# Numerical methods

## 1 | Errors

### Floating-point representation

**Theorem 1.1.** Let $b \in \mathbb{N}$, $b \geq 2$. Any real number $x \in \mathbb{R}$ can be represented of the form

$$x = s\left(\sum_{i=1}^{\infty} \alpha_i b^{-i}\right) b^q$$

where $s \in \{-1, 1\}$, $q \in \mathbb{Z}$ and $\alpha_i \in \{0, 1, \ldots, b-1\}$. Moreover, this representation is unique if $\alpha_1 \neq 0$ and $\forall i_0 \in \mathbb{N}$, $\exists i \geq i_0 : \alpha_i \neq b-1$. We will write

$$x = s(0.\alpha_1\alpha_2\cdots)_b b^q$$

where the subscript $b$ in the parenthesis indicates that the number $0.\alpha_1\alpha_2\alpha_3\cdots$ is in base $b$.

**Definition 1.2 (Floating-point representation).** Let $x$ be a real number. Then, the *floating-point representation of $x$* is:

$$x = s\left(\sum_{i=1}^{t} \alpha_i b^{-i}\right) b^q$$

Here $s$ is called the *sign*; $\sum_{i=1}^{t} \alpha_i b^{-i}$, the *significant* or *mantissa*, and $q$, the *exponent*, limited to a prefixed range $q_{\min} \leq q \leq q_{\max}$. Therefore, the floating-point representation of $x$ can be expressed as:

$$x = smb^q = s(0.\alpha_1\alpha_2\cdots\alpha_t)_b b^q$$

Finally, we say a floating-point number is *normalized* if $\alpha_1 \neq 0$.

| Format | $b$ | $t$ | $q_{\min}$ | $q_{\max}$ | bits |
|---|---|---|---|---|---|
| IEEE simple | 2 | 24 | -126 | 127 | 32 |
| IEEE double | 2 | 53 | -1022 | 1023 | 64 |

Table 1: Parameters of IEEE simple and IEEE double formats.

**Definition 1.3.** Let $x \in \mathbb{R}$ be such that $x = s(0.\alpha_1\alpha_2\cdots)_b b^q$ with $q_{\min} \leq q \leq q_{\max}$. We say the *floating-point representation by truncation of $x$* is:

$$fl_T(x) = s(0.\alpha_1\alpha_2\cdots\alpha_t)_b b^q$$

We say the *floating-point representation by rounding of $x$* is:

$$fl_R(x) = \\ = \begin{cases} s(0.\alpha_1\cdots\alpha_t)_b b^q & \text{if } 0 \leq \alpha_{t+1} < \dfrac{b}{2} \\ s(0.\alpha_1\cdots\alpha_{t-1}(\alpha_t+1))_b b^q & \text{if } \dfrac{b}{2} \leq \alpha_{t+1} \leq b-1 \end{cases}$$

**Definition 1.4.** Given a value $x \in \mathbb{R}$ and an approximation $\tilde{x}$ of $x$, the *absolute error* is:

$$\Delta x := |x - \tilde{x}|$$

If $x \neq 0$, the *relative error* is:

$$\delta x := \frac{|x - \tilde{x}|}{x}$$

If $x$ is unknown, we take:

$$\delta x \approx \frac{|x - \tilde{x}|}{\tilde{x}}$$

**Definition 1.5.** Let $\tilde{x}$ be an approximation of $x$. If $\Delta x \leq \frac{1}{2}10^{-t}$, we say $\tilde{x}$ *has $t$ correct decimal digits*. If $x = sm10^q$ with $0.1 \leq m < 1$, $\tilde{x} = s\tilde{m}10^q$ and

$$u := \max\{i \in \mathbb{Z} : |m - \tilde{m}| \leq \frac{1}{2}10^{-i}\}$$

then we say that $\tilde{x}$ *has $u$ significant digits*.

**Proposition 1.6.** Let $x \in \mathbb{R}$ be such that $x = s(0.\alpha_1\alpha_2\cdots)_b b^q$ with $\alpha_1 \neq 0$ and $q_{\min} \leq q \leq q_{\max}$. Then, its floating-point representation in base $b$ and with $t$ digits satisfy:

$$|fl_T(x) - x| \leq b^{q-t} \qquad |fl_R(x) - x| \leq \frac{1}{2}b^{q-t}$$
$$\left|\frac{fl_T(x) - x}{x}\right| \leq b^{1-t} \qquad \left|\frac{fl_R(x) - x}{x}\right| \leq \frac{1}{2}b^{1-t}$$

**Definition 1.7.** The *machine epsilon* $\epsilon$ is defined as:

$$\epsilon := \min\{\varepsilon > 0 : fl(1 + \varepsilon) \neq 1\}$$

**Proposition 1.8.** For a machine working by truncation, $\epsilon = b^{1-t}$. For a machine working by rounding, $\epsilon = \frac{1}{2}b^{1-t}$.

### Propagation of errors

**Proposition 1.9 (Propagation of absolute errors).** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. If $\Delta x_j$ is the absolute error of the variable $x_j$ and $\Delta f(x)$ is the absolute error of the function $f$ evaluated at the point $x = (x_1, \ldots, x_n)$, we have:

$$|\Delta f(x)| \lesssim \sum_{j=1}^{n} \left|\frac{\partial f}{\partial x_j}(x)\right| |\Delta x_j|$$[1]

The coefficients $\left|\frac{\partial f}{\partial x_j}(x)\right|$ are called *absolute condition numbers of the problem.*

**Proposition 1.10 (Propagation of relative errors).** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. If $\delta x_j$ is the relative error of the variable $x_j$ and $\delta f(x)$ is the relative error of the function $f$ evaluated at the point $x = (x_1, \ldots, x_n)$, we have:

$$|\delta f(x)| \lesssim \sum_{j=1}^{n} \frac{\left|\frac{\partial f}{\partial x_j}(x)\right| |x_j|}{|f(x)|} |\delta x_j|$$

The coefficients $\frac{\left|\frac{\partial f}{\partial x_j}(x)\right||x_j|}{|f(x)|}$ are called *relative condition numbers of the problem.*

[1]The symbol $\lesssim$ means that we are omitting terms of order $\Delta x_j \Delta x_k$ and higher.

### Numerical stability of algorithms

**Definition 1.11.** An algorithm is said to be *numerically stable* if errors in the input lessen in significance as the algorithm executes, having little effect on the final output. On the other hand, an algorithm is said to be *numerically unstable* if errors in the input cause a considerably larger error in the final output.

**Definition 1.12.** A problem with a low condition number is said to be *well-conditioned*. Conversely, a problem with a high condition number is said to be *ill-conditioned*.

## 2 | Zeros of functions

**Definition 1.13.** Let $f:\mathbb{R}\to\mathbb{R}$ be a function. We say $\alpha$ is a *zero* or a *solution to the equation* $f(x)=0$ if $f(\alpha)=0$.

**Definition 1.14.** Let $f:\mathbb{R}\to\mathbb{R}$ be a sufficiently differentiable function. We say $\alpha$ is a *zero of multiplicity* $m\in\mathbb{N}$ if

$$f(\alpha)=f'(\alpha)=\cdots=f^{(m-1)}(\alpha)=0 \quad \text{and} \quad f^{(m)}(\alpha)\neq 0$$

If $m=1$, the zero is called *simple*; if $m=2$, *double*; if $m=3$, *triple*...

### Root-finding methods

For the following methods consider a continuous function $f:I\subseteq\mathbb{R}\to\mathbb{R}$ with an unknown zero $\alpha\in I$. Given $\varepsilon>0$, we want to approximate $\alpha$ with $\tilde{\alpha}$ such that $|\alpha-\tilde{\alpha}|<\varepsilon$.

**Method 1.15 (Bisection method).** Suppose $I=[a_0,b_0]$. For each step $n\geq 0$ of the algorithm we will approximate $\alpha$ by

$$c_n=\frac{a_n+b_n}{2}$$

If $f(c_n)=0$ we are done. If not, let

$$[a_{n+1},b_{n+1}]=\begin{cases}[a_n,c_n] & \text{if} \quad f(a_n)f(c_n)<0\\ [c_n,b_n] & \text{if} \quad f(a_n)f(c_n)>0\end{cases}$$

and iterate the process again[2]. The length of the interval $[a_n,b_n]$ is $\frac{b_0-a_0}{2^n}$ and therefore:

$$|\alpha-c_n|<\frac{b_0-a_0}{2^{n+1}}<\varepsilon \iff n>\frac{\log\left(\frac{b_0-a_0}{\varepsilon}\right)}{\log 2}-1$$

**Method 1.16 (*Regula falsi* method).** Suppose $I=[a_0,b_0]$. For each step $n\geq 0$ of the algorithm we will approximate $\alpha$ by

$$c_n=b_n-f(b_n)\frac{b_n-a_n}{f(b_n)-f(a_n)}=\frac{a_nf(b_n)-b_nf(a_n)}{f(b_n)-f(a_n)}.$$

If $f(c_n)=0$ we are done. If not, let

$$[a_{n+1},b_{n+1}]=\begin{cases}[a_n,c_n] & \text{if} \quad f(a_n)f(c_n)<0\\ [c_n,b_n] & \text{if} \quad f(a_n)f(c_n)>0\end{cases}$$

and iterate the process again.

**Method 1.17 (Secant method).** Suppose $I=\mathbb{R}$ and that we have two different initial approximations $x_0$, $x_1$. Then, for each step $n\geq 0$ of the algorithm we obtain a new approximation $x_{n+2}$, given by:

$$x_{n+2}=x_{n+1}-f(x_{n+1})\frac{x_{n+1}-x_n}{f(x_{n+1})-f(x_n)}$$

**Method 1.18 (Newton-Raphson method).** Suppose $I=\mathbb{R}$, $f\in\mathcal{C}^1$ and that we have an initial approximation $x_0$. Then, for each step $n\geq 0$ we obtain a new approximation $x_{n+1}$, given by:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}$$

**Method 1.19 (Newton-Raphson modified method).** Suppose $I=\mathbb{R}$, $f\in\mathcal{C}^1$ and that we have an initial approximation $x_0$ of a zero $\alpha$ of multiplicity $m$. Then, for each step $n\geq 0$ we obtain a new approximation $x_{n+1}$, given by:

$$x_{n+1}=x_n-m\frac{f(x_n)}{f'(x_n)}$$

**Method 1.20 (Chebyshev method).** Suppose $I=\mathbb{R}$, $f\in\mathcal{C}^2$ and that we have an initial approximation $x_0$. Then, for each step $n\geq 0$ we obtain a new approximation $x_{n+1}$, given by:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}-\frac{1}{2}\frac{f(x_n)^2f''(x_n)}{f'(x_n)^3}$$

### Fixed-point iterations

**Definition 1.21.** Let $g:[a,b]\to[a,b]\subset\mathbb{R}$ be a function. A point $\alpha\in[a,b]$ is *n-periodic* if $g^n(\alpha)=\alpha$ and $g^j(\alpha)\neq\alpha$ for $j=1,\ldots,n-1$[3].

**Theorem 1.22 (Fixed-point theorem).** Let $(M,d)$ be a complete metric space and $g:M\to M$ be a contraction[4]. Then, $g$ has a unique fixed point $\alpha\in M$ and for every $x_0\in M$,

$$\lim_{n\to\infty}x_n=\alpha, \quad \text{where } x_n=g(x_{n-1}) \quad \forall n\in\mathbb{N}$$

**Proposition 1.23.** Let $(M,d)$ be a metric space and $g:M\to M$ be a contraction of constant $k$. Then, if we want to approximate a fixed point $\alpha$ by the iteration $x_n=g(x_{n-1})$, we have:

$$d(x_n,\alpha)\leq\frac{k^n}{1-k}d(x_1,x_0) \qquad \text{(a priori estimation)}$$

$$d(x_n,\alpha)\leq\frac{k}{1-k}d(x_n,x_{n-1}) \qquad \text{(a posteriori estimation)}$$

**Corollary 1.24.** Let $g:\mathbb{R}\to\mathbb{R}$ be a function of class $\mathcal{C}^1$. Suppose $\alpha$ is a fixed point of $g$ and $|g'(\alpha)|<1$. Then, there exists $\varepsilon>0$ and $I_\varepsilon:=[\alpha-\varepsilon,\alpha+\varepsilon]$ such that $g(I_\varepsilon)\subseteq I_\varepsilon$ and $g$ is a contraction on $I_\varepsilon$. In particular, if $x_0\in I_\varepsilon$, the iteration $x_{n+1}=g(x_n)$ converges to $\alpha$.

[2]Note that bisection method only works for zeros of odd multiplicity.
[3]Note that 1-periodic points are the fixed points of $f$.
[4]Remember definitions **??**, **??** and **??**.

**Definition 1.25.** Let $g : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^1$ and $\alpha$ be a fixed point of $g$. We say $\alpha$ is an *attractor fixed point* if $|g'(\alpha)| < 1$. In this case, any iteration $x_{n+1} = g(x_n)$ in $I_\varepsilon$ converges to $\alpha$. If $|g'(\alpha)| > 1$, we say $\alpha$ is a *repulsor fixed point*. In this case, $\forall x_0 \in I_\varepsilon$ the iteration $x_{n+1} = g(x_n)$ doesn't converge to $\alpha$.

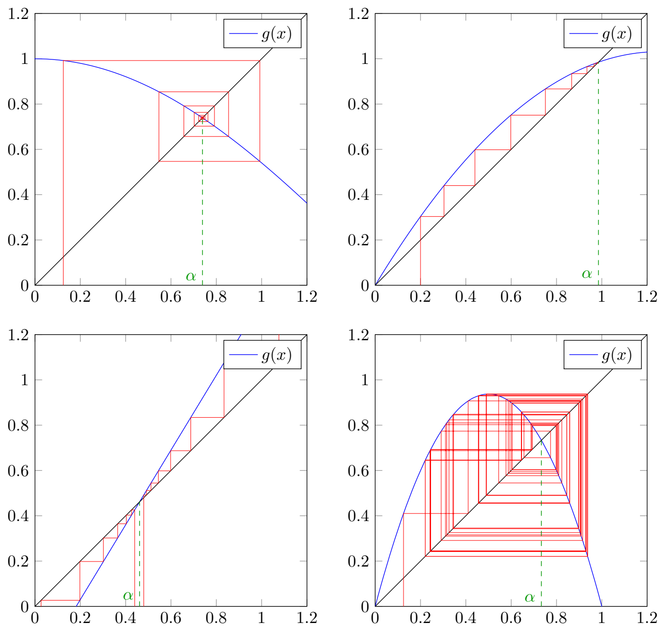


Figure 1: Cobweb diagrams. In the figures at the top, $\alpha$ is a attractor point, that is, $|g'(\alpha)| < 1$. More precisely, the figure at the top left occurs when $-1 < g'(\alpha) \leq 0$ and the figure at the top right when $0 \leq g'(\alpha) < 1$. In the figure at bottom left, $\alpha$ is a repulsor point. Finally, in the figure at bottom right the iteration $x_{n+1} = g(x_n)$ has no limit. It is said to have a *chaotic behavior*.

### Order of convergence

**Definition 1.26 (Order of convergence).** Let $(x_n)$ be a sequence of real numbers that converges to $\alpha \in \mathbb{R}$. We say $(x_n)$ has *order of convergence* $p \in \mathbb{R}^+$ if exists $C > 0$ such that:

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = C$$

The constant $C$ is called *asymptotic error constant*. For the case $p = 1$, we need $C < 1$. In this case the convergence is called *linear convergence*; for $p = 2$, is called *quadratic convergence*; for $p = 3$, *cubic convergence*... If it's satisfied that

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = 0$$

for some $p \in \mathbb{R}^+$, we say the sequence has *order of convergence at least* $p$.

**Theorem 1.27.** Let $g : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^p$ and let $\alpha$ be a fixed point of $g$. Suppose

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0$$

with $|g'(\alpha)| < 1$ if $p = 1$. Then, the iteration $x_{n+1} = g(x_n)$, with $x_0$ sufficiently close to $\alpha$, has order of convergence at least $p$. If, moreover, $g^{(p)}(\alpha) \neq 0$, then the previous iteration has order of convergence $p$ with asymptotic error constant $C = \frac{|g^{(p)}(\alpha)|}{p!}$.

**Theorem 1.28.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^3$ and $\alpha$ be a simple zero of $f$. If $f''(\alpha) \neq 0$, then Newton-Raphson method for finding $\alpha$ has quadratic convergence with asymptotic error constant $C = \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right|$.

If $f \in \mathcal{C}^{m+2}$, and $\alpha$ is a zero of multiplicity $m > 1$, then Newton-Raphson method has linear convergence but Newton-Raphson modified method has at least quadratic convergence.

**Theorem 1.29.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^3$ and let $\alpha$ be a simple zero of $f$. Then, Chebyshev's method for finding $\alpha$ has at least cubic convergence.

**Definition 1.30.** We define the *computational efficiency of an algorithm* as a function $E(p, t)$, where $t$ is the time taken for each iteration of the method and $p$ is the order of convergence of the method. $E(p, t)$ must satisfy the following properties:

1. $E(p, t)$ is increasing with respect to the variable $p$ and decreasing with respect to $t$.
2. $E(p, t) = E(p^m, mt)\ \forall m \in \mathbb{R}$.

Examples of such functions are the following:

$$E(p, t) = \frac{\log p}{t} \qquad E(p, t) = p^{1/t}$$

### Sequence acceleration

**Method 1.31 (Aitken's $\Delta^2$ method).** Let $(x_n)$ be a sequence of real numbers. We denote:

$$\Delta x_n := x_{n+1} - x_n$$
$$\Delta^2 x_n := \Delta x_{n+1} - \Delta x_n = x_{n+2} - 2x_{n+1} + x_n$$

*Aitken's $\Delta^2$ method* is the transformation of the sequence $(x_n)$ into a sequence $y_n$, defined as:

$$y_n := x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$$

with $y_0 = x_0$.

**Theorem 1.32.** Let $(x_n)$ be a sequence of real numbers such that $\lim_{n\to\infty} x_n = \alpha$, $x_n \neq \alpha\ \forall n \in \mathbb{N}$ and $\exists C$, $|C| < 1$, satisfying

$$x_{n+1} - \alpha = (C + \delta_n)(x_n - \alpha) \quad \text{with} \quad \lim_{n\to\infty} \delta_n = 0$$

Then, the sequence $(y_n)$ obtained from Aitken's $\Delta^2$ process is well-defined and

$$\lim_{n\to\infty} \frac{y_n - \alpha}{x_n - \alpha} = 0$$ [5]

[5] This means that Aitken's $\Delta^2$ method produces an acceleration of the convergence of the sequence $(x_n)$.

**Method 1.33 (Steffensen's method).** Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function and suppose we have an iterative method $x_{n+1} = g(x_n)$. Then, for each step $n$ we can consider a new iteration $y_{n+1}$, with $y_0 = x_0$, given by:

$$y_{n+1} = y_n - \frac{(g(y_n) - y_n)^2}{g(g(y_n)) - 2g(y_n) + y_n}$$

**Proposition 1.34.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^2$ and $\alpha$ be a simple zero of $f$. Then, Steffensen's method for finding $\alpha$ has at least quadratic convergence[6].

### Zeros of polynomials

**Lemma 1.35.** Let $p(z) = a_0 + a_1 z + \cdots + a_n z^n \in \mathbb{C}[x]$ with $a_n \neq 0$. We define

$$\lambda := \max \left\{ \left| \frac{a_i}{a_n} \right| : i = 0, 1, \ldots, n-1 \right\}$$

Then, if $p(\alpha) = 0$ for some $\alpha \in \mathbb{C}$, $|\alpha| \leq \lambda + 1$.

**Definition 1.36 (Strum's sequence).** Let $(f_i)$, $i = 0, \ldots, n$, be a sequence of continuous functions defined on $[a, b] \subset \mathbb{R}$ and $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^1$ such that $f(a)f(b) \neq 0$. We say $(f_n)$ is a *Sturm's sequence* if:

1. $f_0 = f$.
2. If $\alpha \in [a, b]$ satisfies $f_0(\alpha) = 0 \implies f_0'(\alpha) f_1(\alpha) > 0$.
3. For $i = 1, \ldots, n-1$, if $\alpha \in [a, b]$ satisfies $f_i(\alpha) = 0 \implies f_{i-1}(\alpha) f_{i+1}(\alpha) < 0$.
4. $f_n(x) \neq 0 \; \forall x \in [a, b]$.

**Definition 1.37.** Let $(a_i)$, $i = 0, \ldots, n$, be a sequence. We define $\nu(a_i)$ as the number of sign variations of the sequence

$$\{a_0, a_1, \ldots, a_n\}$$

without taking into account null values.

**Theorem 1.38 (Sturm's theorem).** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^1$ such that $f(a)f(b) \neq 0$ and with a finite number of zeros. Let $(f_i)$, $i = 0, \ldots, n$, be a Sturm sequence defined on $[a, b]$. Then, the number of zeros of $f$ on $[a, b]$ is

$$\nu(f_i(a)) - \nu(f_i(b))$$

**Lemma 1.39.** Let $p \in \mathbb{C}[x]$ be a polynomial. Then, the polynomial $q = \dfrac{p}{\gcd(p, p')}$ has the same roots as $p$ but all of them are simple.

**Proposition 1.40.** Let $p \in \mathbb{R}[x]$ be a polynomial with $\deg p = m$. We define $f_0 = \dfrac{p}{\gcd(p, p')}$ and $f_1 = f_0'$. If $\deg f_0 = n$, then for $i = 0, 1, \ldots, n-2$, we define $f_{i+2}$ as:

$$f_i(x) = q_{i+1}(x) f_{i+1}(x) - f_{i+2}(x)$$

(similarly to the euclidean division between $f_i$ and $f_{i+1}$). Then, $f_n$ is constant and hence the sequence $(f_i)$, $i = 0, \ldots, n$, is a Sturm sequence.

**Theorem 1.41 (Budan-Fourier theorem).** Let $p \in \mathbb{R}[x]$ be a polynomial with $\deg p = n$. Consider the sequence $(p^{(i)})$, $i = 0, \ldots, n$. If $p(a)p(b) \neq 0$, the number of zeros of $p$ on $[a, b]$ is:

$$\nu\left(p^{(i)}(a)\right) - \nu\left(p^{(i)}(b)\right) - 2k, \quad \text{for some } k \in \mathbb{N} \cup \{0\}$$

**Corollary 1.42 (Descartes' rule of signs).** Let $p = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{R}[x]$ be a polynomial. If $p(0) \neq 0$, the number of zeros of $p$ on $[0, \infty)$ is:

$$\nu(a_i) - 2k, \quad \text{for some } k \in \mathbb{N} \cup \{0\}$$[7]

**Theorem 1.43 (Gershgorin circle theorem).** Let $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ be a complex matrix and $\lambda$ be an eigenvalue of $A$. For all $i, j \in \{1, 2, \ldots, n\}$ we define:

$$r_i = \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}| \qquad R_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\}$$

$$c_j = \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{kj}| \qquad C_j = \{z \in \mathbb{C} : |z - a_{jj}| \leq c_j\}$$

Then, $\lambda \in \bigcup_{i=1}^n R_i$ and $\lambda \in \bigcup_{j=1}^n C_j$. Moreover in each connected component of $\bigcup_{i=1}^n R_i$ or $\bigcup_{j=1}^n C_j$ there are as many eigenvalues (taking into account the multiplicity) as disks $R_i$ or $C_j$, respectively.

**Corollary 1.44.** Let $p(z) = a_0 + a_1 z + \cdots + a_n z^n + z^{n+1} \in \mathbb{C}[x]$. We define

$$r = \sum_{i=1}^{n-1} |a_i| \quad c = \max\{|a_0|, |a_1| + 1, \ldots, |a_{n-1}| + 1\}$$

Then, if $p(\alpha) = 0$ for some $\alpha \in \mathbb{C}$,

$$\alpha \in (B(0, 1) \cup B(-a_n, r)) \cap (B(-a_n, 1) \cup B(0, c))$$

## 3 | Interpolation

**Definition 1.45.** We denote by $\Pi_n$ the *vector space of polynomials with real coefficients and degree less than or equal to n.*

**Definition 1.46.** Suppose we have a family of real valued functions $\mathfrak{C}$ and a set of points $\{(x_i, y_i)\}_{i=0}^n := \{(x_i, y_i) \in \mathbb{R}^2 : i = 0, \ldots, n \text{ and } x_j \neq x_k \iff j \neq k\}$. These points $\{(x_i, y_i)\}_{i=0}^n$ are called *support points*. The *interpolation problem* consists in finding a function $f \in \mathfrak{C}$ such that $f(x_i) = y_i$ for $i = 0, \ldots, n$[8].

[6] Note that the advantage of Steffensen's method over Newton-Raphson method is that in the former we don't need the differentiability of the function whereas in the latter we do.

[7] Note that making the change of variable $t = -x$ one can obtain the number of zeros on $(-\infty, 0]$ of $p$ by considering the polynomial $p(t)$.

[8] Types of interpolation are for example polynomial interpolation, trigonometric interpolation, Padé interpolation, Hermite interpolation and spline interpolation.

## Polynomial interpolation

**Definition 1.47.** Given a set of support points $\{(x_i, y_i)\}_{i=0}^n$, *Lagrange's interpolation problem* consists in finding a polynomial $p_n \in \Pi_n$ such that $p_n(x_i) = y_i$ for $i = 0, 1, \ldots, n$.

**Definition 1.48.** Let $\{(x_i, y_i)\}_{i=0}^n$ be a set of support points. We define $\omega_n(x) \in \mathbb{R}[x]$ as:

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

We define *Lagrange basis polynomials* $\ell_i(x) \in \mathbb{R}[x]$ as:

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

**Proposition 1.49.** Let $\{(x_i, y_i)\}_{i=0}^n$ be a set of support points. Then, Lagrange's interpolation problem has a unique solution and this is:

$$p_n(x) = \sum_{i=0}^{n} y_i \ell_i(x)$$

**Method 1.50 (Neville's algorithm).** Let $\{(x_i, y_i)\}_{i=0}^n$ be a set of support points, $\{i_0, \ldots, i_k\} \subset \{0, \ldots, n\}$ and $P_{i_0,\ldots,i_k}(x) \in \Pi_k$ be such that $P_{i_0,\ldots,i_k}(x_{i_j}) = y_{i_j}$ for $j = 0, \ldots, k$. Then, it is satisfied that:

1. $P_i(x) = y_i$.
2. $P_{i_0,\ldots,i_k}(x) = \dfrac{\begin{vmatrix} P_{i_1,\ldots,i_k}(x) & x - x_{i_k} \\ P_{i_0,\ldots,i_{k-1}}(x) & x - x_{i_0} \end{vmatrix}}{x_{i_k} - x_{i_0}}$.

**Definition 1.51.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be pairwise distinct points. We define the *divided difference of order $k$ of $f$ applied to* $\{x_i\}_{i=0}^k$, denoted by $f[x_0, \ldots, x_k]$, as the coefficient of $x^k$ of the interpolating polynomial with support points $\{(x_i, f(x_i))\}_{i=0}^k$.

**Proposition 1.52.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be pairwise distinct points. Lagrange interpolating polynomial with support points $\{(x_i, f(x_i))\}_{i=0}^n$ is:

$$p_n(x) = f[x_0] + \sum_{j=1}^{n} f[x_0, \ldots, x_j]\omega_{j-1}(x)$$

**Method 1.53 (Newton's divided differences method).** Let $f : \mathbb{R} \to \mathbb{R}$ be a function. For $x \in \mathbb{R}$, we have $f[x] = f(x)$. And if $\{x_i\}_{i=0}^n \subset \mathbb{R}$ are different points, then

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$$

**Theorem 1.54.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^{n+1}$, $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be pairwise distinct points and $p_n \in \mathbb{R}[x]$ be the interpolating polynomial with support points $\{(x_i, f(x_i))\}_{i=0}^n$. Then, $\forall x \in [a, b]$,

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_n(x)$$

where $\xi_x \in \langle x_0, \ldots, x_n, x\rangle$[9].

[9] The interval $\langle a_1, \ldots, a_k\rangle$ is defined as $\langle a_1, \ldots, a_k\rangle := (\min(a_1, \ldots, a_k), \max(a_1, \ldots, a_k))$.

**Lemma 1.55.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^{n+1}$ and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be pairwise distinct points. Then: $\exists \xi \in \langle x_0, \ldots, x_n\rangle$ such that:

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

**Proposition 1.56.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^{n+1}$, $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be pairwise distinct points and $\sigma \in S_n$. Then,

$$f[x_0, \ldots, x_n] = f[x_{\sigma(0)}, \ldots, x_{\sigma(n)}]$$

**Definition 1.57.** Let $\{(x_i, y_i)\}_{i=0}^n$ be support points. The points $\{x_i\}_{i=0}^n$ are *equally-spaced* if

$$x_i = x_0 + ih, \quad \text{for } i = 0, \ldots, n \text{ and with } h := \frac{x_n - x_0}{n}$$

**Proposition 1.58.** Let $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be equally-spaced points such that $x_i = x_0 + ih$, where $h = \frac{x_n - x_0}{n}$. Then:

$$\max\{|\omega_n(x)| : x \in [x_0, x_n]\} \leq \frac{h^{n+1}n!}{4}$$

**Definition 1.59.** Let $f : [a, b] \to \mathbb{R}$ be a function and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be equally-spaced points. We define:

$$\Delta f(x) := f(x + h) - f(x)$$
$$\Delta^{n+1} f(x) := \Delta(\Delta^n f(x))$$

**Lemma 1.60.** Let $f : [a, b] \to \mathbb{R}$ be a function and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be equally-spaced points. Then:

$$\Delta^n f(x_0) = n! h^n f[x_0, \ldots, x_n]$$

**Corollary 1.61.** Let $f \in \mathbb{R}[x]$ with $\deg f = n$. Suppose we interpolate $f$ with equally-spaced nodes. Then, $\Delta^n f(x) \equiv$ constant.

## Hermite interpolation

**Definition 1.62.** Given sets of points $\{x_i\}_{i=0}^m \subset \mathbb{R}$, $\{n_i\}_{i=0}^m \subset \mathbb{N}$ and $\{y_i^{(k)} : k = 0, \ldots, n_i - 1\}_{i=0}^m \subset \mathbb{R}$, *Hermite interpolation problem* consists in finding a polynomial $h_n \in \Pi_n$ such that $\sum_{i=0}^m n_i = n + 1$ and

$$h_n^{(k)}(x_i) = y_i^{(k)} \text{ for } i = 0, \ldots, m \text{ and } k = 0, \ldots, n_i - 1$$

**Proposition 1.63.** Hermite interpolation problem has a unique solution.

**Definition 1.64.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^n$ and $\{x_i\}_{i=0}^n \subset \mathbb{R}$ be points. We define $f[x_i, \overset{(n+1)}{\ldots}, x_i]$ as:

$$f[x_i, \overset{(n+1)}{\ldots}, x_i] = \frac{f^{(n)}(x_i)}{n!}$$

**Theorem 1.65.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^{n+1}$, $\{x_i\}_{i=0}^m \subset \mathbb{R}$ be pairwise distinct points and $\{n_i\}_{i=0}^m \subset \mathbb{N}$ be such that $\sum_{i=0}^m n_i = n + 1$. Let $h_n$ be the Hermite interpolating polynomial of $f$ with nodes $\{x_i\}_{i=0}^m \subset \mathbb{R}$, that is,

$$h_n^{(k)}(x_i) = f^{(k)}(x_i) \text{ for } i = 0, \ldots, m \text{ and } k = 0, \ldots, n_i - 1.$$

Then, $\forall x \in [a, b]$ $\exists \xi_x \in \langle x_0, \ldots, x_n, x\rangle$ such that:

$$f(x) - h_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x - x_0)^{n_0} \cdots (x - x_m)^{n_m}$$

### Spline interpolation

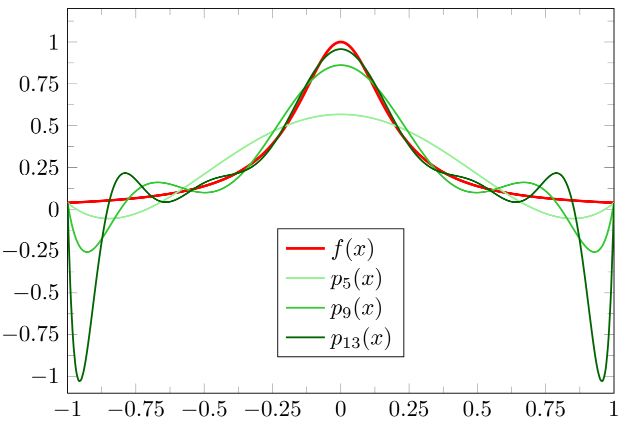


Figure 2: Runge's phenomenon. In this case $f(x) = \frac{1}{1+25x^2}$. $p_5(x)$ is the 5th-order Lagrange interpolating polynomial with equally-spaced interpolating points; $p_9(x)$, the 9th-order Lagrange interpolating polynomial with equally-spaced interpolating points, and $p_{13}(x)$, the 13th-order Lagrange interpolating polynomial with equally-spaced interpolating points.

**Definition 1.66 (Spline).** Let $\{(x_i, y_i)\}_{i=0}^n$ be support points of an interval $[a, b]$. A *spline of degree* $p$ is a function $s : [a, b] \to \mathbb{R}$ of class $\mathcal{C}^{p-1}$ satisfying:

$$s_{|_{[x_i,x_{i+1}]}} \in \mathbb{R}[x] \quad \text{and} \quad \deg s_{|_{[x_i,x_{i+1}]}} = p$$

for $i = 0, \ldots, n-1$ and $s(x_i) = y_i$ for $i = 0, \ldots, n$. The most common case are splines of degree $p = 3$ or *cubic splines*. In this case we can impose two more conditions on their definition in one of the following ways:

1. *Natural cubic spline*:
$$s''(x_0) = s''(x_n) = 0$$
2. *Cubic Hermite spline*: Given $y_0', y_n' \in \mathbb{R}$,
$$s'(x_0) = y_0', \quad s'(x_n) = y_n'$$
3. *Cubic periodic spline*:
$$s'(x_0) = s'(x_n), \quad s''(x_0) = s''(x_n)$$

**Definition 1.67.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. We define the *seminorm*[10] *of* $f$ as:

$$\|f\|^2 = \int_a^b (f''(x))^2 \mathrm{d}x$$

**Proposition 1.68.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^2$ interpolating the support points $\{(x_i, y_i)\}_{i=0}^n \subset \mathbb{R}^2$, $a \leq x_0 < \cdots < x_n \leq b$. If $s$ a spline associated with $\{(x_i, y_i)\}_{i=0}^n$, then:

$$\|f-s\|^2 = \|f\|^2 - \|s\|^2 - 2(f'-s)s''\Big|_{x_0}^{x_n} + 2\sum_{i=1}^n (f-s)s'''\Big|_{x_{i-1}^+}^{x_i^-}$$

**Theorem 1.69.** Let $f : [a, b] \to \mathbb{R}$ a function of class $\mathcal{C}^2$ interpolating the support points $\{(x_i, y_i)\}_{i=0}^n \subset \mathbb{R}^2$, $a \leq x_0 < \cdots < x_n \leq b$. If $s$ is the natural cubic spline associated with $\{(x_i, y_i)\}_{i=0}^n$, then:

$$\|s\| \leq \|f\|$$ [11]

## 4 | Numerical differentiation and integration

### Differentiation

**Theorem 1.70 (Intermediate value theorem).** Let $f : [a, b] \to \mathbb{R}$ be a continuous function, $x_0, \ldots, x_n \in [a, b]$ and $\alpha_0, \ldots, \alpha_n \geq 0$. Then, $\exists \xi \in [a, b]$ such that:

$$\sum_{i=0}^n \alpha_i f(x_i) = \left(\sum_{i=0}^n \alpha_i\right) f(\xi)$$

**Theorem 1.71 (Forward and backward difference formula of order 1).** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. Then, the forward difference formula of order 1 is:

$$f'(a) = \frac{f(a+h) - f(a)}{h} - \frac{f''(\xi)}{2}h$$

where $\xi \in \langle a, a+h \rangle$. Analogously, the backward difference formula of order 1 is:

$$f'(a) = \frac{f(a) - f(a-h)}{h} + \frac{f''(\eta)}{2}h$$

where $\eta \in \langle a-h, a \rangle$.

**Theorem 1.72 (Symmetric difference formula of order 1).** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^3$. Then, the symmetric difference formula of order 1 is:

$$f'(a) = \frac{f(a+h) - f(a-h)}{2h} - \frac{f^{(3)}(\xi)}{6}h^2$$

where $\xi \in \langle a-h, a+h \rangle$.

**Theorem 1.73 (Symmetric difference formula of order 2).** Let $f : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^4$. Then, the symmetric difference formula of order 2 is:

$$f''(a) = \frac{f(a+h) - 2f(a) + f(a-h)}{h^2} - \frac{f^{(4)}(\xi)}{12}h^2$$

where $\xi \in \langle a-h, a, a+h \rangle$.

### Richardson extrapolation

**Theorem 1.74 (Richardson extrapolation).** Suppose we have a function $f$ that approximate a value $\alpha$ with an error that depends on a small quantity $h$. That is:

$$\alpha = f(h) + a_1 h^{k_1} + a_2 h^{k_2} + \cdots$$

with $k_1 < k_2 < \cdots$ and $a_i$ are unknown constants. Given $q > 0$, we define:

$$D_1(h) = f(h) \quad \text{and} \quad D_{n+1}(h) = \frac{q^{k_n} D_n(h/q) - D_n(h)}{q^{k_n} - 1}$$

And we can observe that $\alpha = D_{n+1}(h) + O(h^{k_{n+1}})$.

[10] The term *seminorm* has been used instead of *norm* to emphasize that not all properties of a norm are satisfied with this definition.

[11] We can interpret this result as the natural cubic spline being the configuration that require the least "energy" to be "constructed".

### Integration

**Definition 1.75.** Let $f : [a, b] \to \mathbb{R}$ be a continuous function, $\{x_i\}_{i=0}^n \subset [a, b]$ be a set of nodes and $p_n$ be the Lagrange interpolating polynomial with support points $\{(x_i, f(x_i))\}_{i=0}^n$. We define the *quadrature formula* as:

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b p_n(x)\mathrm{d}x$$

**Lemma 1.76.** Let $f : [a, b] \to \mathbb{R}$ be a continuous function $\{x_i\}_{i=0}^n \subset [a, b]$ be a set of nodes. Then:

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=1}^n a_i f(x_i) \quad \text{where } a_i := \int_a^b \ell_i(x)\mathrm{d}x$$

**Definition 1.77.** The *degree of precision* of a quadrature formula is the largest $m \in \mathbb{N}$ such that the formula is exact for $x^k$ $\forall k = 0, 1, \ldots, m$.

**Lemma 1.78.** Let $p \in \Pi_n$ be a polynomial and $\{x_i\}_{i=0}^n \subset [a, b]$ be a set of nodes. Then:

$$\int_a^b p(x)\mathrm{d}x = \sum_{i=0}^n a_i p(x_i)$$

for some $a_i \in \mathbb{R}$.

### Newton-Cotes formulas

**Theorem 1.79 (Mean value theorem for integrals).** Let $f, g : [a, b] \to \mathbb{R}$ be such that $f$ is continuous and $g$ integrable. Suppose that $g$ does not change the sign on $[a, b]$. Then, $\exists \xi \in [a, b]$ such that:

$$\int_a^b f(x)g(x)\mathrm{d}x = f(\xi)\int_a^b g(x)\mathrm{d}x$$

**Theorem 1.80 (Closed Newton-Cotes Formulas).** Let $f : [a, b] \to \mathbb{R}$ be a function and $\{x_i\}_{i=0}^n \subset [a, b]$ be a set of equally-spaced points. If $I = \int_a^b f(x)\mathrm{d}x$ and $h = \frac{b-a}{n}$, then $\exists \xi \in [a, b]$ such that:

- If $n$ is even and $f \in \mathcal{C}^{n+2}$:

$$I = \sum_{i=0}^n a_i f(x_i) + \frac{h^{n+3} f^{n+2}(\xi)}{(n+2)!}\int_0^n t\prod_{i=0}^n (t-i)\mathrm{d}t$$

- If $n$ is odd and $f \in \mathcal{C}^{n+1}$:

$$I = \sum_{i=0}^n a_i f(x_i) + \frac{h^{n+2} f^{n+1}(\xi)}{(n+1)!}\int_0^n \prod_{i=0}^n (t-i)\mathrm{d}t$$ [12]

**Corollary 1.81 (Trapezoidal rule).** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. Then, $\exists \xi \in [a, b]$ such that:

$$\int_a^b f(x)\mathrm{d}x = \frac{h}{2}(f(a) + f(b)) - \frac{f''(\xi)}{12}h^3$$

where $h = b - a$. This is the case $n = 1$ of closed Newton-Cotes formulas.

**Corollary 1.82 (Simpson's rule).** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^4$. Then, $\exists \xi \in [a, b]$ such that:

$$\int_a^b f(x)\mathrm{d}x = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) - \frac{f^{(4)}(\xi)}{90}h^5$$

where $h = \frac{b-a}{2}$. This is the case $n = 2$ of closed Newton-Cotes formulas.

**Theorem 1.83 (Composite Trapezoidal rule).** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^4$, $h = \frac{b-a}{n}$ and $x_j = a + jh$ for each $j = 0, 1, \ldots, n$. Then, $\exists \xi \in [a, b]$ such that:

$$I = \int_a^b f(x)\mathrm{d}x = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \\ - \frac{f''(\xi)(b-a)}{12}h^2$$

We denote by $T(f, a, b, h)$ the approximation of $I$ by trapezoidal rule.

**Theorem 1.84 (Composite Simpson's rule).** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^4$, $n$ be an even number, $h = \frac{b-a}{n}$ and $x_j = a + jh$ for each $j = 0, 1, \ldots, n$. Then, $\exists \xi \in [a, b]$ such that:

$$I = \int_a^b f(x)\mathrm{d}x = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{n/2-1} f(x_{2j}) + \right. \\ \left. + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{f^{(4)}(\xi)(b-a)}{180}h^4$$

We denote by $S(f, a, b, h)$ the approximation of $I$ by Simpson's rule.

### Romberg method

**Definition 1.85.** We define *Bernoulli polynomials* $B_n(x)$ as $B_0(x) = 1$, $B_1(x) = x - \frac{1}{2}$ and

$$B'_{k+1} = (k+1)B_k \quad \text{for } k \geq 1$$

*Bernoulli numbers* are $B_n = B_n(0)$, $\forall n \geq 0$[13].

**Theorem 1.86 (Euler-Maclaurin formula).** Let $f \in \mathcal{C}^{2m+2}([a, b])$ be a function. Then:

$$T(f, a, b, h) = \int_a^b f(t)\mathrm{d}t + \sum_{k=1}^m \frac{B_{2k}h^{2k}}{(2k)!}\left(f^{(2k-1)}(b) - \right. \\ \left. - f^{(2k-1)}(a)\right) + \frac{(b-a)B_{2m+2}h^{2m+1}}{(2m+2)!}f^{(2m+2)}(\xi)$$

where $h = \frac{b-a}{n}$, $B_n$ are the Bernoulli numbers and $\xi \in [a, b]$.

[12]Note that when $n$ is even, the degree of precision is $n + 1$, although the interpolation polynomial is of degree at most $n$. When $n$ is odd, the degree of precision is only $n$.

[13]Exponential generating function of the sequence $(B_n)$ of Bernoulli numbers is $\frac{x}{\mathrm{e}^x - 1} = \sum_{n=1}^{\infty} \frac{B_n}{n!}x^n$.

**Theorem 1.87 (Romberg method).** Let $f \in \mathcal{C}^{2m+2}([a,b])$ be a function. Then, by Euler-Maclaurin formula, we obtain:

$$T(f,a,b,h) = \int_a^b f(t)\mathrm{d}t + \beta_1 h^2 + \beta_2 h^4 + \cdots$$

where $h = \frac{b-a}{n}$. For $n = 1,2,\ldots$ we define:

$$T_{n,1} = T\left(f,a,b,\frac{b-a}{2^n}\right) \quad T_{n,m+1} = \frac{4^m T_{n+1,m} - T_{n,m}}{4^m - 1}$$

for $m \leq n$. Then, we can observe that:

$$T_{n,m} = \int_a^b f(t)\mathrm{d}t + O\left(\left(\frac{b-a}{2^n}\right)^{2m}\right)$$

### Orthogonal polynomials

**Definition 1.88.** Let $f,g : [a,b] \to \mathbb{R}$ be continuous function and $\omega(x) : [a,b] \to \mathbb{R}^+$ be a weight function. The expression

$$\langle f,g\rangle = \int_a^b \omega(x)f(x)g(x)\mathrm{d}x$$

defines a positive semidefinite dot product in the vector space of bounded functions on $[a,b]$.

**Definition 1.89 (Orthogonal polynomials).** Let $\mathfrak{P} = \{\phi_i(x) \in \mathbb{R}[x] : \deg\phi_i(x) = i, i \in \mathbb{N}\cup\{0\}\}$ be a family of polynomials and $\omega(x) : [a,b] \to \mathbb{R}^+$ be a weight function. We say $\mathfrak{P}$ is *orthogonal with respect to the weight* $\omega(x)$ *on an interval* $[a,b]$ if

$$\langle \phi_i,\phi_j\rangle = \int_a^b \omega(x)\phi_i(x)\phi_j(x)\mathrm{d}x = 0 \iff i \neq j$$

Note that $\langle \phi_i,\phi_i\rangle > 0$ for each $i \in \mathbb{N}\cup\{0\}$.

**Lemma 1.90.** We define $\mathfrak{P}_n$ as $\mathfrak{P}_n = \{\phi_i(x) \in \Pi_n : \deg\phi_i(x) = i \text{ and } \langle\phi_i,\phi_j\rangle = 0 \iff i\neq j, i = 0,\ldots,n\}$. Then, $\mathfrak{P}_n$ is an *orthogonal basis of* $\Pi_n$.

**Lemma 1.91.** Let $\phi_k \in \mathfrak{P}_k$ and $q \in \Pi_n$. Then, $\langle q,\phi_k\rangle = 0$ for each $k > n$.

**Lemma 1.92.** Let $\phi_n \in \mathfrak{P}_n$. Then, $\forall n \in \mathbb{N}\cup\{0\}$, all roots of $\phi_n$ are real, simple and contained in the interval $(a,b)$, where the associated weight function $\omega(x)$ is defined.

**Theorem 1.93 (Existence of orthogonal polynomials).** For each $n \in \mathbb{N}\cup\{0\}$ there exists a unique monic orthogonal polynomial $\phi_n$ with $\deg\phi_n = n$, associated with the weight function $\omega(x)$, defined by:

$$\phi_0 = 1 \quad \phi_1(x) = x - \alpha_0$$
$$\phi_{n+1}(x) = (x-\alpha_n)\phi_n(x) - \beta_n\phi_{n-1}(x)$$

with $\alpha_n = \frac{\langle\phi_n,x\phi_n\rangle}{\langle\phi_n,\phi_n\rangle}$ $\forall n \in \mathbb{N}\cup\{0\}$ and $\beta_n = \frac{\langle\phi_n,\phi_n\rangle}{\langle\phi_{n-1},\phi_{n-1}\rangle}$ $\forall n \in \mathbb{N}$.

**Definition 1.94 (Chebyshev polynomials).** *Chebyshev polynomials* $T_n$ are the orthogonal polynomials defined on $[-1,1]$ with the weight $\omega(x) = \frac{1}{\sqrt{1-x^2}}$. These can be defined recursively as:

$$T_0(x) = 1 \quad T_1(x) = x$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

for $n = 1,2,\ldots$ Moreover $T_n(x) = \cos(n\arccos(x))$ which implies that the roots of $T_n(x)$ are:

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right) \quad \text{for } k = 1,\ldots,n$$

**Definition 1.95 (Laguerre polynomials).** *Laguerre polynomials* $L_n$ are the orthogonal polynomials defined on $[0,\infty)$ with the weight $\omega(x) = \mathrm{e}^{-x}$. These can be defined recursively as:

$$L_0(x) = 1 \quad L_1(x) = 1 - x$$
$$L_{n+1}(x) = \frac{(2n+1-x)L_n(x) - nL_{n-1}(x)}{n+1}$$

for $n = 1,2,\ldots$ The closed form of these polynomials is:

$$L_n(x) = \sum_{k=0}^n \binom{n}{k}\frac{(-1)^k}{k!}x^k$$

**Definition 1.96 (Legendre polynomials).** *Legendre polynomials* $P_n$ are the orthogonal polynomials defined on $[-1,1]$ with the weight $\omega(x) = 1$. These can be defined recursively as:

$$P_0(x) = 1 \quad P_1(x) = x$$
$$P_{n+1}(x) = \frac{(2n+1)xP_n(x) - nP_{n-1}(x)}{n+1}$$

for $n = 1,2,\ldots$ The closed form of these polynomials is:

$$P_n(x) = \frac{1}{2^n}\sum_{k=0}^n \binom{n}{k}^2 (x-1)^{n-k}(x+1)^k$$

### Gaußian quadrature

**Definition 1.97.** Let $f : [a,b] \to \mathbb{R}$ be a function and $\omega(x) : [a,b] \to \mathbb{R}^+$ be a weight function. Given a set of nodes $\{x_i\}_{i=1}^n \subset [a,b]$, the *quadrature formula with weight* $\omega(x)$ *of a function* $f$ is

$$\int_a^b \omega(x)f(x)\mathrm{d}x \approx \sum_{i=1}^n \omega_i f(x_i) \quad \text{with } \omega_i = \int_a^b \omega(x)\ell_i(x)\mathrm{d}x$$

**Lemma 1.98.** Let $f : [a,b] \to \mathbb{R}$ be a function and $\{x_i\}_{i=1}^n$ be the zeros of the orthogonal polynomial $\phi_n \in \mathfrak{P}_n$ with weight $\omega(x)$ on the interval $[a,b]$. Then, the formula

$$\int_a^b \omega(x)f(x)\mathrm{d}x \approx \sum_{i=1}^n \omega_i f(x_i) \quad \text{with } \omega_i = \int_a^b \omega(x)\ell_i(x)\mathrm{d}x$$

is exact for all polynomials in $\Pi_{2n-1}$.

**Proposition 1.99.** Let $f : [a,b] \to \mathbb{R}$ be a function and $\{x_i\}_{i=1}^n$ be the zeros of the orthogonal polynomial $\phi_n \in \mathfrak{P}_n$ with weight $\omega(x)$ on the interval $[a,b]$. Then, in the formula

$$\int_a^b \omega(x)f(x)\mathrm{d}x \approx \sum_{i=1}^n \omega_i f(x_i)$$

the values $\omega_i$ are positive and real for $i = 1,\ldots,n$.

**Theorem 1.100.** Let $f : [a, b] \to \mathbb{R}$ be a function of class $\mathcal{C}^{2n}$ and $\{x_i\}_{i=1}^n$ be the zeros of the orthogonal polynomial $\phi_n \in \mathfrak{P}_n$ with weight $\omega(x)$ on the interval $[a, b]$. Then:

$$\int_a^b \omega(x)f(x)\mathrm{d}x - \sum_{i=1}^n \omega_i f(x_i) = \frac{f^{2n}(\xi)}{(2n)!}\langle\phi_n, \phi_n\rangle$$

where $\xi \in [a, b]$.

# 5 | Numerical linear algebra

## Triangular matrices

**Definition 1.101.** A matrix $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ is *upper triangular* if $a_{ij} = 0$ whenever $i > j$. That is, $\mathbf{A}$ is of the form:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{(n-1)n} \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

**Definition 1.102.** A matrix $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ is *lower triangular* if $a_{ij} = 0$ whenever $j > i$. That is, $\mathbf{A}$ is of the form:

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & a_{n(n-1)} & a_{nn} \end{pmatrix}$$

**Definition 1.103.** A linear system with a triangular matrix associated is called a *triangular system*.

## Matrix norms

**Definition 1.104.** A *matrix norm* on the vector space $\mathcal{M}_n(\mathbb{R})$ is a function $\|\cdot\| : \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ satisfying all properties of a norm[14] and that:

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|, \quad \forall \mathbf{AB} \in \mathcal{M}_n(\mathbb{R})$$

**Definition 1.105.** Let $\|\cdot\|_\alpha$ be a vector norm. We say a matrix norm $\|\cdot\|_\beta$ is *compatible with* $\|\cdot\|_\alpha$ if

$$\|\mathbf{Av}\|_\alpha \leq \|\mathbf{A}\|_\beta\|\mathbf{v}\|_\alpha \quad \forall \mathbf{A} \in \mathcal{M}_n(\mathbb{R}) \text{ and } \forall \mathbf{v} \in \mathbb{R}^n$$

**Definition 1.106.** Let $\|\cdot\|$ be a vector norm and $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. We define a *subordinated matrix norm* $\|\cdot\|$ as:

$$\begin{aligned} \|\mathbf{A}\| = \max\{\|\mathbf{Av}\| : \mathbf{v} \in \mathbb{R}^n \text{ such that } \|\mathbf{v}\| = 1\} = \\ = \sup\left\{\frac{\|\mathbf{Av}\|}{\|\mathbf{v}\|} : \mathbf{v} \in \mathbb{R}^n \text{ such that } \mathbf{v} \neq 0\right\} \end{aligned}$$

**Lemma 1.107.** All subordinated matrix norms are compatible.

**Lemma 1.108.** For all subordinated matrix norm $\|\cdot\|$, we have $\|\mathbf{I}\| = 1$.

**Definition 1.109.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{C})$ be a matrix. We define the *spectrum* $\sigma(\mathbf{A})$ *of* $\mathbf{A}$ as:

$$\sigma(\mathbf{A}) = \{\lambda \in \mathbb{C} : \lambda \text{ is and eigenvalue of } \mathbf{A}\}$$

[14] See definition **??**.

**Definition 1.110.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{C})$ be a matrix. We define the *spectral radius* $\rho(\mathbf{A})$ *of* $\mathbf{A}$ as:

$$\rho(\mathbf{A}) = \max\{|\lambda| \in \mathbb{C} : \lambda \in \sigma(\mathbf{A})\}$$

**Proposition 1.111.** Let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$. Given the vector norms:

$$\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$$
$$\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$$
$$\|\mathbf{v}\|_\infty = \max\{|v_i| : i = 1, \ldots, n\}$$

their subordinated matrix norms are respectively:

$$\|\mathbf{A}\|_1 = \max\left\{\sum_{i=1}^n |a_{ij}| : j = 1, \ldots, n\right\}$$
$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^{\mathrm{T}}\mathbf{A})}$$
$$\|\mathbf{A}\|_\infty = \max\left\{\sum_{j=1}^n |a_{ij}| : i = 1, \ldots, n\right\}$$

**Proposition 1.112 (Properties of matrix norms).**

1. Matrix norms are continuous functions.
2. Given two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, there exist $\ell, L \in \mathbb{R}^+$ such that:
$$\ell\|\mathbf{A}\|_\beta \leq \|\mathbf{A}\|_\alpha \leq L\|\mathbf{A}\|_\beta \quad \forall \mathbf{A} \in \mathcal{M}_n(\mathbb{R})$$
3. For all subordinated matrix norm $\|\cdot\|$ and for all $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$:
$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|$$
4. Given a matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ and $\varepsilon > 0$, there exist a matrix norm $\|\cdot\|_{\mathbf{A},\varepsilon}$ such that:
$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|_{\mathbf{A},\varepsilon} \leq \rho(\mathbf{A}) + \varepsilon$$

**Definition 1.113.** A matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ is *convergent* if $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$.

**Theorem 1.114.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. The following statements are equivalent:

1. $\mathbf{A}$ is convergent.
2. $\lim_{k\to\infty} \|\mathbf{A}^k\| = \mathbf{0}$ for some matrix norm $\|\cdot\|$.
3. $\rho(\mathbf{A}) < 1$.

**Corollary 1.115.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. If there is a matrix norm $\|\cdot\|$ satisfying $\|\mathbf{A}\| < 1$, then $\mathbf{A}$ converges.

**Theorem 1.116.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$.

1. The series $\sum_{k=0}^\infty \mathbf{A}^k$ converges if and only if $\mathbf{A}$ converge.

2. If $\mathbf{A}$ is convergent, then $\mathbf{I}_n - \mathbf{A}$ is non-singular and moreover:
$$(\mathbf{I}_n - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k$$

**Corollary 1.117.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. If there is a subordinated matrix norm $\|\cdot\|$ satisfying $\|\mathbf{A}\| < 1$, then $\mathbf{I}_n - \mathbf{A}$ is non-singular and moreover:
$$\frac{1}{1+\|\mathbf{A}\|} \leq \|(\mathbf{I}_n - \mathbf{A})^{-1}\| \leq \frac{1}{1-\|\mathbf{A}\|}$$

**Matrix condition number**

**Definition 1.118.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$. We define the *condition number $\kappa(\mathbf{A})$ of* $\mathbf{A}$ as:
$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$$

**Theorem 1.119.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{Ax} = \mathbf{b}$ be a system of linear equations and $\|\cdot\|$ be a subordinated matrix norm. Suppose we know $\mathbf{A}$ and $\mathbf{b}$ with absolute errors $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$, respectively. Therefore, we actually have to solve the system:
$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = (\mathbf{b} + \Delta\mathbf{b})$$
If $\|\Delta\mathbf{A}\| < \frac{1}{\|\mathbf{A}^{-1}\|}$, then:
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1 - \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|}\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right)$$

**Theorem 1.120.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ and $\|\cdot\|$ be a subordinated matrix norm. Then:

1. $\kappa(\mathbf{A}) \geq \rho(\mathbf{A})\rho(\mathbf{A}^{-1})$.
2. If $\mathbf{b}, \mathbf{z} \in \mathbb{R}^n$ are such that $\mathbf{Az} = \mathbf{b}$, then:
$$\|\mathbf{A}^{-1}\| \geq \frac{\|\mathbf{z}\|}{\|\mathbf{b}\|}$$
3. If $\mathbf{B} \in \mathcal{M}_n(\mathbb{R})$ is a singular matrix, then:
$$\kappa(\mathbf{A}) \geq \frac{\|\mathbf{A}\|}{\|\mathbf{A} - \mathbf{B}\|}$$

**Iterative methods**

**Definition 1.121.** Suppose we want to solve the system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^n$. We choose a matrix $\mathbf{N} \in \mathrm{GL}_n(\mathbb{R})$ and define $\mathbf{P} := \mathbf{N} - \mathbf{A}$. Then:
$$\mathbf{Ax} = \mathbf{b} \iff \mathbf{x} = \mathbf{N}^{-1}\mathbf{Px} + \mathbf{N}^{-1}\mathbf{b} =: \mathbf{Mx} + \mathbf{N}^{-1}\mathbf{b}$$
The matrix $\mathbf{M}$ is called the *iteration matrix*. This defines a fixed-point iteration in the following way:
$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b} \\ \mathbf{x}^{(0)} \quad \text{(initial approximation)} \end{cases}$$

**Theorem 1.122.** The iterative method $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ is convergent if and only if $\mathbf{M}$ is convergent and if and only if $\rho(\mathbf{M}) < 1$.

**Corollary 1.123.** If $\|\mathbf{M}\| < 1$ for some matrix norm, then the iterative method $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ is convergent.

**Definition 1.124.** We define the *rate of convergence $R$* of an iterative method $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ as:
$$R = -\log(\rho(\mathbf{M}))$$

**Proposition 1.125.** Let $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ be an iterative method to approximate the solution $\mathbf{x}$ of a system of equations $\mathbf{Ax} = \mathbf{b}$. Then, we have the following estimations:
$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \frac{\|\mathbf{M}\|^k}{1-\|\mathbf{M}\|}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \qquad \text{(a priori)}$$
$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \frac{\|\mathbf{M}\|}{1-\|\mathbf{M}\|}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \qquad \text{(a posteriori)}$$

**Definition 1.126.** Let $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$. We say $A$ is *strictly diagonally dominant by rows* if
$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$$
We say $A$ is *strictly diagonally dominant by columns* if
$$|a_{jj}| > \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}|$$

**Definition 1.127 (Jacobi method).** Let $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$ be such that $\prod_{i=1}^n a_{ii} \neq 0$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{Ax} = \mathbf{b}$ be a system of equations. *Jacobi method* consists in defining a matrix $\mathbf{N}$ (and consequently matrices $\mathbf{P}$ and $\mathbf{M}$ as defined above) in the following way:
$$\mathbf{N} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$
$$\mathbf{P} = \mathbf{N} - \mathbf{A} = \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{(n-1)n} \\ -a_{n1} & \cdots & -a_{n(n-1)} & 0 \end{pmatrix}$$
$$\mathbf{M} = \mathbf{N}^{-1}\mathbf{P} = \begin{pmatrix} 0 & \frac{-a_{12}}{a_{11}} & \cdots & \frac{-a_{1n}}{a_{11}} \\ \frac{-a_{21}}{a_{22}} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{-a_{(n-1)n}}{a_{(n-1)(n-1)}} \\ \frac{-a_{n1}}{a_{nn}} & \cdots & \frac{-a_{n(n-1)}}{a_{nn}} & 0 \end{pmatrix}$$
Note that the iterative method $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ can also be written as:
$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right) \quad \text{for } i = 1, \ldots, n.$$

**Theorem 1.128.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be such that $\prod_{i=1}^n a_{ii} \neq 0$ and $\mathbf{b} \in \mathbb{R}^n$. If $\mathbf{A}$ is strictly diagonally dominant by rows or columns, then Jacobi method applied to solve the system $\mathbf{Ax} = \mathbf{b}$ is convergent.

**Definition 1.129 (Gauß-Seidel method).** Let $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$ be such that $\prod_{i=1}^n a_{ii} \neq 0$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{Ax} = \mathbf{b}$ be a system of equations. *Gauß-Seidel method* consists in defining a matrix $\mathbf{N}$ (and consequently matrices $\mathbf{P}$ and $\mathbf{M}$ as defined above) in the following way:

$$\mathbf{N} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & a_{n(n-1)} & a_{nn} \end{pmatrix}$$

$$\mathbf{P} = \mathbf{N} - \mathbf{A} = \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{(n-1)n} \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$\mathbf{M} = \mathbf{N}^{-1}\mathbf{P}$$

Note that the iterative method $\mathbf{x}^{(k+1)} = \mathbf{M}\mathbf{x}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ can also be written as:

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

for $i = 1, \ldots, n$.

**Theorem 1.130.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be such that $\prod_{i=1}^n a_{ii} \neq 0$ and $\mathbf{b} \in \mathbb{R}^n$. If $\mathbf{A}$ is strictly diagonally dominant by rows, then Gauß-Seidel method applied to solve the system $\mathbf{Ax} = \mathbf{b}$ is convergent.

**Method 1.131 (Over-relaxation methods).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{Ax} = \mathbf{b}$ be a system of equations and $\alpha \in \mathbb{R}$ be a parameter (called *relaxation factor*). *Over-relaxation methods* consist in defining matrices $\mathbf{N}(\alpha)$, $\mathbf{P}(\alpha)$ and $\mathbf{M}(\alpha)$ as follows:

$$\mathbf{P}(\alpha) = \mathbf{N}(\alpha) - \mathbf{A} \qquad \mathbf{M}(\alpha) = \mathbf{N}(\alpha)^{-1}\mathbf{P}(\alpha)$$

Then, the iterative method can be written as:

$$\mathbf{x}^{(k+1)} = \mathbf{M}(\alpha)\mathbf{x}^{(k)} + \mathbf{N}(\alpha)^{-1}\mathbf{b}$$

**Method 1.132 (Successive over-relaxation method).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ be such that $\alpha \neq -1$ and $\mathbf{x}^{(k+1)} = \mathbf{N}^{-1}\mathbf{P}\mathbf{x}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ be an iterative method. *Successive over-relaxation method (SOR)* consists in defining

$$\mathbf{N}(\alpha) = (1+\alpha)\mathbf{N} \quad \text{and} \quad \mathbf{P}(\alpha) = \mathbf{P} + \alpha\mathbf{N}$$

because it must be true that $\mathbf{A} = \mathbf{N}(\alpha) - \mathbf{P}(\alpha)$. Then, the previous iteration becomes:

$$\mathbf{x}^{(k+1)} = \mathbf{N}(\alpha)^{-1}\mathbf{P}(\alpha)\mathbf{x}^{(k)} + \mathbf{N}(\alpha)^{-1}\mathbf{b}$$

**Definition 1.133.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ be such that $\alpha \neq -1$ and $\mathbf{x}^{(k+1)} = \mathbf{N}(\alpha)^{-1}\mathbf{P}(\alpha)\mathbf{x}^{(k)} + \mathbf{N}(\alpha)^{-1}\mathbf{b}$ be a SOR method. Since $\mathbf{M}(\alpha) = \mathbf{N}(\alpha)^{-1}\mathbf{P}(\alpha)$, we have that

$$\mathbf{M}(\alpha) = \frac{1}{1+\alpha}(\mathbf{M} + \alpha\mathbf{I}_n)$$

and therefore:

$$\sigma(\mathbf{M}(\alpha)) = \left\{\frac{\lambda+\alpha}{1+\alpha} : \lambda \in \sigma(\mathbf{M})\right\}$$

**Theorem 1.134.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{x}^{(k+1)} = \mathbf{M}\mathbf{x}^{(k)} + \mathbf{N}^{-1}\mathbf{b}$ be an iterative method. Suppose that the eigenvalues $\lambda_i$, $i = 1, \ldots, n$, of $\mathbf{M}$ are all real and satisfy:

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n < 1$$

Then, the associated SOR method given by $\mathbf{N}(\alpha) = (1+\alpha)\mathbf{N}$ and $\mathbf{P}(\alpha) = \mathbf{P} + \alpha\mathbf{N}$ converges for $\alpha > -\frac{1+\lambda_1}{2}$. Moreover, $\rho(\mathbf{M}(\alpha))$ is minimum whenever $\alpha = -\frac{\lambda_1+\lambda_n}{2}$.

### Eigenvalues and eigenvectors

**Definition 1.135.** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be a matrix whose eigenvalues are $\lambda_1, \ldots, \lambda_n$. $\lambda_1$ is called *dominant eigenvalue of* $\mathbf{A}$ if $|\lambda_1| > |\lambda_i|$ for $i = 2, \ldots, n$. The associated eigenvector to $\lambda_1$ is called *dominant eigenvector of* $\mathbf{A}$.

**Definition 1.136.** We say a matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ is *reducible* if $\exists \mathbf{P} \in \mathcal{M}_n(\mathbb{R})$ a permutation matrix, such that

$$\mathbf{PAP}^{-1} = \begin{pmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{F} & \mathbf{G} \end{pmatrix}$$

for some square matrices $\mathbf{E}$ and $\mathbf{G}$ and for some other matrix $\mathbf{F}$. A matrix is *irreducible* if it is not reducible.

**Theorem 1.137 (Perron-Frobenius theorem).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be a non-negative irreducible matrix. Then, $\rho(\mathbf{A})$ is a real number and it is the dominant eigenvalue.

**Method 1.138 (Power method).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. For simplicity, suppose $\mathbf{A}$ is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Suppose $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. The *power method* consists in finding an approximation of the dominant eigenvalue $\lambda_1$ starting from an initial approximation $\mathbf{x}^{(0)}$ of $\mathbf{v}_1$. We define:

$$\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)} \qquad k \geq 0$$

Suppose $\mathbf{x}^{(0)} = \sum_{i=1}^n \alpha_i\mathbf{v}_i$. If we denote by $\mathbf{v}_{i,m}$ the $m$-th component of the vector $\mathbf{v}_i$ and choose $\ell$ such that $\mathbf{v}_{1,\ell} \neq 0$. Then:

$$\lim_{k\to\infty} \frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \mathbf{v}_1 \qquad \lim_{k\to\infty} \frac{\mathbf{x}_\ell^{(k+1)}}{\mathbf{x}_\ell^{(k)}} = \lambda_1$$

provided that $\alpha_1 \neq 0$. More precisely we have:

$$\frac{\mathbf{x}_\ell^{(k+1)}}{\mathbf{x}_\ell^{(k)}} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

**Method 1.139 (Normalized power method).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ and $\|\cdot\|$ be a vector norm[15]. For simplicity suppose $\mathbf{A}$ is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Suppose $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. The *normalized power method* consists in defining

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|} \quad \mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{y}^{(k)} \quad \text{for } k \geq 0$$

Suppose $\mathbf{x}^{(0)} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ such that $\alpha_1 \neq 0$. If we choose $\ell$ such that $\mathbf{v}_{1,\ell} \neq 0$. Then:

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{v}_1 \qquad \lim_{k\to\infty} \frac{\mathbf{x}_\ell^{(k+1)}}{\mathbf{y}_\ell^{(k)}} = \lambda_1$$

More precisely we have:

$$\frac{\mathbf{x}_\ell^{(k+1)}}{\mathbf{y}_\ell^{(k)}} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

**Method 1.140 (Rayleigh quotient).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. Suppose we have a power method $\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)}$ to approximate the dominant eigenvalue $\lambda_1$ of $\mathbf{A}$. Then *Rayleigh quotient* approximates $\lambda_1$ as follows:

$$\lim_{k\to\infty} \frac{\left(\mathbf{x}^{(k+1)}\right)^{\mathrm{T}} \cdot \mathbf{x}^{(k)}}{\left(\mathbf{x}^{(k)}\right)^{\mathrm{T}} \cdot \mathbf{x}^{(k)}} = \lambda_1$$

More precisely:

$$\frac{\left(\mathbf{x}^{(k+1)}\right)^{\mathrm{T}} \cdot \mathbf{x}^{(k)}}{\left(\mathbf{x}^{(k)}\right)^{\mathrm{T}} \cdot \mathbf{x}^{(k)}} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)$$

If instead of a power method, we have a normalized power method $\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}$, $\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{y}^{(k)}$, then:

$$\lim_{k\to\infty} \frac{\left(\mathbf{x}^{(k+1)}\right)^{\mathrm{T}} \cdot \mathbf{y}^{(k)}}{\left(\mathbf{y}^{(k)}\right)^{\mathrm{T}} \cdot \mathbf{y}^{(k)}} = \lambda_1$$

**Method 1.141 (Inverse power method).** Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ and $\mu \in \mathbb{C}$. The *inverse power method* consists in finding an approximation of the eigenvalue $\lambda$ closest to $\mu$ starting from an initial approximation $\mathbf{x}^{(0)}$ of its associated eigenvector $\mathbf{v}$. So we applied the power method to the matrix $(\mathbf{A} - \mu\mathbf{I}_n)^{-1}$. That is, we have the recurrence:

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|} \quad \mathbf{x}^{(k+1)} = (\mathbf{A} - \mu\mathbf{I}_n)^{-1}\mathbf{y}^{(k)} \quad \text{for } k \geq 0$$

Or, equivalently,

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|} \quad (\mathbf{A} - \mu\mathbf{I}_n)\mathbf{x}^{(k+1)} = \mathbf{y}^{(k)} \quad \text{for } k \geq 0$$

Therefore, in each step we have to solve a system of equations to obtain $\mathbf{x}^{(k+1)}$. Finally[16], if we choose $\ell$ such that $\mathbf{v}_\ell \neq 0$, then:

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{v} \qquad \lim_{k\to\infty} \frac{\mathbf{x}_\ell^{(k+1)}}{\mathbf{y}_\ell^{(k)}} = \frac{1}{\lambda - \mu}$$ [17]

### Exact methods

**Method 1.142 (Gaussian elimination).** Let $\mathbf{A} = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$. We define $a_{ij}^{(1)} := a_{ij}$ for $i, j = 1, \ldots, n$ and

$$\mathbf{A}^{(1)} := \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{(n-1)n}^{(1)} \\ a_{n1}^{(1)} & \cdots & a_{n(n-1)}^{(1)} & a_{nn}^{(1)} \end{pmatrix}$$

For $i = 2, \ldots, n$ we define $m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ to transform the matrix $\mathbf{A}^{(1)}$ into a matrix $\mathbf{A}^{(2)}$ defined by $a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}$ for $i = 2, \ldots, n$ and by $a_{ij}^{(1)}$ for $i = 1$. That is, we obtain a matrix of the form:

$$\mathbf{A}^{(1)} \sim \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{(n-1)n}^{(2)} \\ 0 & a_{n2}^{(2)} & \cdots & a_{n(n-1)}^{(2)} & a_{nn}^{(2)} \end{pmatrix} =: \mathbf{A}^{(2)}$$

Proceeding analogously creating multipliers $m_{ij}$, $i > j$, to echelon the matrix $\mathbf{A}$, at the end we will obtain an upper triangular matrix $\mathbf{A}^{(n)}$ of the form:

$$\mathbf{A}^{(n)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} & \cdots & a_{3n}^{(3)} \\ 0 & 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & a_{(n-1)(n-1)}^{(n-1)} & a_{(n-1)n}^{(n-1)} \\ 0 & 0 & 0 & \cdots & 0 & a_{nn}^{(n)} \end{pmatrix}$$

**Method 1.143.** *Partial pivoting* method in gaussian elimination consists in selecting as the pivot element the entry with largest absolute value from the column of the matrix that is being considered.

**Method 1.144.** *Complete pivoting* method in gaussian elimination interchanges both rows and columns in order to use the largest element (by absolute value) in the matrix as the pivot.

**Definition 1.145 (LU descompostion).** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ be a matrix. A *LU decomposition of* $\mathbf{A}$ is an expression $\mathbf{A} = \mathbf{L}\mathbf{U}$, where $\mathbf{L} = (\ell_{ij}), \mathbf{U} = (u_{ij}) \in \mathcal{M}_n(\mathbb{R})$

[15] For power method it is recommended to use $\|\cdot\|_\infty$.

[16] Alternatively, here we could have applied the Rayleigh quotient.

[17] There's another method that applies power method to the matrix $\mathbf{A} - \mu\mathbf{I}_n$ with the same purpose as the inverse power method but without having to solve a system of equations in each iteration. In this case, this method gives the farthest eigenvalue of $\mathbf{A}$ from $\mu$.

are matrices of the form:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \ell_{n1} & \cdots & \ell_{n(n-1)} & 1 \end{pmatrix} \tag{1}$$

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & u_{(n-1)n} \\ 0 & \cdots & 0 & u_{nn} \end{pmatrix} \tag{2}$$

**Lemma 1.146.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{A}\mathbf{x} = \mathbf{b}$ be a system of linear equations. Suppose $\mathbf{A} = \mathbf{L}\mathbf{U}$ for some matrices $\mathbf{L}, \mathbf{U} \in \mathcal{M}_n(\mathbb{R})$ of the form of (1) and (2), respectively. Then, to solve the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ we can proceed in the following way:

1. Solve the triangular system $\mathbf{L}\mathbf{y} = \mathbf{b}$.
2. Solve the triangular system $\mathbf{U}\mathbf{x} = \mathbf{y}$.

**Proposition 1.147.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$. Then:

1. If LU decomposition exists, it is unique.
2. If we can make the gaussian elimination without pivoting rows, then[18]:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & m_{n(n-1)} & 1 \end{pmatrix} \quad \mathbf{U} = \mathbf{A}^{(n)}$$

**Definition 1.148.** A *permutation matrix* is a square binary matrix that has exactly one entry of 1 in each row and each column and 0 elsewhere.

**Proposition 1.149.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$. Then, there exist a permutation matrix $\mathbf{P} \in \mathcal{M}_n(\mathbb{R})$ and matrices $\mathbf{L}, \mathbf{U} \in \mathcal{M}_n(\mathbb{R})$ of the form of (1) and (2), respectively, such that:

$$\mathbf{P}\mathbf{A} = \mathbf{L}\mathbf{U}$$

**Definition 1.150 (QR descompostion).** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ be a matrix. A *QR decomposition of* $\mathbf{A}$ is an expression $\mathbf{A} = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q}, \mathbf{R} \in \mathcal{M}(\mathbb{R})$ are such that $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular.

**Lemma 1.151.** Let $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{A}\mathbf{x} = \mathbf{b}$ be a system of linear equations. Suppose $\mathbf{A} = \mathbf{Q}\mathbf{R}$ for some orthogonal matrix $\mathbf{Q}$ and some upper triangular matrix $\mathbf{R}$, both of size $n$. Then, solve the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is equivalent to solve the triangular system $\mathbf{R}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{b}$.

**Lemma 1.152.** Let $\mathbf{Q}$ be an orthogonal matrix. Then:

1. $\det \mathbf{Q} = \pm 1$.
2. $\|\mathbf{Q}\|_2 = 1$.

[18] In practice, LU decomposition is implemented making gaussian elimination and storing the values $m_{ij}$ in the position $ij$ of the matrix $\mathbf{A}^{(k)}$, where there should be a 0.