number will be very bad!

The condition number of Vandermont matrix will deteriorate as we increase the number of points (simply because, the more points we have the similar all the basis functions will get!)

Numerical Analysis — lec 4 L+H — 15.10.2020

is a linear problem

General interpolation problem from $\mathcal{X} := C^0([a,b])$ (continuous fcts)
to the space $\mathcal{Y} := V := \text{span}\{v_i\}_{i=0}^n$ finite dimensional with dimension $(n+1)$
with fixed support points $\{x_i\}_{i=0}^n$

$$\Pi : C^0([a,b]) \longrightarrow V$$
$$u(x) \longrightarrow p(x) := \sum_{i=0}^{n} p^i v_i(x)$$

the choice of this (#of nmber pts) is important in principle you could have an higher # of points but then it would'nt be possible to satisfy exactly the relationship in which the points coincide at those higher # of points

such that p and u coincides in $\{x_i\}_{i=0}^n$ (higher than the # of basis fcts)

Indeed
$$\sum_{j=0}^{n} p^j v_j(x_i) = u(x_i)$$
this is a linear system
$$\equiv$$
$$\mathbb{V}_{ij} := v_j(x_i) \qquad \mathbb{V}\,\underline{p} = \underline{u}$$
$$\{\underline{p}\}^j = p^j$$
$$\{\underline{u}\}_i = u(x_i)$$

POWER BASIS IN $[0,1]$

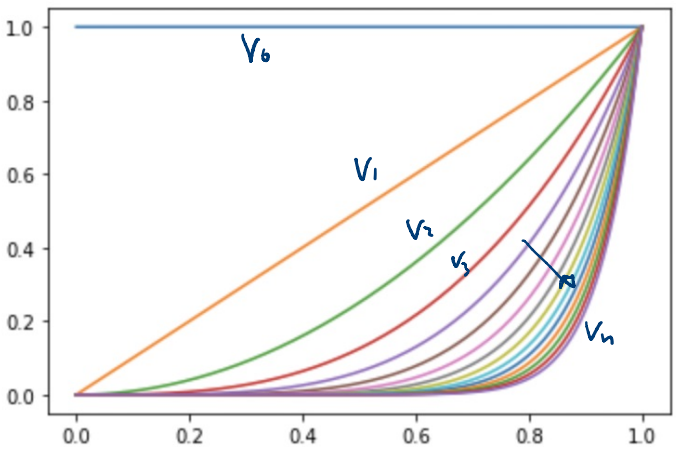
The basis functions $v_i \in C^0([a,b])$
basis fcts can be chosen among
(polynomials, cos/sin, sigmoids)

Example: $[a,b] = [0,1]$

$$v_i = \text{pow}(x,i) = x^i$$

if we fix $n+1 = 5$ $\quad \underline{p} := [1,0,0,2,3]$

$$p(x) = 1 + 2x^3 + 3x^4$$
$$= p^i v_i(x)$$



$$\underline{u} + \underline{v} = (u+v)$$
$$(\underline{u}+\underline{v})_i = u(x_i) + v(x_i)$$

then here we use

Linearity implies
$$\Pi(u+v) = \Pi(u) + \Pi(v)$$
because
$$\Pi(u) = \left(\mathbb{V}^{-1}\underline{u}\right)^i v_i(x) \qquad \Pi(v) = \left(\mathbb{V}^{-1}\underline{v}\right)^i v_i(x)$$
$$\Pi(u+v) = \left(\mathbb{V}^{-1}(\underline{u}+\underline{v})\right)^i v_i(x)$$

Condition number for interpolation problem

We use as norms in $C^0([a,b])$ $\|\cdot\|_\infty$, the same in V:

exploit the fact that the interpolation problem as we defined it is a linear problem

$$\|\Pi(u) - \Pi(\hat{u})\|_\infty = \|\Pi(u-\hat{u})\|_\infty \leq \|\Pi\|_* \|u-\hat{u}\|_\infty$$

by def. — can be seen as $\sim K_{abs}$

What is $\|\Pi\|_*$?

Let's apply the definition of a norm

$$\|\Pi\|_* := \sup_{0 \neq u \in C^0([a,b])} \frac{\|\Pi(u)\|_\infty}{\|u\|_\infty} = \sup_{0\neq u \in V} \frac{\| V^{ij} u(x_j)\, v_i(x) \|_\infty}{\|u\|_\infty}$$

(domain of the function "symbolized" by the problem Π; norm of V; inverse of matrix V; p^i; $p = \Pi u$)

$$\leq \sup_{0 \neq u \in V} \frac{\|V^{-1}\|_\infty \|u\|_\infty \max_i \|v_i\|_\infty}{\|u\|_\infty} \leq \|V^{-1}\|_\infty \max_i \|v_i\|_\infty$$

because for sure the maximum of the absolute value of u is greater than all the values of u at x_j, since x_j is a smaller number in the interval than all points in the interval!

u disappears => can get rid of the sup

in finite dimension all norms are equivalent

$$\text{cond}(V^{-1}) \max_i \|v_i\|_\infty \Rightarrow$$

cond number of the problem deteriorates if matrix deteriorates and deteriorates when basis function becomes too large

To minimize cond number you can play with 2 things: $\{x_i\}_{i=1}^N$, v_i

Two things: $\{x_i\}_{i=1}$, v_i

$$\Rightarrow \min \text{cond}(V^{-1}) \equiv \{v_i\}_{i=0}^n \text{ are}$$

the Lagrange basis functions:

$$\ell_i := \prod_{\substack{j \neq i \\ j=0}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

↳ $V = \mathrm{Id}$ by construction

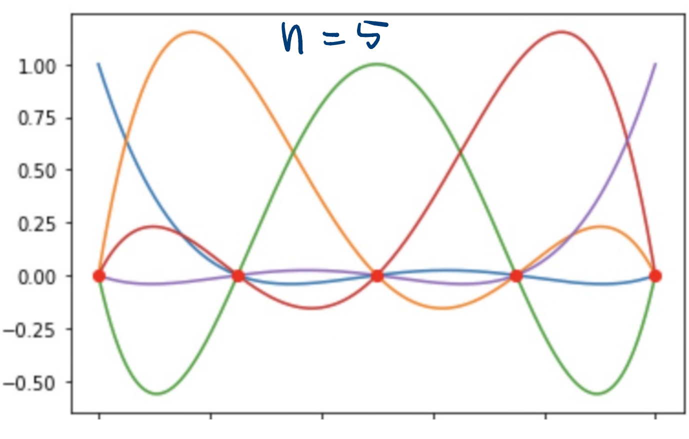
$$\rightsquigarrow V_{ij} := \ell_j(x_i) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

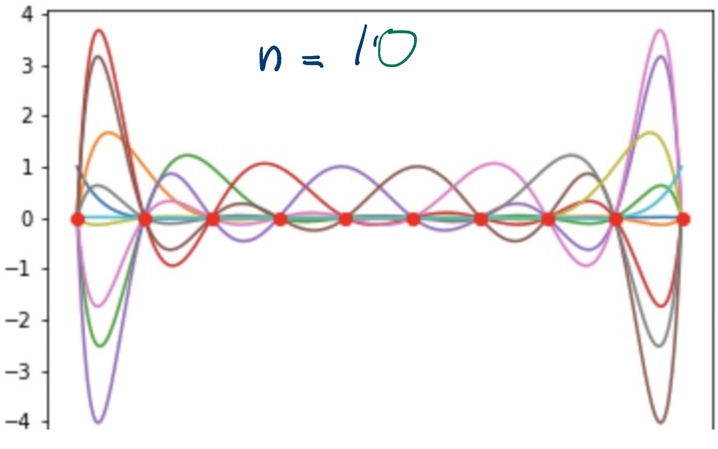
We should choose x_i s.t. $\max_i \|\ell_i\|_\infty$ is small

For power basis:

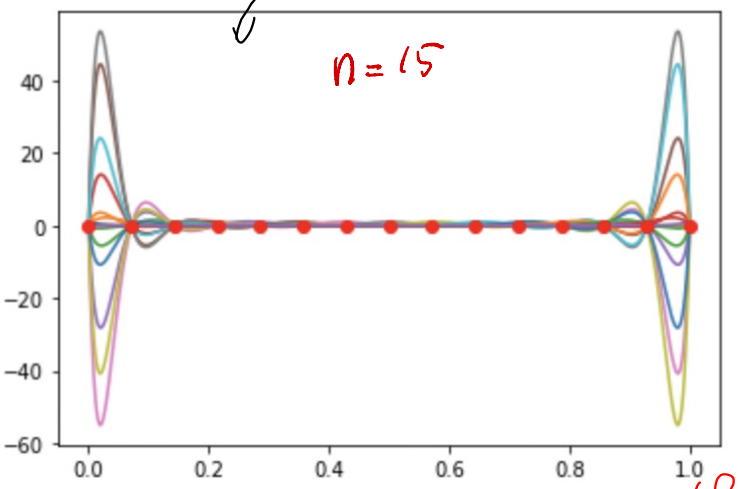
```
n = 1 : 1.0
n = 2 : 2.6180339887498953
n = 3 : 15.099657722502098
n = 4 : 98.86773850722759
n = 5 : 686.4349418185955
n = 6 : 4924.371056611224
n = 7 : 36061.16088021232
n = 8 : 267816.7009075794
n = 9 : 2009396.3800421846
n = 10 : 15193229.677753646
n = 11 : 115575244.54431371
n = 12 : 883478687.0721825
n = 13 : 6780588492.454725
n = 14 : 52214926084.1525
n = 15 : 403234616528.72504
```

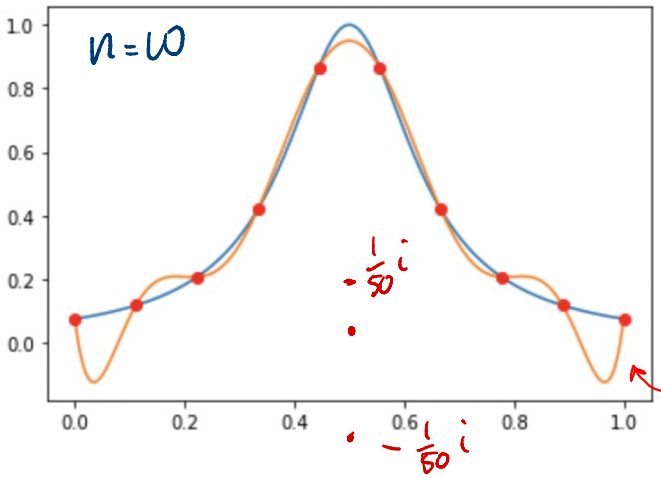
→ cond(V) for power basis

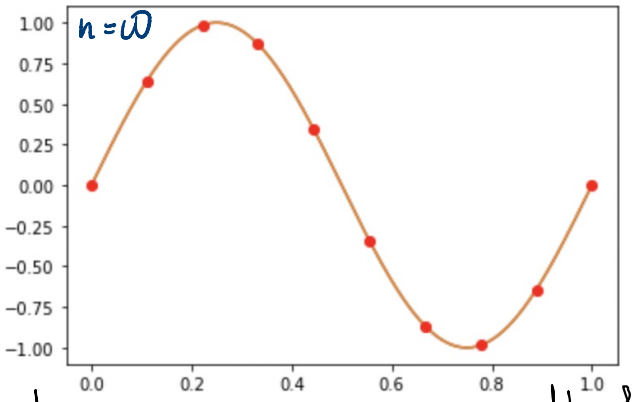




then you start getting bad oscillations







Is interpolation a good strategy for approximating a function?

the orange line represents, using LAGRANGE interpolation $\sum_{i=0}^{n} u(x_i)\, l_i(x) = p(x)$ with lagrange basis

blue line is the interpolated function $f(x) = \frac{1}{1+50(x-\frac{1}{2})^2}$

BAD BEHAVIOR HERE (away from the center) while at the center things are good

but this is FUNCTION DEPENDENT $f(x) = \sin(2\pi x)$

for sine results are perfect

Can we do better?

What can we say for smooth fcts about interpolation pities of pol interpol

Thm $f \in C^{(n+1)}([a,b])$ if $x_i \in (a,b)$, $\forall x \in (a,b)\ \exists \xi \in (a,b)$

(we here n+1 pb derivatives => needs for n+1 derivatives)

$$\left(f - \Pi^n f\right)(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)\, \omega(x)$$

Interpolation of f using n points

This is our error treasurement

where $\omega(x)$ is the characteristic polynomial of $\{x_i\}_{i=1}^{n}$

$$\omega(x) = \prod_{i=0}^{n} (x - x_i)$$

can we make these two numbers small?

First consequence: $\|\Pi f - f\|_\infty \leq \frac{\|\omega\|_\infty}{(n+1)!} \|f^{(n+1)}\|_\infty$

If can control the n+1 derivative of the function ANYWHERE :) you have a bound on the error at every point, simply bound by the norm of the n+1 derivative and the norm of the w function

proof: $\forall x$ define $G(t)$ s.t. it is equal to 0 in all points in which the polynomials and the function coincide (support points) plus the point x

$$G(t) = \big(f(t) - p(t)\big)\,\omega(x) - \big(f(x) - p(x)\big)\,\omega(t)$$

where $p(t) = \sum_{i=0}^{n} f(x_i)\,\ell_i(t) = (\mathbb{I}f)(t)$ ↝ interpolation function of f

$G(t)$ has $n+2$ ($n+1$ support points + x) zeros: $\{x_i\}_{i=0}^{n} \cup \{x\}$

According to Rolle's theorem (applied iteratively): $\exists\, \xi \in (a,b)$ s.t. $\dfrac{d^{n+1} G(\xi)}{dt^{n+1}} = 0$

f with two zeros $\Rightarrow$ $\exists\, \xi$ s.t. $f'(\xi) = 0$

writing explicitly the derivative

$$\frac{d^{n+1}}{dt^{n+1}}\,G(\xi) = f^{(n+1)}(\xi)\cdot\omega(x) - \big(f(x) - p(x)\big)\,(n+1)! = 0$$

$\frac{d^{n+1}}{dt^{n+1}}(\omega(t))$ — just removed

$$\Rightarrow\quad f(x) - (\mathbb{I}f)(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega(x) \qquad \blacksquare \quad \text{QED}$$

How to apply this?

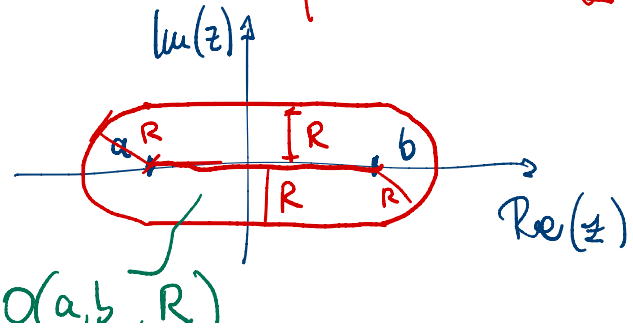
Thm 2: If f is analytically extendible in an ovoid $O(a,b,R)$ with $R > 0$

Then $\|f^{(n+1)}\|_\infty \le \dfrac{(n+1)!}{R^{n+1}}\,\|\tilde{f}\|_{L^\infty(O(a,b,R))}$

Analytically extendible means: it

Take x, and replace with $z \in \mathbb{C}$

we are asking that $f \in C^\infty$ in $O(a,b,R)$

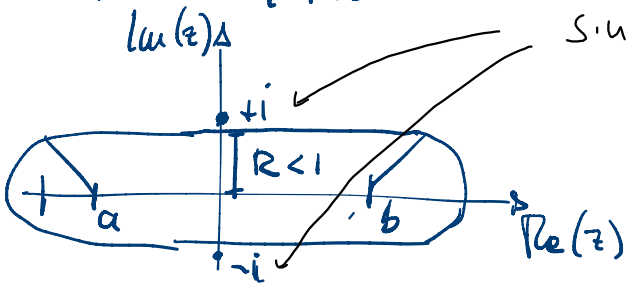


EXAMPLE

Runge function: $f = \frac{1}{1+x^2}$

extension to $\mathbb{C}$

$$f(z) = \frac{1}{1+z^2}$$



f is analytical in $\mathcal{O}(-\infty, +\infty, 1-\varepsilon) \quad \forall \varepsilon > 0$

What are the consequences?

$$\| f^{(n+1)} \| \leq \frac{(n+1)!}{R^{n+1}} \| f \|_\infty$$

$R < 1$ otherwise i get singularities

So, by the previous theorem:

$$\| p - f \|_\infty \leq \| f^{(n+1)} \|_\infty \frac{\| w \|_\infty}{(n+1)!} \leq \frac{(n+1)!}{(n+1)!} \frac{\| w \|_\infty}{R^{n+1}} \| f \|_\infty$$

where

$$\| w \|_\infty = \left\| \prod_{i=0}^{n} (x - x_i) \right\| \leq |b-a|^{(n+1)} \quad x \in (a,b)$$

since $x \in (a,b) \Rightarrow |x - x_i| \leq |b-a|$

$$\| p - f \|_\infty \leq \left(\frac{|b-a|}{R} \right)^{n+1} \| f \|_\infty$$

↳ good only when $|b-a| < R$, otherwise $\| p - f \|_\infty$ IS NOT BOUNDED

$\Rightarrow$ Runge: $\frac{1}{1+x^2}$ in $(-1, 1)$ everything ok

Can we do better?

Best Approximation. Definition:

Restrict ourselves

$\mathcal{X}$ Banach, reflexive, strictly convex

V subspace of $\mathcal{X}$

We call p the best approximation in V of f in $\mathcal{X}$ when:

$$\|f-p\| \leq \|f-q\| \quad \forall q \in V$$

in other words

$$\|f-p\| = E(f) = \inf_{q \in V} \|f-q\|$$

Thm: $\exists! p$ only in these cases (mandatory conditions)

How does polynomial interpolation compare to best approximation?

We can say that

$$\|f - \Pi^n f\|_\infty = \|f - p + p - \Pi^n f\|_\infty =$$

add and subtract best approx

p is a polyn of deg n $\Rightarrow$ its polyn interpolation is itself

$$= \|f - p + \Pi^n(p - f)\|_\infty$$

$$\text{triangular inequality} \leq \|f-p\|_\infty + \|\Pi^n\|_* \|p-f\|_\infty$$

$$\text{property of norms} \leq \left(1 + \|\Pi^n\|_*\right) \|p-f\|_\infty$$

how away we are from best interpolation

For Lagrange interpolation:

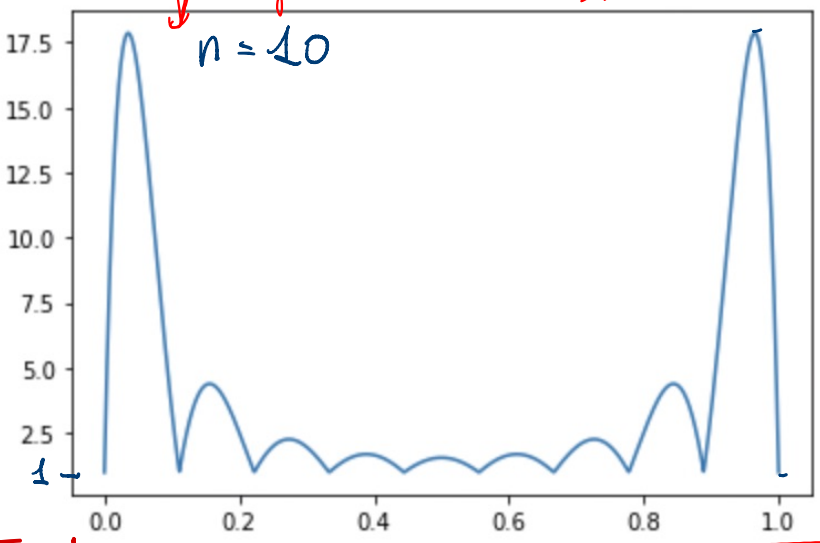
$$\|\Pi^n\|_* := \sup_{u \in C^0([a,b])} \frac{\|\sum_{i=0}^n u(x_i)\, l_i(x)\|_\infty}{\|u\|_\infty} \leq$$

$$= \|\sum_{i=0}^n |l_i(x)|\|_\infty$$

def: $\Lambda(x) := \sum_{i=1}^n |l_i(x)|$ LEBESGUE FUNCTION

$\Rightarrow \| \Pi^n_* \| = \| \Lambda \|_\infty$

Lebesgue functions are such that => bad results. Indeed $\| \Lambda \|_\infty$ for equispaced points grows quickly



$$\| \Lambda \|_\infty \leq \frac{2^{n+1}}{e n \log n}$$

A collection of points: $\{x_i\}_{i=0}^n$ is a list of lists of points with increasing size

Erdos

$\forall$ collection of points $\{x_i\}_{i=0}^n$ $\exists\, c > 0$ s.t.

$$\| \Lambda \|_\infty \geq \frac{2}{\pi} \log(n+1) - c$$

independently of how you choose the points, at best you get logarithmic with the number of the points

Faber

$\forall$ collection of points $\{x_i\}_{i=0}^n$ $\exists f$ s.t.

$\lim_{n \to \infty} \| \Pi^n f - f \|_\infty \to \infty$

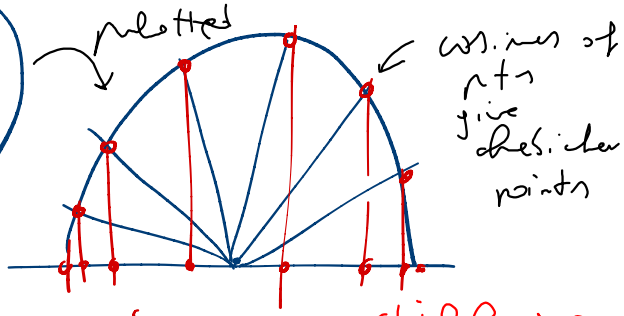
$\Leftrightarrow$ you can always find a f that destroys all

The best is $\alpha := \{x_i\}_{i=0}^n$ s.t. $\| \Lambda^\alpha \|_\infty \leq \| \Lambda^\beta \|_\infty$

$\forall\, \beta = \{\tilde{x}_i\}_{i=0}^n$

such $\alpha := \{x_i\}_{i=0}^n$ are called Chebyshev points: between $[-1, 1]$

defined as $x_i = \cos\left(\frac{(2i+1)\pi}{2n+2}\right)$



There's an estimate that says

$$\| \Lambda^\alpha \|_\infty \leq \frac{2}{\pi} \log(n+1) + 1$$

Best case scenario

still grows logarithmically

$$\frac{2}{\pi} \log(n+1) - c \leq \| \Lambda^\alpha \|_\infty \leq \frac{2}{\pi} \log(n+1) + 1$$

=) π INTERPOLATION IS GOOD BUT NOT π SO GOOD

Nonlinear equations

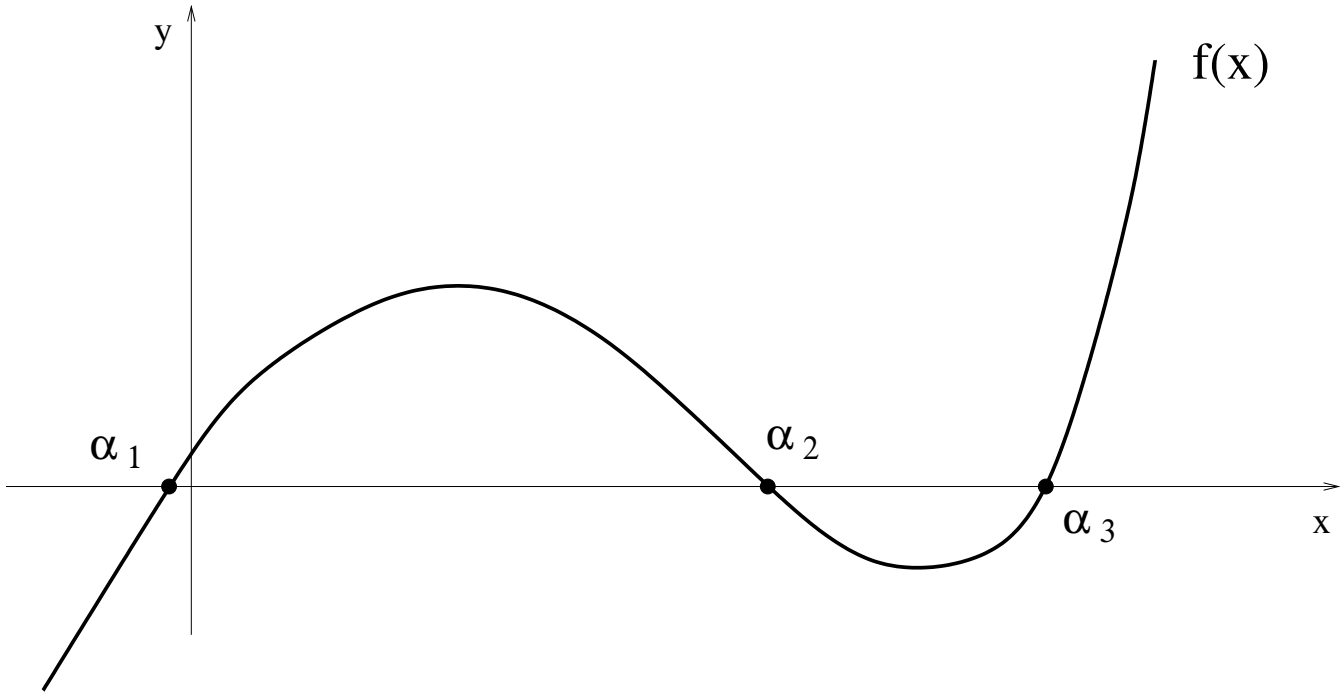
Numerical Analysis

Profs. Gianluigi Rozza - Luca Heltai

2019-SISSA mathLab Trieste

Nonlinear equations

Objective: Find the root of scalar (or vector) non-linear functions, i.e., find $\alpha \in \mathbb{R}$ such that $f(\alpha) = 0$.



Examples and motivation

Example 1 (Interest rates)**.** We want to compute the mean interest rate I of a portfolio over several years. We invest $v = 1000$ Euro every year. After 5 years we end up with $M = 6000$ Euro.
The relation between M, v, I and the number of years n is

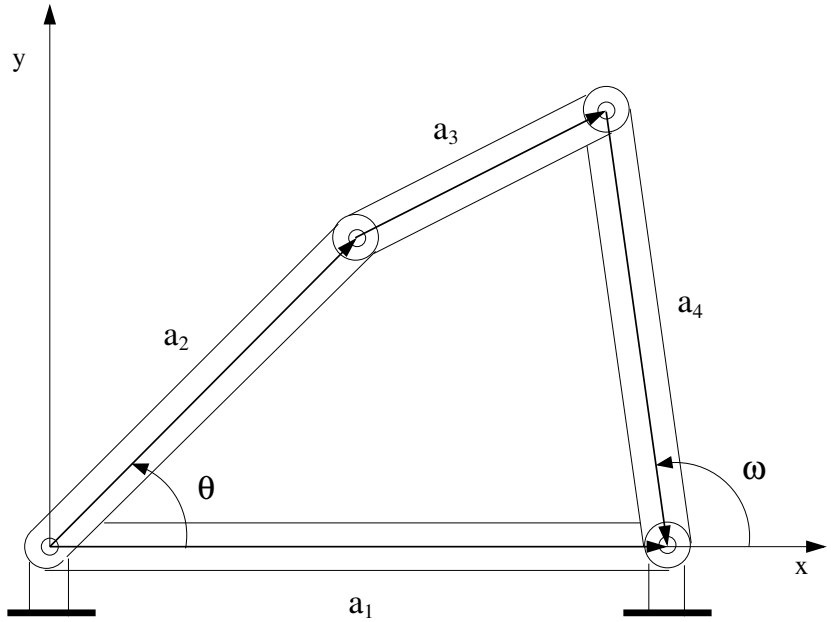
$$M = v \sum_{k=1}^{n} (1+I)^k = v \frac{1+I}{I} \left[(1+I)^n - 1 \right]$$

This can be rewritten as: *find I such that*

$$f(I) = M - v \frac{1+I}{I} \left[(1+I)^n - 1 \right] = 0 \tag{1}$$

Therefore we have to solve a nonlinear equation in I, for which we can't find an analytical solution.

Example 2 (Rods system)**.** Let us consider the mechanical system represented by four rigid rods



For any admissible angle ω, we want to compute the angle θ between $\mathbf{a}_1$ and $\mathbf{a}_2$.
Thanks to the vector identity

$$\mathbf{a}_1 - \mathbf{a}_2 - \mathbf{a}_3 - \mathbf{a}_4 = 0$$

and keeping $\mathbf{a}_1$ on the x-axis, we can derive the following equation: involving ω and θ:

$$\frac{a_1}{a_2}cos(\omega) - \frac{a_1}{a_4}cos(\theta) - cos(\omega - \theta) = -\frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4} \qquad (2)$$

where a_i is the length of the ith rod. Equation (2) is nonlinear and can be solved only for particular values of ω. For a general ω it is not possible to find an analytic solution.

Example 3 (State equation of a gas)**.** We want to determine the volume V occupied by a gas at temperature T and pressure p. The state equation (i.e. the equation that relates p, V et T) is

$$\left[p + a\left(\frac{N}{V}\right)^2\right](V - Nb) = kNT\ ,$$

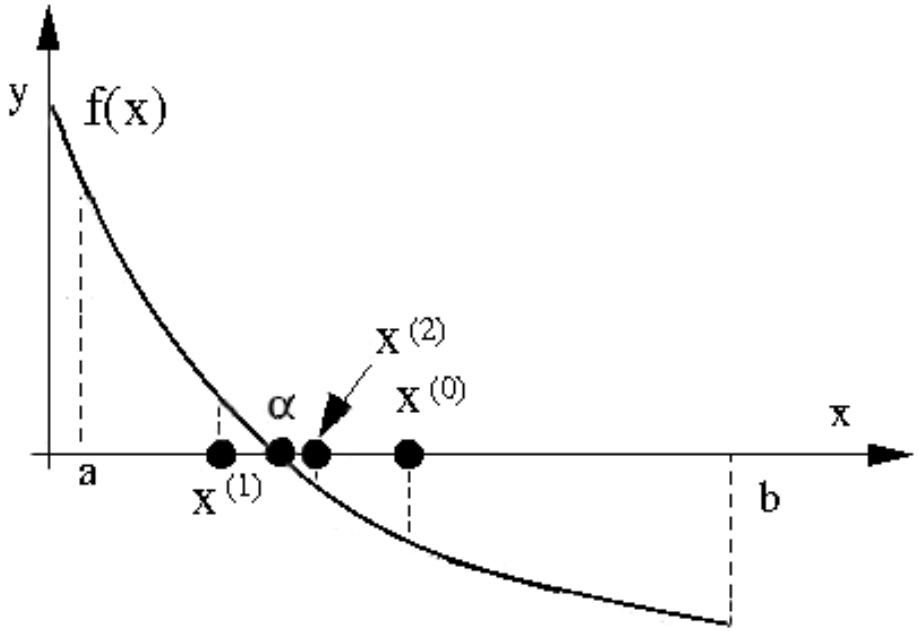
where a and b are two coefficients that depend on the specific gas, N is the number of molecules which are contained in the volume V and k is the Boltzmann constant. We need therefore to solve a nonlinear equation whose root is V.

Bisection method

(Book Sec. 2.2)

This method is used to compute the root of a continuous function f, i.e., the point α such that $f(\alpha) = 0$. We can build a *sequence* $x^{(0)}, x^{(1)}, \ldots, x^{(k)}$, ($x^{(0)}$ such that $\lim_{k\to\infty} x^{(k)} = \alpha$.

We assume that $f : (a, b) \to \mathbb{R}$ and $a < b$. If $f(a)f(b) < 0$, since f is continuous, we know that there exists (at least) one root α of f in the interval (a, b).



Then

1. We set $a^{(0)} = a$, $b^{(0)} = b$ and $x^{(0)} = \frac{a^{(0)}+b^{(0)}}{2}$,
2. if $f(x^{(0)}) = 0$, then $x^{(0)}$ is the zero.
3. if $f(x^{(0)}) \neq 0$, then:

   (a) if $f(x^{(0)})f(a^{(0)}) > 0 \Rightarrow$ the zero $\alpha \in (x^{(0)}, b^0)$ and we define $a^{(1)} = x^{(0)}$, $b^{(1)} = b^{(0)}$ and $x^{(1)} = (a^{(1)} + b^{(1)})/2$

   (b) if $f(x^{(0)})f(a^{(0)}) < 0 \Rightarrow$ the zero $\alpha \in (a^{(0)}, x^{(0)})$ and we define $b^{(1)} = x^{(0)}$, $a^{(1)} = a^{(0)}$ et $x^{(1)} = (a^{(1)} + b^{(1)})/2$

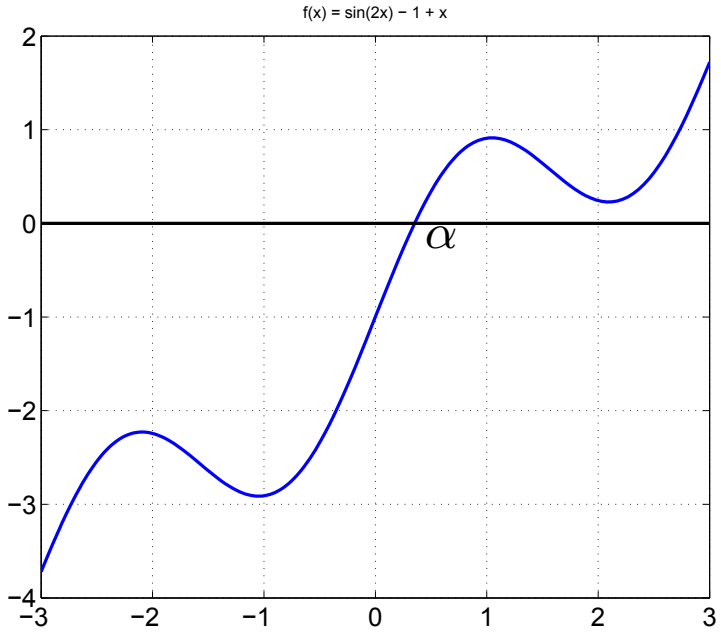
By the divisions of this type, we construct the sequence $x^{(0)}, x^{(1)}, \ldots, x^{(k)}$ that satisfies for all k,

$$|e^{(k)}| = |x^{(k)} - \alpha| \leq \frac{b-a}{2^{k+1}},$$

because we divide the interval by 2 at every step, starting from step 0

Example 4. We want to find the zero of the function $f(x) = \sin(2x) - 1 + x$. We draw the graph of the function f using the following commands in Matlab/Octave:

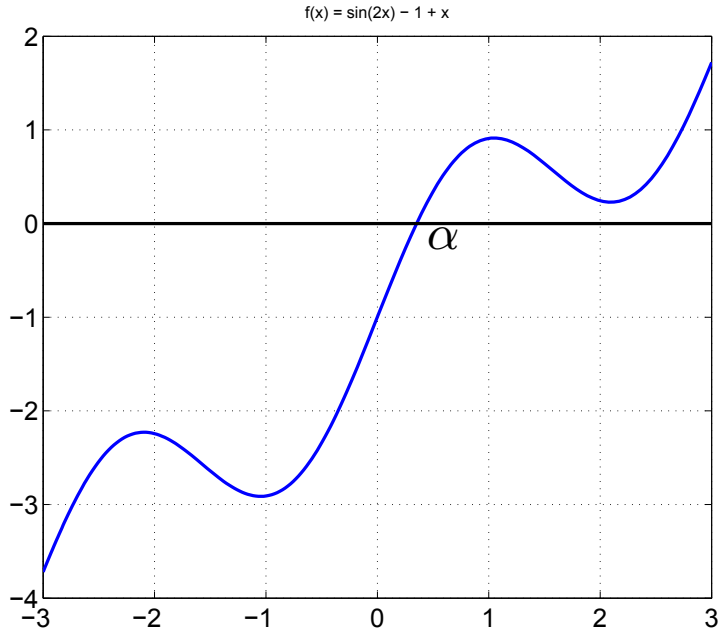
```
>> f = @(x) sin(2*x) -1 + x;
>> x=-3:0.1:3;
>> plot(x,f(x)); grid on;
```



If we appply the bisection method in the interval $[-1, 1]$ with a tolerance 10^{-8} and maximum number of iterations $k_{max} = 1000$

```
>> [zero,res,niter]=bisection(f,-1,1,1e-8,1000);
```

We find the value $\alpha = 0.352288462$ after 27 iterations.



APRIORI KNOWLEDGE OF # OF MAX ITERATION

Assume stopping criterion is based on size of interval.

At it k:

$$\frac{b-a}{2^k} < \varepsilon \quad \text{(} \varepsilon \text{: set tolerance)} \quad \rightarrow \text{solve for } k: \quad k \geq \log_2 \frac{b-a}{\varepsilon} = \frac{\ln(b-a) - \ln(\varepsilon)}{\ln 2}$$

Newton's method

(Chapt. 2.3 of the book)

Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function.
Let $x^{(0)}$ be an initial guess. Let us consider the equation $y(x)$ which passes through the point $(x^{(k)}, f(x^{(k)}))$ and which has the slope $f'(x^{(k)})$:

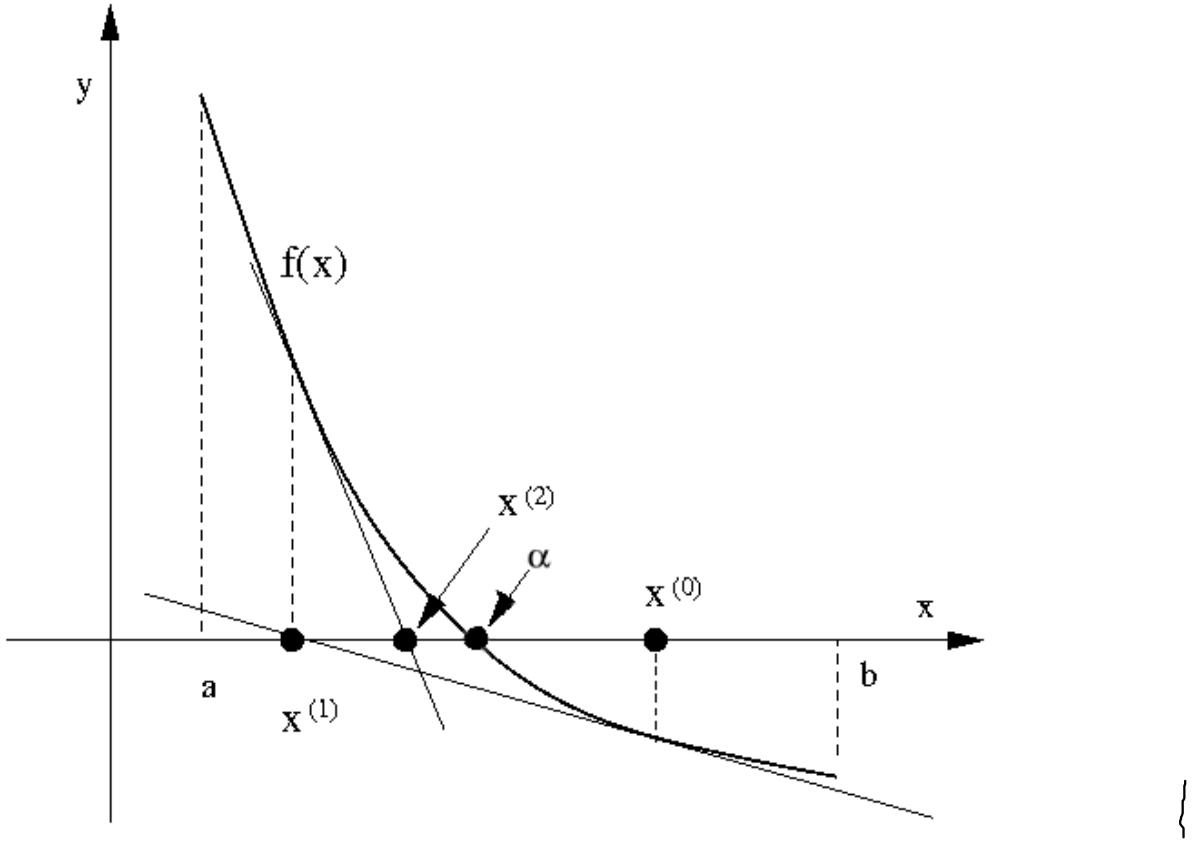
$$y(x) = f'(x^{(k)})(x - x^{(k)}) + f(x^{(k)}).$$

We define $x^{(k+1)}$ by the point where this line intersects the axis x, i.e. $y(x^{(k+1)}) = 0$. We deduce that:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k = 0, 1, 2 \ldots . \tag{3}$$

Newton's method

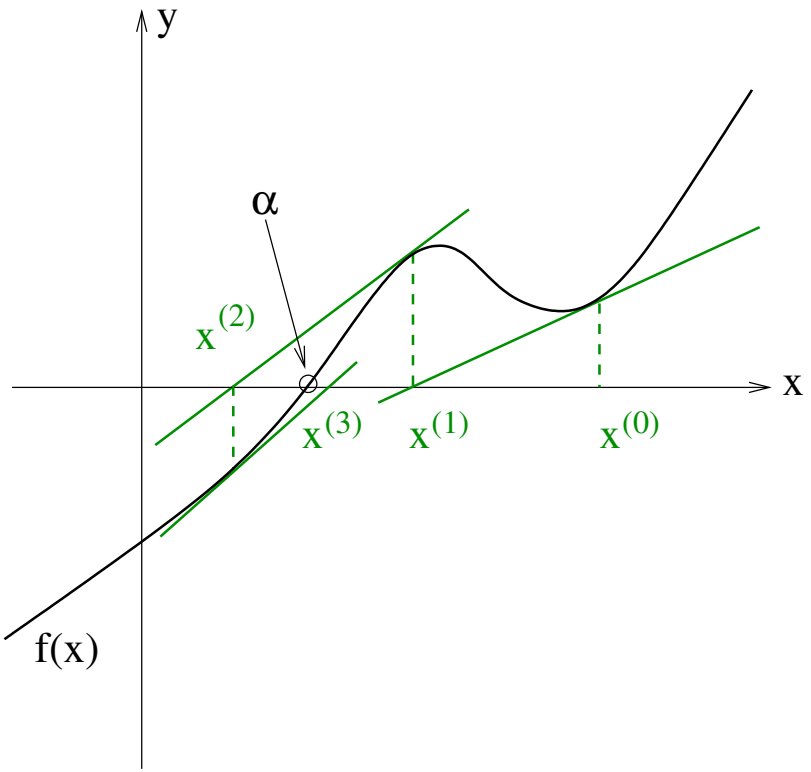
Starting from the point $x^{(0)}$, the sequence $\{x^{(k)}\}$ converges to the root of f

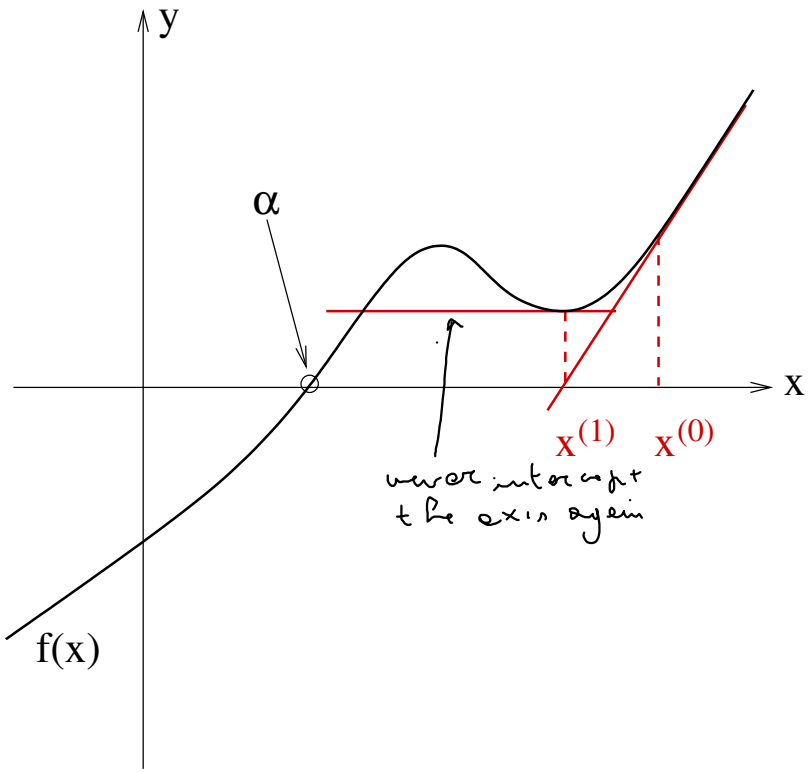


Convergence?

Does this method always converge?

- it depends on the property of the function;
- and on the initial guess.





Fixed point iterations.

(Chapt. 2.4 in the book)

A general method for finding the roots of a nonlinear equation $f(x) = 0$ is the transformation of the problem in an equivalent problem $x - \phi(x) = 0$, where the auxiliary function $\phi : [a, b] \to \mathbb{R}$ must have the following property :

$$\phi(\alpha) = \alpha \qquad \text{if and only if} \qquad f(\alpha) = 0.$$

and ϕ is called CONTRACTION

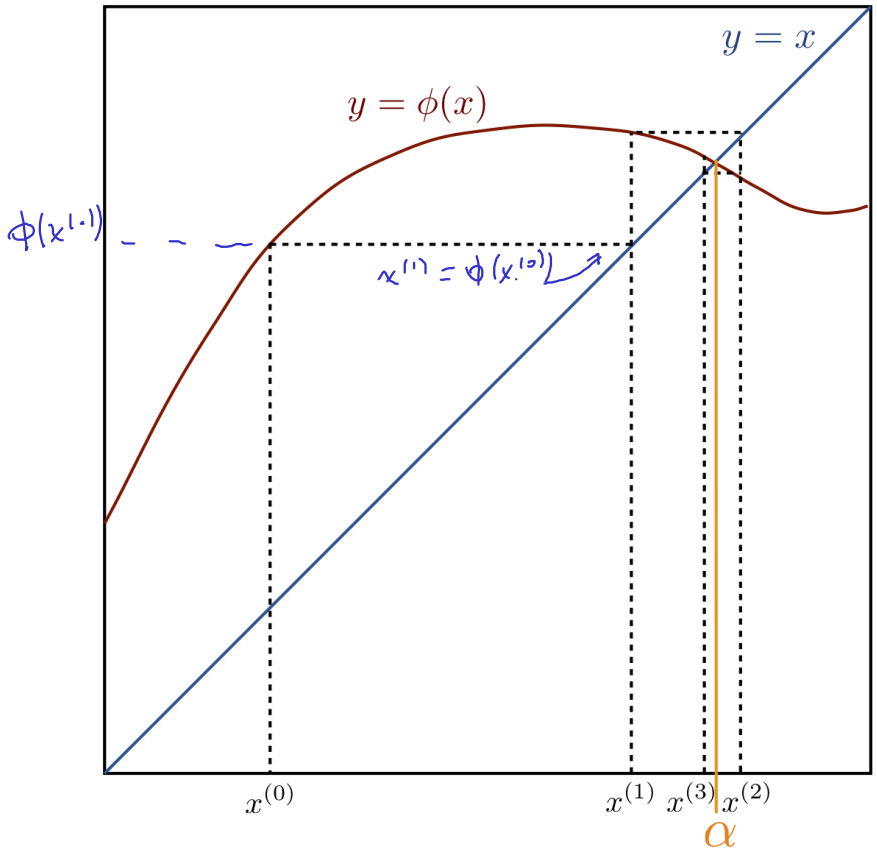
The point α is called *a fixed point* of ϕ. Searching the zeros of f is reduced to the problem of determining the fixed points of ϕ.

Idea : It could be computed by the following algorithm: $x^{(k+1)} = \phi(x^{(k)})$, $k \geq 0$. Indeed, if $x^{(k)} \to \alpha$ and if ϕ is continuous on $[a, b]$, then the limit α satisfies $\phi(\alpha) = \alpha$.

$x^{(k+1)} = \phi(x^{(k)}) \longrightarrow \phi(\alpha) = \alpha$ (follows from the definition of fct ϕ)

$x^{(k)} \to \alpha$

Starting from the point $x^{(0)}$, the sequence $\{x^{(k)}\}$ converges to the fixed point α

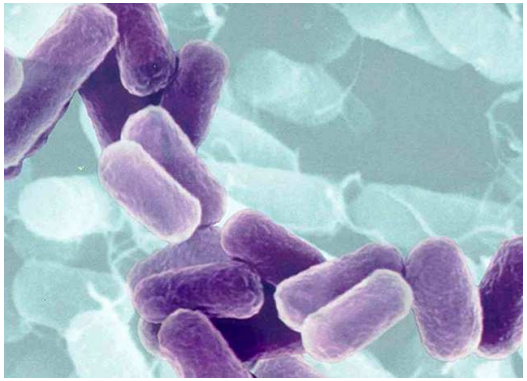


Example 5 (Population dynamics)**.**

In the survey population (e.g. bacteria), we want to establish a link between the number of individuals in the generation x and the number of individuals in the next generation x^+:

$$x^+ = \phi(x) = xR(x), \tag{4}$$

where $R(x)$ represents the rate of growth (or decay) of the considered population.



Several models are available for $R(x)$:

- The Malthusian growth model (Thomas Malthus 1766-1834),

$$x^+ = \phi_1(x) = xR_1(x) \text{ with } R_1(x) = r, \quad r \text{ is a positive constant}$$

- the model of growth with limited resources (Pierre François Verhulst, 1804-1849),

$$x^+ = \phi_2(x) = xR_2(x) \text{ with } R_2(x) = r/(1 + x/K), \quad r > 0, K > 0$$

K may be some kind of measurement for available resources → x/K is the number of members of population per resource

that improve the Malthausian growth model by taking into account the growth of a population is limited by the resources.

- the predator/prey model

$$x^+ = \phi_3(x) = xR_3(x) \text{ with } R_3(x) = rx/(1 + (x/K)^2)$$

that represents the change of the Verhulst model by presence of an antagonist population.

The dynamics of a population is defined by a iterative process, starting from a given initial guess ($x^{(0)}$),

$$x^{(k+1)} = \phi(x^{(k)}), \quad k \geq 0,$$

where $x^{(k)}$ represents the number of individuals in k-th generation.
In addition, the steady states (equilibriums) x^* of a considered population are identified by the following problem,

$$x^* = \phi(x^*), \tag{5}$$

or equivalently,

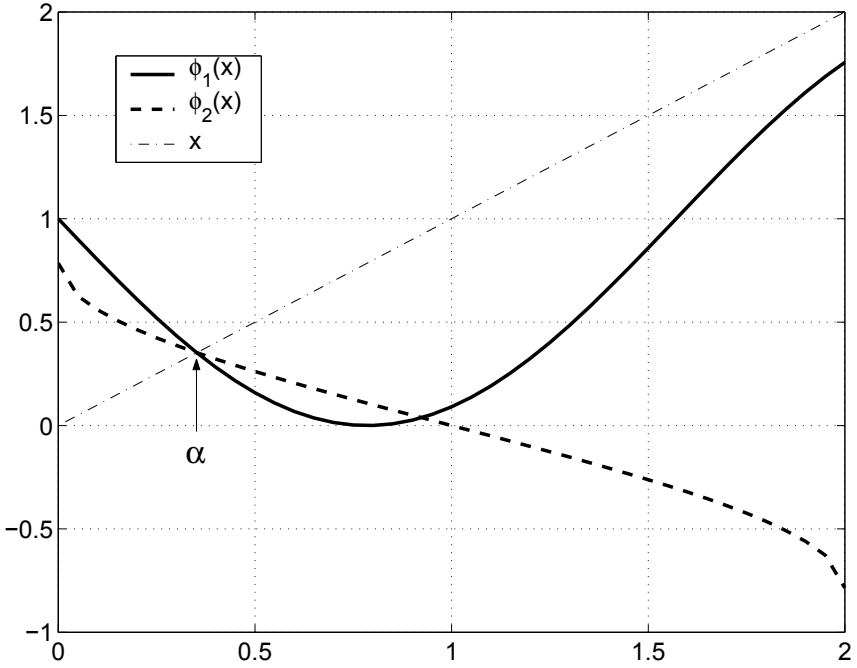
$$x^* = x^* R(x^*), \quad \text{i.e.} \quad R(x^*) = 1. \tag{6}$$

In both cases, we ask to solve a non-linear problem. In particular the problem is called the fixed point problem.

Example 6. We consider the equation $f(x) = \sin(2x) - 1 + x = 0$. We can rewrite it in two different fashions:

$$x = \phi_1(x) = 1 - \sin(2x)$$

$$x = \phi_2(x) = \frac{1}{2}\arcsin(1-x), \qquad 0 \le x \le 1$$



that is: the pair $(x, \phi(x))$ belongs to a square

Proposition 1. *(Global convergence)*

1. Assume that $\phi(x)$ is continuous on $[a, b]$ and such that $\phi(x) \in [a, b]$ for all $x \in [a, b]$; then there exists at least one fixed point $\alpha \in [a, b]$ of ϕ.

2. If $\exists L < 1$ such that $|\phi(x_1) - \phi(x_2)| \leq L|x_1 - x_2|$ $\forall x_1, x_2 \in [a, b]$,

then there exists a unique fixed point $\alpha \in [a, b]$ and the sequence $x^{(k+1)} = \phi(x^{(k)})$, $k \geq 0$ converges to α, for any initial guess $x^{(0)} \in [a, b]$.

Proof.

1. The function $g(x) = \phi(x) - x$ is continuous in $[a, b]$ and, thanks to assumption made on the range of ϕ, it holds $g(a) = \phi(a) - a \geq 0$ and $g(b) = \phi(b) - b \leq 0$. By applying the theorem of zeros of continuous functions, we can conclude that g has at least one zero in $[a, b]$, i.e. ϕ has at least one fixed point in $[a, b]$.

2. Indeed, should two different fixed points α_1 and α_2 exist, then

$$|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq L|\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2|,$$

which cannot be. There exists a unique fixed point $\alpha \in [a, b]$ of ϕ. □

Let $x^{(0)} \in [a,b]$ and $x^{(k+1)} = \phi(x^{(k)})$. We have

$$0 \leq |x^{(k+1)} - \alpha| = |\phi(x^{(k)}) - \phi(\alpha)| \leq L|x^{(k)} - \alpha| \leq \ldots \leq L^{k+1}|x^{(0)} - \alpha|,$$

fixed point; (1); (2); by applying 1 and 2 iteratively

i.e.

$$\frac{|x^{(k)} - \alpha|}{|x^{(0)} - \alpha|} \leq L^k.$$

Because $L < 1$, for $k \to \infty$, we obtain

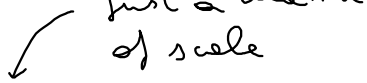
$$\lim_{k\to\infty} |x^{(k)} - \alpha| \leq \lim_{k\to\infty} L^k = 0.$$

ok, because $|x^{(0)} - \alpha|$ is constant

So, $\forall x^{(0)} \in [a,b]$, the sequence $\{x^{(k)}\}$ defined by $x^{(k+1)} = \phi(x^{(k)})$, $k \geq 0$ converges to α when $k \to \infty$.

Remark 1.
If $\phi(x)$ is differentiable in $[a, b]$ and

$\exists K < 1$ *such that* $|\phi'(x)| \leq K \ \forall x \in [a, b]$,



then the condion 2 of the proposal (1) is satisfied. This assumption is stronger, but is more often used in practice because it is easier to check.

Definition 1. *For a sequence of real numbers* $\{x^{(k)}\}$ *that converges,* $x^{(k)} \to \alpha$, *we say that the convergence to* α *is linear if exists a constant* $C < 1$ *such that, for* k *that is large enough*

$$| x^{(k+1)} - \alpha | \leq C | x^{(k)} - \alpha | .$$

If exists a constant $C > 0$ *such that the inequality*

$$| x^{(k+1)} - \alpha | \leq C | x^{(k)} - \alpha |^2$$

is satisfied, we say that convergence is quadratic.
In general, the convergence is with order p, $p \geq 1$, *if exists a constant* $C > 0$ *(with* $C < 1$ *when* $p = 1$*) such that the following inequality is satisfied*

$$| x^{(k+1)} - \alpha | \leq C | x^{(k)} - \alpha |^p .$$

$C > 1$ => doesn't converge

Proposition 2. *(Local convergence - Theorem 2.1 in the book)*
Let ϕ be a continuous and differentiable function on $[a,b]$ and α be a fixed point of ϕ. If $|\ \phi'(\alpha)\ | < 1$, then there exists $\delta > 0$ such that, for all $x^{(0)}$, $|\ x^{(0)} - \alpha\ | \leq \delta$, the sequence $\{x^{(k)}\}$ defined by $x^{(k+1)} = \phi(x^{(k)})$ converges to α when $k \to \infty$.
Moreover, it holds

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'(\alpha).$$

Proof: exp $\phi(x^{(k)})$ around α, proceed as in proof of next prop

Note that, if $0 <|\ \phi'(\alpha)\ | < 1$, then for any constant C such that $|\phi'(\alpha)| < C < 1$, if k is large enough, we have:

$$|\ x^{(k+1)} - \alpha\ | \leq C\ |\ x^{(k)} - \alpha\ |\ .$$

$\frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha}$ will reach $\phi'(\alpha)$ from above or below $\Rightarrow$ in any case, it will be $< C$

Proposition 3. *(Proposition 2.2 in the book) Let ϕ be a twice differentiable on $[a, b]$ and α be a fixed point of ϕ. Let us consider that $x^{(0)}$ converges locally. If $\phi'(\alpha) = 0$ and $\phi''(\alpha) \neq 0$, then the fixed point interations converges with order 2 and*

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{\phi''(\alpha)}{2}.$$

Proof. Using the Taylor series for ϕ with $x = \alpha$, we have

$$x^{(k+1)} - \alpha = \phi(x^{(k)}) - \phi(\alpha) = \underbrace{\phi'(\alpha)}_{0}(x^{(k)} - \alpha) + \frac{\phi''(\eta)}{2}(x^{(k)} - \alpha)^2$$

where η is between $x^{(k)}$ and α. So, we have

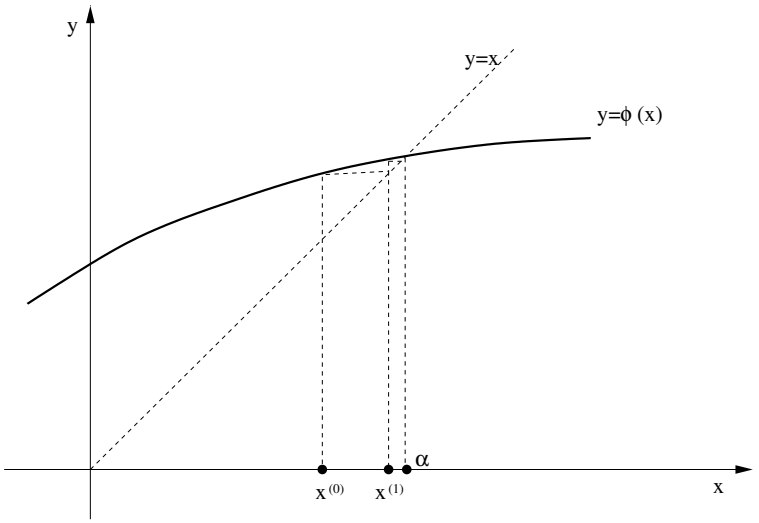
$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \lim_{k\to\infty} \frac{\phi''(\eta)}{2} = \frac{\phi''(\alpha)}{2}.$$

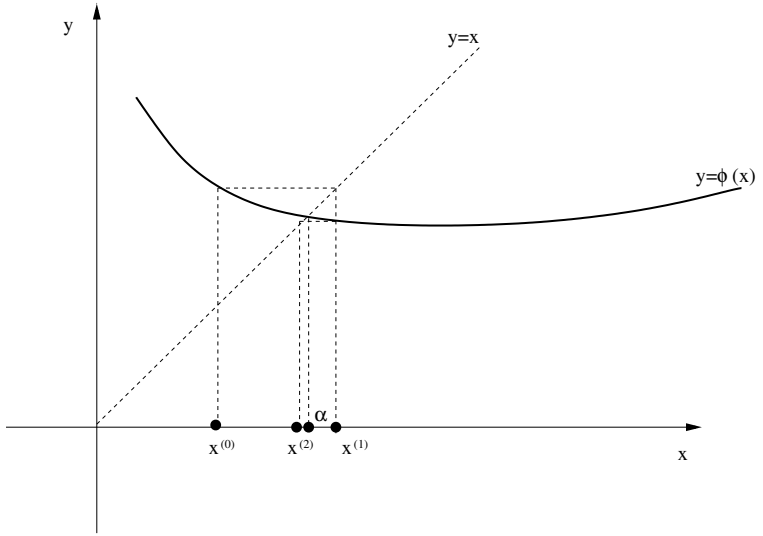
$\square$

Some examples on how the value $|\ \phi'(\alpha)\ |$ influences the convergence

Convergent cases:

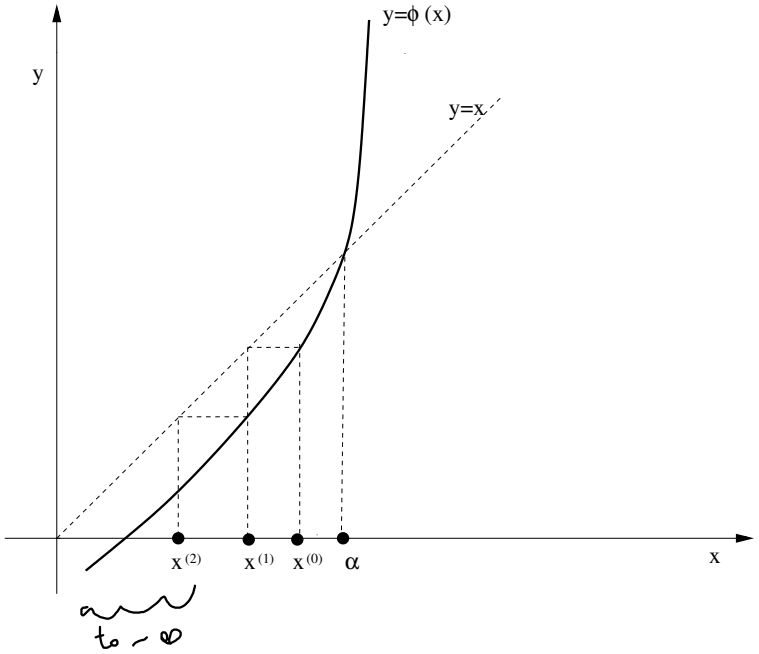
$$0 < \phi'(\alpha) < 1, \qquad -1 < \phi'(\alpha) < 0.$$

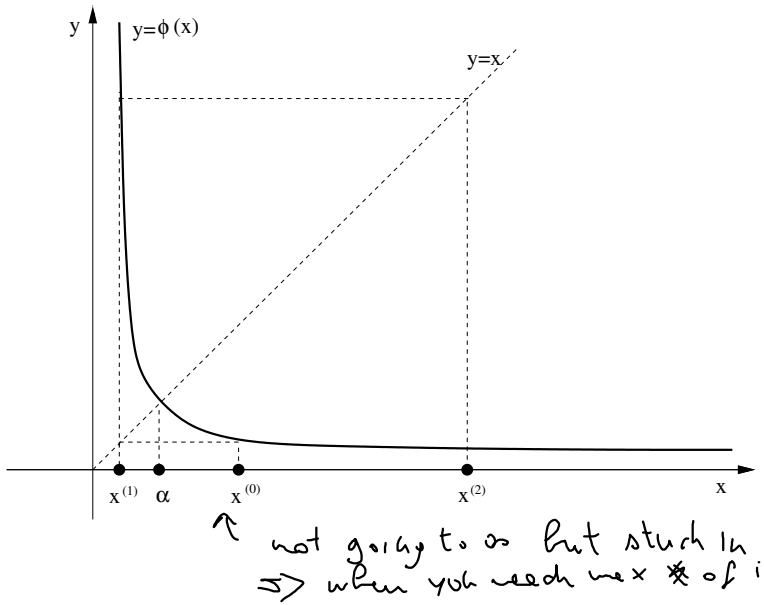




Divergent cases:

$$\phi'(\alpha) > 1, \qquad \phi'(\alpha) < -1.$$



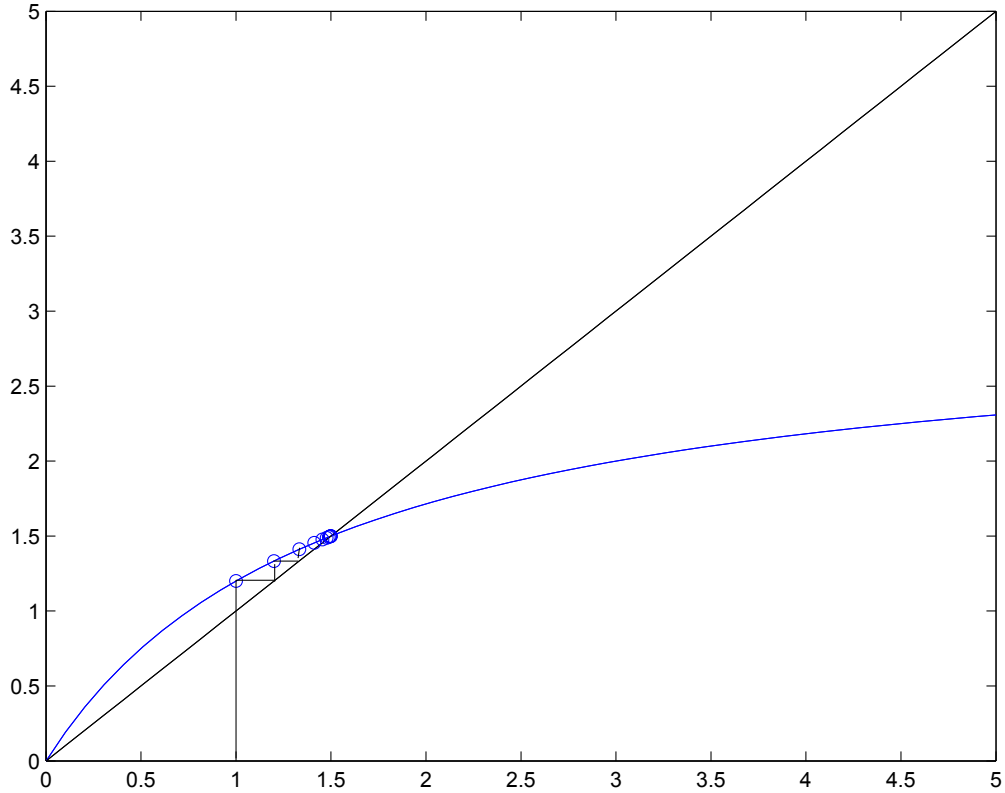


Example. 5 (suite) We apply the fixed point iterations on functions $\phi_2(x) = rx/(1 + x/K)$ and $\phi_3(x) = rx^2/(1 + (x/K)^2)$ that represent the Verhulst model and predator/prey model respectively, with $K = 1.5$ and $r = 2$. We consider the starting point $x^{(0)} = 1.0$.

```
>> phi2=@(x) x.*(2./(1+(x./1.5)));
>> phi3=@(x) x.*(2*x./(1+(x./1.5).^2));
>> x=linspace(0,5,50);
>> figure
>> plot(x,phi2(x),'b',x,x,'k');
>> [p2,res2,niter2]=fixpoint(phi2,1,1e-6,1000);
>> figure
>> plot(x,phi3(x),'b',x,x,'k');
>> [p3,res3,niter3]=fixpoint(phi3,1,1e-6,1000);
```

We find the stationary points $\alpha_2 = 1.5$ and $\alpha_3 = 3.9271$.

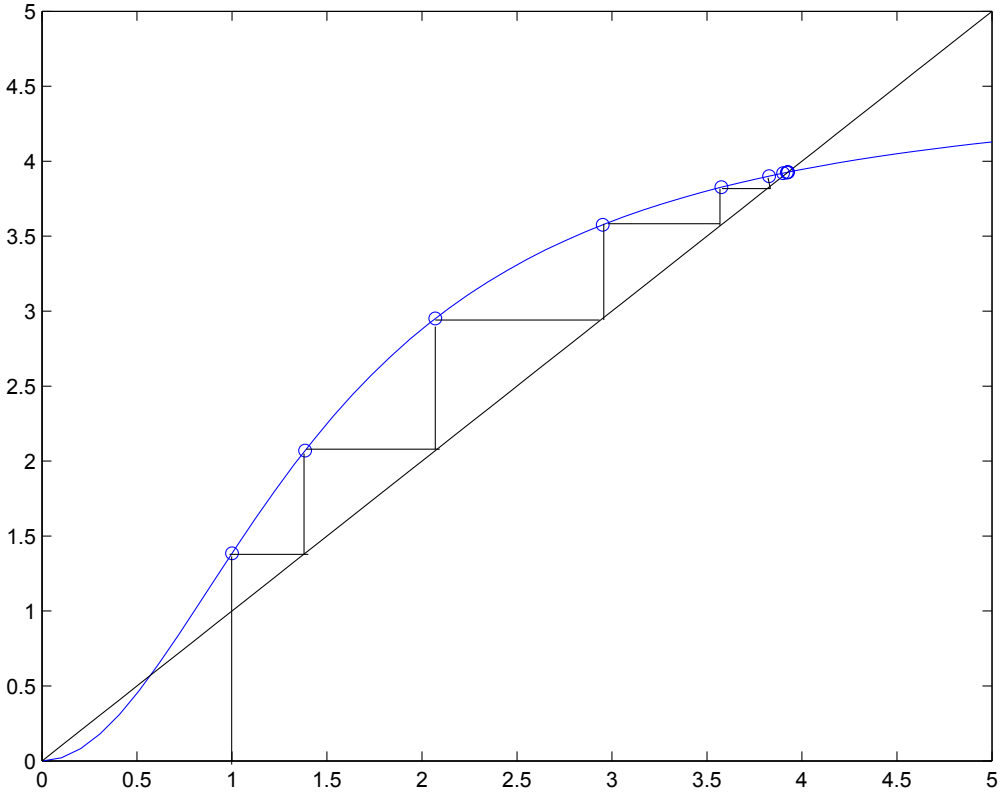
Function $\phi_2(x)$:



Careful to what first index is (0 or 1!)

$$
\begin{aligned}
x^{(0)} &= 1.0000, & |x^{(0)} - \alpha_2| &= 0.5000 \\
x^{(1)} &= 1.2000, & |x^{(1)} - \alpha_2| &= 0.3000 \\
x^{(2)} &= 1.3333, & |x^{(2)} - \alpha_2| &= 0.1667 \\
x^{(3)} &= 1.4118, & |x^{(3)} - \alpha_2| &= 0.0882
\end{aligned}
$$

Function $\phi_3(x)$:



$$
\begin{aligned}
x^{(0)} &= 1.0000, & |x^{(0)} - \alpha_3| &= 2.9271 \\
x^{(1)} &= 1.3846, & |x^{(1)} - \alpha_3| &= 2.5424 \\
x^{(2)} &= 2.0703, & |x^{(2)} - \alpha_3| &= 1.8568 \\
x^{(3)} &= 2.9509, & |x^{(3)} - \alpha_3| &= 0.9761
\end{aligned}
$$

Example. 6 (cont) We have used the fixed point algorithms using the two fonctions ϕ_1 and ϕ_2 with initial value $x^{(0)} = 0.7$. Remember that both have the same fixed point α.

easier, but not working!

$$x = \phi_1(x) = 1 - \sin(2x)$$

because $|\phi_1'(\alpha)| > 1$

! FIRST THING TO CHECK is I deriv!

more difficult but works

$$x = \phi_2(x) = \frac{1}{2}\arcsin(1-x), \qquad 0 \le x \le 1$$

```
>> [p1,res1,niter1]=fixpoint(phi1,0.7,1e-8,1000);
>> [p2,res2,niter2]=fixpoint(phi2,0.7,1e-8,1000);
```

The fixed point algorithm with the first function does not converge, while with the second one it converges to $\alpha = 0.352288459558650$ in 44 iterations. Indeed, $\phi_1'(\alpha) = -1.5237713$ and $\phi_2'(\alpha) = -0.65626645$.

More about the Newton method.

Very good convergence property (QUADRATIC)

Take the rest of Newton method and embed it in the whole fixed point framework

The Newton method is a fixed point method: $x^{(k+1)} = \phi(x^{(k)})$ for the function

$$\phi(x) = x - \frac{f(x)}{f'(x)}.$$

Let α be a zero of f, i.e. such that $f(\alpha) = 0$. Note that $\phi'(\alpha) = 0$, when $f'(\alpha) \neq 0$. Indeed,

So good because we obtain fixed point with II procedure (less manipulation) and it is consistent with Newton method (II order)

use it one algorithm

image fixed point taking the property of Newton method

QUADRATIC CONVERGENCE to the fixed point which is "testified" by obtaining $\phi'(\alpha) = 0$

$$\phi'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2}.$$

if $f'(\alpha) = 0$ UNLUCKY CASE

$\phi'(\alpha) = 1 - \dfrac{[f'(\alpha)]^2 - \overbrace{f(\alpha)}^{=0}\, f''(\alpha)}{[f'(\alpha)]^2}$ (α is root of f)

$= 1 - \dfrac{[f'(\alpha)]^2}{[f'(\alpha)]^2} = 0$

! Important that $\phi''(\alpha) \neq 0$

If this fails to happen failure is of Newton method (it is going to be linear)

↓

Solve by change a little bit the ϕ ft

(DEFLATION METHOD)

↓

modify Newton scheme by making sure that the II derivative $\neq 0$

Theorem 1. *If f is twice differentiable, $f(\alpha) = 0$ and $f'(\alpha) \neq 0$, then there exists $\delta > 0$ such that, if $| x^{(0)} - \alpha | \leq \delta$, the sequence defined by the Newton method converges to α.*
Moreover, the convergence is quadratic; more precisely

$$\lim_{k \to \infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)}.$$

this represents $\phi''(\alpha)$

Proof. The property of convergence comes from the Proposition 2, while the quadratic convergence is a consequence of the Proposition 3, because

$\phi'(\alpha) = 0$ and $\frac{\phi''(\alpha)}{2} = \frac{f''(\alpha)}{2f'(\alpha)}$.

do second derivative of $\phi(x) = x - \frac{f(x)}{f'(x)}$

⇒ We recover quadratic convergence thanks to info coming from the function without the need to compute $\phi, \phi' \ldots$ but only f', f''

Definition 2. *Let α be a zero of f. α is said to have multiplicity m, $m \in \mathbb{N}$, if $f(\alpha) = \ldots = f^{(m-1)}(\alpha) = 0 \quad$ and $f^{(m)}(\alpha) \neq 0$.*
A zero that has multiplicity $m = 1$ is called simple zero.

$\frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha}$ IS UNCOMPUTABLE, because every ft empty related to α is UNCOMPUTABLE, since α IS NOT KNOWN

Remark 2. *If $f'(\alpha) = 0$, the convergence of the Newton method is linear, not quadratic. We can use the modified Newton method:*

$$x^{(k+1)} = x^{(k)} - m\frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k = 0, 1, 2 \ldots. \tag{7}$$

at least = 2 since $f'(\alpha) = 0$

where m is the multiplicity of α.

If the multiplicity m of α is unknown, there are other methods, *the adaptive methods*, which can recover the quadratic order of convergence.

Multiplying f by m or reduce f' by m
=> IMPO: you can try to avoid updating f' at each iteration and delay the update of the most expensive computational ptics
↳ ADAPTIVITIES ↗ not free! you pay in ACCURACY

A stopping criterion for Newton

When to stop the Newton method? A good stopping criterion is the **control of the increment** : the iterations is completed when

$$|x^{(k+1)} - x^{(k)}| < \epsilon \tag{8}$$

where ϵ is a fixed tolerance.

Indeed, if we denote $e^{(k)} = \alpha - x^{(k)}$ is the error of the iteration k, we have

$$e^{(k+1)} = \alpha - x^{(k+1)} = \phi(\alpha) - \phi(x^{(k)}) = \phi'(\xi^{(k)})e^{(k)},$$

where $\xi^{(k)}$ is between $x^{(k)}$ and α, and

$$x^{(k+1)} - x^{(k)} \stackrel{\pm\alpha}{=} \alpha - x^{(k)} - \alpha + x^{(k+1)} = e^{(k)} - e^{(k+1)} = \left(1 - \phi'(\xi^{(k)})\right) e^{(k)}. \quad (9)$$

Assuming that if k is large enough, we have $\phi'(\xi^{(k)}) \approx \phi'(\alpha)$ and knowing that the Newton method for $\phi'(\alpha) = 0$, if α is a simple zero, we find the estimation

$$|e^{(k)}| \approx |x^{(k+1)} - x^{(k)}|.$$

The error that we commit when we adopt the criterion (8) is smaller than the fixed tolerance.

Stopping criteria: the general case

In general, for all discussed methods, we can use two different stopping criteria: the iterations is completed when

$$|x^{(k+1)} - x^{(k)}| < \epsilon \qquad \text{(control of the increment)},$$

or

$$|f(x^{(k)})| < \epsilon \qquad \text{(control of the residual)},$$

where ϵ is a fixed tolerance.

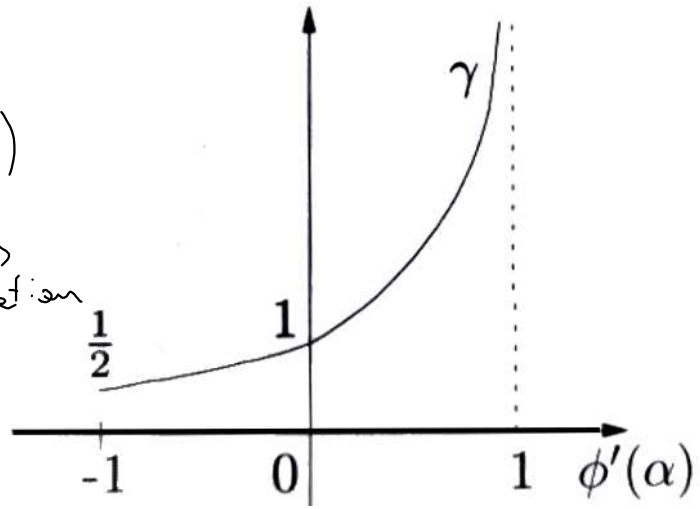
$f(x^{(k)})$ is the RESIDUAL, because $f(\alpha) = 0$, so there we should guarantee the consistency of the methodology

Using fixed point iterations we obtain the following estimation:

$$e^{(k)} \approx \frac{1}{(1-\phi'(\alpha))}(x^{(k+1)} - x^{(k)}).$$

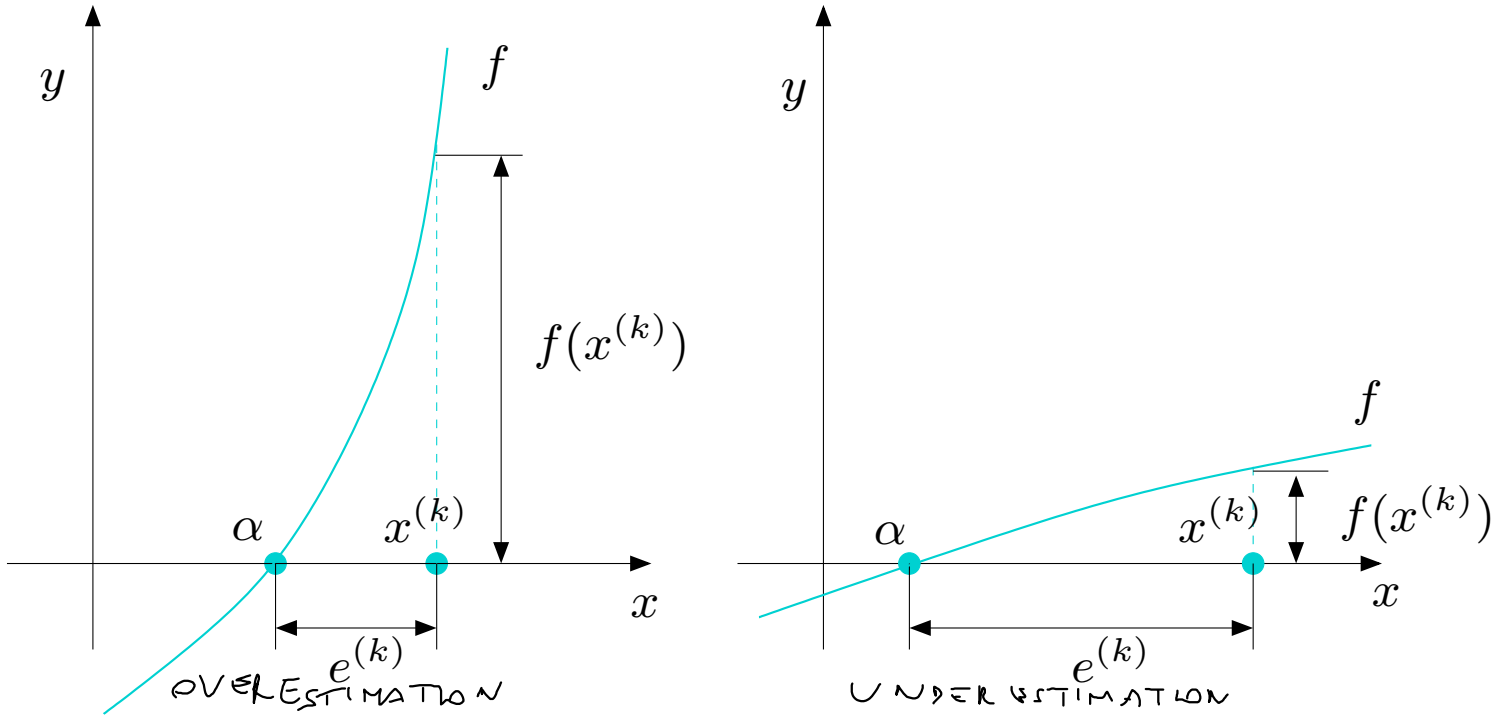
We can plot the graph of $\frac{1}{(1-\phi'(\alpha))}$ and comment on the relevance of the stopping criterium based on the increment:

- if $\phi\prime(\alpha)$ is near to 1 the test is not satisfactory
- for methods of order 2 ($\phi\prime(\alpha) = 0$), the criterium is optimal,
- if $-1 < \phi\prime(\alpha) < 0$ the criterium is still all right.



Stopping Criteria

The stopping criterium based on the control on the residual $|f(x^{(k)})| < \epsilon$ is satisfactory only if $|f'| \simeq 1$ near the root α. Otherwise it is too strong (if $|f'| \gg 1$) or too weak (if $|f'| \ll 1$):



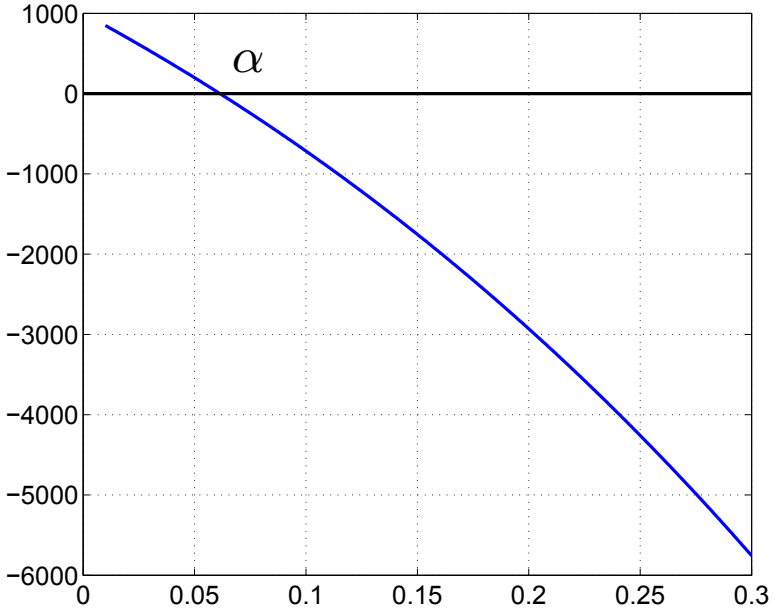
Two cases where the residual is a bad estimator of the error: $|f'(x)| \gg 1$ (left), $|f'(x)| \ll 1$ (right)) with x near to α

MYDEA: if divide $f(x^{(k)})$ by $f'(x^{(k)})$ (normalization)

Applications

Example. 1 (suite) We draw the graph of $f(I) = M - v\frac{1+I}{I}\left[(1+I)^n - 1\right]$ on the interval $[0.010.3]$ with $M = 6000$, $v = 1000$ and $n = 5$:

```
>> f=@(x) 6000-1000*(1+x).*((1+x).^5 - 1)./x;
>> I = [0.01:0.001:0.3];
>> grid on;plot(I,feval(f,x));
```



The root of f is between 0.05 and 0.1.

We can apply the bisection method on the interval $[0.05, 0.1]$ with a tolerance 10^{-5}

```
>> [zero,res,niter]=bisection(f,0.05,0.1,1e-5,1000);
```

The approximate solution after 12 iterations is $\bar{x} = 0.061407470703125$.

We can apply the Newton method with initial guess $x^{(0)} = 0.05$

```
>> df=@(x) 1000*((1+x).^5.*(1-5*x) - 1)./(x.^2);
>> [zero,res,niter]=newton(f,df,.05,1e-5,1000);
```

The result is approximately the same, but we need only 3 iterations

The interest rate is 6.14%.

! If you use Newton Method : you have to give also the I derivative of the nonlinear equation

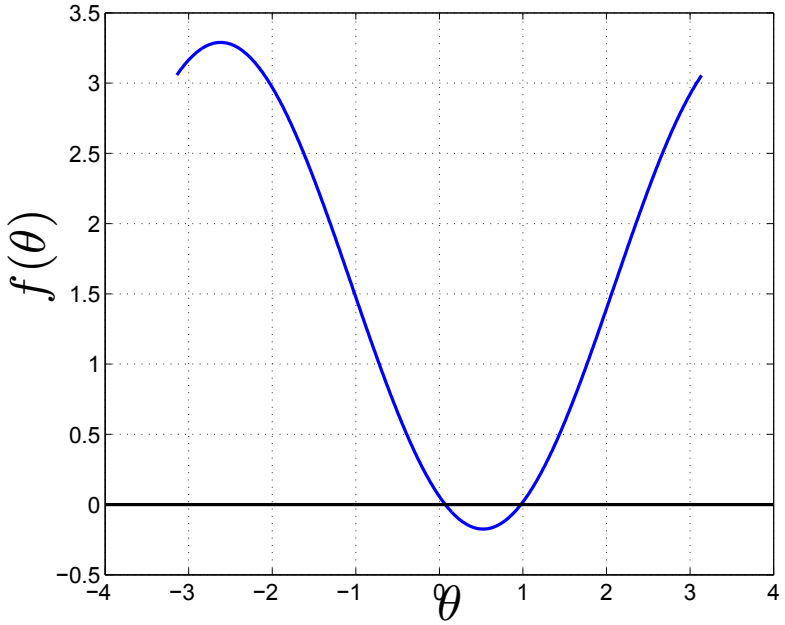
Example. 2 (cont) We would like to plot the angle θ as function of ω for $0 \leq \omega \leq \pi$ with $a_1 = 10\,cm$, $a_2 = 13\,cm$, $a_3 = 8\,cm$, $a_4 = 10\,cm$.
For each ω, we have to solve the nonlinear poblem

$$f(\theta) = \frac{a_1}{a_2} cos(\omega) - \frac{a_1}{a_4} cos(\theta) - cos(\omega - \theta) + \frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4} = 0. \qquad (10)$$

To start with, we plot the graph of $f(\theta)$ for $\omega = \pi/3$:

```
>> F = @(x, a1, a2, a3, a4, omega) ...
       (a1/a2)*cos(omega) - (a1/a4)*cos(x) - cos(omega-x) ...
     + ( (a1.^2 + a2.^2 - a3.^2 + a4.^2) / (2*a2*a4) );
>> a1=10; a2=13; a3=8; a4=10; omega=pi/3;
>> f = @(x) F(x,a1,a2,a3,a4,omega);
>> x = [-pi:0.01:pi];
>> plot( x, f(x) ); grid on;
```

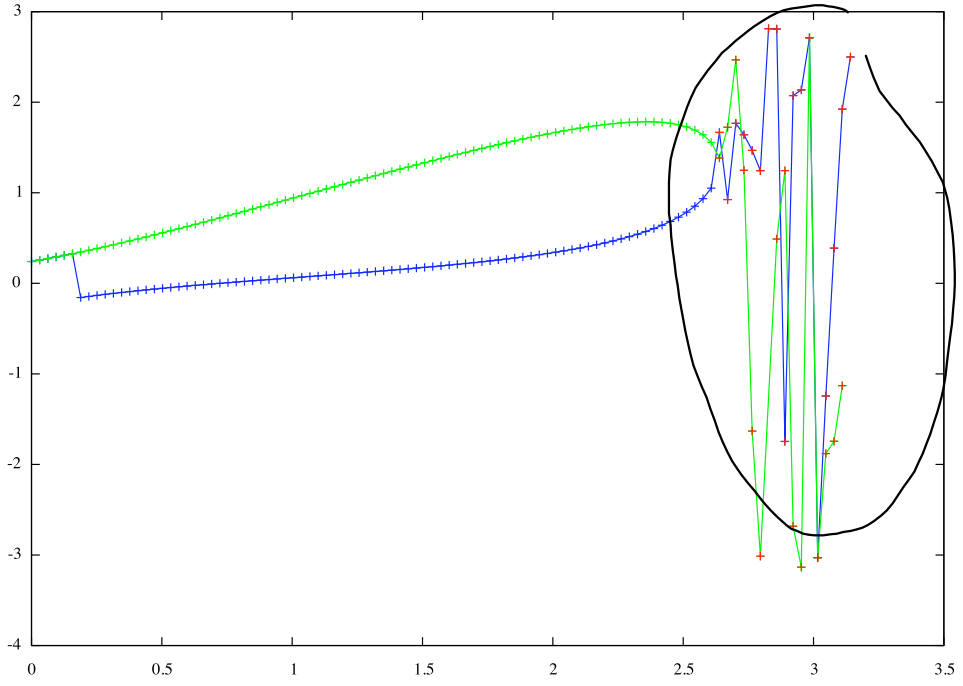
Function f with $\omega = \pi/3$.



We have two roots, which means that we have two possible configurations.

Now we chose 101 different values for ω, $\omega_k = k\frac{\pi}{100}$, $k = 0, \ldots, 100$ and for each of them solve the nonlinear problem (10) with the Newton method. Since we know that we may have two distinct solutions, we give two different initial guesses to our Newton method. For example $\theta_{01} = -0.1$ and $\theta_{02} = 2/3\pi$

The following figure shows the solutions as function of ω.
For $\omega > 2.6358$, the Newton algorithm does not converge anymore. In fact, with these values there is no configuration possible



no solution, but INCONSISTENT BEHAVIOR of MECHANICAL SYSTEM oscillation due to numerical wire : NOT PHYSICAL

Here are the Matlab/Octave commands that we have used:

```
>> n=101; x01=-0.1; x02=2*pi/3; nmax=100;
>> dF = @(x,a1,a2,a3,a4,w) a1/a4*sin(x)-sin(w-x);
>> for k=1:1:n
    omega(k) = (k-1)*pi/100;
    f  = @(x)  F(x,a1,a2,a3,a4,omega(k));
    df = @(x) dF(x,a1,a2,a3,a4,omega(k));
    [theta1(k),res,niter] = newton(f,df,x01,1e-5,nmax);
    [theta2(k),res,niter] = newton(f,df,x02,1e-5,nmax);
   end
>> plot(omega,theta1,'b:',omega,theta2,'g-')
```

Example. 3 (suite) We consider the carbon dioxide (CO_2), for which $a = 0.401\ Pa\,m^6$ and $b = 42.7 \cdot 10^{-6}\ m^3$.
We search the volume occupied by $N = 1000$ molecules of CO_2 in temperature $T = 300\ K$ and pressure $p = 3.5 \cdot 10^7\ Pa$. We know that the Boltzmann constant is $k = 1.3806503 \cdot 10^{-23}\ Joule\,K^{-1}$.

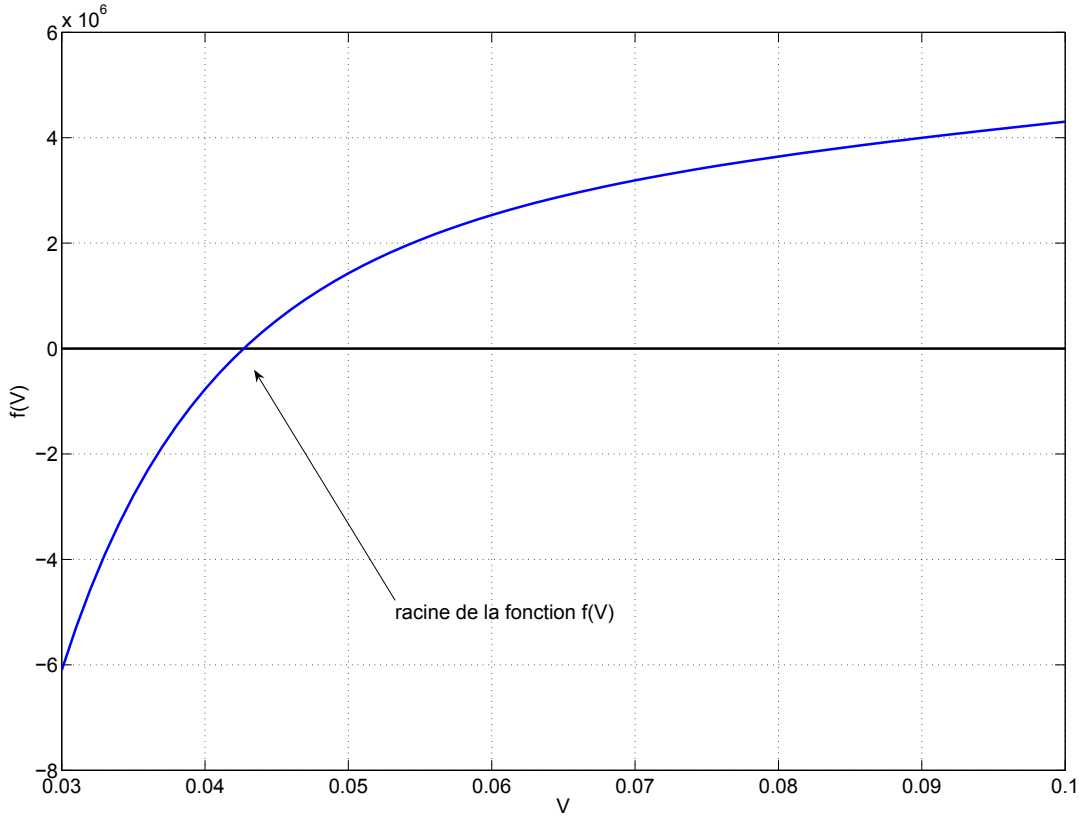
We draw the graph of the function

$$f(V) = \left[p + a\left(\frac{N}{V}\right)^2\right](V - Nb) - kNT$$

for $V > 0$. We do not consider $V < 0$ (it does not have physical meaning), because V is the volume of gas.
We use the commands in Matlab/Octave:

```
>> a=0.401; b=42.7e-6; p=3.5e7; T=300; N=1000; k=1.3806503e-23;
>> f = @(x,p,T,a,b,N,k)(p+a*((N./x).^2)).*(x-N*b)-k*N*T;
>> x=[0.03:0.001:0.1];
>> plot(x,f(x,p,T,a,b,N,k))
>> grid on
```

We obtain the graph of the function $f(V)$:



We see that there is a zero for $0.03 < V < 0.1$. If we apply the bisection method on the interval $[0.03, 0.1]$ with a tolerance 10^{-12}:

```
[zero,res,niter]=bisection(f,0.03,0.1,1e-12,1000,p,T,a,b,N,k);
```

then we find, after 36 iterations, the value $V = 0.0427$.
If we use the Newton method with the same tolerance, starting form the initial point $x^{(0)} = 0.03$,

```
>> df = @(x,p,T,a,b,N,k) -2*a*N^2/(x^3)*(x-N*b)+(p+a*((N./x).^2));
>> [zero,res,niter]=newton(f,df,0.03,1e-12,1000,p,T,a,b,N,k);
```

then we find the same solution after 6 iterations.
The conclusion is that the volume V occupied by the gas is $0.0427\ m^3$.

The rope method

This method is obtained by replacing $f'(x^{(k)})$ by a fixed q in the Newton method:

$$x^{(k+1)} = x^{(k)} - \frac{1}{q} f(x^{(k)}), \quad k = 0, 1, 2 \ldots . \tag{11}$$

We can take, for example, $q = f'(x^{(0)})$ or $q = \dfrac{f(b) - f(a)}{b - a}$, in the case when we search a zero in the interval $[a, b]$.

Example. 6 (suite) We apply the rope method and the Newton method to find the zero of f.

The rope method in the interval $[-1, 1]$, with $x^{(0)} = 0.7$:

```
>>  [zero,res,niter]=chord(f,-1,1,0.7,1e-8,1000)
```

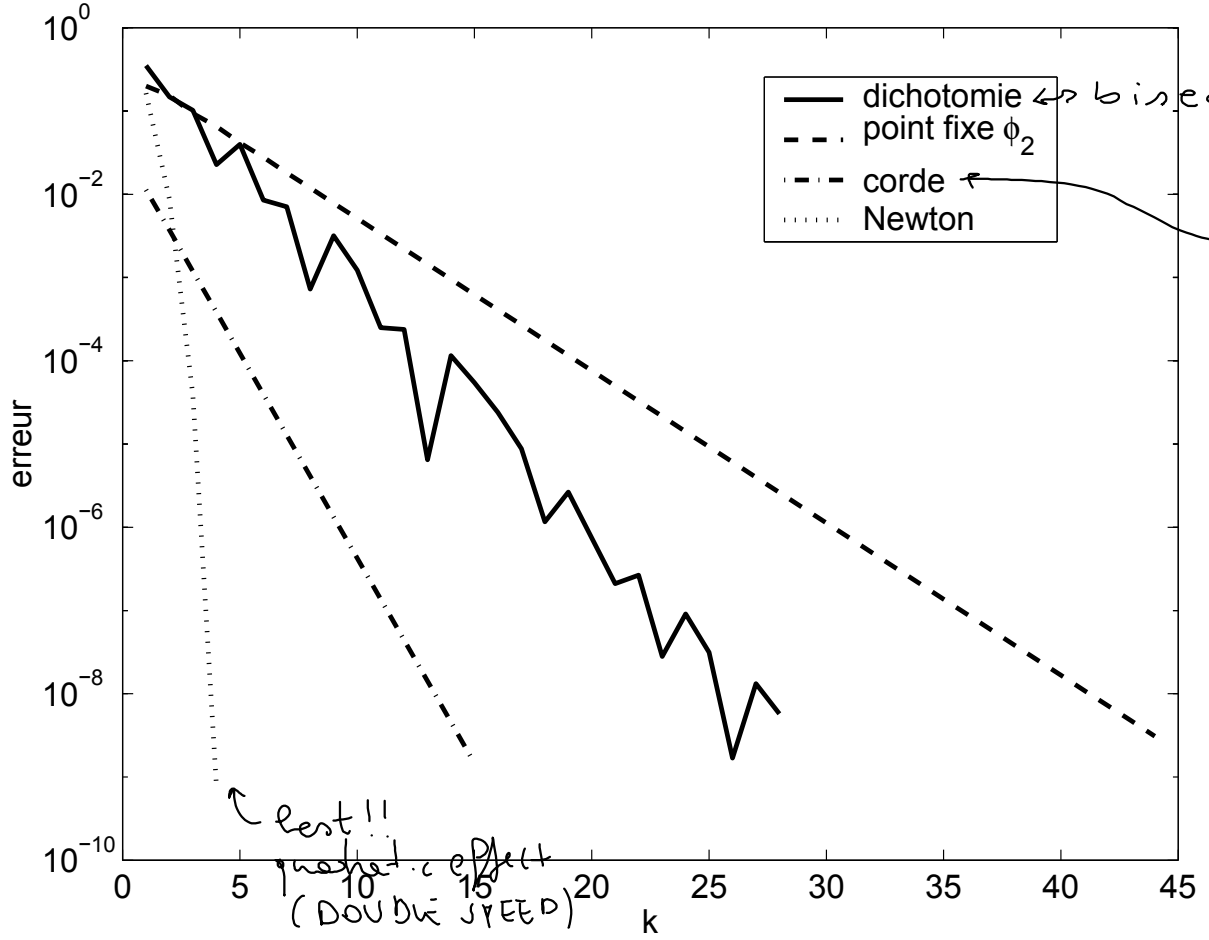
We find the result after 15 iterations.

The Newton method with the same $x^{(0)}$:

```
>>  df = @(x) 2*cos(2*x) + 1;
>>  [zero,res,niter]=newton(f,df,0.7,1e-8,1000);
```

We find the result after 5 iterations. Much faster.

Values of the errors plotted versus the number of iterations for 4 methods: bisection, fixed point ϕ_2, rope and Newton. There is logarithmic scale on the axis y.



Remark 3. *The rope method is also a fixed point method for*

$$\phi(x) = x - \frac{1}{q} f(x).$$

So, we have $\phi'(x) = 1 - \frac{1}{q} f'(x)$ *and thanks to the Proposition 2, we obtain that the method converges if the following condition is satisfied:*

$$| \, 1 - \frac{1}{q} f'(\alpha) \, | < 1 \, .$$

Lecture 11 - LH

5 Nov 2020

we want n to be ~~one~~ of interpolation => # of pts = # of basis fcts

Recap on interpolation: 1) a set of interpolation points $\{x_i\}_{i=0}^n$

2) a set of basis functions $\{v_i\}_{i=0}^n$

Interpolation of $f \in C^0([a,b])$ $x_i \in [a,b]$ $\forall i$

on the space $V = \text{span}\{v_i\}_{i=0}^n$ is defined as the solution to:

find p in V s.t. $p(x_i) = f(x_i)$ $\forall i$

Given $\{x_i\}_{i=0}^n$ best choice for $\{v_i\}_{i=0}^n$ is Lagrange basis

$$\Rightarrow \quad v_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

So the interpolation poly becomes

$$p = \sum_{i=0}^{n} v_i(x) f(x_i) := \mathcal{L}^n f$$

Next step was:

Try to find the "best" set of $\{x_i\}_{i=0}^n$

Remember that L^2 [illegible] monic poly

best approx

$$\|\mathcal{L}^n f - f\|_\infty = \|\mathcal{L}^n(f - p) + p - f\|_\infty \leq \left(\|\mathcal{L}^n\|_* + 1\right) \|p - f\|_\infty$$

when Lagrange basis function is used

$$\|\mathcal{L}^n\|_* \leq \|\Lambda(\{x_i\}_{i=0}^n)\|_\infty$$

We try to min the sum of the absolute values of the basis fct (as before)

$$\Lambda := \sum_{i=0}^{n} |v_i|$$

← Lebesgue function

For Chebyshev points:

$$\frac{2}{\pi} \log(n+1) - c \leq \|\Lambda\|_\infty \leq \frac{2}{\pi} \log(n+1) + 1$$

Erdos — Phelyphen

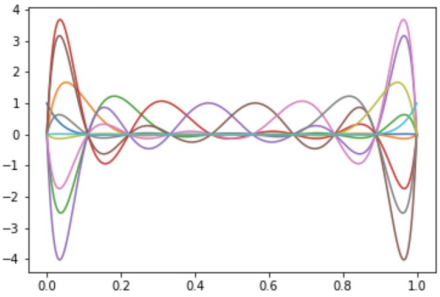
⚠ True only if the function is continuous

If you add more regularity, you get better things (Result for analytically extendable f, with continuous derivatives up to the n-th derivative)

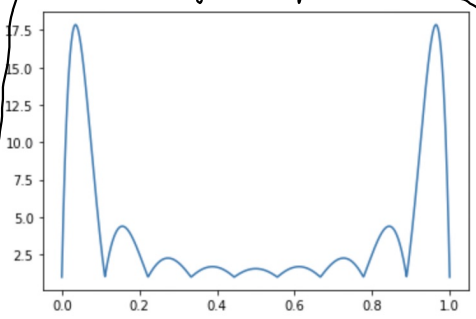
⚠ INTERPOL IS EXACT AT POINT OF INTERPOL

$\Rightarrow$ try to get exact approx every where is not smart way

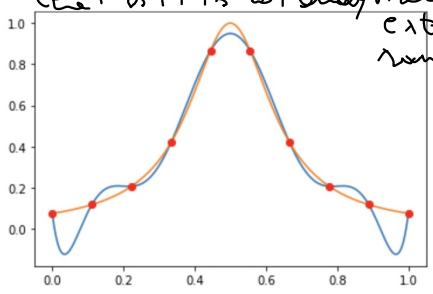
USING EQUISPACED PTS



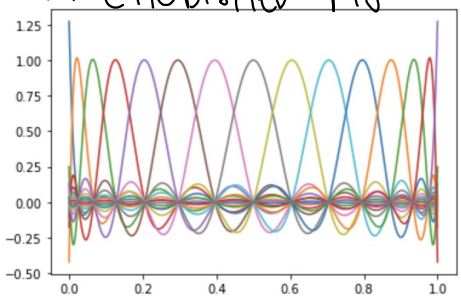
Lebesgue function

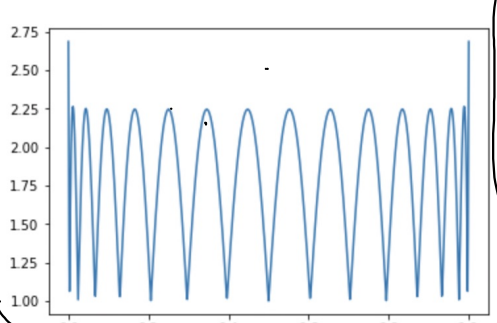


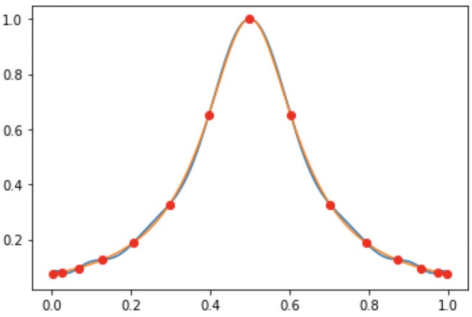
tells us if the pt is wrong that is it is not only totally extendible somehow



USING CHEBISHEV PTS







If interpolation is not good, what can we do in $C^0([a,b])$?

Yes: **Weierstrass Approximation Theorem**

$\forall f \in C^0([a,b])$, $\forall \varepsilon > 0$, $\exists n$, $p \in \mathbb{P}^n$, s.t. $\|f-p\|_\infty < \varepsilon$

~~No matter what f you can always find a p such that p approx f~~

Proof: Let B_n be a sequence of linear operators s.t.

1) B_n positive linear operators $C([a,b]) \to \mathbb{P}^n([a,b])$

2) $B_n q \to q$ pointwise $\forall q \in \mathbb{P}^2([a,b])$ that is only for quadratic polys

Then: **$B_n f$ converges uniformly to f $\forall f \in C^0([a,b])$**

An op is defined positive

N.B: B_n is "positive" if $q(x) \geq 0 \Rightarrow (B_n q)(x) \geq 0$ in $[a,b]$

(if $q(x) \leq 0$ it remains negative if B is positive)

Part 1:

- $\forall f \in C^0([a,b])$, $\forall x_0 \in [a,b]$:
- construct $q^\pm \in \mathbb{P}^2$ s.t. $q^-(x_0) < f(x_0) < q^+(x_0)$ $\forall x$.
- use $B_n q^\pm \to q^\pm$
- use $B_n(f-q^-)(x_0) > 0$ $B_n(q^+-f)(x_0) > 0$

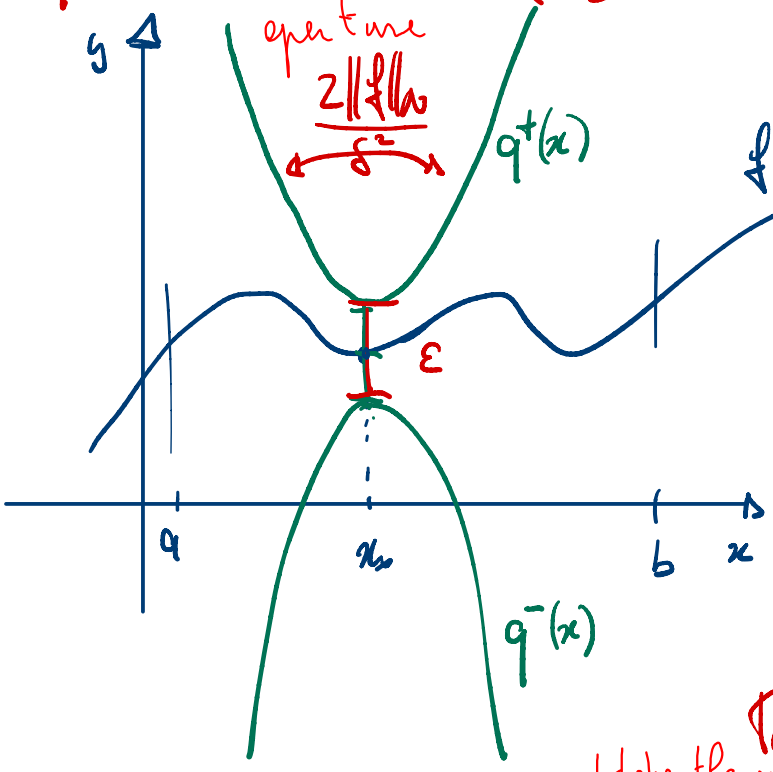
$f \in C^0([a,b]) \Rightarrow \forall \varepsilon > 0, \exists \delta > 0$

$|x_1 - x_2| \le \delta \rightarrow |f(x_1) - f(x_2)| \le \varepsilon$ } def of continuity!

Define

$$q^{\pm} := f(x_0) \pm \left(\frac{\varepsilon}{2} + \frac{2\|f\|_\infty}{\delta^2}(x - x_0)^2 \right)$$

← Constructing $q^\pm$ in this way guarantees that they are always ε/2 far from f



for varying x_0

$q^\pm(x) = a^\pm x^2 + b^\pm x + c^\pm$

$a^\pm, b^\pm, c^\pm$ depend on $x_0, \|f\|_\infty, \delta, \varepsilon$

Take $M = \max_{x_0 \in [a,b]} \left(|a^\pm|, |b^\pm|, |c^\pm| \right)$

(I take the max over the entire interval)

M depends on $\|f\|_\infty, \delta, \varepsilon$, but Not on x_0

Choose N large enough s.t. $\left\| B_n x^i - x^i \right\|_\infty \le \frac{\varepsilon}{6M}$ $\forall n > N$ for $i = 0,1,2$

($\exists N$ s.t. this is possible)

Since we might that this condition must hold for each of the two (HYPOTHESIS!)

$\Rightarrow \forall x_0$ $\left\| B_n q^\pm - q^\pm \right\| \le \frac{\varepsilon}{2}$ $\forall n \ge N$

M controls every coeff => this allows to say that $\frac{\varepsilon}{6M} \le \frac{\varepsilon}{2}$ (!)

in x_0 $f(x_0) - \varepsilon \le q^-(x_0) - \frac{\varepsilon}{2} \le B_n q^-(x) \le B_n f(x_0)$

Definition of q^- — positivity

And repeat for the reversed inequality:

$\forall x_0, \exists N$ s.t. $\forall n \ge N$ $B_n f \le B_n q^+ \le q^+ + \frac{\varepsilon}{2} \le f(x) + \varepsilon$

$$\left\| B_n f - f \right\| \le \varepsilon$$

How to define B_n ? Let $(a,b] = [0,1]$

Trick : Consider $\left((1-x)+x\right)^n = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i}$ polynomial of degree n

$= \sum_{i=0}^{n} v_i(x)$ ($\Leftarrow 1$ defines basis functions)

$$v_i(x) = \binom{n}{i} x^i (1-x)^{n-i}$$

Bernstein polynomials.

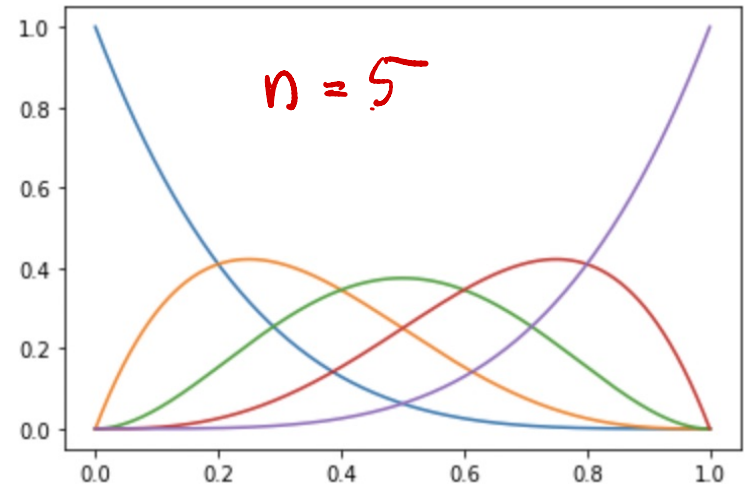
- $B_n f := \sum_{i=0}^{n} v_i(x) f\left(\frac{i}{n}\right)$ (def) — difference between interpolation and this

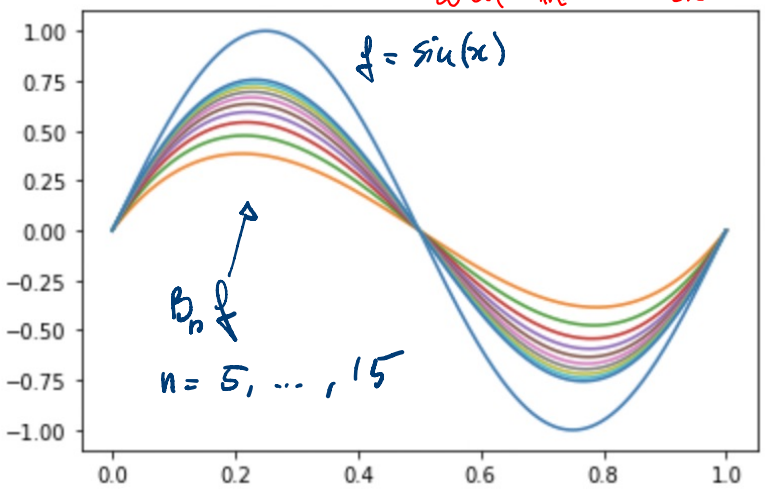
$f \geq 0 \Rightarrow B_n f \geq 0$ because v_i positive

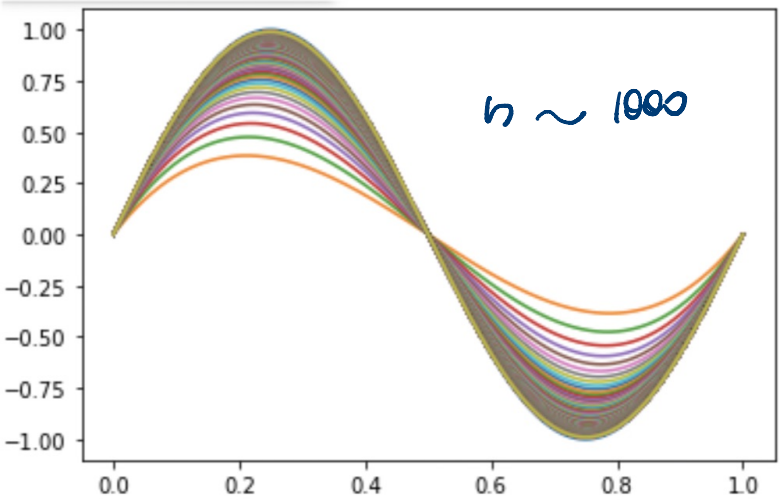
v_i always positive in $(0,1)$ but are also $\neq$ from 0 in all the interval but 0 or 1

You cannot say that the v_i are interlacing at any pt but 0 and 1

- $B_n x = x$ $\quad B_n 1 = 1$
- $B_n x^2 = \left(\frac{n-1}{n}\right) x^2 + \frac{1}{n} x \neq x^2$ for large n still $\neq x^2$ but $n \to \infty$ $B_n x^2 \to x^2$







We don't have a quick convergence

Lecture 12 — PROJECTION BASE APPROX METHOD — 10. Nov. 2020

1) Approximate and measure using L^2 norm (vs L^∞) ← derived from a scalar product

→ Scalar product : $u, v \in L^2([0,1])$,

$$(u,v) := \int_0^1 u \cdot v \, dx$$

$$\|a+b\|^2 = (a+b, a+b) = \|a\|^2 + \|b\|^2 + 2(a,b)$$

↝ $$\|u\|^2_{L^2([0,1])} \equiv \|u\|^2 := (u,u) = \int_0^1 |u|^2 dx$$

Let's assume we have $V := \text{span}\{v_i\}_{i=0}^n$, and $\|\cdot\|$.

The best approximation of $u \in L^2([0,1])$ in V satisfies:

$p \in V$ s.t. $(u-p, v) = 0 \quad \forall v \in V$

best approx

$(u-p,v) = 0 \ \forall v \in V \iff p$ is B.A. of u in V

Proof: Assume p is best approx $\Rightarrow (u-p,v) = 0 \ \forall v \in V$

by def of B.A.

$$\|u-p\|^2 \leq \|u - p + tq\|^2 \quad \forall t > 0, \ \forall q \in V$$

↑ if I move in any direction from p I get bigger distance from f to interpolant

$$\|u - p + tq\|^2 - \|u-p\|^2 \geq 0$$

$(a+b)^2 - (a-b)^2 = 4(a,b)$

$$\|\underbrace{(u-p+\tfrac{t}{2}q)}_{a} + \underbrace{(\tfrac{t}{2}q)}_{b}\|^2 - \|\underbrace{u-p}_{a-b}\|^2 \geq 0$$

$a = u - p + \frac{t}{2}q$

$b = \frac{t}{2}q$

$$2\cancel{4}\left(u - p + \tfrac{t}{2}q, \tfrac{t}{2}q\right) \geq 0$$

Linearity of scalar prod in I argument

$$0 \leq 2(u-p, q)t + t^2(q,q)$$

We have a symmetry for: $q \to -q$

$$-\frac{t}{2}\|q\|^2 \leq (u-p,q) \leq \frac{t}{2}\|q\|^2$$

$\forall t > 0$
$\forall q \in V$

$$\Rightarrow \boxed{(u-p,q) = 0}$$

Assume now $^{*}\ (u-p, v) = 0 \quad \forall v \in V \Rightarrow p$ is B.A.

$$\|u-v\|^2 = \|u-p+p-v\|^2 = \|u-p\|^2 + \|p-v\|^2 + 2(u-p, p-v)$$

(the last term $\to 0$ by * and the fact that $u-p \perp p$)

$$\Rightarrow \|u-p\|^2 \leq \|u-v\|^2 \quad \forall v \in V$$

that is p is B.A. of u in V ($(u-p, p) = 0$: the error is to be orthogonal to the BA, otherwise I can move in the direction of the orthogonal to improve the accuracy)

B.A (ween) V: Find $p \in V$ s.t.

1) $(p, v) = (u, v) \quad \forall v \in V$

you can write $p(x) = \sum p^J v_J(x)$ $\quad (\forall p \in V \ \exists! \{p^i\}_{i=0}^{n}$ s.t. $p = p^i v_i)$ because the space is finite dimensional and generated by basis of the v_i

$\Rightarrow$ 2) $(p^J v_J, v_i) = (u, v_i) \quad i = 0, \dots, n$

MASS MATRIX $\rightarrow$

$$\underline{\underline{M}}\, \underline{p} = \underline{U}$$

$$M_{iJ} p^J = U_i$$

$$M_{iJ} = (v_i, v_J)$$

$$U_i = (u, v_i)$$

compute this $\Rightarrow$ solve the prob

$$p^J = \underline{M^{Ji}}\, U_i$$

Example:

$$V = \operatorname{span}\{\operatorname{pow}(x, i)\}_{i=0}^{n} \qquad p = \sum_J p^J x^{(J)} \quad \text{power basis}$$

then

$$M_{iJ} = \int_0^1 x^i \cdot x^J \, dx = \int_0^1 x^{(i+J)} \, dx = \frac{1}{i+J+1}$$

M here is called Hilbert Matrix: $\operatorname{cond.}(H) \simeq \frac{(1+\sqrt{2})^{4n}}{\sqrt{n}}$ $\leftarrow$ it is Terribly conditioned matrix

(If i or J very large $\Rightarrow$ columns or rows become similar ($i+J+1 \sim i+1+J+1$ ...)

What's the best possible "M" matrix?

Identity: $\quad M_{iJ} = \delta_{iJ} = \begin{cases} 1 & \text{if } i = J \\ 0 & \text{if } i \neq J \end{cases}$

Def B.A. in a space with norm defined from scalar product, select basis function and try to compute mass matrix

⇓

B.S. in a finite dim space (spanned by some basis fct) what you obtain is

$$\underline{\underline{M}}\,\underline{p} = \underline{U}$$

=> best pt is the one in which you don't have to write any matrix => possible if $\underline{\underline{M}} = \underline{\underline{\mathbb{1}}}$

=> $\varphi_i = (u, v_i)$

↓

Simplest procedure

GRAM SCHMIDT PROCEDURE

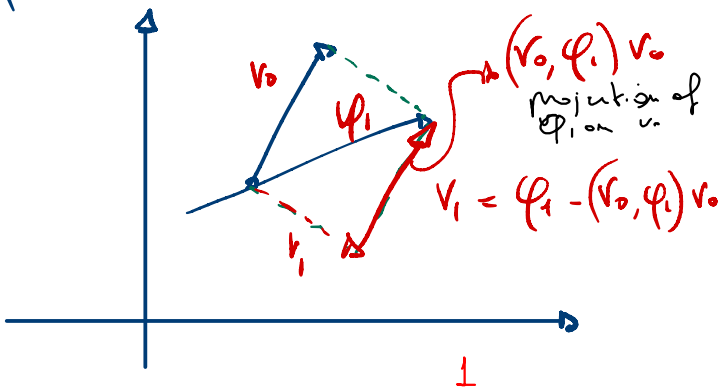
Let's try to find an orthonormal basis

Set $v_0 = 1$ (start with I el of basis) $\longrightarrow$ $\left(\int_0^1 (1)^2 dx\right)^{\frac{1}{2}} = 1$

$$\varphi_{i+1} = \hat{x}\, v_i - \sum_{j=0}^{i-1} \left(\hat{x} v_i, v_j\right) v_j$$

and normalize

$$v_{i+1} = \frac{\varphi_{i+1}}{\|\varphi_{i+1}\|}$$

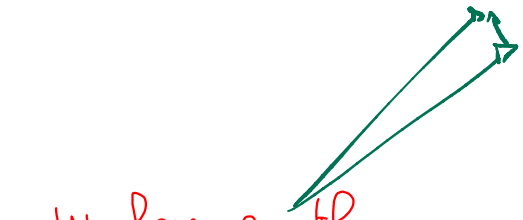


Possible problems:

$$\|\varphi_i\|^2 = \|\hat{x} v_0\|^2 + (\hat{x} v_0, v_0)^2 \underbrace{\|v_0\|^2}_{1} - 2(\hat{x} v_0, v_0)^2$$

$$= \|\hat{x} v_0\|^2 - (\hat{x} v_0, v_0)^2$$

This can be very small!!



We have another solution which does not ask same norm: $\theta_{ij} \neq \delta_{ij}$

but it is diagonal:

Bonnet recursion: $P_0 = 1 \quad P_1 = x$

$$(n+1) P_{n+1}(x) = (2n+1)\, x P_n(x) - n P_{n-1}(x)$$

(Bonnet)

→ Are orthogonal in $[-1, 1]$, and

$$P_n(1) = 1 \quad \forall n$$

(dim (1) (start from fact that P_0, P_1 are ok))

"The only" thing which is left to do is to compute

$$\int_0^1 u\, P_i\, dx$$

Approximate integrals.

Between functions you def the error as $\|p(x) - u(x)\|$ where $\|\cdot\|$ is the norm used.

Lecture 13

12.11.2020

Recap of Lec 12: Given $u \in L^2([0,1])$, find $p \in V = \text{span}\{v_i\}_{i=0}^{n}$ s.t.

$$(p, v) = (u, v) \quad \forall\, v \in V$$

$$\Rightarrow \sum_j \left(p^j v_j, v_i\right) = (u, v_i) \qquad i = 0, \dots, n$$

Rewritable as a system $\Rightarrow \underline{\underline{M}}\,\underline{p} = \underline{U}$

$$M_{ij} := (v_i, v_j)$$

$$U_i = (u, v_i)$$

$$(a,b) := \int_0^1 a \cdot b \, dx$$

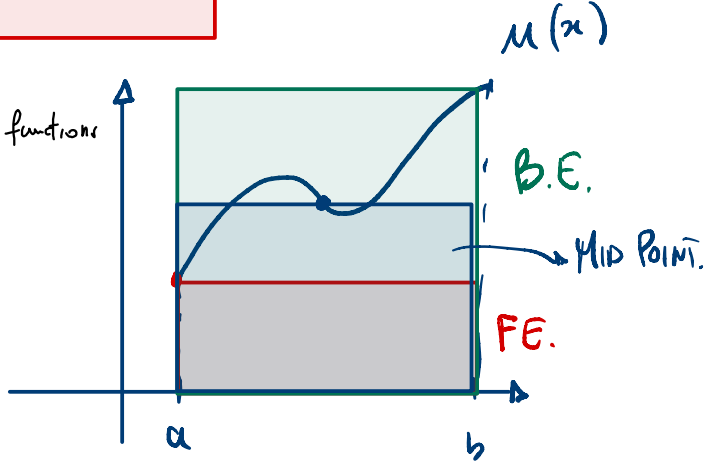
Induces a norm $\Rightarrow \|a\|^2 = (a,a) \Rightarrow \|a\| = \left(\int_0^1 a^2 dx\right)^{\frac{1}{2}}$

Today: Approximate $\int_a^b u(x)\, dx$

For now: $u \in C^0([a,b])$

Rough approx schemes for $C^0([a,b])$ functions

For example:
- FE (Forward Euler)
- BE (Backward Euler)
- MIDPOINT



We have, for each scheme

FE: $\int_a^b u(x) \simeq u(a)(b-a)$

BE: $\int_a^b u(x) \simeq u(b)(b-a)$

MIDPOINT: $\int_a^b u(x) \simeq u\left(\frac{a+b}{2}\right)(b-a)$

You don't use just one of them, but you apply all the three, one for each special case (such as particular sub interval of $[a,b]$)

Interpolatory Quadrature Rules:

$$I(u) := \int_a^b u(x)\, dx$$

$$I_n(u) := \sum_{i=0}^{n} u(q_i)\, w_i \simeq \int_a^b u(x)\, dx = I(u)$$

q_i: quadrature points; w_i: weights

if you choose correctly both q_i and w_i

If we want to integrate we can choose $\{q_i\}_{i=0}^n$ and $\{w_i\}_{i=0}^n$ — $2(n+1)$ unknowns

Most popular choices: Given $\{q_i\}_{i=0}^n$, construct interpolatory quadrature rule

Lagrange interpolation of u function → $\mathcal{L}^n u$ and integrate (exactly!) $\mathcal{L}^n u$

$$\Rightarrow \boxed{I^n(u) = \int_0^1 (\mathcal{L}^n u)(x)\,dx}$$

where $\mathcal{L}^n u = \sum_{i=0}^n u(q_i)\,\ell_i(x)$ ← choose the weights

"Bimi"

$$\ell_i = \prod_{\substack{j=0 \\ j\neq i}}^n \frac{(x-q_j)}{(q_i-q_j)}$$

$$\Rightarrow \boxed{I^n(u) = \int_0^1 \sum_{i=0}^n u(q_i)\,\ell_i(x)\,dx} = \sum_{i=0}^n u(q_i) \underbrace{\int_0^1 \ell_i(x)\,dx}_{w_i}$$

(doesn't depend on x)

number independent on the function I'm trying to integrate

$$= \boxed{\sum_{i=0}^n u(q_i)\,w_i}$$

FE, B.E., Mid Point are all interpolatory quadrature rules with $\mathbb{P}^0$ polynomial approximation

↳ that is with only 1 quadrature pt

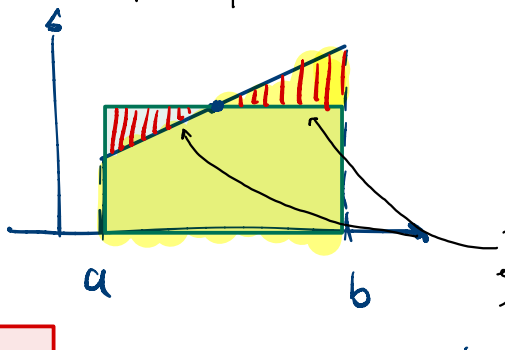
Teo 1 Interpolatory quadrature rules with $n+1$ points are exact at least for polynomials of order $\leq n$

Proof: $\forall p \in \mathbb{P}^n$ (polyn. of ord. n), $\mathcal{L}^n p = p$ by construction

$$\Rightarrow I(p) = I^n(p) \quad \forall p \in \mathbb{P}^n$$

Degree of accuracy of I^n is the maximum integer k s.t. $I(p) = I^n(p)$ $\forall p \in \mathbb{P}^k$

For the midpoint quadrature rule



$I^*(p) = I(p) \quad \forall\, p \in \mathbb{P}^1$

at least one case where

$k > n$

Theo Given $\{q_i\}_{i=0}^{n}$, degree of accuracy is $\leq 2(n+1)$

proof: construct $\omega(x) = \prod_{i=0}^{n} (x - q_i)$ polynomial, $\omega \in \mathbb{P}^{n+1}$

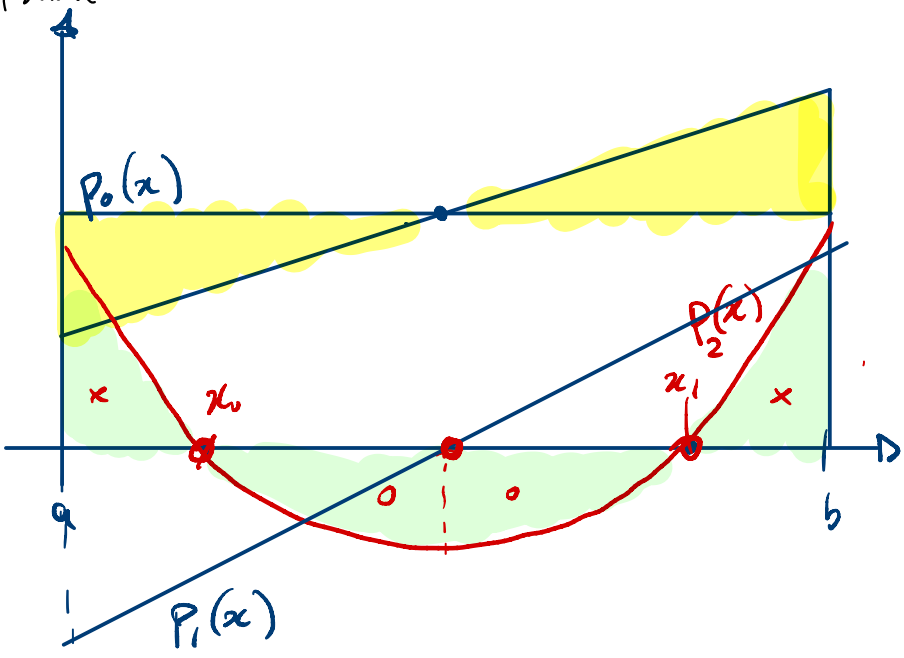
$\omega^2(x) > 0 \quad \forall x \neq q_i, \quad \omega(q_i) = 0 \quad i = 0, \dots, n$

$\Rightarrow I(\omega^2) > 0 \qquad I^n(\omega^2) = 0$ (because) by construction.

$\Rightarrow \exists\, p = \omega^2 \in \mathbb{P}^{2n+2}$ s.t. $I(p) \neq I^n(p)$

Can the degree of accuracy be $k = 2n+1$? Yes

Assume current situation



Areas marked with x and with o must be equal

equivalent of saying $\equiv$

average of $p_2(x)$ is zero

Where

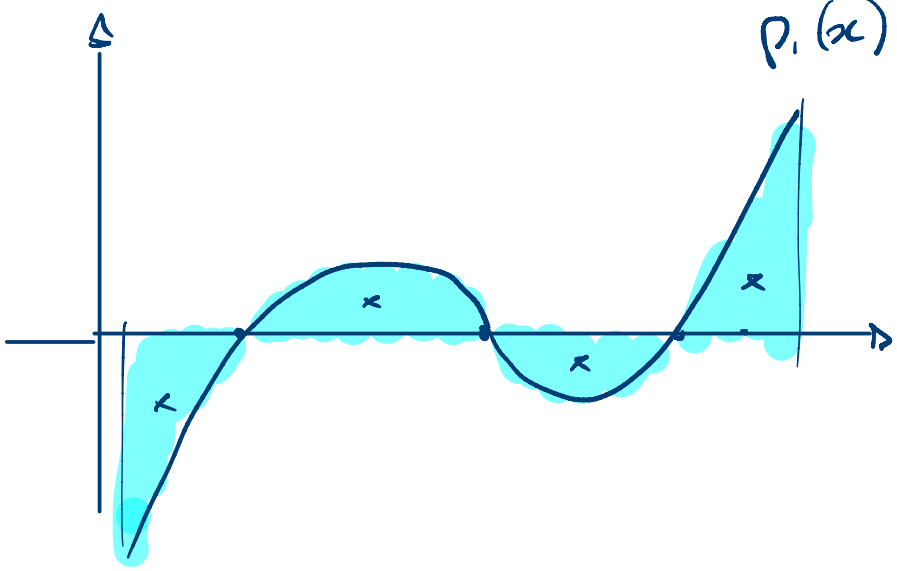
$p_0(x) = 1$

$p_1(x) = (x - 0.5)$ ← $\frac{a+b}{2}$

$p_2(x) = (x - x_0)(x - x_1)$ s.t. $\int p_2(x) = 0$

Now take the pt

$p_1(x) \cdot p_2(x) \in \mathbb{P}^3(x)$



The areas marked with x are equal

Theo

Let $u \in \mathbb{P}^{n+m}$ $\quad m \leq n+1$

then $I^n(u) = I(u)$

$\Leftrightarrow$

I^n has degree of accuracy $k = n+m$

$\Leftrightarrow$

$$\int_a^b \omega(x)\, p = 0 \qquad \forall p \in \mathbb{P}^{m-1}$$

ω can be chosen as $\perp$ to every p we are asking u to be Legendre polynomial

$$\omega = \prod_{i=0}^{n} (x - q_i) \in \mathbb{P}^{n+1}$$

Proof

Any polynomial $p \in \mathbb{P}^{n+m}$ can be written as : $\mathbb{P}^{n+1}$ $\mathbb{P}^{m-1}$ $\mathbb{P}^{m-1}$

$$p(x) = \omega(x)\, \pi(x) + q(x)$$

basically division + rest

by Ruffini's theorem

Since $q(x) \in \mathbb{P}^{m-1} \subseteq \mathbb{P}^n \quad (m \leq n+1)$

Then $\rightarrow I(q) = I^n(q)$

$$I(p) = I(\omega\pi) + I(q)$$ by linearity of interpolation

$$I^n(p) = I^n(\omega\pi) + I^n(q)$$

$I^n(\omega\pi) \rightarrow 0$

$= I(q)$

If $I(\omega\pi) = 0 = \int_0^1 \omega\pi \, dx$ then $I(p) = I^n(p)$

Main menage

Choose $\{q_i\}_{i=0}^{n}$ as the roots of Legendre basis of order $n+1$: $\int_a^b Le_{n+1} \, l_J \, dx = 0 \quad J < n+1$

Le_{n+1} ← Legendre basis of ord $n+1$

The roots of Le_{n+1} are called Gauss quadrature points. And I^n in this case is a Gauss quadrature formula, and it has degree of accuracy $k = 2n+1$

General Strategy: find $\{q_i\}, \{w_i\}$ s.t. ← have $2n+1$ pts ⇒ need $2n+1$ CNs to satisfy

$$\sum_{j=0}^{n} v_i(q_j)\, w_j = \int_a^b v_i(x)\,dx \quad \text{for } \{v_i\}_{i=0}^{2n+1} \text{ known}$$

you impose that the interpolatory quadrature formula is EXACT for a fixed set of arbitrary functions

$2(n+1)$ unknowns, in $2n+2$ equations.

NON LINER SYSTEM OF $2n+2$ equations.

if you can solve this, you win.

for $v_i = x^i$ you obtain the GAUSS QUADRATURE FORMULA