INTRO

REAL NUMBERS REPRESENTATION in $\mathbb{F}$

A computer stores a number in the following way:

$$x = (-1)^s \cdot (0.a_1 a_2 \dots a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t}, \text{ with } a_1 \neq 0$$

- s is a sign bit (1 or 0)
- β is the basis adopted by computer (usually 2)
- m is the MANTISSA of lenght t made by digits a_i, with $0 \leq a_i < \beta - 1$
- e is the exponent

The set of number representable by a machine is characterized by β, t, and the range (L, U) of the exponent. It is denoted as $\mathbb{F}(\beta, t, L, U)$.

The roundoff error when we replace $x \neq 0$ with its representation in $\mathbb{F}$, $\hat{x}$, is

$$\frac{|x - \hat{x}|}{|x|} \leq \frac{1}{2} \epsilon_M, \text{ with } \epsilon_M = \beta^{1-t}$$

ϵ_M is the MACHINE EPSILON, a.k.a. the minimal variation representable by a machine. $\frac{1}{2}\epsilon_M$ is the ROUNDOFF UNIT.

$|x - \hat{x}|$ is an ABSOLUTE ERROR, while $\frac{|x - \hat{x}|}{|x|}$ is a RELATIVE ERROR of our approximation.
The relative error accounts for the order of magnitude of x.

0 is not part of $\mathbb{F}$ and is handled separately. A number exceeding the lower bound is treated as 0, while a number exceeding the upper bound is treated as infinity. $\mathbb{F}$ is more dense near 0, and less near infinity.

In $\mathbb{F}$ associativity and distributivity don't hold in some cases.

There can be a LOSS OF SIGNIFICANT DIGITS. Indeterminate forms as $0/0$ or ∞/∞ produces NaN.

COMPLEX NUMBERS

$$z = x + iy = \rho e^{i\theta} = \rho(\cos\theta + i \sin\theta), \text{ with } i = \sqrt{-1}, \quad x = \mathrm{Re}(z), y(\mathrm{Im}), \quad \rho = \sqrt{x^2 + y^2}$$

z is a complex number ($\in \mathbb{C}$) with real part x and imaginary part y. Its modulus is ρ. Its COMPLEX CONJUGATE is

$$\overline{z} = x - iy = \rho e^{-i\theta} = \rho(\cos\theta - i \sin\theta)$$

It is used in the CONJUGATE TRANSPOSITION of matrices:

$$(A_{ij})^* = \overline{A}_{ji}$$

MATRICES

- $A+B = (a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij})$
- $\lambda A = (\lambda a_{ij})$
- $\underset{m\times n}{C} = \underset{m\times p}{A}\underset{p\times n}{B} = (c_{ij}) = \sum_{k=1}^{p} a_{ik} b_{kj}$
- $AA^{-1} = A^{-1}A = I$

The INVERSE of a matrix exists only if its DETERMINANT $\det(A)$ is $\neq 0$.

$\det(A) \neq 0$ iff column vectors are linearly independent.

If a matrix is DIAGONAL ~~OR TRIANGULAR~~, its determinant is the product of diagonal elements.

A matrix is LOWER/UPPER TRIANGULAR if all elements above/under the main diagonal are zero.

If $A \in \mathbb{R}^{m\times n}$ and its transpose $A^T \in \mathbb{R}^{n\times m}$, A is SYMMETRIC if $A = A^T$. If $A = A^H$, A is HERMITIAN.

VECTORS

A set of vectors $\{y_1, \dots, y_m\}$ is LINEARLY INDEPENDENT if

$$a_1 y_1 + \dots + a_m y_m = 0 \iff a_1, \dots, a_m = 0$$

B is a BASIS for $\mathbb{R}^n$ or $\mathbb{C}^n$ if $B = \{y_1, \dots y_n\}$ and $y_1, \dots y_n$ are all independent vectors. Any vector w in $\mathbb{R}^n$ can therefore be written as

$$w = \sum_{k=1}^{n} a_k y_k$$

a_k are the unique COMPONENTS of w w.r.t. B.

The SCALAR PRODUCT (DOT) of v and w is defined as

$$(v, w) = w^T v = \sum_{k=1}^{n} a_k b_k \text{, with a, b respective components of } v \text{ and } w$$

The MODULUS of a vector v is given by the EUCLIDEAN NORM formula

$$\|v\| = \sqrt{(v,v)} = \sqrt{\sum_{k=1}^{n} v_k^2}$$

The VECTOR PRODUCT (CROSS) of $v, w \in \mathbb{R}^3$ is the vector u, orthogonal to v and w, with modulus $|u| = |v||w|\sin\alpha$

$v \in \mathbb{C}^n$ is an EIGENVECTOR of $A \in \mathbb{C}^{n\times n}$ associated with EIGENVALUE λ if

$$Av = \lambda v$$

The Eigenvalues of diagonal and triangular matrices are the elements on the diagonal.

A matrix is said to be POSITIVE DEFINITE if

$$z^T A z \geq 0 \quad \forall z \in \mathbb{R}^n$$

↳ > DEFINITE

> SEMI-DEFINITE

REAL FUNCTIONS

If $f(\alpha) = 0$ α is a ZERO or ROOT of f. It is SIMPLE if $f'(\alpha) \neq 0$, MULTIPLE otherwise.

The space $\mathbb{P}_n$ of polynomials of degree $\leq n$ is

$$p_n(x) = \sum_{k=0}^{n} a_k x^k, \text{ with } a_k \text{ given coefficients}$$

The number of zeros cannot usually be estimated a priori (except for polynomials, where it's $= n$). The values for p_n zeros cannot be computed with an explicit formula for $n \geq 5$

FUNDAMENTAL THEOREM OF INTEGRATION, for f continuous in $[a,b]$

$$F(x) = \int_a^x f(t)\,dt \quad \forall x \in [a,b) \Rightarrow F'(x) = f(x) \quad \forall x \in [a,b)$$

(F: PRIMITIVE)

FIRST MEAN-VALUE THEOREM FOR INTEGRALS, for f continuous in $[a,b)$ and $x_1, x_2 \in [a,b)$ with $x_1 < x_2$

$$\exists \xi \in (x_1, x_2) \text{ s.t. } f(\xi) = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} f(t)\,dt$$

$f \in [a,b]$ is DIFFERENTIABLE in $x \in (a,b)$ if

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

exists and is finite.

MEAN VALUE THEOREM: if $f \in C^0([a,b])$ and is differentiable in (a,b)

$$\exists \xi \in (a,b) \text{ s.t. } f'(\xi) = \frac{f(b) - f(a)}{b - a}$$

TAYLOR EXPANSION OF p_n: If $f \in C^n([x_0 - c, x_0 + c])$ (neighborhood of x_0), f can be approximated in that interval as

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

ESTIMATING ERRORS

REAL WORLD PROBLEM x_{ph} —(MODELING ERROR, e_m)→ (MATHEMATICAL) MODELIZED PROBLEM $x = \int f(t)dt$ —(TRUNCATION ERROR, e_t)→ NUMERICAL PROBLEM $x_n = \sum_k f(t_k)\alpha_k$ —(ROUNDOFF ERROR, e_r)→ COMPUTER ESTIMATE $\hat{x}$

The sum of TRUNCATION ERROR (from reducing a problem to a finite set of operations) and ROUNDOFF ERROR coming from machine representation, is called COMPUTATIONAL ERROR e_c

$$e_c^{abs} = |x - \hat{x}| \qquad e_c^{rel} = \frac{|x - \hat{x}|}{|x|}$$

To convert a mathematical problem (MAT) in numerical form (NUM), we use a DISCRETIZATION PARAMETER h, positive.

If (NUM) $\to$ (MAT) as $h \to 0$, the numerical process is said to be CONVERGENT

If we can bound e_c as $e_c \leq Ch^p$, we say that the method is CONVERGENT OF ORDER p. If a lower bound $C'h^p \leq e_c$ also exists, we can approximate the final error.

Logarithmic scale is effective for numerical methods since lines slopes represent the order of convergence for methods. The semi-logarithmic scale is also used to visualize functions that span many orders of magnitude in y in a short x interval.

The computational cost is $O(\text{ops})$ and can be constant, linear, polynomial, exponential, factorial, ecc

Numerical approximation can be performed exclusively on WELL-POSED PROBLEMS, problems for which the solution:

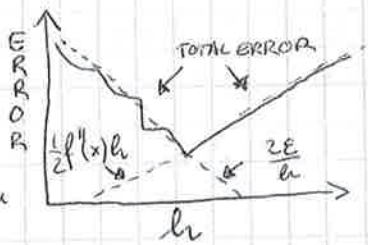
- EXISTS
- IS UNIQUE
- DEPENDS CONTINUOUSLY ON DATA

TOTAL ERROR : $f(x) - \hat{f}(\hat{x}) = \underbrace{\hat{f}(\hat{x}) - f(\hat{x})}_{\text{COMPUTATIONAL ERROR } (e_c = e_+ + e_r)} + \underbrace{f(\hat{x}) - f(x)}_{\text{PROPAGATED DATA ERROR (independent from } f)}$

ex. finite differences approximation ($f'(x) = \lim_{h \to 0} \frac{f(x-h) - f(x)}{h}$)

- TRUNCATION ERROR (obtained through Taylors) $\sim \frac{1}{2}|f''(x)|h + O(h^2)$
- ROUNDING ERROR $\sim \frac{2\varepsilon}{h}$, with ε = machine precision

The optimal h is thus $h = 2\sqrt{\frac{\varepsilon}{|f''(x)|}}$



PROBLEM STABILITY: Small changes in input data produce small variations on the output. Synonym of well-posedness.

Given δ a perturbation in data s.t. $d + \delta d \in D$, and $x + \delta x$ the perturbed solution, then

$$\forall d \in D \quad \exists \eta(d), K \text{ s.t } \|\delta d\|_d < \eta \in D \implies \|\delta x\|_x < K\|\delta d\|_d$$

CONDITION NUMBERS: Can be either relative or absolute, and measure problem sensitiveness w.r.t. input data.

If we define $\Delta y = f(x) - f(\hat{x})$ and $\Delta x = x - \hat{x}$, we have:

RELATIVE CN $K_{rel} = \frac{\Delta y / y}{\Delta x / x} \approx |f'(x)| \frac{|x|}{|f(x)|}$ ABSOLUTE CN $K_{abs} = \frac{\Delta y}{\Delta x}$ (if $f(x)$ or $x = 0$) $\approx |f$

If $K \gg 1$, the problem is ILL-POSED (SENSITIVE, UNSTABLE) and is thus not approximable through numerical methods.

A numerical approximation can be seen as a sequence of simpler approximating problems that converge to the original one

$$\lim_{n\to\infty} \|y_n - y\| = \lim_{n\to\infty} \|x_n - x\| = 0$$

$$\hookrightarrow \lim_{n\to\infty} f_n(x) = f(x)$$

BANACH SPACES

Given $\vec{V}$ over $\mathbb{R}$ or $\mathbb{C}$, a <u>seminorm</u> is a function $|\cdot| : V \to \mathbb{C}$ which satisfies:

- $|cf| = |c||f| \quad \forall c \in \mathbb{C}$ (HOMOGENEITY)
- $|f+g| \le |f| + |g|$ (TRIANGULAR INEQUALITY)

$\vec{V}$ is a VECTOR SPACE and the norm is a LINEAR MAPPING.

If $|f| = 0$ iff $f = 0$ (POSITIVE DEFINITE) is also verified, we have a NORM.

A vector space is said to be COMPLETE if every Cauchy sequence in that space converges to one of the space's elements.

A complete vector space with a norm is called BANACH SPACE

The scalar product is a mapping $V \times V \to \mathbb{C}$ which is:

- LINEAR $(\alpha_1 v_1 + \alpha_2 v_2, w) = \alpha_1 (v_1, w) + \alpha_2 (v_2, w)$
- SYMMETRIC $(v, w) = \overline{(w, v)}$
- POSITIVE DEFINITE $(v_1, v_2) \ge 0 \ \forall v_1, v_2$ and $(v_1, v_2) = 0$ iff $v_1, v_2 = 0$

A Banach space with scalar product and a norm $\|f\| = (f, f)$ induced by the product is called HILBERT SPACE

Some examples of norms in Banach spaces:

- $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
- $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$ (EUCLIDEAN NORM)
- $\|x\|_\infty = \sup_{1 \le i \le n} |x_i|$

In a FINITE-DIMENSIONAL vector space (dimension is given by the number of vectors in the basis), all norms are equivalents:

$$\forall \|\cdot\|_a, \|\cdot\|_b \ \exists\ 0 \le C_1 \le C_2 \text{ s.t. } C_1 \|x\|_b \le \|x\|_a \le C_2 \|x\|_b$$

GENERAL PROBLEM

CASE

$f : X \to Y$, both Banach

$x \to f(x)$

$\Longrightarrow$

APPROXIMATED CASE

$\hat{f} = f_n : X_n \to Y_n$

$x_n \to f_n(x_n) = \hat{x} \to \hat{f}(\hat{x})$

CONVERGE, CONSISTENCY AND LAX-RICHTMEYER

A numerical method is CONVERGENT if the approximation $\hat{f}_n$ of a problem f satisfies:

- $\lim_{n\to\infty} \| x_n - x \|_{\mathbb{X}} = 0$
- $\lim_{n\to\infty} \| \hat{f}_n(x_n) - f(x) \|_{\mathbb{X}} = 0$ ($\hat{f}_n(x_n)$: APPROX.)

A numerical problem is CONSISTENT when, if $x \in \mathbb{X}_n \ \forall n$, we have that

$$\lim_{n\to\infty} \| f_n(x) - f(x) \| = 0$$

($f_n(x)$: EXACT)

Example: Sum of two numbers

- $\mathbb{X} = \mathbb{R}^2$, $\|x\|_{\mathbb{X}} = |x_1| + |x_2| = \|x\|_{\ell^1(\mathbb{R}^2)}$
- $\mathbb{Y} = \mathbb{R}$, $\|y\|_{\mathbb{Y}} = |y| = \|y\|_{\ell^1(\mathbb{R}^1)}$

$$K_{rel} = \underbrace{\frac{|\Delta y|}{|\Delta x|}}_{\leq 1} \cdot \frac{|x|}{|y|} \Rightarrow K_{rel} \leq \frac{|x_1| + |x_2|}{|x_1 + x_2|}$$

RESULT: Unstable in $\mathbb{F}$ when $x_1 \cong -x_2 \Rightarrow K_{rel} \to \infty$.

- A convergent approximation is always stable
- Finite differences are unstable since they are a sum of two numbers with close absolute value and opposite sign
- For integration, $K_{rel} = \frac{\int |x|}{|\int x|}$, so it is ill-posed when $x \sim 0$
- The condition number of a matrix A is $K_{rel} = \|A^{-1}\| \|A\|$. This usually corresponds to

$$K(A) = \frac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|}$$

The LAX-RICHTMYER THEOREM says that if a problem is consistent, then stability and convergence are equivalent.

- STABILITY controls perturbations in data and their impact
- CONSISTENCY controls bad approximations of a problem
- CONVERGENCE controls bad discretizations of the problem space (and includes stability)

A method is CONSISTENT if the residual (error produced by plugging the exact solution in the scheme) $\to 0$ as $h \to 0$

NONLINEAR EQUATIONS

We may want to find the roots of non linear functions ($\alpha \in \mathbb{R}$ s.t. $f(\alpha)=0$) in a computational way. Most common approaches are ITERATIVE, since there is no explicit solving formula for $p \in \mathbb{R}^n$ with $n \geq 5$ (ABEL'S THEOREM).

BISECTION METHOD (Linear convergence)

It is used to compute the root of a function f on interval $[a, b]$

CONSTRAINTS FOR CONVERGENCE:
- f should be continuous on $[a\ b]$
- interval endpoints should have different sign ($f(a)f(b)<0$) to have at least 1 solution (THEOREM OF ZEROS FOR CONTINUOUS FUNCTIONS)

We generate a sequence of intervals whose lenght is halved at each step, with $x^{(k)}$ being the midpoint at step k.

The error of estimation at step k is:

$$|e^{(k)}| = |x^{(k)} - \alpha| < \frac{1}{2}|I^{(k)}| = \left(\frac{1}{2}\right)^{k+1}(b-a)$$

In order to ensure that the error $|e^{(k)}| < \varepsilon$, we carry out k_{min} iterations at least:

$$k_{min} > \log_2\left(\frac{b-a}{\varepsilon}\right) - 1$$

The error does not decrease monotonically. The only possible stopping criterion is controlling the size of $I^{(k)}$

NEWTON'S METHOD (Quadratic or linear convergence)

It is used to compute the root of a function f by using the values of f and f' (more efficient than bisection).

CONSTRAINTS FOR CONVERGENCE:
- $f: \mathbb{R} \to \mathbb{R}$ should be differentiable
- x_0 is sufficiently close to α given f (estimate through graph and bisection)

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k = 0, 1, \dots$$

If $f \in C^2$, we have that $\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)} \Rightarrow$ QUADRATIC CONVERGENCE

If f has zeros with multiplicity $m > 1$, if $f'(x) \neq 0 \ \forall x \in I(\alpha)$ the method converges linearly. To restore quadratic convergence, one can use the MODIFIED NEWTON METHOD or ADAPTIVE NETWON METHODS if m is unknown.

$$x^{(k+1)} = x^{(k)} - m\frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k = 0, 1, \dots$$

(α of f has multiplicity m iff $f(\alpha) = \dots = f^{(m-1)}(\alpha) = 0$ and $f^{(m)}(\alpha) \neq 0$)

STOPPING CRITERION: Control of the increment

$$|x^{(k+1)} - x^{(k)}| < \varepsilon$$

We can also perform a test on the residual which is valid only if $|f'(x)| \simeq 1 \ \forall x \in I(\alpha)$, else it produces an over or underestimation of error.

$$|r^{(k_{min})}| = |f(x^{(k_{min})})| < \varepsilon$$

SECANT METHOD (superlinear convergence)

In case $f'(x)$ is not available we can replace its value with an incremental ratio based on previous values.

$$x^{(k+1)} = x^{(k)} - \left(\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} \right)^{-1} f(x^{(k)})$$

CONSTRAINTS FOR CONVERGENCE:
- α has $m = 1$ (for superlinear)
- $x^{(0)}$ is selected in $I(\alpha)$ suitable
- $f'(x) \neq 0 \ \forall x \in I(\alpha)$

If $m = 1$ and $f \in C^2(I(\alpha))$, $\exists c > 0$ s.t.

$$|x^{(k+1)} - \alpha| \leq c |x^{(k)} - \alpha|^p, \text{ with } p \approx 1.618$$

Else, the method converges linearly.

SYSTEMS OF NONLINEAR EQUATIONS

Given $f_1, \dots, f_n$ nonlinear functions in $x_1, \dots x_n$, we can set $\bar{f} = (f_1, \dots, f_n)^T$ and $\bar{x} = (x_1, \dots, x_n)^T$ to write a system as

$$\bar{f}(\bar{x}) = 0$$

We can extend the Newton method to that system by replacing the f' with the JACOBIAN MATRIX $J_{\bar{f}}$, as

$$(J_{\bar{f}})_{ij} = \frac{\partial f_i}{\partial x_j} \quad i, j = 1, \dots, n$$

The secant method can also be adopted by recursively defining matrices B_k which are suitable approximations of $J_{\bar{f}}(x^{(0)})$ (BROYDEN METHOD). This belongs to the family of QUASI-NEWTON METHODS.

FIXED POINT ITERATIONS

Given a function $\phi : [a, b] \to \mathbb{R}$, we want to find a α s.t.

$$\phi(\alpha) = \alpha$$

If α exists, it is called a FIXED POINT of ϕ and it can be

be computed as follows:

$$x^{(k+1)} = \phi(x^{(k)}), \quad k \geq 0, \text{ with } x^{(0)} \text{ initial guess.}$$

ϕ is called the ITERATION FUNCTION. The Newton method is a special case of fixed point iteration where

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

GLOBAL CONVERGENCE

1. If $\phi(x)$ is continuous in $[a,b]$ and $\phi(x) \in [a,b] \; \forall x \in [a,b]$ then THERE EXISTS AT LEAST 1 $\alpha \in [a,b]$.
2. Moreover, if $\exists L < 1$ s.t. (ASYMPTOTIC CONVERGENCE FACTOR)

$$|\phi(x_1) - \phi(x_2)| \leq L |x_1 - x_2| \quad \forall x_1, x_2 \in [a,b]$$

then $\alpha \in [a,b]$ IS UNIQUE and the ITERATION CONVERGES TO $\alpha \; \forall x^{(0)} \in [a,b]$.

PROOF: 1. From our assumptions we have that $g(x) = \phi(x) - x$ is continuous in $[a,b]$, with:

$$g(a) = \phi(a) - a \geq 0 \quad \text{and} \quad g(b) = \phi(b) - b \leq 0$$

For THEOREM OF Ø FOR C. FUNCTIONS we know that g has at least 1 zero, and thus $\exists \alpha$ for ϕ in $[a,b]$

2. If two fixed points existed, we would have

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

which is absurd for $L < 1$. For x^0 in $[a,b]$ and $x^{(k+1)} = \phi(x^{(k)})$ we have

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

THE SMALLER L, THE FASTER THE CONVERG.

Since $L < 1$ for $k \to \infty$ a.k.a. convergence.

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

LOCAL CONVERGENCE (OSTROWSKI'S THEOREM)

If ϕ a continuous and differentiable function on $[a,b]$ with fixed point α, and $|\phi'(\alpha)| < 1$. $\exists \delta > 0$ s.t.

$$|x^{(0)} - \alpha| \leq \delta \quad \forall x^{(0)} \text{ in } [a,b]$$

for which the $x^{(k)}$ converges to α. It holds

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

$\forall C$ s.t. $0 < |\phi'(\alpha)| < C < 1$, for large k: $|x^{(k+1)} - \alpha| \leq C |x^{(k)} - \alpha|$ q

- If $|\phi'(\alpha)| > 1$ the method diverges. If $|\phi'(\alpha)| = 1$, it depends on the function.

QUADRATIC CONVERGENCE:

If $\phi \in C^2([a,b])$ and α is fixed point of ϕ, with ϕ having local convergence. Then, if $\phi'(\alpha) = 0$ and $\phi''(\alpha) \neq 0$ FPI CONVERGES WITH ORDER 2 and

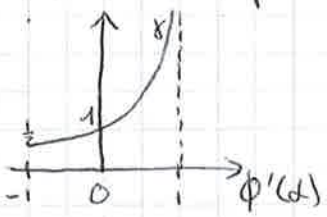
$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{1}{2}\phi''(\alpha)$$

STOPPING CRITERIA: Error estimation at step k is

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

Satisfactory when we have quadratic convergence (since $\phi'(\alpha) = 0$) or when $-1 < \phi'(\alpha) < 0$, problems when $\phi'(\alpha) \simeq 1$.

In that case we can use the control of the residual as described for newton method (page 8).



AITKEN METHOD: If ϕ converges linearly to α, there must be a λ s.t. $\phi(x^{(k)}) - \alpha = \lambda(x^{(k)} - \alpha)$. This allows us to obtain a better estimate of $x^{(k+1)}$ than $\phi(x^{(k)})$

$$\alpha = x^{(k)} + (\phi(x^{(k)}) - x^{(k)})/(1-\lambda), \text{ with}$$

$$\lambda^{(k)} = \frac{\phi(\phi(x^{(k)})) - \phi(x^{(k)})}{\phi(x^{(k)}) - x^{(k)}} \text{ given}$$

$$\lim_{k\to\infty} \lambda^{(k)} = \phi'(\alpha).$$

$$\Rightarrow x^{(k+1)} = x^{(k)} - \frac{(\phi(x^{(k)}) - x^{(k)})^2}{\phi(\phi(x^{(k)})) - 2\phi(x^{(k)}) + x^{(k)}}, \quad k \geq 0$$

(AITKEN'S EXTRAPOLATION FORMULA, STEFFENSON'S METHOD.)

The derived function $\phi_\Delta(x)$ has the same α as $\phi(x)$, but converges faster:

- LINEAR $\phi \longrightarrow$ QUADRATIC ϕ_Δ
- $p \geq 2$ $\phi \longrightarrow 2p-1$ ϕ_Δ
- LINEARLY with $m \geq 2$ $\phi \longrightarrow$ LINEARLY with $L = 1 - 1/m$ ϕ_Δ

It may converge even if normal FPI diverges.

ROPE METHOD

Obtained by modifying the Newton method, replacing $f'(x)$ with a fixed q

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{q}, \quad k = 0, 1, \dots$$

e.g. $q = \frac{f(b)-f(a)}{b-a}$ in $[a, b]$.

Since it is a FPI with $\phi(x) = x - \frac{1}{q} f(x)$, we have convergence when

$$\phi'(x) = \left| 1 - \frac{1}{q} f'(\alpha) \right| \leq 1$$

INTERPOLATION

APPROXIMATION

Approximating a set of data or a function in $[a, b]$ consists in finding a suitable function $\tilde{f}$ that represents them with enough accuracy.

We can use Taylor polynomials to approximate complex functions but they require many computations and have unpredictable behaviors on the sides of the domain.

If X is a Banach space and $M \subseteq X$, $\tilde{f} \in M$ is the BEST APPROXIMATION of a function $f \in X$ when

$$\|f - \tilde{f}\| = E(f) = \inf_{f \in M} \|f - \tilde{f}\|$$

- $\tilde{f} \in M$ is the best approximation of f in M
- If M is a finite-dimensional subspace of X, then $\exists\, \tilde{f}$ B.A. of f in M (EXISTENCE THEOREM)
- If X is STRICTLY CONVEX (any x, y on the unit sphere ∂B are joined by a segment that touches ∂B only in x, y), then $\tilde{f}$ IS UNIQUE (UNIQUENESS THEOREM)

INTERPOLATION

Given $n+1$ points $\{q_i = (x_i, y_i)\}_{i=0}^{n}$ on an interval, we want to find the function p s.t. $p(x_i) = y_i \ \forall i$.

We call this function p INTERPOLANT, and the point NODES. We say that p interpolates y_i in nodes q_i.

Interpolation is a form of approximation that could be used both to simplify a complex function in order to make it easier to derive or to understand data distribution. The interpolants can be POLYNOMIAL, TRIGONOMETRIC, RATIONAL, ecc.

LAGRANGE INTERPOLATION ($\mathbb{P}$ is POLYNOMIAL)

Given $n+1$ couples $\{x_i, y_i\}$, $i=0,\dots n$ with x_i as nodes, we want to find a polynomial of degree $\leq n$ ($\Pi_n \in \mathbb{P}^n$) s.t.

$$\Pi_n(x_i) = y_i \quad \forall i$$

If y_i represent the values of a continuous f, Π_n is the interpolant of f, denoted $\Pi_n f$.

→ In this setting, $\exists!\ \Pi_n \in \mathbb{P}^n$ s.t. $\Pi_n(x_i) = y_i\ \forall i$

In order to obtain an expresion for Π_n, we study a specid case in which $y_j = 0\ \forall j$ except when $y_{j=k} = 1$.

$$\varphi_k \in \mathbb{P}^n,\ \varphi_k(x_j) = \delta_{jk} = \begin{cases} 1 & \text{if } j=k \\ 0 & \text{otherwise} \end{cases}$$

(δ_{jk}: KRONECKER SYMBOL)

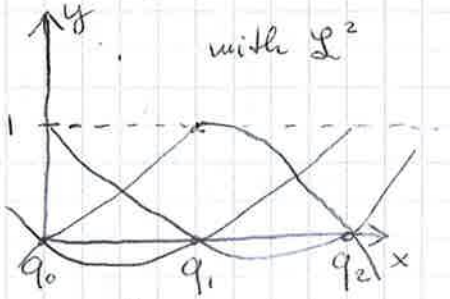
φ_k is called LAGRANGE BASIS since it is a basis for $\mathbb{P}^n$ (all its elements define the whole spoce $\mathbb{P}^n$). It has the following expression:

$$\varphi_k(x) = \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}$$

→ Also called LAGRANGE CHARACTERISTIC POLYNOMIALS

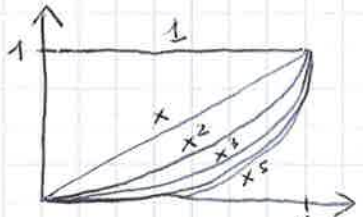
We define the LAGRANGE INTERPOLANT of f on nodes $x_0,\dots,x_n$ the following linear combination of degree n:

$$\mathcal{L}^n f = \sum_{k=0}^{n} f(x_k)\,\varphi_k(x)$$



with $\mathcal{L}^2$

$\mathcal{L}^2 f$ on 3 nodes.



$P^n = \text{span}\{x^i\}_{i=0}^{n}$

The LAGRANGE BASIS is especially fit for good approximations since other polynomial sets may be ill-conditioned for the approximation task

The example P^n generates the VANDERMONDE MATRIX, which is ill conditioned since for n large rightmost columns will be very similar, making the matrix "hardly invertible"

$$B = (B_{ij}) = \begin{bmatrix} 1 & q_0 & q_0^2 & \cdots & q_0^n \\ 1 & q_1 & q_1^2 & \cdots & q_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & q_n & q_n^2 & \cdots & q_n^n \end{bmatrix} = (q_i^j)$$

INTERPOLATION ERROR

$\{x_i, i=0,\dots,n\}$ $(n+1)$ ~~bounded~~ nodes on a bounded interval I.
$f \in C^{n+1}(I)$. Then, $\forall x \in I \quad \exists \xi_x \in I$ s.t.

$$E_n f(x) = f(x) - \mathcal{L}^n f(x) = \frac{f^{(n+1)}(\xi_x)\,\omega(x)}{(n+1)!},$$

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

Since $\|\cdot\|_\infty$ represent the highest value (sup) of a function, we can bound the error as

$$\|f(x) - \mathcal{L}^n f(x)\|_\infty \le \frac{\|f^{(n+1)}(\xi)\|_\infty \, \|\omega(x)\|_\infty}{(n+1)!}$$

If f is analitically extendible in an oval $O(a,b,R)$ with $R>0$

$$\Rightarrow \|f^{(n+1)}\|_{\infty, \overline{O(a,b,R)}} \le \frac{(n+1)!}{R^{n+1}} \|f\|_{\infty, \overline{O(a,b,R)}}$$

Thus, we can control the $(n+1)$ derivative directly with f.

Also, $$\|f - \mathcal{L}^n f\|_{\infty,[a,b]} \le \frac{(n+1)!}{R^{n+1}} \|f\|_{\infty, O(a,b,R)} \left(\frac{|b-a|}{R}\right)^{n+1}$$

$\Rightarrow$ Increasing the degree n of interpolator DOESN'T GUARANTEE a better approximation of f. Indeed, we may have that

$$\lim_{n\to\infty} \|f - \mathcal{L}^n f\|_\infty = \infty$$

RUNGE COUNTEREXAMPLE: $f(x) = \frac{1}{(1+x^2)}$ is interpolated at equispaced nodes in $I = [-5,5]$. The error $\|f - \mathcal{L}^n f\|_\infty$ tends to infinity when $n \to \infty$.

The presence of severe oscillations of $\mathcal{L}f$ w.r.t. f, especially near the endpoints indicates lack of convergence. This is also called RUNGE'S PHENOMENON.

STABILITY OF INTERPOLATION

We want to estimate the impact of perturbed values $\hat{f}(x)$ on the interpolator $\mathcal{L}^n f$. We have that

$$\|\mathcal{L}^n f - \mathcal{L}^n \hat{f}\|_\infty = \Lambda_n(\{x_i\}_{i=0}^n) \|f - \hat{f}\|, \text{ where}$$

$$\Lambda_n(\{x_i\}_{i=0}^n) = \left\| \sum_{i=0}^{n} |\varphi_i(x)| \right\|_\infty$$

is the LEBESGUE'S CONSTANT depending on interpolation nodes

From first formula, we have that small variations in f yield small changes in $\mathcal{L}^n f$ IF Λ is SMALL. Therefore Λ can be regarded as a CONDITION NUMBER FOR INTERPOLATION.

It can be proved that $\| \mathcal{L}^n f \| \leq \Lambda_n(x)$.

For Lagrange interpolation at equispaced nodes, one has

$$\Lambda_n(x) \simeq \frac{2^{n+1}}{e \cdot n(\log n + \gamma)}, \text{ with } \begin{matrix} e \approx 2.718 \\ \gamma \approx 0.547 \end{matrix}$$

For large values of n, this becomes unstable.

DISTANCE FROM B.A. If $p = \inf_{x \in P^n} \| f - x \|_\infty$, wit $\{x_i\}_{i=0}^n$ $n+1$ nodes.

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

$$\leq \| I - \mathcal{L}^n \|_2 \| f - p \|_\infty$$

$$\leq (\| I \|_2 - \| \mathcal{L}^n \|_2) \| f - p \|_\infty$$

$$\leq (1 + \| \mathcal{L}^n \|_2) \| f - p \|_\infty$$

$\Rightarrow$

where I is an operator for $f-p$ and

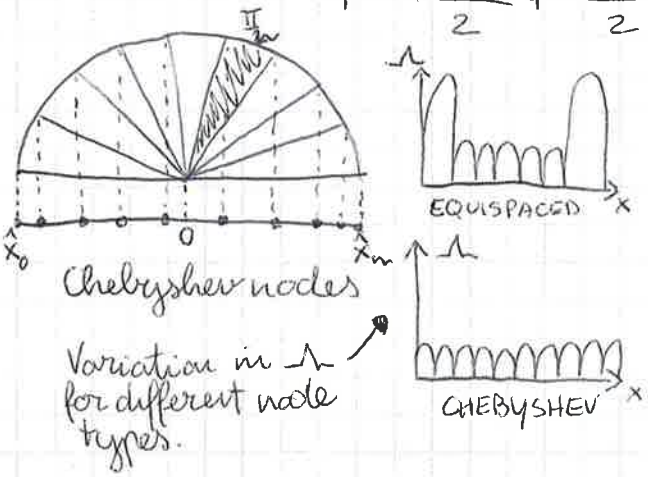
$$\| \mathcal{L}^n \|_2 = \sup_{n \neq 0} \frac{\| \mathcal{L}^n n \|_\infty}{\| n \|_\infty}$$

p is the BEST APPROXIMATION of f on nodes $\{x_i\}_{i=0}^n$.

CHEBYSHEV NODES

In order to minimize Λ and thus avoid Runge's phenomenon we can use CHEBYSHEV NODES on interval $[a, b]$, defined as

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \hat{x}_i, \text{ where } \hat{x}_i = -\cos\left(\frac{\pi i}{n}\right), \; i = 0, \dots, n$$



The nodes are equispaced on the semicircunference of diameter $[a, b]$ and are clustered towards the endpoints of the interval

For Chebyshev nodes, if f is continuously differentiable in $[a, b]$, $\mathcal{L}^n f$ converges to f as $n \to \infty$ $\forall x \in [a, b]$

ERDOS THEOREM $\forall X$, X being an INFINITE TRIANGULAR MATRIX OF INTERPOLATION POINTS, we have:

$$\Lambda_n(X) \geq \frac{2}{\pi} \log(n+1) - c$$

$$X = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ x_0 & 0 & \cdots & \cdots & 0 \\ x_0 & x_1 & 0 & \cdots & 0 \\ x_0 & x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

For CHEBYSHEV NODES we have $\Lambda_n(X) \leq \frac{2}{\pi} \log(n+1) + 1$

For EQUISPACED NODES we have $\Lambda_n(X) \leq \frac{2^{n+1}}{c(n \log n)}$

FABER THEOREM $\forall X \ \exists f \in C^0([a,b]) \text{ s.t. } \| f - \mathcal{L}^n f \|_\infty \not\to 0$

The Faber theorem proves that even on Chebyshev nodes not all continuous functions will converge when used for interpolation.

WEIERSTRASS APPROXIMATION THEOREM

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

It shows that polynomial functions are dense in $C^0([a,b])$ and each polynomial can be uniformly approximated by one with rational coefficients.

BERNSTEIN POLYNOMIALS

The $n+1$ BERNSTEIN BASIS for $\mathbb{P}^n$ are defined as

$$b_{n,i}(x) = \binom{n}{i} x^i (1-x)^{n-i}, \quad i = 0, \dots, n$$

$b_{n,i}$ is the i-th polynomial in the Bernstein basis of degree n.

A linear combination of $b_{n,i}$ is called a BERNSTEIN POLYNOMIAL of degree n based on function f.

$$B_n f(x) = \sum_{k=0}^{n} f\left(\tfrac{k}{n}\right) b_{n,k}(x)$$

$\forall x$, this is a weighted average of the $n+1$ values $f(\frac{k}{n})$, called BERNSTEIN COEFFICIENTS.

PROPERTIES:
- $\sum_{i=0}^{n} b_{i,n} = (1 - x + x)^n = 1^n = 1 \cdot \forall x \in [0,1]$
- $b_{i,n} \geq 0 \quad \forall x \in [0,1]$
- B_n is a LINEAR POSITIVE operator: $B_n f \geq 0$ if $f \geq 0$
- $B_n f(\frac{i}{n}) \neq f(\frac{i}{n})$ and if $f \in C^0([a,b])$ we have that $B_n f(x) \to f(x)$ as $n \to \infty$

$\Rightarrow$ The convergence is not POINTWISE as in interpolation, but UNIFORM:

$$\lim_{n \to \infty} \| f(x) - B_n f(x) \|_\infty = 0 \quad \text{with} \quad 0 \leq x \leq 1$$

QUALITATIVE PROOF OF WEIERSTRASS THEOREM

$\forall f \in C(I)$ and $\forall x_0 \in I$ we can find a quadratic function q s.t. $q > f \ \forall x$, but $q(x_0)$ is close to $f(x_0)$. Same can be done with $q < f$.

$q = f(x) \pm \left(\frac{\varepsilon}{2} + \frac{\| f \|_\infty}{2\delta} (x - x_0)^2 \right)$, with $|x_1 - x_2| \leq \delta \Rightarrow |f(x_1) - f(x_2)| \leq \frac{\varepsilon}{2}$

$$q = a x^2 + b x + c, \quad M = \max_{x_0 \in [a,b]} (a(x_0), b(x_0), c(x_0)).$$

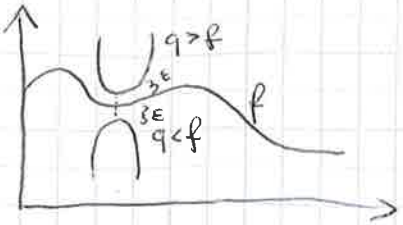
M depends exclusively on $\| f \|$, ε and δ but not on x_0

By choosing a large n we have $\|f_i - B_n f_i\| \leq \frac{\varepsilon}{M}$. With triangle inequality we get $\|q - B_n q\| \leq 3\varepsilon$. We have then

$$B_n f(x_0) \leq B_n q(x_0) \leq q(x_0) + 3\varepsilon = f(x_0) + 4\varepsilon$$

Same can be done below, having that

$$\forall x_0 \quad f(x_0) - \varepsilon \leq B_n f(x_0) \leq f(x_0) + \varepsilon$$
$$\Rightarrow \|B_n f - f\|_\infty \leq \varepsilon$$



MORE ON INTERPOLATION

- We can build a piecewise linear interpolant of f to avoid Runge effect when the number of nodes increases. $\tilde{f}$ is a piecewise linear continuous function also called FINITE ELEMENT INTERPOLANT
- We can perform interpolation by cubic splines, which are piecewise cubic functions $f \in C^2$
- While the MINMAX APPROXIMATION we used so far is based on $\|\cdot\|_\infty$ the LEAST SQUARES APPROXIMATION uses the Euclidean norm $\|\cdot\|_2$ to minimize $MSE = \sum_{i=0}^{n} (y_i - \tilde{f}(x_i))^2$.
- Piecewise linear and splines are well suited to approximate data and functions in several dimensions.
- Trigonometric interpolation is well suited to approximate periodic functions. $\tilde{f}$ is a linear combination of sin and cos functions. FFT and IFFT allow for efficient computation of Fourier coefficients for a trigonometric interpolant from node values.

BEST APPROXIMATION IN HILBERT SPACES

L^2 is an Hilbert space where the norm induced by the scalar product between vectors is $\|x\|_{L^2} = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2} = \sqrt{(x,x)}$

$$(a,b),\ a\&b \in L^2([0,1]) = \int_0^1 a\cdot b,\quad \|a\| = \sqrt{\int_0^1 a^2}$$

BEST APPROXIMATION THEOREM in L^2

Given a function $f \in L^2(\mathbb{P}^n)$, p. is B.A. of f in $\mathbb{P}^n$ iff

$$(f-p, q) = 0 \quad \forall q \in \mathbb{P}^n,\ \forall f \in L^2(\mathbb{P}^n)$$

(RECALL: $\|p-f\| \le \|q-f\|\ \forall q \in \mathbb{P}^n$ if p is B.A. of f w.r.t. chosen norm.

PROOF: Knowing that p. is B.A. $\Rightarrow$

$$\|q-f\|^2 = \|q-p+p-f\|^2 = \|q-p\|^2 + \|p-f\|^2 + \underbrace{2(q-p,\ p-f)}_{0 \text{ since } p-q=q\in\mathbb{P}^n \text{ and } (0,n)=0}$$

$\hookrightarrow (A+B)^2 = A^2+B^2+2AB$

$$\Rightarrow \|p-f\|^2 \le \|q-f\|^2\ \forall q \in \mathbb{P}^n$$

PROOF #2: Knowing that $(f-p,q) = 0 \Rightarrow$

$$\|p-f\|^2 \le \|p-f+tq\|^2,\ \text{with } t \ge 0 \text{ perturbation},\ q \in \mathbb{P}^n$$

splitting tq and adding $\pm \frac{tq}{2}$

$$\|\underbrace{p-f+\frac{tq}{2}}_{A} - \underbrace{\frac{tq}{2}}_{-B}\|^2 \le \|\underbrace{p-f+\frac{tq}{2}}_{A} + \underbrace{\frac{tq}{2}}_{+B}\|^2$$

$(A+B)^2-(A-B)^2 = 4AB$

$$0 \le 4\left(p-f+\frac{tq}{2},\ \frac{tq}{2}\right)$$

$$0 \le t^2\|q\|^2 + 2t(p-f,q)$$

$$\Rightarrow (p-f,q) \ge -\frac{t}{2}\|q\|^2$$

By choosing $-q$ instead, we get $(p-f,q) \le \frac{t}{2}\|q\|^2$

Thus, it is valid $\forall t, \forall q$ that

$$-\frac{t}{2}\|q\|^2 \le (p-f,q) \le \frac{t}{2}\|q\|^2\ \forall t,\ \forall q$$

, which implies that $(p-f,q) = 0$ since t can be chosen to bound it on both sides.

Since $(p-f,q) = 0\ \forall q \Leftrightarrow (p-f, v_i) = 0\ \forall i = 0,1,\dots,n$ with $\mathbb{P} = \text{span}\{v_i\}$

$$\Rightarrow (p, v_i) = (f, v_i) \Rightarrow \left(\sum_{j=0}^{n} p_j v_j,\ v_i\right) = (f, v_i)$$

Computing integrals is easier than performing interpolation, and it yields better results.

MATRIX FORMULATION

We can rewrite $(\sum p_j v_j, v_i) = (f, v_i)$ as a matricial relation $Mp = F$, where $M_{ij} = (v_j, v_i) = \int_0^1 v_j v_i$ and $F_i = (f, v_i) = \int_0^1 f v_i$

If we set $v_i = x^{(i)}$, we obtain the HILBERT MATRIX

$$M_{ij} = \int_0^1 x^{(j)} x^{(i)} = \frac{1}{i+j+1}$$

The condition number of the Hilbert matrix is

$$K(M) = O\left(\frac{(1+\sqrt{2})}{\sqrt{n}}\right)^{4n}$$

When n increases K explodes, which is very bad. M is difficult to invert and very ill-conditioned because of collinear lines. We would like $M_{ij} = I$ so we use the LEGENDRE BASIS FUNCTION to make it orthonormal (perpendicular, a.k.a. diagonal) w.r.t. L^2

We want $v_i \in \mathbb{P}^n$ s.t. $M_{ij} = (v_i, v_j) = \delta_{ij}$. To build it, we use the GRAHAM-SCHMIDT METHOD:

$$\begin{cases} v_0 = 1, \ f \text{ s.t. } \int_0^1 f = 1 \\ f^{i+1} = x^{i+1} - \sum_{j=0}^{i} (x^{i+1}, v_j) v_j \\ v_{i+1} = \dfrac{f^{i+1}}{\|f^{i+1}\|} \end{cases}$$

→ The set of additive basis having unity as first element. This ensures orthogonality between basis functions.

As i (the degree) increases, $v_{i+1} = \dfrac{f^{i+1}}{\|f^{i+1}\|} \longrightarrow \infty$ since $x^{i+1} \longrightarrow \infty$ and thus $f^{i+1} \longrightarrow 0$.

We can avoid unstability by using $v_{i+1} = \dfrac{f_{i+1}}{f_{i+1}(0)}$ instead.

The points created with Graham-Schmidt represent the LEGENDRE BASIS. They make p (best approximation) easy to compute since M becames easy to invert and we have a diagonal matrix formed by orthogonal basis

$$p = M^{-1}F$$

INTEGRATION

Integration is an operation $f[a,b] \to \mathbb{R}$, defined as

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

Integration is very expensive from a numeric point of view if f is complicated. Our purpose is to make it simpler. given $f \in C^{r}([a,b])$

Many possible approaches, called QUADRATURES.

- MIDPOINT FORMULA  $I_{mp}(f) = (b-a)\, f\left(\frac{a+b}{2}\right)$ → DEGREE 1
- TRAPEZOIDAL FORMULA  $I_t(f) = \frac{b-a}{2}\left(f(a)+f(b)\right)$ → DEGREE 1
- SIMPSON FORMULA, using $\mathcal{L}^2 f$: $I_S(f) = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right)$ ↳ DEGREE 3

and their composite variants, using M intervals:

- $I^c_{mp} = H\sum_{k=1}^{M} f(\bar{x}_k)$
- $I^c_t(f) = \frac{H}{2}\sum_{k=1}^{M-1} f(x_k) + \frac{H}{2}\left(f(a)+f(b)\right)$
- $I^c_S(f) = \frac{H}{6}\sum_{k=1}^{M}\left(f(x_{k-1}) + 4f(\bar{x}_k) + f(x_k)\right)$

with $\bar{x}_k = \frac{(x_{k-1}+x_k)}{2}$ and $H = \frac{(b-a)}{M}$

Those are all specific cases of a more general quadrature formula.

$$I_n(f) = \sum_{i=0}^{n} \alpha_i\, f(y_i)$$

- $\{y_i\}$ are the quadrature NODES
- α_i are the quadrature weights.

We can use $\mathcal{L}^n f \in \mathbb{P}^n$ at nodes y_i as approximating function, to get the INTERPOLATORY QUADRATURE FORMULA

$$f_n(x) = \mathcal{L}^n f(x) : \quad I_n(f) = \int_a^b f_n(x)\,dx = \int_a^b \sum_{i=0}^{n} \varphi_i(x)\, f(y_i)\,dx$$
$$= \sum_{i=0}^{n} f(y_i) \int_a^b \varphi_i(x)\,dx \Rightarrow \sum_{i=0}^{n} \alpha_i\, f(y_i)$$

with α_i being $\int_a^b \varphi_i(x)\,dx$.

The DEGREE OF ACCURACY / EXACTNESS of a quadrature is the integer r s.t. quadrature using $\mathbb{P}^n$ doesn't produce errors on $I(p)$:

$$\max_{n \in \mathbb{N}} I_n(f) = \int f \quad \forall f \in \mathbb{P}^r$$

MIDPOINT RULE: For f linear function $\in [a,b]$, we can choose y_i as $\alpha_i = \frac{1}{2}b + \frac{1}{2}a$ to cancel out positive and negative approximation. → We approximate exactly $f \in \mathbb{P}^1$ with a constant function ($\mathbb{P}^0$). We have that

$$\left|\int_a^b f(x)\,dx - f\left(\frac{b-a}{2}\right)\right| \le \frac{\|f''\|_\infty}{3}\left(\frac{b-a}{2}\right)^3 = \frac{\|f''\|_\infty}{24} \quad \text{for } [a,b] = [0,1]$$

Since we can see the error depends on the size of the interval, we usually prefer composite quadrature, pasting together intervals through continuity conditions to keep them small.

LEGENDRE POLYNOMIALS AND MAX ACCURACY

Given $m \in \mathbb{N} > 0$, a quadrature formula $\sum_{i=0}^{n} \bar{\alpha}_i f(\bar{y}_i)$ has degree of accuracy $n+m$ iff it makes use of interpolation and the nodal polynomial $\omega_{n+1} = \prod_{i=0}^{n} (x - \bar{y}_i)$ associated to nodes $\{\bar{y}_i\}$ is s.t.

$$\int_a^b \omega_{n+1}(x)\, p(x)\, dx = 0 \quad \forall p \in \mathbb{P}_{m-1}$$

The maximum value for m is $n+1$, achieved when ω_{n+1} is proportional to $L_{n+1}(x)$, the Legendre polynomial of degree $n+1$. Legendre polynomials can be computed recursively as

$$L_0(x) = 1, \quad L_1(x) = x, \quad L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x)$$

Since L_{n+1} is orthogonal to $\forall L_{\{0,1,\dots,N\}}$ ($\int_a^b L_{n+1}(x) L_j(x)\, dx = 0 \ \forall j < n+1$), we can see why m is bounded at $n+1$. Thus, the highest degree of accuracy is $2n+1$, obtained using the GAUSS-LEGENDRE FORMULA.

$$I_{GL} = \begin{cases} \bar{y}_i = \text{roots of } L_{n+1}(x) \\ \bar{\alpha}_i = \dfrac{2}{(1-y_i^2)(L'_{n+1}(y_i))^2} \end{cases} \qquad i = 0, \dots, n$$

The related GAUSS-LEGENDRE-LOBATTO FORMULA includes interval bounds among quadrature points, and has a D.O.A of $2n-1$.

The interval used for I_{GL} is $\{-1, 1\}$ thus the $\bar{y}_i, \bar{\alpha}_i$. To recover to original values for (a, b), use Chebyshev formula:

$$y_i = \frac{a+b}{2} + \frac{b-a}{2} \bar{y}_i, \quad \alpha_i = \frac{b-a}{2} \bar{\alpha}_i$$

PROOF: Knowing $f \in \mathbb{P}^{n+m}$, we apply quotient theorem for $\mathbb{P}$

$$f(x) = \underbrace{\omega_{n+1}(x)}_{\in \mathbb{P}^{n+1}} \underbrace{p(x)}_{\in \mathbb{P}^{m-1}} + \underbrace{q(x)}_{\in \mathbb{P}^{n}}$$

$$\int_a^b f(x) = \underbrace{\int_a^b \omega_{n+1}(x) p(x) dx}_{(*)} + \int_a^b q(x) dx$$

Assuming $(*) = 0$, we get $\int_a^b f(x) = \int_a^b q(x) = I_n(q)$ (quadrature for $q \in \mathbb{P}^n$ is exact since we took $n+1$ nodes)

→ Knowing $(*) = 0$, we want to prove that if D.O.A. is $n+m$, then $(*) = 0$

$$I_n(f) = \int f \ \forall f \in \mathbb{P}^{n+m} \rightarrow \int_a^b \underbrace{\omega_{n+1}(x) p(x)}_{\in \mathbb{P}^{n+m}} dx = I_n(\omega_{n+1}(x) p(x))$$

Since $I_n(\omega_{n+1}(x)\,p(x)) = 0$ because $\omega_{n+1}(y_i) = 0 \ \forall i$, we proved it.
To prove that m is bound at n+1, we could replace $p \in \mathbb{P}^{m-1}$ with $\omega_{n+1}(x)$, obtaining

$$\int_a^b \omega_{n+1}(x)\,\omega_{n+1}(x)\,dx = 0 \quad \text{for } m \geq n+2$$

$$\Rightarrow \omega_{n+1}(x) = 0 \longrightarrow \text{FALSE because based on false assumptions.}$$

PEANO INTEGRATION KERNEL THEOREM

The PEANO KERNEL represents the error we make when integrating a function $g(x) = (x-\theta)_+^k$ for a given θ.

$$K(\theta) = E_x\left((x-\theta)_+^k\right) = \int_a^b (x-\theta)_+^k\,dx - I_n\left((x-\theta)_+^k\right)$$

$$\text{with } (x-\theta)_+^k = \begin{cases} (x-\theta)^k & \text{for } x > \theta \\ 0 & \text{for } x < \theta \end{cases}$$

Since $\int_a^b (x-\theta)_+^k\,dx = \left.\frac{(x-\theta)^{k+1}}{k+1}\right|_{x=b} - \left.\frac{(x-\theta)^{k+1}}{k+1}\right|_{x=a}$ → 0 since $a \leq \theta$;
we have that it doesn't depend on a.

The PEANO KERNEL THEOREM says that given a quadrature formula of degree d and $f \in C^{k+1}([a,b])$, with $0 \leq k \leq d$ then:

$$E(f) = \frac{1}{k!}\int_a^b f^{(k+1)}(\theta)\,K(\theta)\,d\theta$$

From this, we get the error bound:

$$|E(f)| \leq \frac{1}{k!}\,\|k\|_2\,\|f^{(k+1)}\|_2 \longrightarrow \text{Other norm combinations: } 1-\infty,\ \infty-1$$

PROOF

$p(x)$, Taylor exp. of f of order k around a. $\quad$ $r(x)$, from P.K.T.

$$f(x) = \underbrace{\sum_{i=0}^{k} \frac{f^{i}(a)}{i!}(x-a)^i}_{p(x)} + \underbrace{\frac{1}{k!}\int_a^b f^{(k+1)}(\theta)\,K(\theta)\,d\theta}_{r(x)} = p(x) + r(x)$$

cancel out because $p \in \mathbb{P}^d$, $d \geq k$

$$E(f) = \int f - I_n(f) = \int p - I_n(p) + \int r - I_n(r) = \int r - I_n(r)$$

$$\int_a^b r = \int_a^b \frac{1}{k!}\left(\int_a^b f^{(k+1)}(\theta)\,(x-\theta)_+^k\,d\theta\right)dx = \int_a^b f^{(k+1)}(\theta)\left(\int_a^b \frac{(x-\theta)_+^k}{k!}\,dx\right)$$

$$I_n(r) = \int_a^b I\left(f^{(k+1)}(\theta)\frac{(x-\theta)_+^k}{k!}\right)d\theta = \int_a^b f^{(k+1)}(\theta)\,I\left(\frac{(x-\theta)_+^k}{k!}\right)d\theta$$

$$\Rightarrow E(f) = \int_a^b f^{(k+1)}(\theta)\,E_x\left((x-\theta)_+^k\right)\cdot\frac{1}{k!} = \frac{1}{k!}\int_a^b f^{(k+1)}(\theta)\,K(\theta)\,d\theta$$

MORE ON NUMERICAL INTEGRATION

- SIMPSON ADAPTIVE FORMULA uses different steplenghts to compute the composite interpolant of the integral, reducing the nodes needed.
- MONTE CARLO METHODS approximate the integral of f as a function statistical mean. They usually lead to poor results.

LINEAR SYSTEMS

A linear system of order n, $n > 0$, is constituted by a given matrix $A = (a_{ij})_{n \times n}$, a given vector $b = (b_i)$ and an unknown vector $x = (x_i)$ that should be found by solving the system.

$$Ax = b \Rightarrow \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 0, \dots, n$$

The solution exists and is unique iff A is NON-SINGULAR ($\det(A) \neq 0$) for any vector b.

In principle, we can compute the solution using the CRAMER RULE, where A_i is the matrix obtained by replacing the i-th column of A by b, by applying LAPLACE EXTENSION.

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, \dots, n$$

However, this is computationally unfeasible since it requires $\approx 3(n+1)!$ operations. We can reduce the computational cost by applying a method from one of the approaches:

- DIRECT METHODS, yield system solution in finite steps
- ITERATIVE METHODS, require an infinity of steps (theoretically)

A full matrix linear system cannot be solved in principle under n^2 operations, one for each element of the matrix.

DIRECT METHODS

$U = (u_{ij}) \Rightarrow u_{ij} = 0 \ \forall i,j$ s.t. $1 \leq j < i \leq n$, U is UPPER TRIANGULAR

$L = (l_{ij}) \Rightarrow l_{ij} = 0 \ \forall i,j$ s.t. $1 \leq i < j \leq n$, L is LOWER TRIANGULAR

If A is non-singular and triangular, we have that

$$\det(A) = \prod_{i=1}^{n} \lambda_i(A) = \prod_{i=1}^{n} a_{ii} \Rightarrow a_{ii} \neq 0 \ \forall i$$

LU FACTORISATION

Let $A \in \mathbb{R}^{n \times n}$, and L, U respectively lower and upper triangular s.t.:

$$A = LU \quad \leftarrow \text{LU DECOMPOSITION/FACTORISATION OF } A$$

Instead of solving a full linear system, we can solve two triangular systems:

$$Ax = b \longrightarrow \begin{cases} Ly = b \\ Ux = y \end{cases}$$

Since the two systems are triangular they can be solved applying respectively a FORWARD SUBSTITUTIONS ALGORITHM to get y from L, and a BACKWARD SUBSTITUTIONS ALGORITHM to get x from U.

Both require n^2 operations to complete

FORWARD

$$y_1 = \frac{1}{l_{11}} b_1$$

$$y_i = \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i-1} l_{ij} y_j \right), \quad \forall i = 2, \dots, n$$

BACKWARD

$$x_n = \frac{1}{u_{nn}} y_n$$

$$x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j \right)$$

$$\forall i = n-1, \dots, 1$$

Finding the matrices L, U required for this task takes around $\frac{2n^3}{3}$ operations, and is done as follows:

1. The elements of L and U satisfy the nonlinear system
$$\sum_{r=1}^{\min(i,j)} l_{ir} u_{rj} = a_{ij}, \quad i,j = 1, \dots, n$$
2. The system is undetermined having n^2 equations and $n^2 + n$ unknowns. Consequently, LU factorization is not unique.
3. By forcing $l_{ii} = 1$ (all diagonal elements of $L = 1$), we eliminate n unknowns, obtaining a determined system that can be solved using GAUSS ELIMINATION METHOD

GAUSS ELIMINATION METHOD (GEM)

The GEM transforms a system $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ in an equivalent system $Ux = b$, where U is an upper triangular matrix and b is a properly transformed b, which can be solved by backward substitution.

To perform the transformation, we exploit the fact that adding to an equation a linear combination of other equations will not change the solution.

$$l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \quad \text{(find coefficient that will yield 0 when combined)}$$

$$\longrightarrow a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)} \quad \forall k = 1, \dots, n-1, \ \forall i = k+1, \dots, n, \ \forall j = k+1, \dots, n$$

→ Will set to 0 all a_{ij} below pivot

$$\searrow b_i^{(k+1)} = b_i^{(k)} - l_{ik} b_k^{(k)}$$

→ Adopt b to changes accordingly.

The elements on the main diagonal ($a_{kk}^{(k)}$) are called PIVOTS and have to be non-zero. Updating a takes $2(n-k)^2$ operations, updating b takes $2(n-k)$ ops and l_{ik} takes $(n-k)$ → total of $\frac{3n^2 - n}{2}$, plus n^2 to solve $Ux = b \approx \frac{2}{3} n^3$ operations.

Gauss method is equivalent to LU factorization, but the latter proves to be very effective when we are trying to solve many systems having different b's but same A (reducing operations from $\frac{2}{3} n^3$ of Gauss to the simple solving of LU, $2Mn^2$, with M the number of systems.

Some matrices to which GEM can be applied (PIVOTS $\neq 0$):

- (STRICTLY) DIAGONAL DOMINANT BY ROW: $|a_{ii}| \geq \sum_{j=1..n, j \neq i} |a_{ij}|, \ i = 1, \dots n$
- (S.) D.D. BY COLUMN: $|a_{jj}| \geq \sum_{i=1,\dots,n, i \neq j} |a_{ij}|, \ j = 1, \dots, n$
- SYMMETRIC POSITIVE DEFINITE: $\lambda_i(A) > 0, \ i = 0, \dots, n$

These matrices have all in common that all their principal submatrices A_i of order $i = 1, \dots, n-1$ are non-singular

- If $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite, $\exists!\ R$ s.t.

$$A = R^T R$$

- This procedure is known as CHOLESKY FACTORIZATION and requires about $\frac{n^3}{3}$ operations (against $\frac{2n^3}{3}$ of LU).
- Since the L, U matrices are triangular, $l_{ii} = 1$, we can calculate the determinant of $A = LU$ as $O(n^3)$.

$$\det(A) = \det(L)\det(U) = 1 \cdot \det U = \prod_{k=1}^{n} u_{kk}$$

- We can model matrix inversion of A as a linear system where $x^{(k)}$ corresponds to the k-th column of A^{-1} and $i^{(k)}$ to the k-th column of $I \in \mathbb{R}^{n \times n}$. We solve:

$$A x^{(k)} = i^{(k)}$$

obtaining the inverse matrix A^{-1} in $2n^3$ operations.

MEMORY-SPACE LIMITATIONS

A square matrix of order n is called SPARSE if the number of nonzero entries is of order n (on n^2 total entries).

The PATTERN is the 2D representation of nonzero entries positions.
- LOWER BAND p_1: $a_{ij} = 0$ when $i > j + p_1$
- UPPER BAND p_2: $a_{ij} = 0$ when $j > i + p_2$

The maximum between p_1, p_2 is called MATRIX BANDWIDTH

The FILL-IN PHENOMENON occurs when after an LU decomposition L and U present less sparsity than the original A, leading to a bigger memory usage. To reduce the phenomena we can apply row and column permutations to REORDER A before performing the factorization (PIVOTING)

PIVOTING

If a pivot in A becomes 0 GEM fails. To avoid that we can reorder the rows in a way that no pivot is zero (columns, also). This technique is called PIVOTING.

$$PA = LU$$

P is a permutation matrix initially set = I, changed accordingly to permutations made on A.

It is advised to perform pivoting at each step of LU factorisation to use always the pivot with maximum modulus in the $A^{(k)}$ submatrix, using both rows and columns permutations (P and Q)

$$PAQ = LU$$

↳ TOTAL PIVOTING ($\frac{2n^3}{3}$ ops)

Alternatively we can search the max modulus pivot in the same row or column of the current one → PARTIAL PIVOTING (n^2 ops)

By applying partial pivoting to LU factorization, we have to solve

$$Ax = b \Rightarrow PAx = Pb \Rightarrow \begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

While for complete pivoting we have:

$$Ax = b \Rightarrow \underbrace{PAQ}_{LU}\ \underbrace{Q^{-1}x}_{x^*} = Pb \Rightarrow \begin{cases} Ly = Pb \\ Ux^* = y \end{cases} \Rightarrow x = Q x^*$$

PRECISION OF DIRECT METHODS

- Total pivoting is more stable than partial pivoting.

When a linear system is solved numerically we are looking for the exact solution $\hat{x}$ of a perturbed system

$$(A + \delta A)\hat{x} = b + \delta b$$

where δA and δb depend on the method use to approximate the results

We call CONDITIONING of a matrix M (symmetric positive definite) the constant

$$K(M) = \frac{\lambda_{max}(M)}{\lambda_{min}(M)}$$

also called the SPECTRAL CONDITION NUMBER of M.

From the perturbed system formula, we get $x - \hat{x} = -A^{-1}\delta b$, and thus:

$$\|x - \hat{x}\| = \|A^{-1}\delta b\|$$

We can set a bound for the relative error, given previous relations, as:

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq K(A)\frac{\|\delta b\|}{\|b\|} = K(A)\frac{\|r\|}{\|b\|}, \quad \text{with residual } r = b - A\hat{x}$$

The bigger the K the worse the solution provided by a direct method. If $K \approx 1$, the matrix is WELL CONDITIONED. r is an estimator of the error $x - \hat{x}$. If K(A) is small then error is small when $\|r\|$ is small. Vice versa, if K(A) is large we can't use r as measure for the error.

OTHER ~~ITERATIVE~~ DIRECT METHODS

- THOMAS ALGORITHM is used to perform an optimized LU-fact. of a tridiagonal matrix in n operations.
- The solution of an overdetermined system $Ax = b$, $A \in \mathbb{R}^{m \times n}$ $m > n$ can be computed using QR factorization or singular value decomposition.

ITERATIVE METHODS

Solving $Ax = b$ iteratively implies building a series of vector $x^{(k)} \in \mathbb{R}^n$ s.t

$$\lim_{k \to \infty} x^{(k)} = x \quad \forall x^{(0)} \in \mathbb{R}$$

This can be achieved through recursion, as

$$x^{(k+1)} = Bx^{(k)} + g, \text{ s.t. } k \geq 0 \quad B \text{ well chosen depending on } A, \; g \text{ vector satisfying } x = Bx + g$$

B is the ITERATION MATRIX, which helps defining error at step k as:

$$e^{(k)} = x - x^{(k)} = B^k e^{(k-1)}, \text{ with } \lim_{k \to \infty} e^{(k)} = 0 \; \forall e^{(0)}$$

The error goes to 0 iff $\rho(B) < 1 = \max |\lambda_i(B)|$. ρ is called the SPECTRAL RADIUS of B, the max. modulus of its eigenvalues.

$\rho(B) < 1$ is necessary for convergence. The smaller $\rho(B)$ the less iterations are needed to reduce $e^{(0)}$ under a threshold ε; we would require at least k_{min} iterations, where

$$\min(k_{min}) \text{ s.t. } \rho(B)^{k_{min}} \leq \varepsilon$$

PRECONDITIONING makes convergence faster and smoother.

CONSTRUCTING AN ITERATIVE METHOD

We usually split A s.t. $A = P - (P - A)$, where P is an invertible matrix called PRECONDITIONER of A.

$$Ax = b \Rightarrow Px = (P - A)x + b \quad (\text{similar to } x = Bx + g)$$

$$\Rightarrow B = P^{-1}(P - A) = I - P^{-1}A \qquad \Rightarrow g = P^{-1}b$$

We can thus define the RICHARDSON METHOD as:

$$P(x^{(k+1)} - x^{(k)}) = r^{(k)} = b - Ax^{(k)} \quad (r^{(k)} \text{ is the residual at iteration } k)$$

it can also be generalized by adding a parameter α_k before the $r^{(k)}$ which is used to improve the convergence of series $x^{(k)}$. Thus is 'equal to' solve the linear system

$$Pz^{(k)} = r^{(k)}, \text{ with } x^{(k+1)} = x^{(k)} + \alpha_k z^{(k)} \quad \left(P\left(\frac{x^{(k+1)} - x^{(k)}}{\alpha_k}\right) = r^{(k)}, \; \frac{x^{(k+1)} - x^{(k)}}{\alpha_k} = z^{(k)} \right)$$

where $z^{(k)}$ is called the PRECONDITIONED RESIDUAL at step k. P should be either diagonal, triangular or tridiagonal to reduce the number of operations required to compute $z^{(k)}$.

JACOBI METHOD

If, given $A \in \mathbb{R}^{n \times n} = (a_{ij})$, we have that $a_{ii} \neq 0 \; \forall i, \; 0 \leq i \leq n$, we can set

$$P = D = \text{diag}(a_{11}, a_{22}, \dots, a_{nn}) \text{ and } \alpha_k = 1 \; \forall k$$

It can be written under the form $x^{(k+1)} = Bx^{(k)} + g$, with

$$B_J = D^{-1}(D - A) = I - D^{-1}A$$

If the matrix $A \in \mathbb{R}^{n \times n}$ is strictly diagonally dominant by row, then the Jacobi method always converges

GAUSS-SEIDEL METHOD

In order to obtain a faster convergence we can include the newly computed components of vector $x_j^{(k+1)}$, $j = 1, \dots, i-1$ to the previous $x_j^{(k)}$, $j \geq i$, to compute $x_i^{(k+1)}$.

In this case the update is SEQUENTIAL rather than SIMULTANEOUS, but leads to faster convergence. It corresponds to

$$P = D - E \text{ and } \alpha_k = 1, \text{ with } E = \begin{cases} e_{ij} = -a_{ij} & \text{if } i > j \\ e_{ij} = 0 & \text{if } i \leq j \end{cases}$$

↳ lower triangular

Then we have:

$$B_{GS} = (D-E)^{-1}(D-E-A)$$

If A is strictly diagonally dominant by row, Gauss-Seidel converges. If A is symmetric positive definite, then Gauss Seidel converges.

If A is tridiagonal whose diagonals are non null and invertible then Jacobi and Gauss-Seidel are either both divergent or both convergent. If they converge, we have that $\rho(B_{GS}) = \rho(B_J)^2$

RICHARDSON METHOD

If $\alpha_k = \alpha \ \forall k$ the method is called STATIONARY, else it is called DYNAMIC.

If A and P are s.p.d., there are two optimal criteria to choose α:

- STATIONARY CASE: $\alpha_k = \alpha_{opt} = \dfrac{2}{\lambda_{min} + \lambda_{max}}$, $k \geq 0$, $\lambda_{\substack{min \\ max}}$ eig. of $P^{-1}A$

  ↳ If $P = I$, we get the STATIONARY RICHARDSON METHOD:

  $$B = I - \alpha A, \quad g = b \text{ with } \alpha = \frac{2}{\lambda_{min}(A) + \lambda_{max}(A)}$$

- DYNAMIC CASE: $\alpha_k = \dfrac{(z^{(k)})^T r^{(k)}}{(z^{(k)})^T A z^{(k)}}$, $k \geq 0$, where $z^{(k)} = P^{-1} r^{(k)}$

  ↳ If $P = I$, we get the GRADIENT METHOD: $r^{(0)} = b - Ax$, $r^{(k+1)} = r^{(k)} - \alpha_k A r^{(k)}$

  $$B = I - \alpha_k A, \quad g = b \text{ with } \alpha_k = \frac{(r^{(k)})^T r^{(k)}}{(r^{(k)})^T A r^{(k)}}, \quad k \geq 0$$

In both cases, the convergence is s.t.:

$$\|x^{(k)} - x\|_A \leq \left(\frac{K(P^{-1}A) - 1}{K(P^{-1}A) + 1} \right)^k \|x^{(0)} - x\|_A, \quad k \geq 0 \text{ where}$$

$\|v\|_A = \sqrt{v^T A v}$ → ENERGY NORM of A

$K(P^{-1}A)$ condition number

The gradient method converges faster, followed by GS and J. If a is a generic matrix, keeping low both K and number of operations is hard. $(\cdot)^{K^2}$

If A is s.p.d. instead, we have that for the gradient method the optimal α_k is

$$\alpha_k = \frac{(r^{(k)}, z^{(k)})}{(Az^{(k)}, z^{(k)})}, \text{ with } z^{(k)} = P^{-1} r^{(k)}$$

and $K(P^{-1}A) = \frac{\lambda_{max}(P^{-1}A)}{\lambda_{min}(P^{-1}A)}$

CONJUGATE GRADIENT METHOD

When A and P are both s.p.d. we can apply the CONJUGATE GRADIENT METHOD, which converges even faster than the gradient (at most n steps) → DIRECT METHOD

→ Given $x^{(0)}$, $r^{(0)} = Ax^{(0)} + b$, $z^{(0)} = P^{-1} r^{(0)}$, $p^{(0)} = z^{(0)}$, we have

$$\alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}} \Rightarrow x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)} \Rightarrow$$

$$\Rightarrow r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)} \Rightarrow P z^{(k+1)} = r^{(k+1)} \Rightarrow$$

$$\Rightarrow \beta_k = \frac{(Ap^{(k)})^T z^{(k+1)}}{(Ap^{(k)})^T p^{(k)}} \Rightarrow p^{(k+1)} = z^{(k+1)} + \beta_k p^{(k)}$$

The error estimate then becomes:

$$\|e^{(k)}\|_A = \|x^{(k)} - x\|_A \leq \frac{2\epsilon^k}{1+\epsilon^{2k}} \|x^{(0)} - x\|_A, \text{ where } \epsilon = \frac{\sqrt{K(P^{-1}A)} - 1}{\sqrt{K(P^{-1}A)} + 1}$$

CONVERGENCE CRITERIA

As for direct methods, we have that

$$\frac{\|x^{(k)} - x\|}{\|x\|} \leq K(A) \frac{\|r^{(k)}\|}{\|b\|}$$

or, if A is preconditioned:

$$\frac{\|x^{(k)} - x\|}{\|x\|} \leq K(P^{-1}A) \frac{\|P^{-1} r^{(k)}\|}{\|P^{-1} b\|}$$

STOPPING CONDITIONS

As for nonlinear systems, we can choose to control the r or the increment

- $\|r^{(k_{min})}\| \leq \varepsilon \|b\| \Rightarrow \|e^{(k_{min})}\| / \|x\| \leq \varepsilon K(A)$, meaningful only if $K(A)$ is reasonably small.
- $\delta^{(k)} = x^{(k+1)} - x^{(k)} \Rightarrow \|\delta^{(k_{min})}\| \leq \varepsilon$ → better if $P(B) \gg 1$

CHOOSING THE METHOD

The choiche of the method is particularly important for large A, and depends largely on context (A properties, resources.)
Direct methods are usually more effective in absence of a good P, but more sensitive to ill-conditioning and require large storage.

LEAST SQUARES

Having $n+1$ points $x_0, \dots, x_n$ and $n+1$ values $y_0, \dots, y_n$ the interpolating polynomial may show large oscillations for large values of n.

We can instead define a polynomial $\tilde{f}_m(x)$ of degree $m < n$ that approximates the data "at best":

$$\sum_{i=0}^{n} |y_i - \tilde{f}_m(x_i)|^2 \leq \sum_{i=0}^{n} |y_i - p_m(x_i)|^2 \quad \forall p_m(x) \in \mathbb{P}$$

If the values of y_i where those of a function f, then $\tilde{f}_m$ is called the LEAST SQUARES APPROXIMATION of f.

We can determine the coefficients of $\tilde{f}_m$ as:

$$\frac{\partial \Phi}{\partial a_k} = 0, \quad k = 0, \dots, m \qquad \text{with} \quad \tilde{f}_m = a_0 + a_1 x_i + \dots + a_m x_i^m$$
$$\text{and} \quad \Phi = \sum_{i=0}^{n} |y_i - \tilde{f}_m|$$

While $\tilde{f}_m$ is a polynomial we can generalize the formula for functions of a space V_m obtained by linearly combining $m+1$ independent functions. $\{\psi_j, j = 0, \dots, m\}$.

The choiche of ψ is dictated by the conjectured behavior of the function underlying the current data distribution.

$$\tilde{f}(x) = \sum_{j=0}^{m} a_j \psi_j(x) \implies B^T B a = B^T y$$

(a can be obtained by solving)

where $B = b_{ij} = \psi_j(x_i)$, a are the unknown coefficients and y are the data.

EIGENVALUES AND EIGENVECTORS

Given $A \in \mathbb{C}^{n \times n}$ the EIGENVALUE PROBLEM consists in finding a scalar λ and a nonnull vector x s.t.

$$Ax = \lambda x$$

Any such λ is called EIGENVALUE of A, while x is the associated EIGENVECTOR. All multiples αx, $\alpha \neq 0$, are also eigenvectors of λ.

If x is known, we can recover λ using the RAYLEIGH QUOTIENT:

$$\frac{x^H A x}{\|x\|^2} \quad (x^H = \bar{x}^T)$$

The eigenvalues of A are the roots of the characteristic polynomial of A:

$$p_A(\lambda) = \det(A - \lambda I).$$

A $n \times n$ matrix has exactly n eigenvalues (with or without multiplicity). A is diagonalizable if $\exists\, U \in \mathbb{C}^{n \times n}$ s.t. $\det(U) \neq 0$ and

$$U^{-1} A U = \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

The columns of U are the eigenvectors of A.

If A is diagonal or triangular, λ's are its diagonal entries. Otherwise, if A is a general large matrix, seeking the zeros of p_A is hard.

POWER METHOD

If $A \in \mathbb{R}^{n \times n}$ and its eigenvalues are ordered as

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \dots > |\lambda_n|$$

then we can compute λ_1 and x_1 iteratively using the POWER METHOD. Given an arbitrary $x^{(0)} \in \mathbb{C}^n$ and setting $y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$, we can compute for $k = 1, 2, \dots$

$$x^{(k)} = A y^{(k-1)}, \quad y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}, \quad \lambda^{(k)} = (y^{(k)})^H A y^{(k)}$$

$y^{(k)}$ can also be expressed as a power, thus the name:

$$y^{(k)} = \beta^{(k)} A^k x^{(0)} \quad \text{where} \quad \beta^{(k)} = \left(\prod_{i=0}^{k} \|x^{(i)}\| \right)^{-1}$$

This method generate a sequence of unitary $\{y^{(k)}\}$ s.t. for $k \to \infty$ they align in the direction of eigenvector x_1. In all cases, we have $\lambda^{(k)} \to \lambda_1$ for $k \to \infty$.

The stopping condition is $|\lambda^{(k)} - \lambda^{(k-1)}| < \varepsilon |\lambda^{(k)}|$

CONVERGENCE OF POWER METHOD

Since $x_1, \dots, x_n$ are assumed to be linearly independent, they are a basis of $\mathbb{C}^n$. We can thus expand them as

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i x_i, \quad y^{(0)} = \beta^{(0)} \sum_{i=1}^{n} \alpha_i x_i, \text{ with } \beta^{(0)} = \frac{1}{\|x^{(0)}\|} \text{ and } \alpha_i \in \mathbb{C}$$

because $x^{(1)} = A y^{(0)} \| x^{(0)} \|$

At step k we have $y^{(k)} = \beta^{(k)} \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i$, $\beta^{(k)} = \frac{1}{\prod_{i=0}^{k} \|x^{(i)}\|}$

therefore $y^{(k)} = \lambda_1^k \beta^{(k)} \left(\alpha_1 x_1 + \sum_{i=2}^{n} \alpha_i \frac{\lambda_i^k}{\lambda_1^k} x_i \right)$

We see that $y^{(k)}$ tends to align to x_1 since $\frac{\lambda_i}{\lambda_1} < 1 \ \forall i \geq 2$

INVERSE POWER METHOD

As the previous, but if A is nonsingular we can use A^{-1} which eigenvalues are the reciprocal of those of A, to obtain the eigenvalue of A with minimum modulus

$$x^{(k)} = A^{-1} y^{(k-1)}, \quad y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}, \quad \mu^{(k)} = (y^{(k)})^H A^{-1} y^{(k)}$$

$$\Rightarrow \lim_{k \to \infty} \mu^{(k)} = \frac{1}{\lambda_n}$$

We can use LU or Cholesky factorization to compute $x^{(k)}$ in $A x^{(k)} = y^{(k-1)}$.

POWER METHOD WITH SHIFT

If we use $A_\mu = A - \mu I$ (SHIFT) whose eigenvalues are $\lambda(A_\mu) = \lambda(A) - \mu$

$$x^{(k)} = A_\mu^{-1} y^{(k-1)}, \quad y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}, \quad \lambda_\mu^{(k)} = \frac{1}{(y^{(k)})^H A_\mu^{-1} y^{(k)}}$$

The λ closest to μ.

The searched eigenvalue is approximately $\lambda(A) = \lambda_\mu + \mu$

GERSHGORIN CIRCLES - COMPUTING THE SHIFT

Let $A \in \mathbb{C}^{n \times n}$, the GERSHGORIN CIRCLES are $C_i^{(r)}$ (ROW CIRCLE), $C_i^{(c)}$ (COLUMN CIRCLE) associated with i-th row and column such that:

$$C_i^{(r)} = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j \neq i}^{n} |a_{ij}| \right\}$$

$$C_i^{(c)} = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j \neq i}^{n} |a_{ji}| \right\}$$

All eigenvalues of A belong to the region of $\mathbb{C}^n$ defined by the intersection of $C_i^{(r)}$ and $C_i^{(c)} \forall i$ (the union of all row circles and column circles). There is no guarantee that a circle contains eigenvalues, unless it's isolated. The circle provide a guess for the shift. All the eigenvalues of a strictly diagonally dominant matrix are non-null

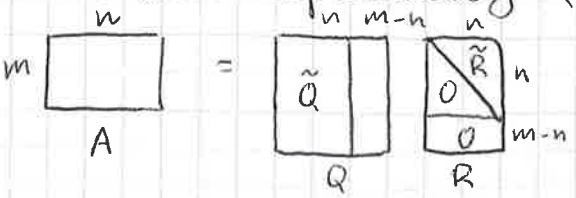
QR METHOD

If A and B are SIMILAR ($P^{-1}AP = B$) then $\lambda_A = \lambda_B$

$$BP^{-1}x = P^{-1}Ax = \lambda P^{-1}x$$

A method to compute all the eigenvalues of A is transforming it in a similar diag/triangular matrix.

The QR method uses repeatedly QR factorization to compute A.



$$\Rightarrow Q^{(k+1)}R^{(k+1)} = A^{(k)} \quad \Rightarrow A^{(k+1)} = R^{(k+1)}Q^{(k+1)}$$

$A^{(k)}$ and $A^{(k+1)}$ are similar and the rate of decay to zero of lower triangular coefficients in $A^{(k)}$ depends on $\max_i |\lambda_{i+1}/\lambda_i|$. If A is symetric, $A^{(k)}$ for $k \to \infty$ is diagonal.

ORDINARY DIFFERENTIAL EQUATIONS

A DIFFERENTIAL EQUATION involves one or more derivatives of an unknown function. If those derivatives are taken w.r.t. a single variable it is called ORDINARY DIFFERENTIAL EQUATION, whereas it is a PARTIAL DIFFERENTIAL EQUATION if partial derivatives are present. The ODE or PDE has order p, where p is the maximum order of differentiation.

Any equation of order $p > 1$ can always be reduced to a system of p equations of order 1.

An ODE admits infinite solution. We formulate a CAUCHY PROBLEM by adding a BOUNDARY CONDITION on initial data to the ODE, ensuring the unicity of the solution.

→ Find $y : I \subset \mathbb{R} \to \mathbb{R}$ s.t.

$$\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \quad \text{(ODE)} \\ y(t_0) = y_0 & \text{(BOUNDARY C.)} \end{cases}$$

- A function is said to be LIPSCHITZ-CONTINUOUS w.r.t. x if $\exists L > 0$ s.t.

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \quad \forall x_1, x_2 \in \mathbb{R}$$

UNIFORMLY Lipschitz-continuous means "on the whole interval". Lipshitz continuity gives more regularity than normal continuity because incremental quotients are bounded (a.k.a. f cannot peak anywhere).

EXISTENCE AND UNICITY (CAUCHY-LIPSCHITZ THEOREM)

If $f(t,y)$ is continuous w.r.t. t and y, and uniformly Lipschitz continuous w.r.t. y, then the solution of the Cauchy problem EXISTS, IS UNIQUE and BELONGS to $C^1(I)$

Solutions of the Cauchy problem are seldom explicit and often cannot be represented even in an implicit form. Numerical methods allows for the approximation of every ODE family for which solutions exist.

The common approach is to divide $I = [t_0, T]$ into N_h intervals of lenght $h = (T - t_0)/N_h$. h is called the DISCRETIZATION STEP. Each $t_n = t_0 + n \cdot h$ is a NODE on which we compute $u_n \approx y_n = y(t_n)$. $\{u_0 = y_0, u_1, \dots, u_{N_h}\}$ is the NUMERICAL SOLUTION of the Cauchy problem.

NUMERICAL DIFFERENTIATION

We aim to approximate, given a function $f : [a,b] \to \mathbb{R}$ continuously differentiable on $[a,b]$, its derivative at a generic $\bar{x} \in [a,b]$ (in case of ODE we call $f \to y$ and $\bar{x} \to t_n$). The derivative $y'(t_n)$ is given by

$$y'(t_n) \begin{cases} = \lim_{h \to 0^+} \dfrac{y(t_n + h) - y(t_n)}{h} \\ = \lim_{h \to 0^+} \dfrac{y(t_n) - y(t_n - h)}{h} \\ = \lim_{h \to 0} \dfrac{y(t_n + h) - y(t_n - h)}{2h} \end{cases}$$

If Dy_n is an approximation of $y'(t_n)$, we then have three possible approaches:

- FORWARD FINITE DIFFERENCE $Dy_n^F = \dfrac{y(t_{n+1}) - y(t_n)}{h}$
- BACKWARD FINITE DIFFERENCE $Dy_n^B = \dfrac{y(t_n) - y(t_{n-1})}{h}$
- CENTERED FINITE DIFFERENCE $Dy_n^C = \dfrac{y(t_{n+1}) - y(t_{n-1})}{2h}$

all for $n = 1, \dots, N_h - 1$, $h = t_{n+1} - t_n = t_n - t_{n-1}$

For both FFD and BFD we have that approximation error is

$$T_n = |y'(t_n) - Dy_n^{F/B}| \le Ch, \text{ where } C = \frac{1}{2} \max_{t \in [t_n, t_{n+1} \text{ or } t_{n-1}]} |y''(t)|$$

while for CFD it is

$$T_n = |y'(t_n) - Dy_n^C| \le Ch^2, \text{ where } C = \frac{1}{6} \max_{t \in [t_{n-1}, t_{n+1}]} |y'''(t)|$$

We call T_n the TRUNCATION ERROR in t_n. T_n is of order $p > 0$ if

$$T_n(h) \le Ch^p, \text{ for } C \ge 0$$

a.k.a. T_n has order 1 for FFD and BFD, and order 2 for CFD

FINITE DIFFERENCE METHOD FOR ODEs

In the Cauchy problem we can approximate the derivative $y'(t_n)$ in t_n using finite differences, obtaining $U_n \approx y(t_n)$

- FORWARD EULER (FE) → EXPLICIT METHOD
$$\begin{cases} \frac{U_{n+1} - U_n}{h} = f(t_n, U_n), \quad n = 0, \dots, N_h - 1 \\ U_0 = y_0 \end{cases}$$
- BACKWARD EULER (BE) → IMPLICIT METHOD
$$\begin{cases} \frac{U_{n+1} - U_n}{h} = f(t_{n+1}, U_{n+1}), \quad n = 0, \dots, N_h \\ U_0 = y_0 \end{cases}$$
- CENTERED EULER (CE)
$$\begin{cases} \frac{U_{n+1} - U_{n-1}}{2h} = f(t_n, U_n) \text{ for } n = 1, 2, \dots, N_h - 1 \\ U_0 = y_0, \quad U_1 \text{ (+. b.d.)} \end{cases}$$

FE is explicit since U_{n+1} depends explicitly on U_n ($U_{n+1} = U_n + h f(t_n, U_n)$) while BE is implicit since U_{n+1} is implicitly defined in terms of U_n ($U_{n+1} = U_n + h f(t_{n+1}, U_{n+1})$). ↑ USE NEWTON OR F.P.I.

FE formula is a simple computation while BE is a nonlinear problem. However BE is generally more stable. Since CE requ U_1 to be applied, is generally preceded by a single pass of FE or BE.

STABILITY (ON UNBOUNDED INTERVALS)

The choiche of h is not arbitrary. If h is not small enough, stability problems may arise

Given the model problem:
$$\begin{cases} y'(t) = \lambda y(t) + 6 \quad (0, \infty) \\ y(0) = 1 \end{cases} \quad \text{where } \lambda < 0 \in \mathbb{R}$$
the exact solution is $y(t) = e^{\lambda t}$ with $y(t) \to 0$ as $t \to \infty$.

The property that $\lim_{n \to \infty} U_n = 0$ is called ABSOLUTE STABILITY

- If we apply FE, we obtain $U_{n+1} = (1 + \lambda h) U_n = (1 + \lambda h)^{n+1}$
If $1 + \lambda h < -1$, then $|U_n| \to \infty$ as $n \to \infty$ $\Rightarrow$ FE is unstable.
We have thus to limit h by imposing $|1 + \lambda h| < 1$ ($h < 2|\lambda|$)

This condition is required on unbounded intervals since N_h (the number of t_n) may $\to \infty$ even if $h \not\to 0$, in order to ensure stability

- If we apply BE to model, we get $U_{n+1} = \left(\frac{1}{1 - \lambda h}\right) U_n = \left(\frac{1}{1 - \lambda h}\right)^{n+1}$
Since $\lim_{n \to \infty} U_n = 0 \ \forall h$, we say that BE is UNCONDITIONALLY STABLE.

ABSOLUTE STABILITY IN PERTURBATION CONTROL

Given a generalized model problem: $\begin{cases} y'(t) = \lambda(t)y(t) + r(t) \\ y(0) = 1 \end{cases}$
on an unbounded interval with λ and r two continuous functions.

If λ and r are constant we get $y(t) = \left(1 + \frac{r}{\lambda}\right) e^{\lambda t} - \frac{r}{\lambda}$ which tends to $-\frac{r}{\lambda}$ as $t \to \infty$, thus a method would not λ-be absolutely stable on it. However it is possible to prove that a method which is absolutely stable on the original model problem keeps perturbations under control even when applied to the generalized model problem as $t \to \infty$.

If we introduce a method to compute z_n which is perturbed by ρ_k at each time step k, representing truncation and numerical errors, we can compute $e_n = |z_n - v_n|$. We find that e_n is bounded by

$$|e_n| \leq \varphi(\lambda)|\rho| \quad \text{where } \varphi(\lambda) = 1 + \left|\frac{z}{\lambda}\right|$$

We also have $\lim_{n\to\infty} |e_n| = \frac{|\rho|}{|\lambda|}$, so the error caused by perturbation doesn't depend neither on n nor h.

e_n is called the PERTURBATION ERROR at step n.

In cases where $\lambda_{min} > 0$ and $\lambda_{max} < \infty$, we can extend the control of perturbation of model problem to normal Cauchy problems if

$$-\lambda_{max} < \frac{\partial f}{\partial y}(t, y) < -\lambda_{min} \quad \forall t \geq 0, \ \forall y \in D_y$$

(possible values of v_n)

In this case, the steplenght h should be chosen as function of $\partial f/\partial y$, depending on the case

→ if h is constant:

$$0 < h < 2 / \max_{t \in [t_0, T]} \left| \frac{\partial f}{\partial y}(t, y(t)) \right|$$

→ if h depends on the step:

$$0 < h_n < 2\frac{\alpha}{|f_y(t_n, v_n)|} \quad \text{for } \alpha < 1$$

CONVERGENCE OF FORWARD EULER

A numerical method is convergent if:

$$\forall n = 0, \dots, N_n \quad |y_n - v_n| \leq C(h), \text{ where } C(h) \to 0 \text{ when } h \to 0$$

Moreover, if $\exists p > 0$ s.t. $C(h) = O(h^p)$ ($\exists c > 0$ s.t. $C(h) \leq ch^p$ for max p), then the method converges WITH ORDER p.

In the case of FE, we have that if $y \in C^2([0,T])$ and f uniformly Lipschitz continuous on y, then

$$|y_{ex}(t_n) - v_n| \leq c(t_n)\, h, \text{ with } \forall n \geq 0$$

where $c(t_n) = \frac{e^{L t_n} - 1}{2L} \max_{t \in [0,T]} |y''(t)|$, with L Lipschitz constant.

The method converges with order $p = 1$

The LOCAL TRUNCATION ERROR of a method represents the error that would be generated by forcing the exact solution to satisfy that specific numerical scheme. For FE we have:

$$T_{n+1}(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - y'(t_n)$$

The GLOBAL TRUNCATION ERROR is $T(h) = \max_n |T_n(h)|$.
For FE this corresponds to:

$$T(h) = \frac{1}{2} \max_{t \in [t_0, T]} |y''(t)|\, h$$

The same results can be applied to BE. If f also satisfies $\frac{\partial f}{\partial y}(t,y) \leq 0 \ \forall t \in [0,T], \forall y \in (-\infty, \infty)$, we have the more precise estimate:

$$|y(t_n) - v_n| \leq h t_n \frac{1}{2} \max_{t \in [0,T]} |y''(t)|$$

CONSISTENCY

Consistency is necessary in order to achieve convergence, since it fullfills the basic assumption that e_n is infinitesimal w.r.t. h. If violated, it would inhibit the global error $\to 0$ when $h \to 0$.

The error follows $O\left(\frac{1}{h}\right)$ when h approaches 0 so it can blow up due to roundoff errors if h is too small.

CRANK-NICOLSON METHOD (TRAPEZOIDAL METHOD)

CN belongs to the family of RUNGE-KUTTA METHODS, which use a single step h but evaluate $f(t,y)$ several times per interval $[t_n, t_{n+1}]$. The number of evaluation at each step is called the ORDER w.r.t. h.

CN is of order 2, obtained by applying the fundamental theorem of integration to the cauchy problem on $[t_n, t_{n+1}]$:

$$\int_{t_n}^{t_{n+1}} y'(t)\,dt = \int_{t_n}^{t_{n+1}} f(t, y(t))\,dt \longrightarrow y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} f(t, y(t))\,dt$$

Then, we use the trapezoidal method to approximate $\int$

$$v_{n+1} - v_n = \frac{h}{2}\left(f(t_n, v_n) + f(t_{n+1}, v_{n+1})\right) \quad \forall n \geq 0$$

The method is unconditionally stable when applied to the model problem and is implicit. Its explicit variant is called Heun method, still of order 2:

$$u_{n+1} - u_n = \frac{h}{2}\left(f(t_n, u_n) + f(t_{n+1}, u_n + h f(t_n, u_n)) \right)$$

IMPROVED EULER METHOD (MIDPOINT METHOD)

If we integrate the ODE but use the midpoint formula instead of the trapezoidal one, we get

$$u_{n+1} - u_n = h f\left(t_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}\right), \text{ where } u_{n+\frac{1}{2}} = u_n + \frac{h}{2} f(t_n, u_n)$$

$$\Rightarrow u_{n+1} - u_n = h f\left(t_{n+\frac{1}{2}}, u_n + \frac{h}{2} f(t_n, u_n)\right)$$

Both Heun and improved Euler require the same conditional stability of FE ($h < \frac{2}{|\lambda|}$).

RUNGE-KUTTA OF ORDER 4 (SIMPSON METHOD)

The RK of O=4 is obtained by approximating $\int$ using the Simpson method

$$u_{n+1} - u_n = \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4), \text{ where } \begin{cases} K_1 = f(t_n, u_n) \\ K_2 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_1) \\ K_3 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_2) \\ K_4 = f(t_{n+1}, u_n + hK_3) \end{cases}$$

It is explicit, still with conditional stability

SYSTEMS OF ODE

Given a system of ODE, each having its initial condition, we can write it in the form

$$\begin{cases} y'(t) = Ay(t) + b(t) & t > 0 \\ y(0) = y_0 & \text{with } A \in \mathbb{R}^{p \times p} \text{ and } b \in \mathbb{R}^p \end{cases}$$

We can apply the same methods as before on the whole system at once. If $b = 0$ and $\lambda_i (\in A) < 0 \ \forall i$, then FE is stable if $h < \frac{2}{\max|\lambda_i|} = \frac{2}{\rho(A)}$. BE stays unconditionally stable.

In the case of a nonlinear problem system of the form $y'(t) = F(t, y(t))$, the stability of explicit methods is

$$h < \frac{2}{\rho(J)}, \text{ where } \rho(J) = \max_i |\lambda_i(J)|, \text{ with } J(t,y) = \frac{\partial F}{\partial y} \text{ for } \lambda_i(J) < 0 \ \forall i$$

J is the JACOBIAN, defined as $J_f = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_n} \\ \vdots & & \\ \frac{\partial f_n}{\partial y_1} & \cdots & \frac{\partial f_n}{\partial y_n} \end{bmatrix}$. The Newton method can be applied on J.

OTHER NOTIONS OF ODEs:

- LOTKA-VOLTERRA EQUATIONS are used to model predator-prey systems in population dynamics. Their form is

$$\frac{dy_1}{dt} = C_1 y_1 (1 - b_1 y_1 - d_2 y_2) \text{ and } \frac{dy_2}{dt} = -C_2 y_2 (1 - b_2 y_2 - d_1 y_1)$$

where C is the growth, d is population interaction and b are nutrients availability.

- ZERO-STABILITY is stability inside a bounded interval. For one-steps methods, this derives from uniform Lipschitz continuity. The LAX-RICHTMEYER EQUIVALENCE THEOREM says that any consistent method is convergent iff it is zero-stable.

- The REGION OF ABSOLUTE STABILITY A is the set of $z (\in \mathbb{C}) = h\lambda$ for which a method is absolutely stable. Methods that are unconditionally absolutely stable are called A-STABLE

- STEP ADAPTIVITY allows to vary time-step h at each time level to match stability constraints and achieve desired accuracy.

- MULTISTEP METHODS achieve higher order of accuracy in general

- Heun method ~~is called~~ belongs to the PREDICTOR-CORRECTOR METHOD family since it requires an explicit step (PREDICTOR) and an implicit one (CORRECTOR) which gives the order of accuracy. Being explicit, they are not adequate on unbounded intervals.

FINITE ELEMENTS AND BOUNDARY-VALUE PROBLEMS

BOUNDARY-VALUE PROBLEMS are differential problems set either in an UNIDIMENSIONAL ($d=1$) or MULTIDIMENSIONAL ($d=2,3$) space for which the value of the unknown solution is given at ENDPOINTS / BOUNDARY (1D / 2-3D).

For the unidimensional case we have a problem set on an interval (a, b) of the real line, where a, b are the endpoints.

$$\begin{cases} -u''(x) = f(x), & x \in [a, b] \\ u(a) = u(b) = 0 \end{cases}$$

For the multidimensional case we have a multidimensional region $\Omega \in \mathbb{R}^d$ instead, with boundary $\partial\Omega$. In this case the differential equation involves the use of PARTIAL DERIVATIVES w.r.t. spatial coordinates.

$$\begin{cases} -\Delta u(x) = f(x), & x \in \Omega \\ u(x) = 0 & x \in \partial\Omega \end{cases}, \quad \text{where } \Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$$

(LAPLACE OPERATOR)

The equation $-u''(x) = f(x)$ and $-\Delta u = f$ are called POISSON EQUATION. Other settings for boundary-value problems are the HEAT and WAVE EQUATIONS. More specifically, a boundary-value problem using the poissan equation with prescribed boundary values is called a DIRICHCET BOUNDARY-VALUE PROBLEM. In this setting, $\exists!\ u \in C^2([a,b])$.

In the NEUMANN PROBLEM instead of the regular boundary conditions of the DIRICHEET PROBLEM we use $u'(a) = \gamma$ and $u'(b) = \delta$ s.t. $\gamma - \delta = \int_a^b f(x)dx$. The equivalent for multidimensional case is prescribing $\frac{\partial u}{\partial n} = \nabla u(x) \cdot n(x) = h(x)$ for $h \in \partial\Omega$, where h is a function s.t. $\int_{\partial\Omega} h = -\int_\Omega f$ and n is the normal direction to the boundary $\partial\Omega$.

We can use either FINITE DIFFERENCES or FINITE ELEMENTS to solve these types of problems, partitioning $[a,b]$ into intervals $I_j = [x_j, x_{j+1}]\ \forall j = 0, \dots, N$ of lenght $h = (b-a)/(N+1)$, where all x_j are called NODES.

FD FOR 1D POISSON PROBLEM

By following the same approach we used to approximate $u'(x)$ through finite differences we apply the Taylor expansion up to the fourth derivative of $u(x+h)$ and $u(x-h)$, and we sum the two (x is x_0 in the formula, while $x \pm h$ is the x).

$$u(x+h) = u(x) + hu'(x) + h^2\frac{u''(x)}{2} + h^3\frac{u'''(x)}{6} + O(h^4)$$

$$+$$

$$u(x-h) = u(x) - hu'(x) + h^2\frac{u''(x)}{2} - h^3\frac{u'''(x)}{6} + O(h^4)$$

$$= u(x+h) + u(x-h) = 2u(x) + h^2 u''(x)$$

$$\Rightarrow u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

This result is valid if $u: [a,b] \to \mathbb{R}$ is sufficiently smooth in a neighborhood of $x \in [a,b]$. The poissan problem thus becomes

$$\begin{cases} \frac{-u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \quad j = 1, \dots, N \\ u_0 = \alpha \text{ and } u_{N+1} = \beta \end{cases}$$

where u_j is an approximation of $u(x_j) = u(x_0 + j \cdot h)$. We can rewrite the problem as a linear system

$$Au_h = h^2 f, \text{ where } u_h = (u_1, \dots, u_N)^T \text{ are the unknowns,}$$

$$f_i = (f(x_1), f(x_2), \dots, f(x_{N-1}), f(x_N)) \text{ and}$$

$$A = \text{tridiag}(-1, 2, -1) \cdot \frac{1}{h^2}$$

Since a is s.p.d., $\exists!$ solution, can be solved with Thanos algo.

For small h (large N), A is ill conditioned since $K(A) = \frac{\lambda_{max}}{\lambda_{min}} = Ch^{-2}$. Thus appropriate methods and preconditioners should be used.

In general, FD requires too much regularity ($v \in C^4$ and $f \in C^2$) so more flexible methods as FINITE ELEMENTS are used.

FINITE ELEMENTS AND THE GALERKIN METHOD

The FINITE ELEMENTS METHOD is an alternative to FD for B-V problems, derived from a reformulation of the poisson problem:

- We multiply both sides of the poisson equation, called STRONG FORMULATION, by a TEST FUNCTION $v \in C^1([a,b])$.
- We integrate the resulting equality on $[a,b]$

$$-\int_a^b u''(x)\, v(x)\, dx = \int_a^b f(x)\, v(x)\, dx \quad \forall v \in V$$

- Using integration by parts, we obtain

$$+\int_a^b u'(x)\, v'(x) - [u'(x)\, v(x)]_a^b = \int_a^b f(x) v(x) dx$$

- By assuming that v vanishes at endpoints, since $v \in V$ follows boundary conditions, we get:

$$\begin{cases} \int_a^b u'(x)\, v'(x)\, dx = \int_a^b f(x)\, v(x)\, dx \quad \forall v \in C^1(a,b) \\ v(a) = v(b) = 0 \end{cases}$$

This last equation is called WEAK FORMULATION of the poisson problem. In this case, both u and v can be less regular than C^1

V is a space of continuous functions and $v'(x) \in V$ is a piecewise continuous function. It is an Hilbert space where the integral of the square is finite (L^2).

To solve the weak formulation, we build $V_h \subset V$ s.t.

$$V_h = \text{span}\{v_i\}_{i=0}^{N_h}, \quad \dim(V_h) = N_h + 1$$

and project the problem in that space, called the FINITE ELEMENTS SPACE of degree 1.

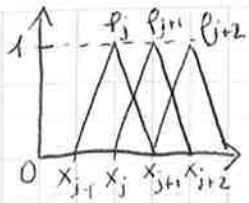
→ Find $u_h \in V_h$ s.t. $u_h(a) = \alpha$, $u_h(b) = \beta$ and:

$$\forall v_h \in V_h^0 \quad \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} u_h'(x)\, v_h'(x)\, dx = \int_a^b f(x) \cdot v_h(x)\, dx$$

Functions in V_h are piecewise polynomial (linear in V_h^0 else of order n) which can be expressed thanks to the function basis φ as:

$$v_h(x) = \sum_{j=1}^{N} v_h(x_j)\, \varphi_j(x) \quad \text{where} \quad \varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & \text{if } x_{j-1} \le x \le x_j \\ \dfrac{x_j - x_{j+1}}{x_j - x_{j+1}} & \text{if } x_j \le x \le x_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

The functions φ_j, $j = 1, \dots, N$ are called SHAPE OR HAT FUNCTIONS and provide a basis for V_h^0.



Shape functions.

Since $|x_{j-1} - x_j| = h$ and the derivative of $\varphi(x)$ corresponds to the slope of the line we have that

$$\varphi_j' = \begin{cases} \frac{1}{h} & x \in [x_{j-1}, x_j] \\ -\frac{1}{h} & x \in (x_j, x_{j+1}] \\ 0 & \text{otherwise} \end{cases}$$

We can rewrite the weak formulation as a system

$A_{FE} u = f$, where $u \to$ vector of unknowns u_j

$f \to$ vectors of $f_i = \int_0^1 f(x) v_{x_i}(x) dx$

$$A_{FE} = \int_0^1 v_j' v_i' dx = \int_0^1 \varphi_j' \varphi_i'$$

→ func $\in V_h$ expressed as basis linear combinations.

Then we have that

$$A_{FE} = \begin{cases} 0 & \text{when } |i-j| \geq 2 \quad \wedge \\ \int_{x_{i-1}}^{x_{i+1}} \varphi_i'(x)^2 dx = \frac{1}{h^2} \int_{x_{i-1}}^{x_{i+1}} dx = \frac{2h}{h^2} = \frac{2}{h} & \text{when } i = j \quad \wedge \\ \int_{x_i}^{x_{i+1}} \varphi_i(x) \varphi_{i-1}(x) = -\frac{1}{h^2} \int_{x_i}^{x_{i+1}} dx = -\frac{2h}{h^2} = -\frac{1}{h} & \text{when } |i-j|=1 \quad \wedge \end{cases}$$

$$\Rightarrow \frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

The matrix is similar to the one of FD, but with $\frac{1}{h}$ instead of $\frac{1}{h^2}$.

The final system has different right-hand side and different solution than the FD one, but have the same accuracy w.r.t. h.

FD works (converges) for $f \in C^2([a,b])$ while FE converges if $\int_a^b f^2(x) dx < \infty$. Using polynomials with $d > 1$ allows for greater convergence, and leads to different matrices

FD FOR 2D POISSON PROBLEM

FD approximate the partial derivatives in PDEs as incremental ratios on a COMPUTATIONAL GRID of finite nodes.

Given our space Ω $(a,b) \times (c,d)$, we partition both intervals in two sets of endpoints Δx and Δy, having cartesian product equal to the grid $\Delta h = \Delta x \times \Delta y$. We look for values $v_{i,j}$ to approximate $v(x_i, y_i)$ on uniformly spaced nodes.

$$\Delta u^h = \delta_x^2 v_{i,j} + \delta_y^2 v_{i,j}, \text{ with } \delta_x^2 v_{i,j} = \frac{v_{i-1,j} - 2v_{i,j} + v_{i+1,j}}{h_x^2}$$

Second order accurate w.r.t. h to replace $\partial^2 v / \partial x^2$ and $\partial^2 v / \partial y^2$ at (x_i, x_j)

$$\delta_y^2 v_{i,j} = \frac{v_{i,j-1} - 2v_{i,j} + v_{i,j+1}}{h_y^2}$$

Replacing it in the 2D Poisson problem, we get:

$$-(\delta_x^2 v_{i,j} + \delta_y^2 v_{i,j}) = f_{i,j} \quad i = 1, \dots, N_x, \; j = 1, \dots, N_y$$

with boundary $v_{i,j} = g_{i,j}$ s.t. $(x_i, y_i) \in \partial\Delta\Omega_h$ ∀ i

If nodes are uniformly spread ($h_x = h_y$) we get:

$$-\frac{1}{h^2}\left(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i,j+1} + u_{i+1,j}\right) = f_{i,j}$$

This scheme is called FIVE POINT SCHEME since it involves five unknown nodal values for Δ. We can adopt the LEXICOGRAPHIC ORDER (left to right, bottom to top) to obtain a tridiagonal matrix form $A \in \mathbb{R}^{n\times n}$

$$A = \text{tridiag}(D, T, D) \text{ with } T = \text{tridiag}\left(-\frac{1}{h_x^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_x^2}\right)$$
$$\text{and } D = \text{diag}\left(-\frac{1}{h_y^2}\right)$$

A is s.p.d. so non-singular, and the system $Au = F$ admits a single solution u_h, which can be found through direct or iterative methods. A is ill-conditioned as for the 1D case: K is of $O(h^{-2})$ as $h \to 0$.

Similarly, we may apply FE by decomposing Ω in polygons called ELEMENTS. φ_k will now look like a pyramid, =1 at k-th vertex and 0 in others.

LAX-MILGRAM THEOREM

Let $V = H_0^1$ (normed and $\in$ scalar product), given $a(\cdot,\cdot): V\times V \to \mathbb{R}$ that is:

- BILINEAR: $\forall v$ $a(v_1, \cdot) \wedge a(\cdot, v_2)$ are linear, aka linear in both variables ($\int_\Omega \nabla v_1 \nabla v_2$) $\to \alpha$ or γ
- CONTINUOUS: $\exists\, c > 0$ s.t. $|a(u,v)| \le C\|u\|\|v\|$ $\forall u, v \in V$
- COERCIVE: $\exists\, \alpha > 0$ s.t. $a(u,u) \ge \alpha\|u\|^2$ $\forall u \in V$

then $\exists!$ u to $a(u,v) = \langle f, v\rangle = f(v)$ $\forall v \in V$, where f is a bounded linear functional $f: V \to \mathbb{R}$ ($\int_\Omega f v\, d\Omega$).

Lax-Milgram is used to prove existence and unicity for both strong and weak formulations of poisson problem

COROLLARY: Solution u is bounded w.r.t. data f

$$\|u\|_V = \frac{1}{\alpha}\|f\|_{V^*}$$

Also, on a finite-dimensional subspace $V_h \subset V$, $V_h = \text{span}\{v_i\}_{i=0}^N$
$\exists!$ solution $u_h \in V_h$ to $(u_h, v) = f(v)$ $\forall v \in V_h$

$\Rightarrow Au = F$, $A_{ij} = a(v_j, v_i)$, $F_i = f(v_i)$

CÉA'S LEMMA $\|u - u_h\| \le \frac{\gamma}{\alpha}\|u - v\|$ $\forall v \in V_h$, a.k.a. u_h is the best approximation of u in V_h up to γ/α.

PROOF 1) $a(u,v) = f(v) = a(u_h, v) \; \forall v \in V_h \Rightarrow a(u - u_h, v) = 0$

2) $\alpha \| u - u_h \|^2 \leq a(u - u_h, u - u_h)$ (coercivity)

$a(u - u_h, u - u_h) = a(u - u_h, u - v) + a(u - u_h, v - u_h)$ (bilinearity)

$= a(u - u_h, u - v)$ since $v - u_h$ for $v = u_h$

$\leq \gamma \| u - u_h \| \| u - v \| \; \forall v \in V_h$ (continuity)

$\Rightarrow \alpha \| u - u_h \|^2 \leq \gamma \| u - u_h \| \| u - v \| \; \forall v \in V_h$

$$\| u - u_h \| \leq \frac{\gamma}{\alpha} \| u - v \| \; \forall v \in V_h$$

→ CONVERGENCE is proven by CÉA'S LEMMA
→ STABILITY is proven by LAX-MILGRAM corollary
→ CONSISTENCY is proven since both sides of the poisson equation vanish when $h \to 0$, thus

$$\lim_{h \to 0} T_h(x_i, y_i) = 0, \quad (x_i, y_i) \in \Delta_h / \partial\Delta_h$$

$\Rightarrow$ The Galerkin method is valid.