University of Trieste

International School for Advanced Studies

The Abdus Salam International Centre for Theoretical Physics

Numerical Analysis

Lectures Notes

Author:
Marco Sciorilli

Gennaio 2021

Abstract

This document contains my notes on the course of Numerical Analysis held by Prof. Luca Heltai and Prof. Gianluigi Rozza for the Master Degree in Data Science and Scientific Computing at SISSA in the year 2020/2021. As they are a work in progress, every correction and suggestion is welcomed. Please, write me at: marco.sciorilli@gmail.com. A special thanks to Gabriele Sarti, as the basis for this work comes from his notes on the same course.

Contents

Chapter 1

Rounding/truncation error, conditional error

1.1 Floating-point representation

Real number representation in $\mathbb{F}$ (the set of the floating-point numbers):

$$x = (-1)^s \cdot (0.a_1 a_2 ... a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t} \quad with\ a \neq 0 \tag{1.1}$$

- s is a sign bit (1 or 0)
- β is the basis adopted by computer (usually it is 2)
- m is the mantissa of length t made by digits a, with $0 \leq a_i \leq \beta - 1$
- 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 commonly denoted as $\mathbb{F}(\beta, t, L, U)$.
The roundoff error occur when we replace $x \neq 0$ with its $\mathbb{F}$ representation, $\hat{x}$, and is defined as

$$\frac{|x - \hat{x}|}{|x|} \leq \frac{1}{2}\epsilon_M \quad with\ \epsilon_M = \beta^{1-t} \tag{1.2}$$

Where ϵ_M is the machine epsilon: the minimal variation representable by a machine

$$\epsilon_M \text{ the largest number } | \quad fl(1 + \epsilon_M) = fl(1) \tag{1.3}$$

Where $fl()$ is the floating-point representation of a number. $\frac{1}{2}\epsilon_M$ is the roundoff unit, $|x - \hat{x}|$ is the absolute error, and $\frac{|x-\hat{x}|}{|x|}$ is the relative error of the approximation operated. The relative error accounts for the order of magnitude of x.
0 is not part of $\mathbb{F}$ and is therefore handled separately. A number exceeding the lower bound is treated as 0 while numbers exceeding the upper bound is treated as inf. $\mathbb{F}$ is not homogeneously dense, but it is denser near 0, and less dense near infinity.
In $\mathbb{F}$ associativity and distributivity are not always respected, as for the case of the loss of significant digits. Indeterminate forms as $\frac{0}{0}$ and $\frac{inf}{inf}$ produces error flagged as NaN.

1.2 Complex numbers

The classic representation of complex number is

$$z = x + iy = \varphi e^{i\theta} = \varphi(cos\theta + isen\theta) \tag{1.4}$$

Where $i = \sqrt{-1}$, $x = Re(z)$, $y(Im)$ and $\varphi = \sqrt{x^2 + y^2}$.
z is a complex number ($\in \mathbb{C}$) with a real part x and an imaginary part y, both represented by two floating-point numbers. Its modulus is φ, and its complex conjugate is

$$\overline{z} = x - iy = \varphi e^{-i\theta} = \varphi(cos\theta - isen\theta) \tag{1.5}$$

The complex conjugate is used in the conjugate transposition of matrices

$$(A_{ij})^* = \overline{A_{ij}} \tag{1.6}$$

1.3 Matrices

Some properties of matrices are

- $A + B = (a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij})$
- $\lambda A = (\lambda a_{ij})$
- $C_{m\times n} = A_{m\times p} B_{p\times n} = (c_{ij}) = \sum_{k=1}^{p} a_{ik} b_{kj}$

If a matrix is diagonal, its determinant is the product of diagonal elements. A matrix is lower/upper triangular if all the 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 symmetrical if $A = A^t$. If $A = A^H = \overline{A}^t$, A is hermitian.

1.4 Vectors

A set of vectors $y_1, ..., y_m$ is linearly independent if

$$a_1 y_1 + ... + a_m y_m = 0 \Leftrightarrow a_1, ..., a_m = 0 \tag{1.7}$$

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

$$w = \sum_{k=1}^{n} a_k y_k \tag{1.8}$$

a_k are unique components of w in relation to B.
The scalar dot product of v and w is defined as

$$(v, w) = w^t v = \sum_{k=1}^{n} a_k b_k \tag{1.9}$$

with a and b respectively components of v and w.
The modulus of a vector v is given by the euclidean norm formula

$$\|v\| = \sqrt{(v,v)} = \sqrt{\sum_{k=1}^{n} v_k^2} \tag{1.10}$$

The vector product (cross product) 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 $A \in \mathbb{C}^{n\times m}$ associated with eigenvalue λ if

$$Av = \lambda v \tag{1.11}$$

The eigenvalues of diagonal and triangular matrices are the elements on the diagonal.
A matrix is said to be positive definite if

$$z^t Az \geq 0 \ \forall \ z \in \mathbb{R}^n \tag{1.12}$$

1.5 Real Functions

If $f(\alpha) = 0$, α is a zero or root of f. It is called simple if $f'(\alpha) \neq 0$, multiple otherwise.
The space $\mathbb{P}_n$ of polynomials of degree $\leq n$ is defined as

$$p_n(x) = \sum_{k=0}^{n} a_k x^k \tag{1.13}$$

with a_k given coefficients.
The number of zeros cannot usually be estimated a priori (except for polynomials, where its $= n$). The value for p_n zeros cannot be computed with an explicit formula for $n \geq 5$.
Foundamental theorem of integration, for f continuous in $[a, b]$

$$F(x) = \int_a^x f(x)dt \ \forall x \in [a,b) \ \Rightarrow \ F'(x) = f(x) \ \forall \ x \in [a,b) \tag{1.14}$$

First mean-value theorem for integrals, for f continuous in $[a, b)$ and $x_1, x_2 \in [a, b)$ with $x_1 < x_2$

$$\exists \xi \in (x_1, x_2) \ s.t. \ f(\xi) = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} f(t)dt \tag{1.15}$$

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

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{1.16}$$

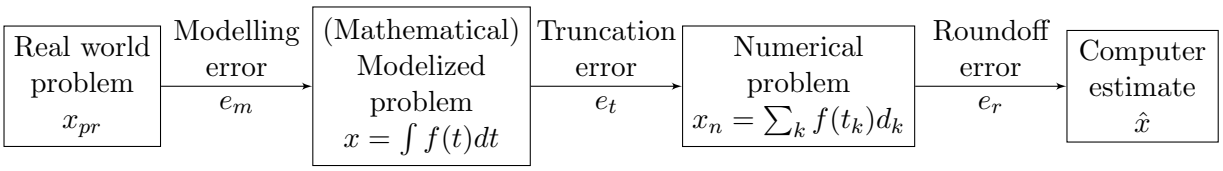
exist and is finite.
Mean value theorem: if $f \in C^\circ([a, b])$ and is differentiable in (a, b)

$$\exists \xi \in (a,b) \ s.t. \ f'(\xi) = \frac{f(b) - f(a)}{b - a} \tag{1.17}$$

Taylor expansion of p_n: if $f \in C^\circ([x_0 - c, x_0 + c])$ (a neighbovarphiod of x_0), f can be approximated in that interval as

$$T_n(x) = f(x_0) + (x - x_0)f'(x_0) + ... + \frac{1}{n!}(x - x_0)^n f^{(n)}(x_0) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{K!} f^{(k)}(x_0) \tag{1.18}$$

1.6 Estimating errors



The sum of truncation error derived from reducing a problem to a finite set of operations and roundoff error coming from a machine representation is called computational error e_c

$$e_c^{abs} = |x - \hat{x}| \quad e_c^{rel} = \frac{|x - \hat{x}|}{|x|} \tag{1.19}$$

To convert a mathematical problem in numerical form we use discretization parameter h, positive.
If $(num) \to (mat)$ as $h \to 0$ the numericl 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 each method. 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 (ops, operations) and can be constant, linear, polynomial, exponential, factorial, ecc.
Numerical approximation can be performed exclusively on well-posed problems, thats to say problems for which the solution:

- Exists
- Is unique
- Depends continuously on data

The total error is:

$$f(x) - \hat{f}(\hat{x}) = \underbrace{\hat{f}(\hat{x}) - f(\hat{x})}_{\substack{\text{computation} \\ \text{error} \\ (e_c = e_t + e_r)}} + \underbrace{f(\hat{x}) - f(x)}_{\substack{\text{propagated data} \\ \text{error} \\ \text{(independent from f)}}} \tag{1.20}$$

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 $\sim \frac{2\epsilon}{h}$ with ϵ =machine precision

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

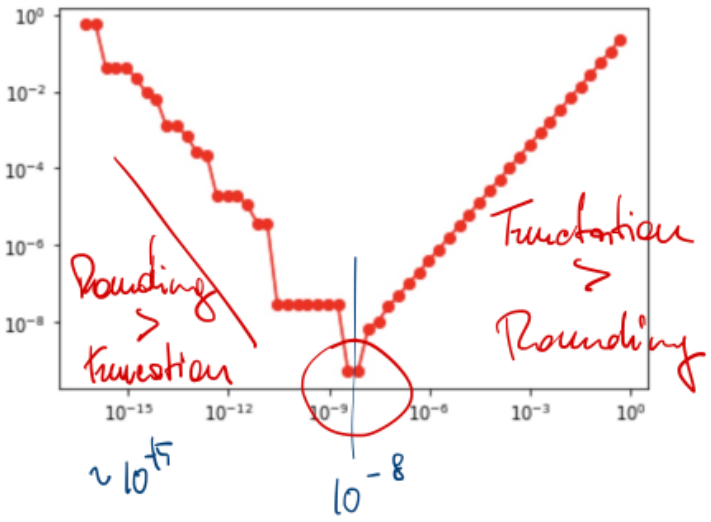


Figure 1.1: Plot of total error vs h. The optimal value of h which minimize the total error is reached when $h \sim \sqrt{\epsilon}$

Problem stability: small changes in input data produce small variation on the output. It is a synonymous of well-posedness.
Given δ a perturbation in data s.t. $d + \delta d \in D$, and $x + \delta x$ the perturbed solution, then

$$\forall d \in D \ \exists \eta(d) \text{ and } K \ s.t. \ \|\delta d\|_d < \eta \in D \Rightarrow \|\delta x\|_x < K \|\delta d\|_d \tag{1.21}$$

Condition numbers: it can be either relative or absolute and measure problem sensitiveness with regards to input data.
Given $f : \mathbb{X} \to \mathbb{Y}$, if we define $\Delta y = \|f(x) - f(\hat{x})\|_{\mathbb{Y}}$ and $\Delta x = \|x - \hat{x}\|_{\mathbb{X}}$, we have that the relative condition numbers is

$$K_{rel} : \sup_{\substack{x,\hat{x}\in\mathbb{X} \\ x-\hat{x}\neq 0 \\ f(x)\neq 0}} \frac{\frac{\Delta y}{\|f(x)\|_{\mathbb{Y}}}}{\frac{\Delta x}{\|x\|_{\mathbb{X}}}} \approx \left|f'(x)\right| \frac{|x|}{|f(x)|} \tag{1.22}$$

$$\leq K_{abs} \sup_{\substack{x,\hat{x}\in\mathbb{X} \\ x-\hat{x}\neq 0 \\ f(x)\neq 0}} \frac{\left\|f^{-1}(f(x))\right\|_{\mathbb{X}}}{\|f(x)\|_{\mathbb{Y}}} = K_{abs} \left\|f^{-1}\right\|_{*} \tag{1.23}$$

And the absolute condition number is:

$$K_{abs} \geq \frac{\Delta y}{\Delta x} \quad (\text{if } f(x) \text{ or } x = 0\ ) \approx \left|f'(x)\right| \quad \Rightarrow K_{abs} : \sup_{\substack{x,\hat{x}\in\mathbb{X} \\ x-\hat{x}\neq 0 \\ f(x)\neq 0}} \frac{\Delta y}{\Delta x} \tag{1.24}$$

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

$$\lim_{n\to\infty} \|y_n - y\| = \lim_{n\to\infty} \|x_n - x\| = 0 \text{ is as } \lim_{n\to\infty} f_n(x) = f(x) \tag{1.25}$$

1.7 Banach Spaces

Given u over $\mathbb{R}$ or $\mathbb{C}$, a seminorm is a function $\|\cdot\| : V \to \mathbb{R}$ which satisfy:

- $\|cu\| = |c|\|u\| \ \forall c \in \mathbb{R}(\text{or } \mathbb{C})$ (homogeneity)
- $\|u + v\| \leq \|u\| + \|v\|$ (triangular inequality)

V is a vector space.
From this two properties, it can be shown that

- $\|u\| \geq 0 \ \forall u \in V$ (non-negativity)

Also, if

- $\|u\| = 0 \Leftrightarrow u = 0$ (positive definite)

is also verified, we have a norm, which is a linear mapping.
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: $(v_1, v_2) \geq 0 \ \ \forall v_1, v_2$ 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.

1.8 Norms

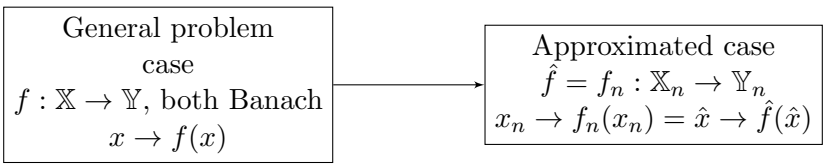
Let's consider three types of norms:

- l_p norms on $\mathbb{R}^n$
  - $\|u\|_p \coloneqq \left(\sum_{i=1}^n |u_i|^p\right)^{\frac{1}{p}}$
  - $\|u\|_\infty \coloneqq \max_i |u_i|$
- L_p norms on $\Omega \subset \mathbb{R}^d$
  - $\|u\|_p \coloneqq \left(\int_\Omega |u|^p\right)^{\frac{1}{p}}$
  - $\|u\|_\infty \coloneqq \sup_{x\in\Omega} |u(x)|$

- $\|\cdot\|_*$ operatorial norms induced by $\|\cdot\|_V$, given $A : V \to W$
  - $\|A\|_* \coloneqq \sup_{0 \neq x \in V} \frac{\|A(x)\|_W}{\|x\|_V}$

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

$$\forall \|\cdot\|_a, \|\cdot\|_b \; \exists 0 \leq c_1 \leq c_2 \; s.t. \; c_1\|x\|_b \leq \|x\|_a \leq c_2\|x\|_b \tag{1.26}$$



1.9 Converge, consistency and Lax-Richtmyer

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

- $\lim_{n\to\infty} \|x_n - x\|_{\mathbb{X}} = 0$
- $\lim_{n\to\infty} \left\| \hat{f}_n(\underbrace{x_n}_{\text{approx.}}) - f(x) \right\|_{\mathbb{X}} = 0$

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

$$\lim_{n\to\infty} \left\| f_n(\underbrace{x}_{\text{exact}}) - f(x) \right\| = 0 \tag{1.27}$$

Example 1. *Sum of two numbers*

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

$$K_{rel} = \frac{|\Delta y|}{|\Delta x|} \cdot \frac{|x|}{|y|} \Rightarrow K_{rel} \leq \frac{|x_1| + |x_2|}{|x_1 + x_2|} \tag{1.28}$$

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, $K_{\text{rel}} = \frac{\int |x|}{|\int x|}$, so it is ill-posed when $x \sim 0$.
- The condition number of a matrix A is $K_{\text{rel}} = \left\|A^{-1}\right\| \|A\|$
  This usually corresponds to

$$K(A) = \frac{|\lambda_{\text{MAX}}(A)|}{|\lambda_{\text{MIN}}(A)|} \tag{1.29}$$

The **Lax-Richtmyer theorem** says that if a problem is consistent, then stability and convergence are equivalent.

- Stability controls perturbation in data and their impact.
- Consistency controls bad approximation 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) goes to 0 as $h \to 0$

Chapter 2

Nonlinear equation

We may want to find the roots of the 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**).

2.1 Bisection method (linear convergence)

It is used to compute the root of a function f on interval $[a, b]$.
Constrains for convergence:

- f should be continuous on $[a, b]$.
- Interval end points 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 length is halved at each step, with $x^{(k)}$ being the midpoint at step k.
The error of estimation at step k is:

$$\left|e^{(k)}\right| = \left|x^{(k)} - \alpha\right| < \frac{1}{2}\left|I^{(k)}\right| = \left(\frac{1}{2}\right)^{k+1}(b-a) \tag{2.1}$$

In order to ensure that the error $\left|e^{(k)} < \epsilon\right|$, we carry out K_{mm} iterations at least:

$$K_{mm} > \log_2\left(\frac{b-a}{\epsilon}\right) - 1 \tag{2.2}$$

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

2.2 Fixed point iterations

Given a function $\phi : [a, b] \to \mathbb{R}$, we want to find an α so that

$$\phi(\alpha) = \alpha \tag{2.3}$$

If α exists, it is called a **fixed point** of ϕ and it could be computed as follows:

$$x^{(k+1)} = \phi(x^{(k)}),\ k \geq 0 \quad \text{with } x^{(0)} \text{ initial guess} \tag{2.4}$$

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

$$\phi_N(x) = x - \frac{f(x)}{f'(x)} \tag{2.5}$$

2.3 Global convergence

1. Iff $\phi(x)$ is continuous in $[a, b]$ and $\phi(x) \in [a, b]\ \forall x \in [a, b]$ then there exists at least one $\alpha \in [a, b]$.
2. Moreover, if $\exists L < 1$ (**Asymptotic convergence factor**) s.t.

$$|\phi(x_1) - \phi(x_2)| \leq L|x_1 - x_2|\ \forall x_1, x_2 \in [a, b] \tag{2.6}$$

   then exists $\alpha \in [a, b]$, unique and the iteration converges to $\alpha\ \forall x^{(0)} \in [a, b]$ (for any initial guess).

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

$$g(a) = \phi(a) - a \geq 0 \quad \text{and} \quad g(b) = \phi(b) - b \leq 0 \tag{2.7}$$

For theorem of zeroes for c functions, we know that g has at least 1 zeros, and thus $\exists\alpha$ for ϕ in $[a, b]$.

2. If two fixed points existed, we would have

$$|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq L|\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2| \tag{2.8}$$

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

$$0 \leq \left|x^{(k+1)} - \alpha\right| = \left|\phi(x^{(k)}) - \phi(\alpha)\right| \leq L\left|x^{(k)} - \alpha\right| \Rightarrow \frac{\left|x^{(k)} - \alpha\right|}{x^{(0)} - \alpha} \leq L^k \tag{2.9}$$

The smaller the L, the faster the convergence.
Since $L < 1$, for $k \to \infty$, we notice

$$\lim_{k\to\infty} \left|x^{(k)} - \alpha\right| \leq \lim_{k\to\infty} L^k = 0 \tag{2.10}$$

Which is convergence.

□

2.4 Convergence

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

$$\left|x^{(k+1)} - \alpha\right| \leq C\left|x^{(k)} - \alpha\right| \tag{2.11}$$

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

$$\left|x^{(k+1)} - \alpha\right| \leq C\left|x^{(k)} - \alpha\right|^2 \tag{2.12}$$

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

$$\left|x^{(k+1)} - \alpha\right| \leq C\left|x^{(k)} - \alpha\right|^p \tag{2.13}$$

2.5 Local convergence (Ostrowski's theorem)

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

$$\left|x^{(0)} - \alpha\right| \leq \delta \quad \forall x^{(0)} \text{ in } [a, b] \tag{2.14}$$

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

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'(\alpha) \tag{2.15}$$

$\forall c$ s.t. $0 < |\phi'(\alpha)| < c < 1$, for large k: $\left|x^{(k+1)} - \alpha\right| \leq c\left|x^{(k)} - \alpha\right|$
If $|\phi'(\alpha) > 1|$, the method diverges. If $|\phi'(\alpha)| = 1$, it depends on the function.

2.6 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 $\phi''(\alpha) \neq 0$, fixed point converges with order 2 and

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{1}{2}\phi''(\alpha) \tag{2.16}$$

2.7 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).
Constrains 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).

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})^2},\ k = 0, 1, ... \tag{2.17}$$

If $f \in \mathcal{C}^2$, we have that $\lim_{k\to\infty} \frac{x^{(k+1)}}{(x^{(k)}-\alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)}$ than we have 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 Newton methods** if m is unknown.

$$x^{(k+1)} = x^{(k)} - m\frac{f(x^(k))}{f'(x^{(k)})}, \quad k = 0, 1, ... \tag{2.18}$$

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

2.8 Stopping criterion

: Control of the increment

$$\left|x^{(k+1)} - x^{(k)}\right| < \epsilon \tag{2.19}$$

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

$$\left|r^{(k_{\min})}\right| = \left|f(x^{(k_{\min})})\right| < \epsilon \tag{2.20}$$

Using fixed point iterations error estimator at step k is

$$\alpha - x^{(k)} = e^{(k)} \approx \frac{1}{(1-\phi'(\alpha))}(x^{(k+1)} - x^{(k)}) \tag{2.21}$$

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 central of the residual as described for newton method.

2.9 Secant method (sublinear convergence)

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

$$x^{(k+1)} = x^{(k)} - \Big(\frac{f(x^{(k)}) - f(x^{k-1})}{x^{(k)} - x^{(k-1)}}\Big)^{-1} f(x^{(k)}) \tag{2.22}$$

Constrains 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 $f \in \mathcal{C}^2(I(\alpha))$, $\exists c > 0$ s.t.

$$\left|x^{(k+1)} - \alpha\right| \leq c\left|x^{(k)} - \alpha\right|^P \quad \text{with } p \approx 1.618 \tag{2.23}$$

Else the method converges linearly.

2.10 Systems of nonlinear equations

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

$$\overline{f}(\overline{x}) = 0 \tag{2.24}$$

We can extend the Newton method to that system by replacing the f' with the Jacobian Matrix $J_{\overline{f}}$, as

$$(J_{\overline{f}})_{ij} = \frac{\partial f_i}{\partial x_j} \quad i, j = 1, ..., n \tag{2.25}$$

The secant method can also be adopted by recursively defining matrices B_k which are suitable approximation of $J_{\overline{f}}(x^0)$ (**Broyden Method**). This belongs to the family of quasi-newton methods.

2.11 Aitken method

If ϕ converges linearly to α, there must be a X s.t. $\phi(x^{(k)}) - \alpha = X(x^{(k)} - \alpha)$. This allows us to obtain a better estimate of $x^{(k+1)}$ than $\phi(x^{(k)})$ (the **Aktken's extrapolation formula, Stefferson's method**)

$$\begin{aligned}
&\alpha = x^{(k)} + \frac{(\phi(x^{(k)} - x^{(k)})}{(1 - \lambda)} \quad \text{with} \\
&\lambda^{(k)} = \frac{\phi(\phi(x^{(k)})) - \phi(x^{(k)})}{\phi(x^{(k)}) - x^{(k)}} \quad \text{given} \\
&\lim_{k \to \infty} \lambda^{(k)} = \phi'(\alpha) \\
&\Rightarrow x^{(k+1)} = x^{(k)} - \frac{(\phi(x^{(k)}) - x^{(k)})^2}{\phi(\phi(x^{(k)})) - 2\phi(x^{(k)}) + x^{(k)}} \quad k \geq 0
\end{aligned}$$

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

- Linear $\phi \to$ Quadratic ϕ_Δ
- $p \geq 2$ $\phi \to 2p - 1$ ϕ_Δ
- Linearly with $m \geq 2$ $\phi \to$ Linearly with $L = 1 - \frac{1}{m}\phi_\Delta$

H may converge even if normal FPI diverges.

2.12 Rope method

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

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{q} \quad k = 0, 1, ... \tag{2.26}$$

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

$$\phi'(x) = \left| 1 - \frac{1}{q} f'(\alpha) \right| < 1 \tag{2.27}$$

Chapter 3

Interpolation

3.1 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 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

$$\left\| f - \tilde{f} \right\| = E(f) = \inf_{\tilde{f} \in M} \left\| f - \tilde{f} \right\| \tag{3.1}$$

- $\tilde{f} \in M$ is the best approximation of the 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), than $\tilde{f}$ is unique (uniqueness theorem).

3.2 Interpolation

Given $n + 1$ points $\{q_i = (x_i, y_i)\}_{i=0}^{n}$ on an interval, we want to find the function φ s.t. $\varphi(x_i) = y_i \forall i$.
We call this function φ **interpolant**, and the point nodes. We say that φ 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 distributions. The iterpolants can be polynomial, trigonometric, rational, ecc.

3.3 Lagrange interpolation (φ is polynomial)

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

$$\pi_n(x_i) = y_i \quad \forall i \text{ (n+1 conditions)} \tag{3.2}$$

If y_i represent the values of a continuous $f(x)$, π_n is the interpolant of $f(x)$. In this setting

$$\forall \pi \in \mathbb{P}^n([a,b]) \ \exists! \ \{\pi_i\}_i^n \in \mathbb{R}^{n+1} \tag{3.3}$$

s.t.

$$\pi(x) = \sum_{i=0}^{n} \pi^i v_i(x) \tag{3.4}$$

$$\pi^i = 0 \quad i = 0, ..., n \Leftrightarrow \pi(x) = 0 \quad \forall x \in [a,b] \tag{3.5}$$

The problem can than be rephrased as

$$\sum_{j}^{n} \pi^j v_j(x_i) = y_i (= f(x_i)) \Leftrightarrow \sum_{j=0}^{n} V_{ij}\pi^j = y_i \tag{3.6}$$

Where $V_{ij} = V_j(x_i)$ is the **Vandermonde matrix** (which is ill conditioned since for n large, rightmost columns will be very similar, making the matrix hardly invertible)

$$V_{ij} = \begin{bmatrix} 1 & q_0 & q_0^2 & ... & q_0^n \\ 1 & q_1 & q_1^2 & ... & q_1^n \\ ... & ... & ... & ... & ... \\ 1 & q_n & q_n^2 & ... & q_n^n \end{bmatrix} = (q_i^s) \tag{3.7}$$

where, following the Einstein notation, we remember that $(V_{ij})^{-1} = V^{ij}$ and

$$V^{ji}V_{ik} = \underbrace{\delta_{jk}}_{\substack{\text{Kronecker} \\ \text{symbol}}} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

So, we obtain

$$\pi^j = V^{ji}y_i \tag{3.9}$$

which in vectorial form is

$$\underline{p} = \mathbb{V}^{-1}\underline{y} \tag{3.10}$$

It is a polynomial that is zero on every point x_k except on x_j where it is 1.
If we consider a reduced version of the problem, we can see how the make it better conditioned.
Let's consider

$$\underbrace{\mathbb{X} = \mathbb{R}^n}_{\text{Set of values of } y \text{ in } x_i} \quad \rightarrow \quad \underbrace{\mathbb{Y} = \mathbb{R}^n}_{\text{Coefficients of the polynomial}} \tag{3.11}$$

with norm $\|y\|_2 \coloneqq (\sum_{i=0}^{n} y_i^2)^{\frac{1}{2}}$.
So

$$\|\pi - \hat{\pi}\|_2 \leq K_{abs}\|y - \hat{y}\|_2 \tag{3.12}$$

$$\|\pi - \hat{\pi}\|_2 \leq \left\|\mathbb{V}^{-1}\right\|_* \|y - \hat{y}\|_2 \tag{3.13}$$

$$\left\|\mathbb{V}^{-1}\right\|_* \coloneqq \max_i \left|\lambda_i(\mathbb{V}^{-1})\right| \tag{3.14}$$

So to make the problem better conditioned, we need to minimize

$$K_{abs} = \left\|\mathbb{V}^{-1}\right\|_2 \to \mathbb{V} = \mathcal{I} \tag{3.15}$$

For this reason, we introduce φ_k, called **Lagrange basis**. It has the following expression (also called Lagrange characteristic polynomials)

$$\varphi_k(x) = \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j} \tag{3.16}$$

We define the Lagrange interpolant of f on nodes $x_0, ..., x_n$ the following linear combination of degree n

$$\mathcal{L}^n f = \sum_{k=0}^{n} f(x_k)\varphi_k(x) \tag{3.17}$$

The Lagrange basis is especially fit for good approximation since other polynomial sets may be ill-conditioned for the approximation task.

3.4 Interpolation error

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

$$E_n f(x) = f(x) - \mathcal{L}^n f(x) = \frac{f^{(n+1)}(\xi_x)w(x)}{(n+1)!}, \quad \text{with } w(x) = \prod_{i=0}^{n} x - x_i \tag{3.18}$$

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

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

$$\|f(x) - \mathcal{L}^n f(x)\|_\infty \leq \frac{\left\|f^{(n+1)}(\xi)\right\|_\infty \|w(x)\|_\infty}{(n+1)!} \tag{3.19}$$

Proof. $\forall x$, we define $G(t)$ s.t.

$$G(t) = (f(t) - \pi(t))w(x) - (f(x) - \pi(x))w(t) \tag{3.20}$$

where

$$\pi(t) = \sum_{i=0}^{n} f(x_i)\varphi_i(t) = (\mathcal{L}f)(t) \tag{3.21}$$

$G(t)$ has $n+2$ zeros:$\{x_i\}_{i=0}^n \cup \{x\}$.
For the Rolle's theorem (if f is continuos and differentiable and $f(a) = f(b)$ then $\exists \xi \in (a,b)$ $s.t.$ $f'(\xi) = 0$)

$$\exists \xi \in (a,b) \ s.t. \ \frac{d^{n+1}G(\xi)}{dt^{n+1}} = 0 \tag{3.22}$$

Then

$$\frac{d^{n+1}}{dt^{n+1}}G(\xi) = f^{(n+1)}(\xi)w(x) - (f(x) - \pi(x))(n+1)! = 0 \tag{3.23}$$

$$\Rightarrow f(x) - (\mathcal{L}f)(x) = \frac{f^{n+1}(\xi)}{(n+1)!}w(x) \tag{3.24}$$

□

Theorem 2. *If f is analytically extendable in an oval $O(a,b,R)$ with $R > 0$, then*

$$\Rightarrow \left\| f^{(n+1)} \right\|_\infty \leq \frac{(n+1)!}{R^{n+1}} \left\| \tilde{f} \right\|_{L^\infty(O(a,b,R))} \tag{3.25}$$

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

$$\left\| f^{(n+1)} \right\| \leq \frac{(n+1)!}{R^{n+1}} \left\| \tilde{f} \right\|_\infty \tag{3.26}$$

$$\|\pi - f\| \leq \left\| f^{(n+1)} \right\|_\infty \frac{\|w\|_\infty}{(n+1)!} \leq \frac{\|w\|_\infty}{R^{(n+1)}} \|f\|_\infty \tag{3.27}$$

$$\|w\|_\infty = \left\| \prod_{i=0}^{n} \right\| (x - x_i) \leq |b-a|^{(n+1} \quad x \in (a,b) \tag{3.28}$$

$$\|\pi - f\|_\infty \leq \Big(\frac{|b-a|}{R} \Big)^{n+1} \|f\|_\infty \tag{3.29}$$

Then increasing the degree n of the interpolator does not guarantee a better approximation of n. Indeed, we may have that

$$\lim_{n \to \infty} \|f - \mathcal{L}^n f\|_\infty = \infty \tag{3.30}$$

3.5 Runge counterexample

$f(x) = \frac{1}{(1+x^2)}$ is interpolated at equispaced nodes in $I = [-s, s]$. 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**.

3.6 Distance from B.A.

If $p = \inf_{x\in\mathbb{P}^n} \|f - x\|_\infty$ with $\{x_i\}_{i=0}^n$ $(n+1)$ nodes, then

$$\begin{aligned}\|f - \mathcal{L}^n f\| =& \|f - p - \mathcal{L}^n(f-p)\| \\ \leq& \|f-p\|_\infty + \|\mathcal{L}^n(f-p)\|_\infty \\ \leq& \|I - \mathcal{L}^n\|_* \|f-p\|_\infty \\ \leq& (\|I\|_* - \|\mathcal{L}^n\|_*)\|f-p\|_\infty \\ \leq& (1 + \|\mathcal{L}^n\|_*)\|f-p\|_\infty\end{aligned}$$

where I is an operator for $f-p$ and $\|\mathcal{L}^n\|_* = \sup_{n\neq 0} \frac{\|\mathcal{L}^n n\|_\infty}{\|n\|_\infty}$.
p is the best approximation of f on nodes $\{x_i\}_{i=0}^n$.

3.7 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

$$\left\|\mathcal{L}^n f - \mathcal{L}^n \hat{f}\right\|_\infty = \Lambda_n(\{x_i\}_{i=0}^n)\left\|f - \hat{f}\right\| \tag{3.31}$$

where

$$\Lambda_n(\{x_i\}_{i=0}^n) = \left\|\sum_{i=0}^n |\varphi_i(x)|\right\|_\infty = \|\Lambda(x)\|_\infty \tag{3.32}$$

is the **Labesgue's constant** depending on interpolation nodes.
From the 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.
We can see

$$\|\mathcal{L}_*^n\| \coloneqq \sup_{v\in C^\circ([a,b])} \frac{\|\sum_{i=0}^n v(x_i)\varphi_i(x)\|}{\|v\|_\infty} \leq \left\|\sum_{i=0}^n |\varphi_i(x)|\right\|_\infty = \Lambda_n(\{x_i\}_{i=0}^n) \tag{3.33}$$

For Lagrange interpolation at equispaced nodes

$$\Lambda_n(x) \simeq \frac{2^{n+1}}{e\cdot n(\log n + \gamma)} \leq \frac{2^{n+1}}{e\cdot n(\log n)}, \quad \text{with } e \approx 2.718 \text{ and } \gamma \approx 0.547 \tag{3.34}$$

For large values of n, this becomes unstable.

3.8 Erdos theorem

$\forall X$, where X is an infinite triangular matrix of interpolation points ($\forall$collection of points $\{x_i\}_{i=0}^n$)

$$X = \begin{bmatrix} 0 & 0 & 0 & ... & 0 \\ x_0 & 0 & 0 & ... & 0 \\ x_0 & x_1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ x_0 & x_1 & x_2 & ... & x_n \end{bmatrix} \tag{3.35}$$

we have

$$\Lambda_n(X) \geq \frac{2}{\pi}\log(n+1) - c \tag{3.36}$$

3.9 Faber theorem

$\forall$ collection of points $\{x_i\}_{i=0}^n$

$$\exists f \in C^0([a,b]) \ s.t. \ \lim_{n\to\infty} \|f - \mathcal{L}^n f\|_\infty \to \infty \tag{3.37}$$

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

3.10 Chebyshev nodes

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

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}\hat{x}_i \tag{3.38}$$

where $\hat{x}_i = -\cos\left(\frac{\pi i}{n}\right),\ i = 0, ..., n.$
Or, equivalently (between $[-1,1]$)

$$x_i = \cos\left(\frac{(2i+1)\pi}{2n+2}\right) \tag{3.39}$$

The nodes are equispaced on the semicirconference of diameter $[a,b]$ and are clustered towards the endpoints of the interval.
For Chebyshev nodes if f is continuously differentiabe in $[a,b]$, $\mathcal{L}^n f$ converges to f as $n \to \infty\ \forall x \in [a,b]$.
For Chebyshev nodes we have $\Lambda_n(X) \leq \frac{2}{\pi}\log(n+1) + 1$.
For equispaced nodes we have $\Lambda_n(X) \leq \frac{2^{n+1}}{c(n\log n)}$.
In this way we get

$$\frac{2}{\pi}\log(n+1) - c \leq \|\Lambda_n\|_\infty \leq \frac{2}{\pi}\log(n+1) + 1 \tag{3.40}$$

3.11 Weierstrass approximation theorem

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

Theorem 3. *Suppose $f \in C^0([a,b])$. Then,*

$$\forall \epsilon > 0\ \exists p \in \mathbb{P}^n \ s.t. \ \|f - p\| < \epsilon,\ \forall x \in [a,b] \tag{3.41}$$

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

3.12 Bernstein coefficients

The $n+1$ Bernstein basis for $\mathbb{P}^n$ are defined as

$$b_{n,i}(x) = \binom{n}{i} x^i (1-x)^{n-i},\ i = 0, ..., n \tag{3.42}$$

$b_{n,i}$ is the $i-th$ polynomial in the Bernstein basis of degree n.
A linear combination of $b_{n,i}$ is called a Bernestein polynomial of degree n based on function f

$$B_n f(x) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big) b_{n,k}(x) \tag{3.43}$$

$\forall x$ this is a weighted average of the $n+1$ values $f(\frac{k}{n})$ called bernstein coefficients.
Properties:

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

The convergence is not pointwise as in interpolation, but uniform

$$\lim_{n\to\infty} \|f(x) - B_n f(x)\|_\infty = 0 \quad \text{with} \quad 0 \leq x \leq 1 \tag{3.44}$$

3.13 Qualitative proof of Weierstrass theorem

$\forall f \in C^\circ(I)$ 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 $f(x_0)$. The same can be done with $q < f$.

$$q^\pm \coloneqq f(x_0) \pm \Big(\frac{\epsilon}{2} + \frac{2\|f\|_\infty}{\delta^2}(x-x_0)^2\Big)$$

with

$|x_1 - x_2| \leq \delta \to |f(x_1) - f(x_2)| \leq \epsilon$

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

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

M depends exclusively on $\|f\|$, ϵ and δ but not on x_0.
By choosing a large N we have $\|f_i - B_n\|_\infty \leq \frac{\epsilon}{6M}$.
Using the triangle inequality we get $\forall x_0,\ \forall n > N\ \|q^\pm - B_n q^\pm\|_\infty \leq \frac{\epsilon}{2}$
We have then in x_0

$$\underbrace{f(x_0) - \epsilon \leq q^-(x_0) - \frac{\epsilon}{2}}_{\text{Definition of } q^-} \underbrace{\leq B_n q^-(x)}_{\text{the last relation}} \underbrace{\leq B_n f(x_0)}_{\text{Positivity}} \tag{3.45}$$

Then

$$\forall x_0,\ \exists N\ s.t.\ \forall n \geq N \quad B_n f \leq B_n q^+ \leq q^+ + \frac{\epsilon}{2} \leq f(x_0) + \epsilon \tag{3.46}$$

Same can be done below, having that

So

$$\rightarrow \|B_n f - f\|_\infty \leq \epsilon \tag{3.47}$$

3.14 More on interpolation

- We can build a piecewise linear interpolant of f to avoid Burge effect when the number of nodes increases. f is a piecewise line or continuous function also called **finite element interpolant**.
- We can perform interpolation bu cubic splines, which are piecewise cubic function $f \in C^2$
- 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.

Chapter 4

Best Approximation in Hilbert spaces

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

$$(a,b),\ a,b \in \mathcal{L}^2([0,1]) = \int_0^1 a \cdot b, \quad \|a\| = \sqrt{\int_0^1 a^2} \tag{4.1}$$

4.1 Best approximation theorem in $\mathcal{L}^2$

Let's assume we have $V \coloneqq span\{v_i\}_{i=0}^n$ and $\|\cdot\|$.
Given a function $f \in \mathcal{L}^2([0,1])$, p is B.A. of f in $[0,1]$ iff

$$(f-p,q) = 0 \quad \forall q \in [0,1],\ \forall f \in \mathcal{L}^2(\mathbb{P}^n) \tag{4.2}$$

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

Proof. Knowing that p is B.A.

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

$$\Rightarrow \|p-f\|^2 \leq \|q-f\|^2 \quad \forall q \in [0,1]$$

□

Or alternatively, we can

Proof. Knowing that $(f-p,q) = 0 \Rightarrow \|p-f\|^2 \leq \|p-f+tq\|^2$ with $t \geq 0$ perturbation,

$q \in \mathbb{P}^n$

$$\begin{aligned}
\left\| \underbrace{p - f + \frac{tg}{2}}_{A} \underbrace{- \frac{tg}{2}}_{-B} \right\|^2 \leq & \left\| \underbrace{p - f + \frac{tg}{2}}_{A} + \underbrace{\frac{tg}{2}}_{+B} \right\|^2 \\
0 \leq & 4\Big(p - f + \frac{tg}{2}, \frac{tg}{2}\Big) \\
0 \leq & t^2 \|q\|^2 + 2t(p - f, q) \\
\Rightarrow (p - f, q) \geq & -\frac{t}{2}\|q\|^2
\end{aligned}$$

By choosing $-q$ instead, we get $(p - f, q) \leq \frac{t}{2}\|q\|^2$.
Thus, it is valid $\forall t, \forall q$ that

$$-\frac{t}{2}\|q\|^2 \leq (p - f, q) \leq \frac{t}{2}\|q\|^2 \tag{4.3}$$

which implies that $(p - f, q) = 0$ since a t can be chosen to bound it on both sides.
Since $(p - f, q) = 0 \ \forall q \Leftrightarrow (p - f, v_i) = 0 \ \forall i = 0, 1, ..., n$ with $\mathbb{P} = span\{v_i\}$

$$\Rightarrow (p, v_i) = (f, v_i) \Rightarrow (\sum_{j=0}^{n} p_j v_j, v_i) = (f, v_i) \tag{4.4}$$

□

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

4.2 Matrix formulation

We can rewrite $(\sum p_j v_j, v_i) = (f, v_i)$ as a matricial relation

$$Mp = F \text{ where } M_{ij} = (v_J, v_i) = \int_0^1 v_j v_i \text{ and } F_i = (f, v_i) = \int_0^1 f v_i \tag{4.5}$$

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

$$M_{ij} = \int_0^1 x^{(J)} x^{(i)} = \frac{1}{i + j + 1} \tag{4.6}$$

The conditional number of the Hilbert matrix is

$$K(M) = O\Big(\frac{(1 + \sqrt{2})}{\sqrt{n}}\Big)^{4n} \tag{4.7}$$

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 w.r.t. $\mathcal{L}^2$.

We want $v_i \in \mathcal{P}^n$ s.t. $M_{ij} = (v_i, v_j) = \delta_{ij}$. To build it, we use the Graham-Schmidt method

$$\begin{cases} v_0 = 1, \quad f \ s.t. \ \int_0^1 f = 1 \\ f^{i+1} = x^{i+1} - \sum_{j=0}(x^{(i+1)}, v_J)v_J \\ v_{i+1} = \frac{f^{i+1}}{\|f^{i+1}\|} \end{cases} \tag{4.8}$$

The first line is set of additive basis having unity as first element. This ensures orthogonality between basis function.
As i (the degree) increases, $v_{i+1} = \frac{f^{i+1}}{\|f^{i+1}\|} \to \infty$ since $x^{i+1} \to \infty$
We can avoid instability by using $v_{i+1} = \frac{f_{i+1}}{f_{i+1}(0)}$ instead.
The points created with Graham-Schmidt represent the Legendre Basis. They make p (best approximation) easy to compute, since M becomes easy to invert and we have a diagonal matrix formed by orthogonal basis

$$p = M^{-1}F \tag{4.9}$$

Chapter 5

Integration(Quadrature)

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

$$I(f) = \int_a^b f(x)dx \tag{5.1}$$

Integration is very expensive from a numeric point of view if f is complicated. Our purpose is to make it simpler, given $f \in C^0([a,b])$.
Many possible approaches called quadratures

- Midpoint formula (degree 1)

$$I_{mp}(f) = (b-a)f\Big(\frac{a+b}{2}\Big) \tag{5.2}$$

- Trapezoidal formula (degree 1)

$$I_t(f) = \frac{(b-a)}{2}(f(a)+f(b)) \tag{5.3}$$

- Simpson formula (degree 3), using $\mathcal{L}^2 f$

$$I_s(f) = \frac{(b-a)}{6}\Big(f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big) \tag{5.4}$$

and their composite variations, using M intervals:

- $I^c_{mp} = H\sum_{k=1}^{M} f(\overline{x}_k)$
- $I^c_t(f) = \frac{H}{2}\sum_{k=1}^{M-1} f(x_k) + \frac{H}{2}(f(a)+f(b))$
- $I^c_s(f) = \frac{H}{6}\sum_{k=1}^{M}(f(x_{k-1}) + 4f(\overline{x}_k) + f(x_k))$
  with $\overline{x}_k = \frac{(x_{k-1}+x_k)}{2}$ and $H = \frac{(b-a)}{M}$

Those are all specific cases of a more general quadrature formula

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(y_i) \tag{5.5}$$

- $\{y_i\}_{i=0}^n$ are quadrature nodes.
- α_i are the quadrature weights.

We can use $\mathcal{L}^n f \in \mathbb{P}^n$ at nodes y_i as approximation function, to get the interpolatory quadrature formula

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

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

Theorem 4. *Interpolatory quadrature rules with $n+1$ points are exact at least for polynomials of order n*

Proof.

$$\forall p \in \mathbb{P}^n, \quad \mathcal{L}^n p = p \tag{5.6}$$
$$\Rightarrow I(p) = I_n(p) \quad \forall p \in \mathbb{P}^n \tag{5.7}$$

□

The degree of accuracy/exactness of a quadrature is the integer r s.t. quadrature using $\mathbb{P}^n$ doesn't produce errors on $I(p)$

$$\max_{M \in \mathbb{N}} I_n(p) = \int p \quad s.t. \quad I(p) = I_n(p) \ \forall p \in \mathbb{P}^r \tag{5.8}$$

Theorem 5. *Given $\{y_i\}_{i=0}^n$ the degree of accuracy is $< 2(n+1)$*

Proof. We construct $w = \prod_{i=0}^n (x - y_i)$, $w \in \mathbb{P}^{n+1}$

$$w^2(x) > 0 \ \forall x \neq y_i, \ w(y_i) = 0 \quad i = 0, .., n \tag{5.9}$$
$$\Rightarrow I(w^2) > 0 \quad I_n(w^2) = 0 \quad \text{by construction} \tag{5.10}$$
$$\exists p = w^2 \ \in \mathbb{P}^{2n+2} \quad s.t. \quad I(p) \neq I_n(p) \tag{5.11}$$

□

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. So we approximate exactly $f \in \mathbb{P}^1$ with a constant function ($\mathbb{P}^0$). We have that

$$\left| \int_a^b f(x)dx - f\Big(\frac{b-a}{2}\Big) \right| \leq \frac{\|f''\|_\infty}{3} \Big(\frac{b-a}{2}\Big)^3 = \frac{\|f''\|_\infty}{24} \quad \text{for } [a, b] = [0, 1] \tag{5.12}$$

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.

5.1 Legendre polynomials and max accuracy

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

$$\int_a^b w_{n+1}(x)p(x)dx = 0 \quad \forall p \in \mathbb{P}_{m-1} \tag{5.13}$$

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

$$L_0(x) = 1, \quad L_1(x) = x, \quad L_{k+1}(x) = \frac{2k+1}{k+1}xL_k(x) - \frac{k}{k+1}L_{k-1}(x) \tag{5.14}$$

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**

$$I_{GL} = \begin{cases} \overline{y}_i = \text{roots of } L_{n+1}(x) \\ \overline{\alpha}_i = \frac{2}{(1-y_i^2)(L'_{n+1}(y_i))^2} \end{cases} \qquad i = 0, \dots, n \tag{5.15}$$

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 $\overline{y}_i, \overline{\alpha}_i$ reconvert to original values for (a, b), use Chebyshev formula.

$$y_i = \frac{a+b}{2} + \frac{b-a}{2}\overline{y}_i, \quad \alpha_i = \frac{b-a}{2}\overline{\alpha}_i \tag{5.16}$$

Theorem 6. *Let $f \in \mathbb{P}^{n+m}$ with $m \leq n+1$.*
Then

$$I_n(f) = T(f) \tag{5.17}$$

iff I_n has degree of accuracy $k = n+m$
Iff

$$\int_a^b w(x)p = 0 \quad \forall p \in \mathbb{P}^{m-1} \quad w = \prod_{i=0}^{n}(x - y_i) \in \mathbb{P}^{n+1} \tag{5.18}$$

Proof. Knowing $f \in \mathbb{P}^{n+m}$, we apply quotient theorem for $\mathbb{P}$ (Ruffini's theorem)

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

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

Assuming that $(*) = 0$, 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 $(*) = 0$

$$I_n(f) = \int f \ \forall f \in \mathbb{P}^{n+m} \rightarrow \int_a^b \underbrace{w(x)p(x)}_{\in \mathbb{P}^{n+m}} dx = I_n(w(x)p(x)) \tag{5.19}$$

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

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

Which is false, because based on false assumption. □

5.2 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.

$$K(\theta) = E_x((x-\theta)_+^k) = \int_a^b (x-\theta)_+^k dx - I_n((x-\theta)_+^k) \tag{5.20}$$

with

$$(x-\theta)_+^k = \begin{cases} (x-\theta)^k & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{5.21}$$

Since

$$\int_a^b (x-\theta)_+^k dx = \frac{(x-\theta)^{k+1}}{k+1}\Big|_{x=b} - \underbrace{\frac{(x-\theta)^{k+1}}{k+1}\Big|_{x=a}}_{0 \text{ since } a \leq \theta} \tag{5.22}$$

we have that it doesn't depend on a.
The Peano kernel theorem says that given a quadrature formula of degree α and $f \in C^{k+1}([a,b])$, with $0 \leq k \leq \alpha$ then

$$|E(f)| \leq \frac{1}{k!}\|k\|_2 \left\| f^{(k+1)} \right\|_2 \tag{5.23}$$

(where other norms combination can be $1-\infty$ and $\infty - 1$)

Proof.

$$f(x) = \underbrace{\sum_{i=0}^{k} \frac{f^i(a)}{i!}(x-a)^i}_{\substack{p(x) \text{ Taylor exp. of } f \\ \text{of order } k \text{ around } a}} + \underbrace{\frac{1}{k!}\int_a^b f^{(k+1)(\theta} K(\theta)d\theta}_{r(x) \text{ from P.K.T.}} = p(x) + r(x)$$

$$E(f) = \int f - I_n(f) = \underbrace{\int p - I_n(p)}_{\text{cancel out because } p \in \mathbb{P}^d, d<k} + \int r - I_n(r) = \int r - I_n(r)$$

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

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

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

□

5.3 More on numerical integration

- Simpson adaptive formula uses different steplenghts to compute the composite interpolant on 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.

Chapter 6

Linear Systems

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

$$A\boldsymbol{x} = \boldsymbol{b} \quad \Rightarrow \sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 0, ..., n \tag{6.1}$$

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 $\boldsymbol{b}$, by applying **Laplace extension**

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, ..., n \tag{6.2}$$

However, this is computationally infeasible since it requires $\approx (n+1)!$ 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 a (theoretically) infinity of steps.

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

6.1 Direct methods

Let's define

$$U = (u_{ij}) \Rightarrow u_{ij} = 0 \ \forall i, j \ s.t. \ 1 \leq j < i \leq n, \ U \text{ is upper triangular}$$
$$L = (l_{ij}) \Rightarrow l_{ij} = 0 \ \forall i, j \ s.t. \ 1 \leq i < j \leq n, \ L \text{ is lower triangular}$$

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

$$\det(A) = \prod_{i=1}^{n} \lambda_i(A) = \prod_{i=1}^{n} a_{ii} \quad \Rightarrow a_{ii} \neq 0 \ \forall i \tag{6.3}$$

6.1.1 LU factorisation

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

$$A = LU \quad \text{LU decomposition/factorisation of } A \tag{6.4}$$

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

$$A\boldsymbol{x} = \boldsymbol{b} \;\Leftrightarrow\; LU\boldsymbol{x} = \boldsymbol{b} \;\Leftrightarrow\; \begin{cases} L\boldsymbol{y} = \boldsymbol{b} \\ U\boldsymbol{x} = \boldsymbol{y} \end{cases} \tag{6.5}$$

Since the two systems are triangular, they can be solved applying respectively a forward substitutions algorithm to get x from U.
Both require n^2 operations to complete.

FORWARD	BACKWARD
$y_1 = \frac{1}{l_{11}} b_1$	$x_n = \frac{1}{u_{nn}} y_n$
$y_i = \frac{1}{l_{ii}}(b_i - \sum_{j=1}^{i-1} l_{ij} y_j),\ \forall i = 2, ..., n$	$x_i = \frac{1}{u_{ii}}(y_i - \sum_{j=i+1}^{n} u_{ij} x_j),\ \forall i = n-1, ..., 1$

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

$$\sum_{r=1}^{\min(i,j)} l_{ir} u_{rj} = a_{ij}, \quad i, j = 1, ..., n \tag{6.6}$$

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.

6.1.2 Gauss elimination method (GEM)

The GEM transforms a system $A\boldsymbol{x} = \boldsymbol{b}$ with $A \in \mathbb{R}^{n\times n}$ in a equivalent system $U\boldsymbol{x} = \hat{\boldsymbol{b}}$, where U is an upper triangular matrix, and $\hat{\boldsymbol{b}}$ is a properly transformed $\boldsymbol{b}$ 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.

$$l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \tag{6.7}$$

We want to find coefficient that will yield 0 when combined.

$$\rightarrow a_{ij}^{(k+1)} = a_{ij}^{(k)} - \underbrace{l_{ik}}_{\substack{\text{Will set to 0 all} \\ a_{ij} \text{ below pivot}}} a_{kj}^{(k)}$$

$$\rightarrow b_i^{(k+1)} = \underbrace{b_i^{(k)}}_{\substack{\text{Adopt } b \text{ to} \\ \text{changes accordingly}}} - l_{ik} b_k^{(k)}$$

$$\forall k = 1, ..., n-1 \quad \forall i = k+1, ..., n \quad \forall j = k+1, ..., n$$

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 $2(n-k)$ operation and l takes $(n-k)$. The total is $\frac{3n^2-n}{2}$, plus n^2 to solve $Ux = \hat{b} \approx \frac{2}{3}n^3$ 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 $\frac{2}{3}Mn^3$ of Gauss to the simple solving of LU, $2Mn^2$, where M is the number of systems).
Same matrices to which GEM can be applied (pivots $\neq 0$):

- (Strictly) diagonal dominant by row: $|a_{ii}| \geq \sum_{\substack{j=1,..,n \\ j \neq i}} |a_{ij}|$ with $i = 1, ..., n$
- (Strictly) Diagonally dominant by column: $\left| a_{jj} \geq \sum_{i=1,...,n i \neq j} \right|$ with $j = 1, ..., n$
- Symmetric positive definite: $\lambda_i(A) > 0$ with $i = 0, ..., n$

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

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

$$A = R^T R \tag{6.8}$$

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

$$\det(A) = \det(L)\det(U) = 1 \cdot \det U = \prod_{k=1}^{n} u_{kk} \tag{6.9}$$

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

$$A\boldsymbol{x}^{(k)} = \boldsymbol{i}^{(k)} \tag{6.10}$$

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

6.1.3 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 phenomenon, we can apply row and column permutations to reorder A before performing the factorisation (pivoting).

6.1.4 Pivoting

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

$$PA = LU \tag{6.11}$$

P is a permutation matrix initially set equal to 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 int he $A^{(k)}$ submatrix, using both rows and columns permutation (P and Q).

$$PAQ = LU \tag{6.12}$$

That's **complete pivoting** and requires $\frac{2n^3}{3}$ operations.
Alternatively we can search the maximum modulus pivot in the same row or column of the current one: **partial pivoting** require n^2 operations.
By applying partial pivoting to LU factorization, we have to solve

$$A\boldsymbol{x} = \boldsymbol{b} \;\Rightarrow PA\boldsymbol{x} = P\boldsymbol{b} \;\Rightarrow \begin{cases} L\boldsymbol{y} = P\boldsymbol{b} \\ U\boldsymbol{x} = \boldsymbol{y} \end{cases} \tag{6.13}$$

While for complete pivoting we have:

$$A\boldsymbol{x} = \boldsymbol{b} \;\Leftrightarrow \underbrace{PAQ}_{LU}\underbrace{Q^{-1}\boldsymbol{x}}_{\boldsymbol{x}^*} = P\boldsymbol{b} \;\Rightarrow \begin{cases} L\boldsymbol{y} = P\boldsymbol{b} \\ U\boldsymbol{x}^* = \boldsymbol{y} \end{cases} \Rightarrow \boldsymbol{x} = Q\boldsymbol{x}^* \tag{6.14}$$

6.1.5 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

$$(A + \delta A)\hat{x} = b + \delta b \tag{6.15}$$

where δA and δb depend on the method used to approximate the results.
We call conditioning of a matrix M (symmetric positive definite) the constant

$$K(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \tag{6.16}$$

also called the spectral condition number of M. From the perturbed system formula, we get $\boldsymbol{x} - \hat{\boldsymbol{x}} = -A^{-1}\delta\boldsymbol{b}$, and thus

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = \left\|A^{-1}\delta\boldsymbol{b}\right\| \tag{6.17}$$

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

$$\frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq K(A)\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} = K(A)\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} \tag{6.18}$$

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

6.1.6 Other direct methods

- **Thomas algorithm** is used to perform an optimised LU-factorization of a triangular matrix in n operations.
- The solution of an overdetermined system $Ax = b$, $A \in \mathbb{R}^{m\times n}$ with $m > n$ can be computed using QR factorization or singular value decomposition.

6.2 Iterative methods

Solving $A\boldsymbol{x} = \boldsymbol{b}$ iteratively implies building a series of vector $\boldsymbol{x}^{(k)\in\mathbb{R}^n}, k \geq 0$ s.t.

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} \quad \forall \boldsymbol{x}^{(0)} \in \mathbb{R} \tag{6.19}$$

This can be achieved through recursion, as

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{g},\ s.t.\ k \geq 0 \tag{6.20}$$

B well chosen depending on A, $\boldsymbol{g}$ vector satisfying $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{g}$.
B is the **iteration matrix**, which helps defining error at step k as

$$\boldsymbol{e}^{(k)} = \boldsymbol{x} - \boldsymbol{x}^{(k)} \tag{6.21}$$

Obtaining with recursion

$$\boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)} \tag{6.22}$$

$\lim_{k\to\infty} \boldsymbol{e}^{(k)} = 0\ \forall \boldsymbol{e}^{(0)}$ (the error goes to 0) iff $\rho(B) < 1 = \max|\lambda_1(B)|$. 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 iteration are needed to reduce $\boldsymbol{e}^{(0)}$ under a threshold ϵ. We require at least $k_{\min}$ iteration, where

$$\min(k_{\min})\ s.t.\ \rho(B)^{k_{\min}} \leq \epsilon \tag{6.23}$$

6.2.1 Constructing an iterative method

We usually split A s.t. $A = P - (P - A)$ where P is an invertible matrix called **preconditioner** of A (preconditioning makes convergence faster and smoother)

$$A\boldsymbol{x} = \boldsymbol{b} \Leftrightarrow P\boldsymbol{x} = (P_A)\boldsymbol{x} + \boldsymbol{b}$$
$$\Rightarrow B = P^{-1}(P - A) = I - P^{-1}A \Rightarrow \boldsymbol{g} = P^{-1}\boldsymbol{b}$$

We can thus define the **Richardson method** as:

$$P(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = \boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} \tag{6.24}$$

where $\boldsymbol{r}^{(k)}$ is the residual at iteration k.
It can also be generalize by adding a parameter α_k before the $\boldsymbol{r}^{(k)}$ which is used to improve the convergence of series $\boldsymbol{x}^{(k)}$. This is equal to solve the linear system

$$Pz^{(k)} = r^{(k)}, \text{ with } x^{(k+1)} = x^{(k)} + \alpha_k z^{(k)} \quad \Big(P\underbrace{\Big(\frac{x^{(k+1)} - x^{(k)}}{\alpha_k}\Big)}_{z^{(k)}} = r^{(k)}\Big) \tag{6.25}$$

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

6.2.2 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

$$P = D = diag(a_{11}, a_{22}, ..., a_{nn}) \text{ and } \alpha_k = 1 \ \forall k \tag{6.26}$$

We deduce then

$$D\boldsymbol{x}^{(k+1)} = -\boldsymbol{b} - (A - D)\boldsymbol{x}^{(k)} \quad k \geq 0 \tag{6.27}$$

By component

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\Big), \quad i = 1, ..., n \tag{6.28}$$

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

$$B_J = D^{-1}(D - A) = I - D^{-1}A \quad \boldsymbol{g} = \boldsymbol{g}_J = D^{-1}\boldsymbol{b} \tag{6.29}$$

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

6.2.3 Gauss-Seidel method

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

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big), \quad i = 1, ..., n \tag{6.30}$$

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

$$P = D - E \text{ and } \alpha_k = 1, \quad \text{with } E \begin{cases} E_{ij} = -a_{ij} & \text{if } i > j \\ E_{ij} = 0 & \text{if } i \leq j \end{cases} \tag{6.31}$$

(E lower triangular)
Then we have

$$B_{GS} = (DE)^{-1}(D - E - A) \qquad \boldsymbol{g}_{GS} = (D - E)^{-1}\boldsymbol{b} \tag{6.32}$$

If A is strictly diagonally dominant by row, Gauss-Seidel converges.
If A is symmetric positive definite, then Gauss-Seidel converges.
If A is triangular whose diagonal are non null and invertible, then Jacobi and Gauss-Seibel are either both divergent or both convergent. If they converge, we have that $\rho(B_{GS}) = \rho(B_J)^2$

6.2.4 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 optional criteria to choose α:

- Stationary case:

$$\alpha_k = \alpha_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}}, \ k \geq 0 \tag{6.33}$$

  λ_{min} eig. of $P^{-1}A$.
  λ_{max}
  If $P = I$, we get the stationary Richardson method:

$$B = I - \alpha A, \ g = b \ \text{ with } \ \alpha = \frac{2}{\lambda_{min}(A) + \lambda_{max}(A)} \tag{6.34}$$

- Dynamic case:

$$\alpha_k = \frac{(\boldsymbol{z}(k))^T \boldsymbol{r}^{(k)}}{(\boldsymbol{z}^{(k)})^T A \boldsymbol{z}^{(k)}}, \quad k \geq 0 \tag{6.35}$$

  where $\boldsymbol{z}^{(k)=P^{-1}\boldsymbol{r}^{(k)}}$ so if $\boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)}$

$$P\boldsymbol{z}^{(k)} = \boldsymbol{r}^{(k)} \tag{6.36}$$
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{z}^{(k)} \tag{6.37}$$
$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A \boldsymbol{z}^{(k)} \tag{6.38}$$

  If $P = I$, we get the gradient method

$$B = I\alpha_k A, \ \boldsymbol{g} = \boldsymbol{b} \ \text{ with } \ \alpha_k = \frac{(\boldsymbol{r}^{(k)})^T \boldsymbol{r}^{(k)}}{(\boldsymbol{r}^{(k)})^T A \boldsymbol{r}^{(k)}}, \ k \geq 0 \tag{6.39}$$

In both cases, the convergence is s.t.

$$\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\right\|_A \leq \left(\frac{K(P^{-1}A)-1}{K(P^{-1})+1}\right)^k \left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}\right\|_A,\ k\geq 0 \ \text{ where } \underbrace{\|\boldsymbol{v}\|_A}_{\substack{\text{energy norm}\\ \text{of } A}} = \sqrt{\boldsymbol{v}^T A\boldsymbol{v}} \text{ and } \underbrace{K(P^{-1}A)}_{\text{condition number}} \tag{6.40}$$

The gradient method converges faster, followed by GS and J. If a is a generic matrix, keeping low both K and the number of operations is hard.

Theorem 7. *If A is s.p.d. instead, we have that for gradient method the optimal α_k is*

$$\alpha_k = \frac{(\boldsymbol{r}^{(k)},\boldsymbol{z}^{(k)})}{(A\boldsymbol{z}^{(k)},\boldsymbol{z}^{(k)})}, \quad \textit{with } \boldsymbol{z}^{(k)} = P^{-1}\boldsymbol{r}^{(k)} \tag{6.41}$$

and

$$K(P^{-1}A) = \frac{\lambda_{max}(P^{-1}A)}{\lambda_{min}(P^{-1}A)} \tag{6.42}$$

6.2.5 Conjugate gradient method

When A and P are both s.p.d., we can apply the conjugate gradient method, which converges even faster then the gradient (at most n steps) $\rightarrow$ direct method.
Given $\boldsymbol{x}^{(0)}$, $\boldsymbol{r}^{(0)} = \boldsymbol{b}A\boldsymbol{x}^{(0)}, \boldsymbol{z}^{(0)} = P^{-1}\boldsymbol{r}^{(0)}, \boldsymbol{p}^{(0)} = \boldsymbol{z}^{(0)}$, we have

$$\begin{aligned}
\alpha_k &= \frac{\boldsymbol{p}^{(k)T}\boldsymbol{r}^{(k)}}{\boldsymbol{p}^{(k)}A\boldsymbol{p}^{(k)}}\\
\boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{p}^{(k)}\\
\boldsymbol{r}^{(k+1)} &= \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{p}^{(k)}\\
P\boldsymbol{z}^{(k+1)} &= \boldsymbol{r}^{(k+1)}\\
\beta_k &= \frac{(A\boldsymbol{p}^{(k)})^T\boldsymbol{z}^{(k+1)}}{(A\boldsymbol{p}^{(k)})^T\boldsymbol{p}^{(k)}}\\
\boldsymbol{p}^{(k+1)} &= \boldsymbol{z}^{(k+1)}\beta_k\boldsymbol{p}^{(k)}
\end{aligned}$$

The error estimate then becomes:

$$\left\|e^{(k)}\right\|_A = \left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\right\|_A \leq \frac{2c^k}{1+c^{2k}}\left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}\right\|_A,\ k\geq 0 \qquad \text{where } c = \frac{\sqrt{K(P^{-1}A)-1}}{\sqrt{K(P^{-1}A)+1}} \tag{6.43}$$

6.2.6 Convergence criteria

As for direct methods, we have that

$$\frac{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\right\|}{\|\boldsymbol{x}\|} \leq K(A)\frac{\left\|\boldsymbol{r}^(k)\right\|}{\|\boldsymbol{b}\|} \tag{6.44}$$

or, if A is preconditioned

$$\frac{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\right\|}{\|\boldsymbol{x}\|} \leq K(P^{-1}A)\frac{\left\|P^{-1}\boldsymbol{r}^{(k)}\right\|}{\|P^{-1}\boldsymbol{b}\|} \tag{6.45}$$

6.2.7 Stopping conditions

- $\left\|r^{(K_{min})}\right\| \leq \epsilon\|b\| \Rightarrow \frac{\left\|e^{(K_{min})}\right\|}{\|x\|} \leq \epsilon K(A)$, which is meaningful only if $K(A)$ is reasonably small.
- $\delta^{(k)} = x^{(k+1)} - x^{(k)} \Rightarrow \left\|\delta^{(K_{min})}\right\| \leq \epsilon$, which is better if $P(B) \gg 1$

6.2.8 Choosing the method

The choice 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.

Chapter 7

Least squares

Having $n+1$ points $x_0, ..., x_n$ and $n+1$ values, $y_0, ..., y_n$, 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":

$$\sum_{i=0}^{n} \left| y_i - \tilde{f}_m(x_i) \right|^2 \leq \sum_{i=0}^{n} |y_i - p_m(x_i)|^2 \quad \forall p_m(x) \in \mathbb{P} \tag{7.1}$$

If the values of y_i were 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:

$$\frac{\partial \phi}{\partial a_k} = 0, \ k = 0, ..., m \quad \text{with } \tilde{f}_m = a_0 + a_1 x_i + ... a_m x_i^m \ \text{ and } \phi = \sum_{i=0}^{n} \left| y_i - \tilde{f}_m \right| \tag{7.2}$$

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 ($\{\psi_i, j = 0, 1, ..., m\}$).
The choice of ψ is dictated by the conjectured behaviour of the function underlying the current data distribution

$$\tilde{f}(x) = \sum_{j=0}^{m} a_j \psi_j(x) \xrightarrow[\substack{a \text{ can be} \\ \text{obtained} \\ \text{by solving}}]{} B^T B a = B^T y \tag{7.3}$$

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

Chapter 8

Eigenvalues and eigenvectors

Given $A \in \mathbb{C}^{n\times n}$, the eigenvalue problem consists in finding a scalar λ and a non-null vector x s.t.

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{8.1}$$

Any such λ is called eigenvalue of A, while $\boldsymbol{x}$ is the associated eigenvector. All multiples $\alpha\boldsymbol{x}, \alpha \neq 0$, are also eigenvectors of λ.
If $\boldsymbol{x}$ is known, we can recover λ using the Rayleigh quotient ($\overline{\boldsymbol{x}}^T = \boldsymbol{x}^H$)

$$\frac{\boldsymbol{x}^H A\boldsymbol{x}}{\|\boldsymbol{x}\|^2} \tag{8.2}$$

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

$$p_A(\lambda) = \det(A - \lambda I) \tag{8.3}$$

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

$$U^{-1}AU = \Lambda = diag(\lambda_1, ..., \lambda_n) \tag{8.4}$$

The columns of U are the eigenvectors of A.
It 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.

8.1 Power method

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

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > ... > |\lambda_n| \tag{8.5}$$

then we can compute λ, and $\boldsymbol{x}$, iteratively using the **power method**.
Given an arbitrary $\boldsymbol{x}^{(0)} \in \mathbb{C}^n$ and setting $\boldsymbol{y}^{(0)} = \frac{\boldsymbol{x}^{(0)}}{\|\boldsymbol{x}^{(0)}\|}$, we can compute for $k = 1, 2, ...$

$$\boldsymbol{x}^{(k)} = A\boldsymbol{y}^{(k-1)}, \quad \boldsymbol{y}^{(k)} = \frac{\boldsymbol{x}^{(k)}}{\|\boldsymbol{x}^{(k)}\|}, \quad \lambda^{(k)} = (\boldsymbol{y}^{(k)})^H A\boldsymbol{y}^{(k)} \tag{8.6}$$

Until $\left|\lambda^{(k)} - \lambda^{(k-1)}\right| < \epsilon\left|\lambda^{(k)}\right|$, where ϵ is the desired tolerance.

8.2 Convergence of power method

Since $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are assumed to be linearly independent, they are a basis of $\mathbb{C}^n$. We can thus expand them as

$$\boldsymbol{x}^{(0)} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \quad \boldsymbol{y}^{(0)} = \beta^{(0)} \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \quad \text{with } \beta^{(0)} = \frac{1}{\left\|\boldsymbol{x}^{(0)}\right\|} \text{ and } \alpha_i \in \mathbb{C} \tag{8.7}$$

At step k we have

$$\boldsymbol{y}^{(k)} = \beta^{(k)} \sum_{i=1}^{n} \alpha_i \underbrace{\lambda_i^k}_{\substack{\text{because} \\ \boldsymbol{x}^{(1)} = A\boldsymbol{y}^{(0)}}} \boldsymbol{x}_i, \qquad \beta^{(k)} = \frac{1}{\prod_{i=0}^{k} \left\|\boldsymbol{x}^{(i)}\right\|} \tag{8.8}$$

therefore

$$\boldsymbol{y}^{(k)} = \lambda_1^k \beta^{(k)} \Big(\alpha_1 \boldsymbol{x}_1 + \sum_{i=2}^{n} \alpha_i \frac{\lambda_i^k}{\lambda_1^k} \boldsymbol{x}_i \Big) \tag{8.9}$$

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

8.3 Inverse power method

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

$$\boldsymbol{x}^{(k)} = A^{-1}\boldsymbol{y}^{(k-1)}, \quad \boldsymbol{y}^{(k)} = \frac{\boldsymbol{x}^{(k)}}{\left\|\boldsymbol{x}^{(k)}\right\|}, \quad \mu^{(k)} = (\boldsymbol{y}^{(k)})^H A^{-1} \boldsymbol{y}^{(k)} \tag{8.10}$$

$$\Rightarrow \lim_{k \to \infty} \mu^{(k)} = \frac{1}{\lambda_n} \tag{8.11}$$

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

8.4 Power method with shift

If we use $A_\mu = A - \underbrace{\mu}_{\text{shift}} I$ whose eigenvalues are $\lambda(A_\mu) = \lambda(A) - \mu$

$$\boldsymbol{x}^{(k)} = A\mu^{-1}\boldsymbol{y}^{(k-1)}, \quad \boldsymbol{y}^{(k)} = \frac{\boldsymbol{x}^{(k)}}{\left\|\boldsymbol{x}^{(k)}\right\|}, \quad \lambda_\mu^{(k)} = \frac{1}{(\boldsymbol{y}^{(k)})^H A^{-1} \boldsymbol{y}^{(k)}} \tag{8.12}$$

The searched eigenvalue is approximately $\underbrace{\lambda(A)}_{\substack{\text{The } \lambda \\ \text{closest to} \\ \mu}} = \lambda_\mu + \mu$

8.5 Gershgorin circles-computing the shift

Let $A \in \mathbb{C}^{n\times n}$, the Gerhgorin circles are $c_i^{(n)}$ (row circle), $c_i^{(c)}$(column circle), associated with $i-th$ row and column such that

$$c_i^{(r)} = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\} \tag{8.13}$$

$$c_i^{(c)} = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}|\} \tag{8.14}$$

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 if it's isolated. The circle provide a guess for the shift. All the eigenvalues of a strictly diagonally dominant matrix are non-null.

8.6 QR method

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

$$BP^{-1}x = P^{-1}Ax = \lambda P^{-1}x \tag{8.15}$$

A method to compute all the eigenvalues of A is transforming it in a similar diagonal/-triangular matrix.
The QR method uses repeatecly QR factorization to compute

$$Q^{(k+1)}R^{(k+1)} = A^{(k)} \Rightarrow A^{(k+1)} = R^{(k+1)}Q^{(k+1)} \tag{8.16}$$

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

Chapter 9

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 equation 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.
We want to find $y : I \subset \mathbb{R} \to \mathbb{R}$ s.t.

$$\begin{cases} y'(t) = f(t, y(y)) \ \forall t \in I & \text{(ODE)} \\ y(t_0) = y_0 & \text{(Boundary c.)} \end{cases} \tag{9.1}$$

A function is said to be **Lipschitz-continuous** w.r.t. x is $\exists L > 0$ s.t.

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \quad \forall x_1, x_2 \in \mathbb{R} \tag{9.2}$$

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

9.1 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 belong to $C^1(I)$.
Solution of the Cauchy problem are seldom explicit and often cannot be represented even in a an implicit form. Numerical methods allow for the approximation of every ODE family for which solutions exist.
The common approach it to divide $I = [t_0, T]$ into N_h intervals of length $h = \frac{(T-t_0)}{N_h}$. h is called the discretization step. Each $t_n = t_0 + n \cdot h$ is a node on which we compute $u_n \approx y_n = y(t_n)$. Lastly, $\{u_0 = y_0, u_1, ..., u_{N_h}\}$ is the numerical solution of the Cauchy problem.

9.2 Numerical differentiation

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

$$y'(t_n) = \begin{cases} = \lim_{h\to 0^+} \frac{y(t_n+h)-y(t_n)}{h} \\ = \lim_{h\to 0^+} \frac{y(t_n)-y(t_n-h)}{h} \\ = \lim_{h\to 0} \frac{y(t_n+h)-u(t_n-h)}{2h} \end{cases} \tag{9.3}$$

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

- Forward finite difference
$$Dy_n^P = \frac{y(t_{n+1}) - y(t_n)}{h} \tag{9.4}$$
- Backward finite difference
$$Dy_n^R = \frac{y(t_n) - y(t_{n-1})}{h} \tag{9.5}$$
- Centered finite difference:
$$Dy_n^C = \frac{y(t_{n+1}) - u(t_{n-1})}{2h} \tag{9.6}$$

all for $n = 1, ..., N_h - 1$, $h = t_{n+1} - t_n = t_n - t_{n-1}$.
For both FFD and BFD we have that approximation error is

$$\tau_n = \left| y'(t_n) - Dy_n^{P/R} \right| \leq Ch, \quad \text{where } C = \frac{1}{2} \max_{t\in[t_n,t_{n+1} \text{ (or } t_{n-1})]} \left| y''(t) \right| \tag{9.7}$$

while for CFD it is

$$\tau_n = \left| y'(t_n) - Dy_n^{C} \right| \leq Ch^2, \quad \text{where } C = \frac{1}{6} \max_{t\in[t_n,t_{n+1}]} \left| y'''(t) \right| \tag{9.8}$$

We call τ_n the truncation error in t_n. τ_n is of order $p > 0$ if

$$\tau_n(h) \leq Ch^p \quad \text{for } C \geq 0 \tag{9.9}$$

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

9.3 Finite difference method for ODEs

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

- Forward Euler(FE): explicit method
$$\begin{cases} \frac{u_{n+1}-u_n}{h} = f(t_n, u_n), \quad n = 0, ..., N_h - 1 \\ u_0 = y_0 \end{cases} \tag{9.10}$$

- Backward Euler(BE): implicit method

$$\begin{cases} \frac{u_{n+1}-u_n}{h} = f(t_{n+1}, u_{n+1}), \quad n = 0, ..., N_h - 1 \\ u_0 = y_0 \end{cases} \tag{9.11}$$

- Centered Euler(CE)

$$\begin{cases} \frac{u_{n+1}-u_n}{2h} = f(t_n, u_n), \quad n = 0, ..., N_h - 1 \\ u_0 = y_0 \\ u_1(t.b.d) \end{cases} \tag{9.12}$$

FE is explicit since u_{n+1} depends explicitly on u_n

$$u_{n+1} = u_n + hf(t_n, u_n) \tag{9.13}$$

while BE is implicit since u_{n+1} is implicitly defined in terms of u_n

$$u_{n+1} = u_n + hf(t_{n+1}, u_{n+1}) \tag{9.14}$$

FE formula is a simple computation, while BE is a nonlinear (use Newton or F.P.I) problem. However BE is generally more stable. Since CE require u_1 to be applied, it generally preceded by a single pass of FE or BE.

9.4 Stability (on unbounded intervals)

The choice of h is not arbitrary. If h is not small enough, stability problems may arise. Given the model problem

$$\begin{cases} y'(t) = \lambda y(t) & t \in (0, \infty) \\ y(0) = 1 & \text{where } \lambda < 0 \in \mathbb{R} \end{cases} \tag{9.15}$$

the exact solution is $y(t) = e^{\lambda t}$ with $\lim y(t)_{t\to\infty} = 0$.
We say that that a numerical scheme associated to the model problem is **absolute stability** if $\lim_{n\to\infty} u_n = 0$.
If we apply FE, we obtain

$$u_{n+1} = (1 + \lambda h) \qquad \text{where } u_n = (1 + \lambda h)^{n+1} \tag{9.16}$$

. If $1 + \lambda h < -1$, then $|u_n| \to \infty$ as $n \to \infty$, so FE is unstable. We have thus to limit h by imposing

$$|1 + \lambda h| < 1 \qquad \text{hence } h < 2/|h| \tag{9.17}$$

called **stability condition**.
This condition is required on unbounded intervals since N_h (the number of t_n) may $\to \infty$ even if $h \nrightarrow 0$, in order to ensure stability.
If we apply BE to model, we get

$$u_{n+1} = \left(\frac{1}{1 - \lambda h}\right) \qquad \text{and therefore } u_n = \left(\frac{1}{1 - \lambda h}\right)^{n+1} \tag{9.18}$$

. Since $\lim_{n\to\infty} u_n = 0 \ \forall h$, we say that BE is unconditionally stable.

9.5 Absolute stability in perturbation control

Given a generalized model problem

$$\begin{cases} y'(t) = \lambda(t)y(t) + r(t) \\ y(0) = 1 \end{cases} \tag{9.19}$$

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 be absolutely stable on it. However it is possible to prove that a method which is absolutely stable on the original mode problem keeps perturbations under control even when applied to the generalized 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 $e_n = |z_n - u_n|$.Supposing that $h < h_0(\lambda) = 2/|\lambda|$, i.e. h ensure the absolute stability of FE applied to the problem. Therefore $(1 + h\lambda)^n < 1 \ \forall n$, and we find that e_n is bounded by

$$|e_n| \leq \varphi(\lambda)|\rho| \quad \text{where } \varphi(\lambda) = 1 + \left|\frac{2}{\lambda}\right| \tag{9.20}$$

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

$$-\lambda_{max} < \frac{\partial f}{\partial y(t,y)} < -\lambda_{min} \quad \forall t \geq 0, \forall y \in D_y \tag{9.21}$$

In this case, the steplength h should be chosen as function of $\frac{\partial f}{\partial y}$, depending on the case

- If h is constant
$$0 < h < 2 \max_{t\in[t_0,T]} \frac{\partial f}{\partial y}(t, y(t)) \tag{9.22}$$
- if h depends on the step
$$0 < h_n < 2\frac{\alpha}{|f_y(t_n, u_n)|} \quad \text{for } \alpha < 1 \tag{9.23}$$

Where D_y is a set that contains the trajectory of $y(t)$

9.6 Convergence of forward Euler

A numerical method is convergent if

$$\forall n = 0, .., N_h \quad |u_n - y(t_n)| \leq C(h) \tag{9.24}$$

where $C(h) \to 0$ when $h \to 0$.
Moreover, if $\exists p > 0$ s.t. $C(h) = O(h) = O(h^P)$ ($\exists c > 0$ s.t. $C(h) \leq ch^p$ for $\max p$), then the method converges with order p.

Theorem 8. *In the case of FE, we have that if $y \in C^2([0,T])$ and f uniformly Lipschitz continuous on y, then*

$$|y(t_n) - u_n| \leq c(t_n)h, \quad \forall n \geq 0 \tag{9.25}$$

where

$$c(t_n) = \frac{e^{Lt_n} - 1}{2L} \max_{t\in[0,T]} \left|y''(t)\right| \tag{9.26}$$

with L Lipschitz constant.
The method converges with order $p = 1$.

The local truncation error of the method represents the error that would be generated by forcing the exact solution to satisfy that specific numerical scheme. For FE we have

$$\tau_{n+1}(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - y'(t_n) \tag{9.27}$$

The global truncation error is $\tau(h) = \max_n |\tau_n(h)|$. We know there exists $\xi_n \in (t_n, t_n + h)$ such that

$$\tau_{n+1}(h) = \frac{1}{2}y''(\xi_n)h \tag{9.28}$$

For FE this corresponds to

$$\tau(h)l \leq \frac{1}{2} \max_{t\in[t_0,T]} \left|y''(t)\right|h \tag{9.29}$$

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

$$|y(t_n) - u_n| \leq ht_n \frac{1}{2} \max_{t\in[0,T]} \left|y''(t)\right| \tag{9.30}$$

9.7 Consistency

Consistency is necessary in order to achieve convergence, since it fulfills the basic assumption that e_n in 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 round-off errors if h is too small.

9.8 Crank-Nicholson 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 $[t_n, t_{n+1}]$

$$\int_{t_n}^{t_{n+1}} y'(t)dt = \int_{t_n}^{t_{n+1}} f(t, y(t) = dt \to y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{9.31}$$

Then, we use the trapezoidal method to approximate the integral

$$u_{n+1} - u_n = \frac{h}{2}(f(t_n, u_n) + f(t_{n+1}, u_{n+1})) \quad \forall n \geq 0 \tag{9.32}$$

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

$$u_{n+1} - u_n = \frac{h}{2}(f(t_n, u_n) + f(t_{n+1}, u_n + hf(t_n, u_n))) \tag{9.33}$$

9.9 Improved Euler method (midpoint method)

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

$$\begin{aligned} &u_{n+1} - u_n = hf(t_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}) \quad \text{where } u_{n+\frac{1}{2}} = u^n + \frac{h}{2}f(t_n, u_n) \\ &\Rightarrow u_{n+1} - u_n = hf(t_{n+\frac{1}{2}}, u_n + \frac{h}{2}f(t_n, u_n)) \end{aligned}$$

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

9.10 Runge-Kutta of order 4 (Simpson method)

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

$$u_{n+1} - u_n = \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \quad \text{where } \begin{cases} k_1 = f(t_n, u_n) \\ k_2 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}k_1) \\ k_3 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}k_2) \\ k_4 = f(t_{n+1}, u_n + hk_3) \end{cases} \tag{9.34}$$

It is explicit, still with conditional stability.

9.11 Systems of ODEs

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

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t) & t > 0 \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases} \tag{9.35}$$

with $A \in \mathbb{R}^{p \times p}$ and $\boldsymbol{b} \in \mathbb{R}^p$, and A has got p distinct eigenvalues.
We can apply the same methods as before on the whole system at once. If $\boldsymbol{b} = 0$ 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

$$h < \frac{2}{p(J)}, \quad \text{where } p(J) = \max_i |\lambda_i(J)| \tag{9.36}$$

with $J(t,y) = \frac{\partial F}{\partial y}$ for $\lambda_j(J) < 0$.
J is the jacobian, defined as

$$J_f = \begin{bmatrix} \frac{\partial f_i}{y_i} & \dots & \frac{\partial f_i}{\partial y_n} \\ \dots & \dots & \dots \\ \frac{\partial f_n}{\partial y_1} & \dots & \frac{\partial f_n}{\partial y_n} \end{bmatrix} \tag{9.37}$$

The Newton method can be applied on J.

9.12 Other notions of ODEs

- Lotka-Volterra equations are used to model predator-pray systems in population dynamics. Their form is

$$\frac{dy_1}{dt} = C_1 y_1(1 - b_1 y_1 - d_2 y_y) \quad \text{and} \quad \frac{dy_2}{dt} = -C_2 y_2(1 - b_2 y_2 - d_1 y_1) \tag{9.38}$$

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

- Zero-Stability is stability inside a bounded interval. For one-step 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 constrains and achieve desired accuracy.

- Multistep methods achieve higher order accuracy in general.

- Heun method 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.

Chapter 10

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

$$\begin{cases} -u''(x) = f(x) & x \in [a, b] \\ u(a) = u(b) = 0 \end{cases} \tag{10.1}$$

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

$$\begin{cases} -\Delta u(x) = f(x) & x \in \Omega \\ u(x) = 0 & x \in \partial\Omega \end{cases} \quad \text{where} \quad \underbrace{\Delta u}_{\substack{\text{Laplace} \\ \text{operator}}} = \sum_{i=1}^{\alpha} \frac{\partial^2 u}{\partial x_i^2} \tag{10.2}$$

The equation $-u''(x) = f(x)$ and $-\Delta u = f$ are called **Poisson equation**.
Other settings for boundary-value problem are the heat and wave equations. More specifically, a boundary-value problem using the Poisson equation with prescribed boundary values is called **Dirichlet boundary-value problem**. In this settings $\exists! u \in C^2([a, b])$.
In the **Neumann problem**, instead of the regular boundary conditions of the Dirichlet problem, we use $u'(a) = \gamma$ and $u'(b) = \lambda$ s.t. $\gamma - \lambda = \int_a^b f(x)dx$. The equivalent for multidimensional case is prescribing $\frac{\partial u}{\partial n} = \nabla u(x)x)$ 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 $I_j = [x_j, x_{j-1}]$ $\forall j = 0, ..., N$ of length $h = \frac{(b-a)}{(N+1)}$ where all x_j are called nodes.

10.1 Finite differences 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).

$$
\begin{aligned}
&u(x+h) = u(x) + hu'(x) + h^2\frac{u''(x)}{2} + h^3\frac{u'''(x)}{6} + O(h^4) \\
&\quad + \\
&u(x-h) = u(x) - hu'(x) + h^2\frac{u''(x)}{2} - h^3\frac{u'''(x)}{6} + O(h^4) \\
&\quad = \\
&u(x+h) + u(x-h) = 2u(x) + h^2u''(x) \\
&\Rightarrow \frac{u''''(\xi)}{12}h^2 + u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}
\end{aligned} \tag{10.3}
$$

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

$$
\begin{cases}
-\frac{u_{j+1}-2u_j+u_{j-1}}{h^2} = f(x_j) \quad j = 1, ..., N \\
u_0 = \alpha \\
u_{N+1} = \beta
\end{cases} \tag{10.4}
$$

where u_j is an approximation of $u(x_j) = u(x_0 + j \cdot h)$, $h = \frac{1}{(N-1)}$. We can rewrite the problem as a linear system

$$
A_{FD}\boldsymbol{u} = h^2\boldsymbol{f} \tag{10.5}
$$

where $\boldsymbol{u} = (u_1, ..., u_n)^T$ are unknowns, $f_i = (f(x_1), f(x_2), ..., f(x_{N-1}), f(x_N)$ and $A = tridiag(-1, 2, -1) \cdot \frac{1}{h^2}$.
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 ($u \in C^2$ and $f \in C$), so more flexible methods as finite elements are used.
If we plot the error $\|u - u_{exact}\|_\infty = \max_{i\in[0,N-1]} |u_i - u(x_i)|$ we obtain a straight line

$$
E = ch^k = c\Big(\frac{1}{N-1}\Big)^k \tag{10.6}
$$

But how do we measure the rate k?

1. Case 1: we know the exact solution. Hypothesis: C is independent of u_h and h.

$$
E_1 = \|u_{h_1} - u_{exact}\|_\infty \simeq Ch_1^k \tag{10.7}
$$

$$
E_2 = \|u_{h_2} - u_{exact}\|_\infty \simeq Ch_2^k \tag{10.8}
$$

$$
\tag{10.9}
$$

   Doing at least two measurements we can obtain

$$
\frac{E_1}{E_2} \simeq \Big(\frac{h_1}{h_2}\Big)^k \tag{10.10}
$$

So we obtain k as

$$k = \frac{\log\left(\frac{E_1}{E_2}\right)}{\log\left(\frac{h_1}{h_2}\right)} = \frac{\log(E_1) - \log(E_2)}{\log(h_1) - \log(h_2)} \tag{10.11}$$

2. Case 2: we do not know the exact solution. We have to make sure to reduce h by a constant factor θ ($h_{i+1} = \theta h_i$), and use at least 3 approximate solutions.

$$\begin{aligned} \|u_{h_2} - u_{h_1}\|_\infty \leq & \|u_{h_2} - u_{exact}\|_\infty + \|u_{h_1} - u_{exact}\|_\infty && (10.12)\\ \sim & Ch_1^k + Ch_2^k && (10.13)\\ \sim & Ch_1^k + C\theta^k h_1^k && (10.14)\\ \sim & C(1+\theta^k)h_1^k && (10.15) \end{aligned}$$

As $h_2 = \theta h_1, h_3 = \theta h_2$

$$\frac{\|u_{h_1} - u_{h_2}\|_\infty}{\|u_{h_2} - u_{h_2}\|_\infty} \sim \frac{C_2 h_1^k}{C_2 h_2^k} = \theta^{-k} \tag{10.16}$$

So, extracting k

$$k = \frac{\log(\|u_{i+2} - u_{i+1}\|_\infty) - \log(\|u_{i+1} - u_i\|_\infty)}{\log(\theta)} \tag{10.17}$$

10.2 Finite elements and the Galerkin method

The finite elements method is an alternative of 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 $v \in C'([a,b])$ smooth enough.
- We integrate the resulting equality on $[a,b]$

$$-\int_a^b u''(x)v(x)dx = \int_a^b f(x)v(x) \quad \forall v \in V \tag{10.18}$$

- Using integration by parts, we obtain

$$\int_a^b u'(x)v'(x) - [u'(x)v(x)]_a^b = \int_a^b f(x)v(x)dx \tag{10.19}$$

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

$$\begin{cases} \int_a^b u'(x)v'(x)dx = \int_a^b f(x)v(x)dx \quad \forall v \in C'(a,b) \\ v(a) = v(b) = 0 \end{cases} \tag{10.20}$$

This last equation is called weak formulation of the Poisson problem. In this case, both u and v can be less regular than C'.

V is a space of continuous function. It is an Hilbert space where the integral of the square is finite(L^2). It is the space of function whose first derivative is in L^2.

$$V \equiv H_0'([a,b]) \coloneqq \{v \in L^2, v' \in L^2, v(a) = v(b) = 0\} \tag{10.21}$$

As a matter of facts

$$\int_a^b fv < +\infty \quad \forall v \in H_0'([a,b]) \tag{10.22}$$

is true if $f \in L^2$, but also if f is less regular than L^2, but satisfy the same condition $f \in V^*$.
To solve the weak formulation, we build $v_h \subset V$ s.t.

$$v_h = span\{v_i\}_{i=0}^{N_h}, \quad \dim(V_h) = N_h + 1 \tag{10.23}$$

and project the problem in that space, called the **finite elements space** of degree 1.
We want to find $u_h \in V_h$ s.t. $u_h(a) = \alpha$, $u_h(b) = \beta$ and (Galerkin approximation)

$$\forall v_h \in V_h^\circ \quad \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} u_h(x) v_h'(x) dx = \int_a^b f(x) v_h(x) dx \tag{10.24}$$

with $\alpha = \beta = 0$. Functions in V_h are piecewise polynomial (linear in V_h°, else of order n) which can be expressed thanks to the function basis φ as:

$$v_h(x) = \sum_{j=1}^{N} v_h(x_j)\varphi_j(x) \quad \text{where} \quad \varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & \text{if } x_{j-1} \le x \le x_j \\ \frac{x_j - x_{j+1}}{x_j - x_{j+1}} & \text{if } x_j \le x \le x_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{10.25}$$

The generic φ_j is then 0 in all the nodes other then $\varphi_j(x_j) = 1$. The functions φ_j with $j = 1, ..., N$ are called **shape** or **hat functions** and provide basis for V_h°, where we define

$$V_h^k :== \{v \in \mathbb{P}^k([x_i, x_{i+1}]) | v(a) = v(b) = 0, v \in C^\circ([a,b])\} \tag{10.26}$$

In our case the polinomial are of order 1. Since $|x_{j-1} - x_j| = h$ and the derivative of φ corresponds to the slope of the line, we have that

$$\varphi_j' = \begin{cases} \frac{1}{h} & x \in [x_{j-1}, x_j] \\ -\frac{1}{h} & x \in [x_j, x_{j+1}] \\ 0 & \text{otherwise} \end{cases} \tag{10.27}$$

We can rewrite the weak formulation as a system

$$A_{FE}\boldsymbol{u} = \boldsymbol{f}_{FE} \tag{10.28}$$

where $\boldsymbol{u} \to$ vector of unknowns u_j, $F \to$ vectors of $F_i = \int_0^1 f(x)v_i(x)dx$

$$A_{FE} = \int_0^1 v_j' v_i' dx = \int_0^1 \varphi_j' \varphi_i' \tag{10.29}$$

Then we have that

$$A_{FE} = \begin{cases} 0 & \text{when } |i-j| \geq 2 \\ \int_{x_{i-1}}^{x_{i+1}} \varphi_i'(x)^2 dx = \frac{1}{h^2} \int_{x_{i-1}}^{x_{i+1}} dx = \frac{2h}{h^2} = \frac{2}{h} & \text{when } i = j \\ \int_{x_i}^{x_{i+1}} \varphi_i(x)\varphi_{i-1}(x) = -\frac{1}{h^2} \int_{x_i}^{x_{i+1}} dx = -\frac{2h}{h^2} = -\frac{1}{h} & \text{when } |i-j| = 1 \end{cases} \tag{10.30}$$

$$\Rightarrow \frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \tag{10.31}$$

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 $\alpha > 1$ allows for greater convergence, and leads to different matrices.

10.3 Finite differences for 2D Poisson problem

FD approximate the partial derivatives in PDEs as incremental ratios on a computational grid of finite nodes.
Given our space $\Omega(a,b) \times (x,d)$(simple and regular), 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 $u_{i,j}$ to approximate $u(x_i, y_i)$ on uniformly spaced nodes

$$\Delta u = \partial_x^2 u_{i,j} + \partial_y^2 u_{i,j} = \sum_{d=0}^{N-1} \frac{\partial^2 u_{i,j}}{\partial x_d^2} \tag{10.32}$$

with

$$\delta_x^2 u_{i,j} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h_x^2} \tag{10.33}$$

$$\delta_y^2 u_{i,j} = \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h_y^2} \tag{10.34}$$

Using second order accuracy w.r.t. h to replace $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ at (x_i, x_j).
Replacing it in the 2D Poisson problem, we get

$$-(\delta_x^2 u_{i,j} + \delta_y^2 u_{i,j}) = f_{i,j} \quad i = 1, ..., N_x \quad j = 1, ..., N_y \tag{10.35}$$

with boundary $u_{i,j} = g_{i,j}$ s.t. $(x_i, y_i) \in \partial\Delta h$.
If nodes are uniformly spread ($h_x = h_y$) we get

$$-\frac{1}{h^2}\Big(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i,j+1} + u_{i+1,j}\Big) = f_{i,j} \tag{10.36}$$

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 $A \in \mathbb{R}^{n \times n}$

$$A = tridiag(D, T, D) \tag{10.37}$$

with

$$T = tridiag\Big(-\frac{1}{h_x^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_y^2}\Big) \quad \text{and} \quad D = diag\Big(-\frac{1}{h_y^2}\Big) \tag{10.38}$$

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.

10.4 Lax-Milgram theorem

Let V be Hilbert (normed and with internal scalar product), given $a(\cdot,\cdot) : V \times V \to \mathbb{R}$) that is:

- Bilinear and bounded: $\forall v\ a(v_i, \cdot) \wedge a(\cdot, v_2)$ are linear, a.k.a. linear in both variables ($\int_\Omega \nabla v_1 \nabla v_2$).
- Continuous: $\exists x > 0$ s.t. $|a(u,v)| \leq C\|u\|\|v\|\ \forall u, v \in V$
- Coercive: $\exists \alpha > 0$ s.t. $a(u,u) \geq \alpha\|u\|^2\ \forall u \in V$

then

$$\forall f \in v^*, \quad \exists! \text{ solution } u \in V \text{ s.t. } a(u,v) =< f, v >= f(v)\ \forall v \in V \tag{10.39}$$

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 formulation of Poisson problem.

Corollary 8.1. *Corollary: Solution u is bounded w.r.t. data f*

$$\|u\|_V = \frac{1}{\alpha}\|f\|_{V^*} \tag{10.40}$$

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

$$\Rightarrow Au = F, \quad A_{ij} = a(v_j, v_i), \quad F_i = f(v_i) \tag{10.41}$$

Lemma 9. *Cea's lemma(orthogonality of error):*

$$a(u_h - u, v_h) = 0 \Rightarrow \quad \|u - u_h\| \leq \frac{c}{\alpha} \inf_{v_h \in V_h} \|u - v_h\| \quad \forall v \in V_h \tag{10.42}$$

a.k.a u_h is the best approximation of u in V_h up to $\frac{c}{\alpha}$

Proof. 1. $a(u,v) = f(v) = a(u_h, v)\ v \in v_h \quad \Rightarrow a(u - u_h, v) = 0$

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

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

$$\Rightarrow \alpha \|u-u_h\|^2 \leq \gamma \|u-u_h\| \|u-v\| \quad \forall v \in V_h \tag{10.43}$$

$$\|u-u_h\| \leq \frac{\gamma}{\alpha} \|u-v\| \quad \forall v \in V_h \tag{10.44}$$

Convergence is proven by Cea'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

$$\lim_{h \to 0} T_h(x_i, y_i) = 0 \qquad (x_i, y_i) \in \Delta h \setminus \partial \Delta h \tag{10.45}$$

So, the Galerkin method is valid. □