APPLICABLE APPROXIMATION THEORY

LUISA FERMO

Abstract. Approximation Theory. Algebraic approximation: Taylor's polynomial, Newton's polynomial, Lagrange's polynomial. Uniform approximation by Lagrange polynomial. Spline approximation: cubic spline. Numerical Integration. Newton-Cotes formulae: rectangle rule, trapezoidal rule, Simpson rule. Weighted interpolatory formula: Gaussian formula. Integral equations. Fredholm integral equations of the second kind. Numerical methods.

Contents

Version January 20, 2023

1. Approximation Theory

The Approximation Theory is the branch of the mathematical analysis studying methods for approximating some mathematical objects by others, questions related to the research of such objects and estimation of the error that arises here. Many areas of mathematics itself make use of quantities that are too complicated, too difficult, and even too abstract to work with directly. Hence, the goal of the approximation theory is to discover and analyze simple, easy to work with, concrete quantities that can do a good, efficient job in their place - for example, polynomials, splines and so on. As we might guess from its name, the approximation theory has both a theoretical side, which is more often concerned with existence and uniqueness questions, and a practical side, which is concerned largely with computational practicalities and precise estimations of error.

In this notes we are going to study how a given function f belonging to an infinite dimension space F can be approximated by another simpler one f_n belonging to a sequence of finite dimensional subspace of $\{\mathbb{F}_n\}_n \subset \mathbb{F}$ such that f_n is close in some sense to f. More precisely, we will consider the case where the function f belongs to the space $C^k([a,\, b])$, $k \geq 0$ of k-times continuous-differentiable functions on $[a,\, b] \subset \mathbb{R}$.

In order to approximate it we have at first to identify a class $\mathbb{F}_n = \{f_n(x)\}$ of all possible approximating functions. In this context we will consider the following subspaces $\mathbb{F}_n$:

(a) the set of all algebraic polynomials of degree at most n

$$\mathbb{P}_n = \{f_n(x) = a_0 + a_1 x + \cdots + a_n x^n\}.$$

(b) the set of all spline functions of order d

$$\mathbb{S}_{n,d} = \{f_n = \text{piecewise-defined by polynomial functions of local degree } d\}.$$

Once this class is fixed, we have to choose an element $f_n \in \mathbb{F}_n$. The criterion we will adopt is the classical approach of the interpolation that is prescribe that the values of f_n in a certain number of distinct points x_i coincide with those of the functions to be approximated y_i that is

$$f_n(x_i) = y_i, \qquad i = 0, 1, \ldots, m.$$

Finally, we have to "measure the goodness" of such approximation by estimating the following norm

$$\|f - f_n\|_\infty := \max_{x \in [a,\, b]} |f(x) - f_n(x)| \tag{1.1}$$

and verifying that

$$\lim_{n \to \infty} \|f - f_n\|_\infty = 0.$$

If this last condition is satisfied, we will say that the sequence $\{f_n\}_n$ converges to f in $C^k([a,\, b])$ in uniform norm.

1.1. **Algebraic approximation.**

The approximation of functions by means of algebraic polynomials is based on the well-known Weierstrass's Theorem which states that any continuous function f can be approximated uniformly by polynomials, no matter how badly behaved f may be on $[a,\, b]$.

Theorem 1.1 (Weierstrass's Theorem)**.** *Let $f \in C^0([a,\, b])$. For every $\epsilon > 0$, there exists a polynomial P such that*

$$|f(x) - P(x)| < \epsilon. \tag{1.2}$$

Definition 1.2. *We define the* **error of best polynomial approximation** *of f by means of algebraic polynomials of degree at most n the quantity*

$$E_n(f) := \inf_{P_n \in \mathbb{P}_n} \|f - P_n\|_\infty. \tag{1.3}$$

Let us note that according to Theorem 1.1, if $f \in C^0([a,\, b])$ we can deduce that $\lim_{n \to \infty} E_n(f) = 0$. Moreover, if $f \in C^0([a,\, b])$ there exists only one polynomial $P_n^* \in \mathbb{P}_n$ such that

$$E_n(f) = \inf_{P_n \in \mathbb{P}_n} \|f - P_n\|_\infty = \|f - P_n^*\|_\infty,$$

and P_n^* is said the **polynomial of best approximation**.

Given a function f is not simple to find its polynomial of best approximation. However the estimate of the error of best approximation could be a useful tool to state which is the best approximation that we could aspect.

The error of best approximation depends on the smoothness properties of the function f we would like to approximate. Indeed the following theorem holds true.

Theorem 1.3. *Let $f \in C^k([a,\, b])$, $k \geq 1$. The following estimate for the error of best approximation holds true*

$$E_n(f) \leq \frac{\mathcal{C}}{n^k} \tag{1.4}$$

where $\mathcal{C}$ is a constant independent of n and k.

Let us remark that if $f \in C^0([a,b])$, we only know that $\lim_{n\to\infty} E_n(f) = 0$ but we do not have an estimate. However, it is possible to prove that if $f \in C^{k+\alpha}([a,b])$ where $C^{k+\alpha}([a,b])$ is the set of all k-times continuous-differentiable functions on $[a,\, b] \subset \mathbb{R}$ such that $f^{(k)}$ is a Lipschitz continuous function of order α with $0 < \alpha \leq 1$ [1], then

$$E_n(f) \leq \frac{\mathcal{C}}{n^{k+\alpha}}, \tag{1.5}$$

where $\mathcal{C}$ is a constant independent of n, k and α.

Example 1.4. Let us consider $f(x) = e^{|x|^{3/2}} \in C^{1+\frac{1}{2}}([-1,1])$. Then, if we want to approximate such a function with seven correct digits, according to (1.5), we have to choose an integer n such that

$$E_n(f) \leq \frac{\mathcal{C}}{n^{3/2}} \leq \frac{\mathcal{C}}{10^8}, \quad \Longrightarrow \text{approximately} \quad n \geq 10^{16/3}.$$

Example 1.5. Let us consider $g(x) = |\cos x|^{\frac{7}{2}} e^{-\cos^2 x} \in C^{3+\frac{1}{2}}([-\pi,\pi])$. Thus, in order to get an approximation of such a function with five correct digits, we have to choose a polynomial of degree n such that

$$E_n(f) \leq \frac{\mathcal{C}}{n^{7/2}} \leq \frac{\mathcal{C}}{10^6}, \quad \Longrightarrow \text{approximately} \quad n \geq 10^2.$$

1.1.1. *Taylor polynomial.* Let us consider a function $f \in C^{n+1}([a,\, b])$ and a point $x_0 \in [a,\, b]$.

It is well-known that a first approximation of this function could be furnished by means of the Taylor polynomial of degree n centered at x_0

$$f(x) = T_n(x) + R_{n+1}(f,x)$$

where

$$T_n(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \tag{1.6}$$

is the Taylor polynomial and

$$R_n(f,x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}, \qquad \xi \in (x_0,\, x)$$

is the remainder term in the Lagrange form.

About this latter term, if the derivatives of f are such that

$$|f^{(k)}(x)| \leq \ell, \qquad \forall x \in [a,\, b],\, \forall k = 0, 1, \ldots, n+1$$

then we have

$$|R_{n+1}(f,x)| \leq \ell \frac{|x-x_0|^{n+1}}{(n+1)!}, \qquad \forall x \in [a,\, b] \setminus \{x_0\}$$

[1] Let $g \in C^0([a,\, b])$. We say that g is a Lipschitz continuous function of order $\alpha \geq 0$ if there exists a constant $M > 0$ such that

$$|f(x) - f(y)| \leq M|x-y|^\alpha, \qquad \forall x, y \in [a,\, b].$$

and consequently

$$\lim_{n\to+\infty} |R_{n+1}(f,x)| = 0, \qquad \forall x \in [a,\, b] \setminus \{x_0\}.$$

Example 1.6. Let us consider the function $f(x) = e^x$. The Taylor expansion of f centered at $x = x_0$ is

$$e^x = e^{x_0} + e^{x_0}(x - x_0) + \cdots + e^{x_0}\frac{(x-x_0)^n}{n!} + R_n(f,x), \qquad R_n(f,x) = e^{\xi}\frac{(x-x_0)^{n+1}}{(n+1)!}. \tag{1.7}$$

Assume that we want to know the approximated value of the Euler number e. By using (1.7) with $x_0 = 0$ and taking into account that for $0 < \xi < 1$ we have

$$|f(1) - T_n(1)| = |R_n(f,1)| = \frac{e^{\xi}}{(n+1)!} < \frac{e}{(n+1)!} < \frac{3}{(n+1)!}$$

we can deduce that if $n = 10$ we get our approximation with seven correct digits being

$$R_n(f,1) \le \frac{3}{11!} \simeq 7.5 \times 10^{-8}.$$

Indeed we have

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{10!} = 2.7182818$$

which is the exact value of the Euler number up to the seven correct digits.

Example 1.7. Let us consider the function $f(x) = \cos(x)$ whose Taylor expansion of f centered at $x = 0$ is

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n\frac{x^{2n}}{(2n)!} + R_{2n+1}(f,x) \tag{1.8}$$

where

$$|R_{2n+1}(f,x)| = \left|\sin\xi\frac{x^{2n+1}}{(2n+1)!}\right| \le \frac{|x^{2n+1}|}{(2n+1)!}. \tag{1.9}$$

Polynomial (1.8) is a good approximation for f in a neighborhood of the initial point $x_0 = 0$. Indeed, if we fix $n = 4$ and $x = 1$, we get

$$\cos(1) = 1 - \frac{1}{2} + \frac{1}{4!} - \frac{1}{6!} + \frac{1}{8!} = 0.5403025793650793.$$

Then, as the exact value is $f^*(1) = 0.5403023058681398$, we can see that we get five correct digits. Such numerical result confirms the theoretical estimate since by (1.9) we can deduce that $|R_9(f,1)| \le 2.48 \times 10^{-5}$. On the other hand, if we fix $n = 4$ and $x = 3$, we get

$$\cos(3) = -0.9747767857142857$$

that is a bad approximation for $\cos(3)$, as the exact value is$f^*(3) = -0.9899924966004454$.

The Taylor polynomial is a good approximation from a theoretical point of view but practically unusable for a practical point of view. In order to approximate a function with such a polynomial we have to compute all the $n+1$ derivatives and then evaluate it at the initial point x_0. This could be simple if we have an elementary function but if we want to approximate for instance $g(x) = \dfrac{(1+x^2+3x)^2\cos(3x+1)}{1+x+e^x}$ such computation could be very hard. Moreover often we would like to approximate functions which are only continuous for example $f(x) = \sqrt{1-x^2}$ or functions whose analytical expression is not known and then it is impossible to use the Taylor polynomial.

1.1.2. *Algebraic interpolation.* Let f be a continuous function on $[a,b]$ and assume we know the values $f(x_1)$, $f(x_2)$, ..., $f(x_n)$. We seek a polynomial $P(x)$ such that the following **interpolation conditions** are satisfied

$$P(x_i) = f(x_i), \quad i = 1, 2, ..., n. \tag{1.10}$$

The following theorem allow us to state that such a polynomial is unique.

Theorem 1.8. *Given n distinct points $x_1, x_2, ..., x_n$ and arbitrary values $f(x_1), f(x_2), ..., f(x_n)$ there is at most one polynomial P of degree less or equal to $n-1$ such that the interpolation conditions (1.10) hold true. A polynomial that satisfies these conditions is called* **interpolating polynomial** *and the points x_i are called* **interpolation points or interpolation nodes**.

Proof. Let us consider a polynomial of degree $n-1$

$$P(x) = a_0 + a_1 x + ... + a_{n-1} x^{n-1}.$$

By requiring conditions (1.10) we get

$$\begin{cases} a_0 + a_1 x_1 + ... + a_{n-1} x_1^{n-1} = f(x_1) \\ a_0 + a_1 x_2 + ... + a_{n-1} x_2^{n-1} = f(x_2) \\ ..... \\ a_0 + a_1 x_n + ... + a_{n-1} x_n^{n-1} = f(x_n) \end{cases}$$

which can be rewritten as

$$\begin{pmatrix} 1 & x_1 & .... & x_1^{n-1} \\ 1 & x_2 & .... & x_2^{n-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}. \tag{1.11}$$

The matrix of the system is the so-called Vandermonde matrix whose determinant is given by

$$det \begin{pmatrix} 1 & x_1 & .... & x_1^{n-1} \\ 1 & x_2 & .... & x_2^{n-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{pmatrix} = \prod_{j=1}^{n-1} \left(\prod_{i=j+1}^{n} (x_i - x_j) \right) \neq 0 \quad \text{if and only if} \quad x_i \neq x_j. \tag{1.12}$$

Then, system (1.11) has a unique solution. □

The proof of the previous theorem allow us to construct an interpolating polynomial in the canonical basis $\{1, x, x^2, ..., x^{n-1}\}$ of the form

$$P(x) = \sum_{i=0}^{n-1} a_i x^i.$$

It is sufficient solve system (1.11) to get the coefficients. By the way, system (1.11) generally require about $\mathcal{O}(n^3)$ operations. In addition, the Vandermonde matrix is notorious for being challenging to solve (especially with Gauss elimination) and prone to large errors in the computed coefficients a_i when n is large and/or $x_i \neq x_j$. Several authors have therefore proposed algorithms that allow us to construct an interpolating polynomial without solve any system.

1.1.3. *Newton's polynomial.* In 1670 Isac Newton wrote the uniquely determined interpolation polynomial of Theorem 1.8 in the basis

$$\left\{ e_j = \prod_{k=1}^{j} (x - x_k) \right\}.$$

More precisely, the **Newton form of the interpolation polynomial** of degree $n-1$ is

$$N_n(x) = a_1 + \sum_{j=2}^{n} a_j \prod_{k=1}^{j-1} (x - x_k) \tag{1.13}$$

where the coefficients a_j are given by

$$a_1 = f(x_1) \quad \text{and} \quad a_j = \frac{f(x_j) - N_{j-1}(x_j)}{\prod_{k=1}^{j-1} (x_j - x_k)}, \quad j = 2, 3, ..., n. \tag{1.14}$$

Thus, according to (1.13), the polynomial of degree 0 which interpolates $(x_1, f(x_1))$ is

$$N_1(x) = f(x_1);$$

while the polynomial of degree 1 which interpolates the points $(x_1, f(x_1)), (x_2, f(x_2))$ reads as

$$N_2(x) = f(x_1) + \frac{f(x_2) - f(x_1)}{(x_2 - x_1)} (x - x_1).$$

In general, the jth coefficient a_j is also said **jth order divided difference** and it depends on the values $x_1, ..., x_j$ and the values $f(x_1), ..., f(x_j)$. To emphasize such dependence, sometimes the coefficients are also written in the following form

$$\begin{aligned}
a_1 &= [x_1; f] = f(x_1) \\
a_2 &= [x_1, x_2; f] = \frac{f(x_2) - f(x_1)}{(x_2 - x_1)} \\
a_3 &= [x_1, x_2, x_3; f] = \frac{[x_2, x_3; f] - [x_1, x_2; f]}{(x_3 - x_1)} \\
&\vdots \qquad\qquad\qquad \vdots \\
a_n &= [x_1, ..., x_n; f] = \frac{[x_2, ..., x_n; f] - [x_1, ..., x_{n-1}; f]}{(x_n - x_1)}.
\end{aligned}$$

Then, according to this notation, the so-called **Newton's divided difference interpolation polynomial** is given by

$$N_n(f, x) = f(x_1) + \sum_{j=2}^{n} [x_1, ..., x_j; f] \prod_{k=1}^{j} (x - x_k). \tag{1.15}$$

Let us note that when two arguments are equal we have

$$[x_0, x_0; f] = \lim_{x \to x_0} [x_0, x; f] = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

and in general

$$\lim_{x_i \to x_0} [x_0, ..., x_i; f] = \frac{f^{(i)}(x_0)}{i!}.$$

Then we get

$$\lim_{x_i \to x_1} N_n(f,x) = f(x_1) + \sum_{i=1}^{n-1} (x-x_1)^i \frac{f^{(i)}(x_1)}{i!}$$

that is the Taylor's polynomial of degree $n-1$ centered at x_1.

1.1.4. *Lagrange's polynomial.* In 1795 Joseph Louis Lagrange discovered the base of the so-called **fundamental Lagrange polynomials**

$$\left\{ l_k(x) = \prod_{\substack{j=1 \\ j \neq k}}^{n} \frac{(x-x_j)}{(x_k-x_j)} \right\}, \tag{1.16}$$

which are polynomials of degree $n-1$ such that for each $j,k = 1,2,...,n$

$$l_k(x_j) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} \tag{1.17}$$

Alternatively, the fundamental Lagrange polynomials can also be written as

$$l_k(x) = \frac{\pi_n(x)}{\pi_n'(x_k)(x-x_k)} \tag{1.18}$$

where $\pi_n(x) = \prod_{k=1}^{n} (x-x_k)$ is a polynomial of degree n having all the interpolation nodes as zeros. Then, he introduced the so-called **Lagrange interpolation polynomial**

$$L_n(f,x) = \sum_{k=1}^{n} l_k(x) f(x_k), \tag{1.19}$$

that is a polynomial of degree $n-1$ interpolating the function f at the interpolation nodes $x_1, x_2, ..., x_n$.

The Lagrange polynomial can be seen as a linear operator

$$\begin{aligned} L_n : C^0([a,b]) &\to \mathbb{P}_{n-1} \\ f \quad &\to L_n(f) \end{aligned}$$

such that

$$L_n(P_{n-1}, x) = P_{n-1}(x), \quad \text{for each} \quad P_{n-1} \in \mathbb{P}_{n-1}.$$

In other words the Lagrange polynomial project a continuous function on the space of the polynomials of degree at most $n-1$.

From a computational point of view, the Lagrange polynomial and the Newton polynomial are equivalent. In both cases the computational cost is of the order $\mathcal{O}(n^2)$ but the Lagrange polynomial is numerical stable even if the nodes are very close.

1.2. **Uniform Approximation by Lagrange polynomial.** Let f be a continuous function on $[-1, 1]$ and assume we have a sequence of monic polynomials $\{q_n\}_{n \in \mathbb{N}}$ such that for each $n \geq 1$ the polynomial q_n has n distinct zeros $x_{n,k}$, $k = 1,2,...,n$ in $[a,b]$, i.e.

$$-1 \leq x_{n,1} < x_{n,2} < ... < x_{n,n} \leq 1.$$

The generalization to the case of a generic interval $[a,b]$ is straightforward.

Let X be the corresponding infinite triangular array of these zeros

$$X = \begin{pmatrix} x_{1,1} & & & & \\ x_{2,1} & x_{2,2} & & & \\ x_{3,1} & x_{3,2} & x_{3,3} & & \\ \cdots & \cdots & \cdots & & \\ x_{n,1} & \cdots & \cdots & x_{n,n} & \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

i.e. a matrix in which the n-th row consists of the zeros of the polynomial q_n. We will call the matrix X **the interpolation matrix** or **the system of interpolation nodes**.

We associate to $\{q_n\}_{n\in\mathbb{N}}$ and then to the matrix X, a sequence of Lagrange polynomial $\{L_n(X,f)\}_{n\in\mathbb{N}}$ defined by

$$L_n(X,f,x) = \sum_{k=1}^{n} l_{n,k}(x) f(x_{n,k}), \quad k = 1,2,...,n.$$

Obviously $L_n(X,f) \in \mathbb{P}_{n-1}$ and the index n in $L_n(X,f)$ denotes the number of the nodes. From now on, to simplify the notation, for a fixed n, we will set $x_{n,k} := x_k$ and $l_{n,k}(X,x) := l_k(x)$. In definitive, for a given interpolation array X, we define a sequence operators $L_n(X) : C^0([-1,1]) \to \mathbb{P}_{n-1}$ called **Lagrange interpolatory process** such that $L_n(X,f,x) = f(x)$ if $f \in \mathbb{P}_{n-1}$.

The main question is the convergence $L_n(X,f) \to f$ when $n \to \infty$ and more precisely what kind of interpolation array X provides this convergence. Denoting by P_{n-1} the polynomial of best approximation in uniform norm we have

$$\begin{aligned} |(f(x) - L_n(X,f,x))| &\leqslant |f(x) - P_{n-1}(x)| + |(L_n(X,f,x) - P_{n-1}(x))| \\ &\leqslant |(f(x) - P_{n-1}(x))| + |L_n(X, f - P_{n-1}, x)| \\ &\leqslant E_{n-1}(f) \left(1 + \sum_{k=1}^{n} |l_k(X,x)|\right) \end{aligned}$$

from which we can deduce that

$$|(f(x) - L_n(X,f,x))| \leq E_{n-1}(f)\,(1 + \Lambda_n(X,x)) \tag{1.20}$$

where

$$\Lambda_n(X,x) = \sum_{k=1}^{n} |l_k(X,x)| \tag{1.21}$$

is the so-called **Lebesgue function**. Then, taking the maximum in (1.20) over $[-1,1]$, we get

$$\|f - L_n(X,f)\| \leq E_{n-1}(f)\,(1 + \|\Lambda(X)\|) \tag{1.22}$$

where

$$\|\Lambda_n(X)\| = \max_{x\in[-1,1]} \sum_{k=1}^{n} |l_k(X,x)| \tag{1.23}$$

are the **Lebesgue constants of the Lagrange interpolation**.

Hence, according to estimate (1.20) and (1.22), the Lebesgue constant and more in general the Lebesgue function play an important role in the convergence of the Lagrange polynomials

In 1914 Faber [5] proved that

$$\|\Lambda_n(X)\| \geq \frac{1}{12} \log n, \quad n \geq 1, \tag{1.24}$$

for any interpolation matrix X. Therefore according to (1.24) the Lebesgue constants are unbounded and for particular choices of the interpolation array X they can take very large values. This is the case when the interpolatory matrix consists of the equidistant nodes on $[-1,1]$,

$$x_k = -1 + 2\frac{k-1}{n-1} \quad k = 1, 2, ..., n.$$

In fact in 1916 Bernstein [2,3] showed by an example the pointwise divergence properties of the Lagrange interpolation on these type of nodes (in 1901 Runge already proved that this choice was bad), and in 1917 Tureckii [17] proved the following asymptotic estimate for the Lebesgue constants

$$\|\Lambda_n(X)\| \sim \frac{2^n}{e\, n\, \log n}.$$

This circumstance can strongly influence the numerical computation, where we handle with perturbed values of f and we compute the polynomial $L_n(X, f+\eta)$, where η is a perturbation of the function f deriving from the evaluation of $f(x_{n,k})$. In this case the actual error is

$$\|(f - L_n(X, f+\eta))\| \leq \|f - L_n(X, f)\| + \|\Lambda_n(X)\|\eta_n$$

where $\eta_n = \max\limits_{1\leq k\leq n} |\eta(x_{n,k})|$. The first term is the theoretical error while the second term represents the numerical error that can be very large even if f is computed with the machine precision.

According to Faber inequality, we will say that the interpolatory array X is an **optimal system of nodes** if and only if there exists a constant $\mathcal{C} \neq \mathcal{C}(n)$ such that

$$\|\Lambda_n(X)\| \leq \mathcal{C} \log n, \quad n > 1. \tag{1.25}$$

Then in such case by (1.22) we get the following error estimate

$$\|f - L_n(X, f)\| \leq \mathcal{C} \log n E_{n-1}(f)$$

and thus under the assumption that $f \in C^k([-1,1])$ we have by applying Theorem (1.5)

$$\|f - L_n(X, f)\| \leq \mathcal{C}\, \frac{\log n}{n^k}.$$

Now we will show some optimal system of nodes.

To this end we consider as sequence of polynomial $\{q_n\}_{n\in\mathbb{N}}$, the sequence of orthonormal polynomials on $(-1,1)$ with respect to the Jacobi weight

$$v^{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$$

i.e. $\{p_n(v^{\alpha,\beta})\}$ where $p_n(w) = \gamma_n x^n + \gamma_{n-1} x^{n-1} + ...$ with $\gamma_n > 0$. Then instead of $L_n(X, f)$ and $\Lambda_n(X)$ we will use the notation $L_n(v^{\alpha,\beta}, f)$ and $\Lambda(v^{\alpha,\beta})$, respectively.

The following classical result of Szegö, whose proof can be found in [15], describes the behaviour of the Lebesgue constants.

Theorem 1.9. *For all $n \in \mathbb{N}$*

$$\|\Lambda_n(v^{\alpha,\beta})\| \sim \begin{cases} \log n, & \text{if } -1 < \alpha, \beta \leq -1/2 \\ n^{\max\{\alpha,\beta\}+\frac{1}{2}}, & \text{otherwise.} \end{cases} \tag{1.26}$$

where the constants in $\sim$ are independent of m.

So by this result we can deduce that only a restricted class of Jacobi polynomials $P_n(v^{\alpha,\beta})$ with $\alpha, \beta \leq -1/2$ gives an optimal interpolation process. Then, for instance, the zeros of the Legendre polynomial ($\alpha = \beta = 0$) and those of Chebyschev polynomial of the second kind ($\alpha = \beta = 1/2$) are not an optimal system. Nevertheless, it is possible to overcome this problem modifying the system.

To this end we construct two sequences of polynomials $\{Z_r = Z_{n,r}\}$ and $\{Y_s = Y_{n,s}\}$ of fixed degree r and s, respectively. Thus, denoting by $x_1 < x_2 < ... < x_n$, the zeros of $P_n(v^{\alpha,\beta})$ we define

$$y_j = -1 + (j-1)\frac{1+x_1}{s} \quad j = 1, 2, ..., s \tag{1.27}$$

$$z_i = x_n + i\frac{1-x_n}{r} \qquad i = 1, 2, ..., r \tag{1.28}$$

and we set

$$Y_s(x) = Y_{n,s}(x) = \prod_{j=1}^{s}(x - y_j), \quad Z_r(x) = Z_{n,r}(x) = \prod_{i=1}^{r}(x - z_i).$$

Hence introduce the polynomial

$$Q_{n+r+s}(x) = Y_{n,s}(x)P_n(v^{\alpha,\beta}, x)Z_{n,r}(x),$$

with zeros $\{y_j\}_{j=1}^{s} \cup \{x_j\}_{j=1}^{n} \cup \{z_j\}_{j=1}^{r}$ and denote by $L_{n,r,s}(v^{\alpha,\beta}, f) \in P_{n+r+s-1}$ the Lagrange polynomial interpolating f at the points

$$-1 < y_1 < ... < y_j < x_1 < ... < x_n < z_1 < ... < z_r < 1.$$

This polynomial can be written as

$$\begin{aligned} L_{n,r,s}(v^{\alpha,\beta}, f, x) &= (Y_sZ_r)(x)\sum_{k=1}^{n} l_k(x)\frac{f(x_k)}{(Y_sZ_r)(x_k)} \\ &+ Y_s(x)p_n(v^{\alpha,\beta}, x)\sum_{k=1}^{r}\bar{l}_k(x)\frac{f(z_k)}{Y_s(z_k)p_m(v^{\alpha,\beta}, z_k)} \\ &+ Z_r(x)p_n(v^{\alpha,\beta}, x)\sum_{k=1}^{s}\bar{\bar{l}}_k(x)\frac{f(y_k)}{Z_r(y_k)p_n(v^{\alpha,\beta}, y_k)}. \end{aligned}$$

The following theorem shows the behaviour of the Lebesgue constants $\|\Lambda_{n,r,s}(v^{\alpha,\beta})\|$ corresponding to the interpolation process $L_{n,r,s}(v^{\alpha,\beta}, f)$ [11, 14].

Theorem 1.10. *Let $\alpha, \beta > -1$ and r, s be non negative integers. The following inequality*

$$\|L_{n,r,s}(v^{\alpha,\beta}, f)\| \leq \mathcal{C}\log n\|f\| \quad \mathcal{C} \neq \mathcal{C}(n, f)$$

holds true if and only if

$$\frac{\alpha}{2} + \frac{1}{4} \leq r \leq \frac{\alpha}{2} + \frac{5}{4} \tag{1.29}$$

$$\frac{\beta}{2} + \frac{1}{4} \leq s \leq \frac{\beta}{2} + \frac{5}{4}. \tag{1.30}$$

Thus, we can see that if $\alpha = \beta = 0$ or $\alpha = \beta = 1/2$, by (1.29) we have that the zeros of $(1-x^2)P_n(v^{0,0}, x)$ or the zeros of $(1-x^2)P_n(v^{1/2,1/2}, x)$ are an optimal system of nodes. In Figure 1 have been represented the Lebesgue constants related to the Lagrange polynomial of degree 5 with different interpolation matrix.

Example 1.11. Let us consider the so-called Runge function $f(x) = \frac{1}{1+x^2}$. Figure 2 shows the behaviour of the Lagrange polynomial in the case when the interpolation nodes are the equispaced nodes (to the left) and the Chebyshev nodes of the first kind (to the right).

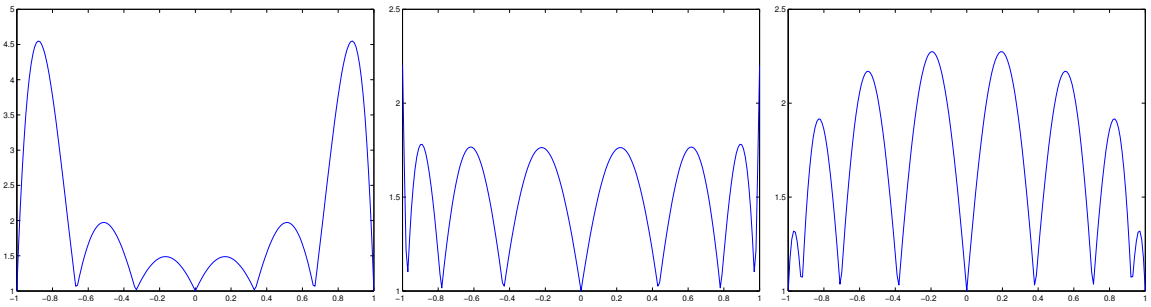

FIGURE 1. The Lebesgue function $\|\Lambda_6(X,x)\|$ in the case when the interpolation nodes are the equispaced nodes (to the left), the zeros of $p_6(v^{-1/2,-1/2})$ (to the center) and the zeros of $(1-x^2)p_6(v^{1/2,1/2})$ (to the right).

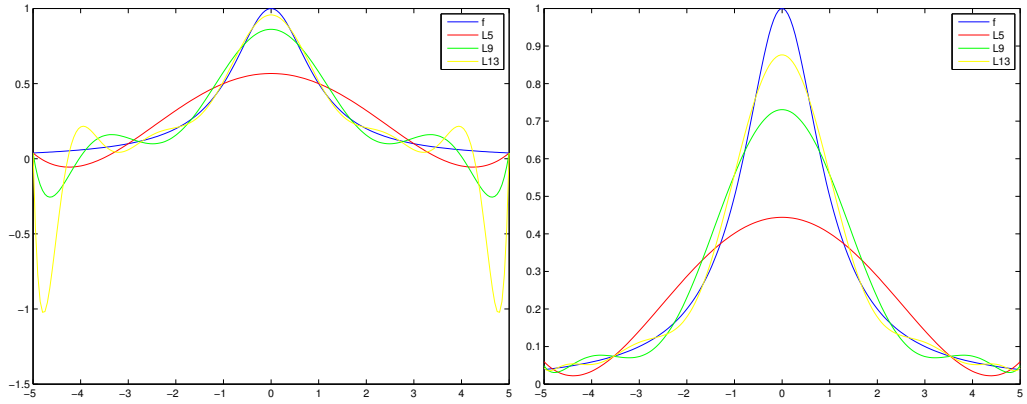

FIGURE 2. The Lagrange polynomial L_m for the Example 1.11 based on the equispaced nodes (to the left) and on the Chebyshev nodes of the first kind (to the right)

n	$\|\Lambda_n(\mathcal{E})\|$	$E_n(\mathcal{E},f)$	$\|\Lambda_n(v^{-1/2,-1/2})\|$	$E_n(v^{-1/2,-1/2},f)$
8	1.0938e+01	8.3640e-02	2.3618	1.9561e-02
16	9.2892e+02	6.2789e-01	2.7663	3.0459e-03
32	2.2944e+07	1.3179e+03	3.1885	4.3153e-04
64	3.7553e+16	1.7183e+11	3.6200	5.7819e-05
128	6.2741e+34	3.4065e+28	4.0563	7.5177e-06
256	2.1466e+71	1.4451e+64	4.9352	1.1742e-07

TABLE 1. Numerical results for Example 1.12

Example 1.12. Let us consider $f(x) = |x|^{5/2}$ with $x \in [-1,1]$. Table 1 shows the Lebesgue constant and the absolute error

$$E_n(X) = \max_{x\in[-1,1]} |f(x) - L_n(X,f)|$$

we have in the case when we approximate the function f by using the Lagrange polynomial based on the equidistant nodes $L_n(\mathcal{E},f)$ and the Lagrange polynomial based on the zeros of the Chebychev polynomial of the fist kind $L_n(v^{-1/2,-1/2},f)$.

1.3. **Spline Interpolation.** In the previous paragraphs we have seen that if we know a function $f \in C^k([a,b])$, we can choose an interpolation matrix X whose Lebesgue constant are as in (1.25) and we can construct a Lagrange interpolation process $L_n(X)$ such that

$$L_n(X,f) \to f, \qquad n \to \infty, \quad \forall x \in [a,b].$$

However, in different applications we do not have the analytical expression of the function f but we only know the values that it has in different points $x_1, x_2, ..., x_n$ of the interval $[a,b]$. Then, in such cases, it is useful to adopt the interpolation with piecewise polynomial function.

Definition 1.13. *Let $a = x_0 < x_1 < \cdots < x_n = b$ be a subdivision of the interval $[a,\, b]$. A function S_n is called* **piecewise polynomial function** *if the restriction of S_n to each subinterval $[x_{i-1},\, x_i]$ reduces to a polynomial of degree at most k_i. Usually, $k_i \equiv k$ for each i i.e. the polynomials have the same degree.*

In other words, a piecewise polynomial function is a function that is a polynomial on each of its sub-domains, but possibly a different one on each. The most simple example of piecewise polynomial function is the piecewise linear function S_n defined as

$$S_n(x) = \frac{(x_{i+1}-x)f(x_i) + (x-x_i)f(x_{i+1})}{x_{i+1}-x_i}, \qquad x \in [x_i,\, x_{i+1}].$$

It is very simple and stable but it furnishes a bad graphic representation. Indeed, as no conditions are established between the derivatives of polynomials, the connection between two different polynomials in general presents a corner point.

The most used piecewise polynomial function are the so-called **spline**. They are constructed by imposing conditions which assure the continuity of the derivatives at the interior points x_i without using the values that such derivatives have in these point that could not be available.

Definition 1.14. *Let $a = x_0 < x_1 < \cdots < x_n = b$ be a subdivision of the interval $[a,\, b]$ and $d \geq 1$. A function $S_d(f,x)$ is called a* **spline** *of degree d with respect to this subdivision if*

- *S_d is a polynomial of degree d in each subinterval $[x_{i-1}, x_i]$, $i = 1, \ldots, n$;*
- *$S_d^{(k)}$ is a continuous function on $[a,\, b]$ for each $k = 0, 1, \ldots, d-1$.*

Practically, the spline are $d-1$ times continuous differentiable on $[a,\, b]$ and coincide with a polynomial of degree d in each subinterval. Next paragraph deals with the most used spline that is the so-called **cubic spline**.

1.3.1. *Cubic Spline.* Let $x_k = a + hk$, $k = 0, 1, \ldots, n$ be $n+1$ equidistant points of the interval $[a,b]$ and let us denote by $S_3(f,x)$ the cubic spline interpolating the function f in the nodes x_i that is

$$S_3(f,x_i) = f(x_i), \qquad i = 0, 1, \ldots, n. \tag{1.31}$$

According to Definition 1.14, $S_3(f,x)$ is a function such that

$$S_3(f,x) = a_i + b_i x + c_i x^2 + d_i x^3, \qquad x \in [x_{i-1}, x_i], \qquad i = 1, \ldots, n \tag{1.32}$$

$$S_3^{(k)}(f, x_i^+) = S_3^{(k)}(f, x_i^-), \quad i = 1, \ldots, n-1, \quad k = 0, 1, 2. \tag{1.33}$$

Conditions (1.31) and (1.33) get a linear system of $4n-2$ unknowns in the $4n$ unknown a_i, b_i, c_i, d_i. Then, in order to obtain a square system we need to add two additional conditions. However, we can get $S_3(f,x)$ by solving a linear system of order $n+1$ instead of $4n$.

To this end, we introduce the following unknowns

$$M_i = S_3''(f, x_i), \qquad i = 0, 1, \ldots, n. \tag{1.34}$$

We know that S_3 is a polynomial of degree 3 in each interval $[x_{i-1}, x_i]$. Hence, we can write

$$S_3''(f, x) = \frac{(x_i - x)M_{i-1} + (x - x_{i-1})M_i}{h_i}, \qquad i = 1, 2, \ldots, n \tag{1.35}$$

where $h_i = x_i - x_{i-1}$. Let us note that $S_3''(f, x) \in C^0([a, b])$.

Then, for each $x \in [x_{i-1}, x_i]$, by integrating (1.35) we get

$$S_3'(f, x) = \frac{-(x_i - x)^2 M_{i-1} + (x - x_{i-1})^2 M_i}{2h_i} + C_i, \qquad i = 1, 2, \ldots, n \tag{1.36}$$

and integrating again we get an expression for the function S_3 we are looking for

$$S_3(f, x) = \frac{(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i}{6h_i} + C_i(x - x_{i-1}) + D_i, \qquad i = 1, 2, \ldots, n \tag{1.37}$$

where C_i and D_i are constants which we can recover by using (1.31). Indeed, since

$$S_3(f, x_{i-1}) = f(x_{i-1}), \qquad \text{and} \qquad S_3(f, x_i) = f(x_i),$$

we find

$$C_i = \frac{f(x_i) - f(x_{i-1})}{h_i} - \frac{h_i(M_i - M_{i-1})}{6} \qquad \text{and} \quad D_i = f(x_{i-1}) - \frac{h_i^2}{6} M_{i-1}.$$

Consequently, if $x \in [x_{i-1}, x_i]$, we can rewrite (1.36) and (1.37) as

$$S_3'(f, x) = \frac{-(x_i - x)^2 M_{i-1} + (x - x_{i-1})^2 M_i}{2h_i} + \frac{f(x_i) - f(x_{i-1})}{h_i} - \frac{h_i(M_i - M_{i-1})}{6}$$

$$S_3(f, x) = \frac{(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i}{6h_i} + \left[\frac{f(x_i) - f(x_{i-1})}{h_i} - \frac{h_i(M_i - M_{i-1})}{6}\right](x - x_{i-1})$$

$$+ f(x_{i-1}) - \frac{h_i^2}{6} M_{i-1}, \qquad i = 1, 2, \ldots, n. \tag{1.38}$$

By proceeding in this way we have required to S_3 and S_3'' to be continuous at the points x_{i-1} and x_i. Now we have to impose that S_3' is continuous too. We can do it by determining the quantities M_i such that the following conditions hold true

$$\lim_{x \to x_i^-} S_3'(f, x) = \lim_{x \to x_i^+} S_3'(f, x), \qquad i = 1, 2, \ldots, n - 1. \tag{1.39}$$

By (1.36) we can deduce that if $x \in [x_{i+1}, x_i]$ it has the following form

$$S_3'(f, x) = \frac{-(x_{i+1} - x)^2 M_i + (x - x_i)^2 M_{i+1}}{2h_{i+1}} + \frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{h_{i+1}(M_{i+1} - M_i)}{6}. \tag{1.40}$$

Then by replacing (1.40) and (1.36) into (1.39) we get the following linear system

$$h_i M_{i-1} + 2(h_i + h_{i+1})M_i + h_{i+1} M_{i+1} = \frac{6}{h_{i+1}}(f(x_{i+1}) - f(x_i)) - \frac{6}{h_i}(f(x_i) - f(x_{i-1})). \tag{1.41}$$

of $n - 1$ equations in the $n + 1$ unknowns $M_0, M_1, ..., M_n$. Hence, in order to get a linear system we need to add two conditions. There are various ways of specifying these two additional constraints such as the following.

Complete cubic spline. We add conditions at the endpoints $x_0 = a$ and $x_n = b$ of the interval. More precisely, we impose that

$$S_3'(f, x_0) = f'(x_0) \qquad S_3'(f, x_n) = f'(x_n) \tag{1.42}$$

from which we have

$$2h_1 M_0 + h_1 M_1 = 6\left[\frac{f(x_1) - f(x_0)}{h_1} - f'(x_0)\right]$$
$$h_n M_{n-1} + 2h_n M_n = 6\left[f'(x_n) - \frac{f(x_n) - f(x_{n-1}}{h_n}\right].$$

These last conditions together with (1.41) lead to the following symmetric tridiagonal system

$$\begin{pmatrix} 2h_1 & h1 & & & \\ h_1 & 2(h_1+h_2) & h_2 & & \\ & \ddots & & & \\ & & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ & & & h_n & 2h_n \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = 6 \begin{pmatrix} \frac{f(x_1)-f(x_0)}{h_1} - f'(x_0) \\ \frac{f(x_2)-f(x_1)}{h_2} - \frac{f(x_1)-f(x_0)}{h_1} \\ \vdots \\ \frac{f(x_n)-f(x_{n-1})}{h_n} - \frac{f(x_{n-1})-f(x_{n-2})}{h_{n-1}} \\ f'(x_n) - \frac{f(x_n)-f(x_{n-1})}{h_n} \end{pmatrix}$$

Then by solving this system we obtain the so-called **complete cubic spline**.

Natural cubic spline. We impose the conditions $M_0 = M_n = 0$. In this case we obtain the so-called **natural cubic spline**. We could also use the correct second derivatives values

$$M_0 = f''(x_0), \quad M_n = f''(x_n). \tag{1.43}$$

In this case, these last conditions together with (1.41) lead to the following tridiagonal symmetric system of order $n-1$

$$\begin{pmatrix} 2(h_1+h_2) & h_2 & & & \\ h_2 & 2(h_2+h_3) & h_3 & & \\ & \ddots & & & \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ & & & h_{n-1} & 2(h_{n-1}+h_n) \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-2} \\ M_{n-1} \end{pmatrix}$$
$$= 6 \begin{pmatrix} \frac{f(x_2)-f(x_1)}{h_2} - \frac{f(x_1)-f(x_0)}{h_1} - \frac{h_1}{6} f''(x_0) \\ \frac{f(x_3)-f(x_2)}{h_3} - \frac{f(x_2)-f(x_1)}{h_2} \\ \vdots \\ \frac{f(x_{n-1})-f(x_{n-2})}{h_{n-1}} - \frac{f(x_{n-2})-f(x_{n-3})}{h_{n-2}} \\ \frac{f(x_n)-f(x_{n-1})}{h_n} - \frac{f(x_{n-1})-f(x_{n-2})}{h_{n-1}} - \frac{h_n}{6} f''(x_n) \end{pmatrix}$$

Periodic cubic spline. If we are dealing with a periodic function we can impose the periodic conditions

$$S_3'(f, x_0) = S_3'(f, x_n) \tag{1.44}$$
$$S_3''(f, x_0) = S_3''(f, x_n). \tag{1.45}$$

Condition (1.45) leads to the following equation

$$2h_1 M_0 + h_1 M_1 + 2h_n M_n + h_n M_{n-1} = 6\left[\frac{f(x_1) - f(x_0)}{h_1} - \frac{f(x_n) - f(x_{n-1}}{h_n}\right], \tag{1.46}$$

while condition (1.45) allow us to state that

$$M_0 = M_n.$$

Then by replacing this last equation in (1.47) and taking into account that $f(x_0) = f(x_n)$ equation (1.47) reads as

$$h_1 M_1 + h_n M_{n-1} + 2M_n(h_1 + h_n) = 6\left[\frac{f(x_1) - f(x_n)}{h_1} - \frac{f(x_n) - f(x_{n-1})}{h_n}\right], \tag{1.47}$$

In this case this last condition together with (1.41) lead to the following diagonal dominant symmetric system

$$\begin{pmatrix} 2(h_1+h_2) & h_2 & & & & h_1 \\ h_2 & 2(h_2+h_3) & h_3 & & & \\ & \ddots & & & & \\ & & & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ h_1 & & & & h_n & 2(h_1+h_n) \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix}$$

$$= 6 \begin{pmatrix} \frac{f(x_2)-f(x_1)}{h_2} - \frac{f(x_1)-f(x_n)}{h_1} \\ \frac{f(x_3)-f(x_2)}{h_3} - \frac{f(x_2)-f(x_1)}{h_2} \\ \vdots \\ \frac{f(x_n)-f(x_{n-1})}{h_n} - \frac{f(x_{n-1})-f(x_{n-2})}{h_{n-1}} \\ \frac{f(x_1)-f(x_n)}{h_1} - \frac{f(x_n)-f(x_{n-1})}{h_n} \end{pmatrix}.$$

This procedure lead to the so-called **periodic spline**.

In all the previous case the system we have to solve is non singular and since the matrix of coefficient is a diagonal dominant matrix we can solve it by using the Gauss method. Once we have determined the array $M_0, M_1, ..., M_n$, in order to compute the cubic spline in a point $x \in [a, b]$, we have at first determine the subinterval $[x_{i-1}, x_i]$ which contain x and then use the formula (1.38).

About the convergence the following theorem holds true.

Theorem 1.15. *Let $S_3(f, x)$ be a cubic spline with the additional conditions (1.43) or (1.44) and (1.45). If $f \in C^2([a, b])$ then*

$$\lim_{h\to 0} \frac{\|f^{(p)} - S_3^{(p)}(f)\|_\infty}{h^{2-p}} = 0, \quad p = 0, 1, 2, \quad h = \max_{1\le i\le n} h_i.$$

If $f \in C^k([a, b])$, $k = 3, 4$ and there exists a constant $\mathcal{C}$ such that $\frac{h}{h_i} \le \mathcal{C} < \infty$ then for $p = 0, 1, 2, 3$

$$\begin{cases} \lim\limits_{h\to 0} \dfrac{\|f^{(p)} - S_3^{(p)}(f)\|_\infty}{h^{3-p}} = 0, & k = 3 \\ \lim\limits_{h\to 0} \dfrac{\|f^{(p)} - S_3^{(p)}(f)\|_\infty}{h^{4-p}} = 0, & k = 4. \end{cases} \tag{1.48}$$

By the way it was proved that the maximum order of convergence we can get with the cubic spline is h^4 that is

$$\|f(x) - S_3(x)\|_\infty = \mathcal{O}(h^4)$$

even if $f \in C^k([a, b]), k > 4$.

The natural cubic spline are the most interesting spline. By the way if $f''(a)$ and $f''(b)$ are not zero the order of convergence is not the optimal one. In fact, if $f \in C^4([a, b])$, the order of

convergence is $\mathcal{O}(h^4)$ in each subinterval of $[a,b]$ but in a neighbord of the endpoints it reduces to $\mathcal{O}(h^2)$.

It is also possible to determine a cubic spline without specifying any extra conditions at the end points (other than that the spline interpolates the data points there). This is the case of the so-called **not-a-knot spline** which requires that the third derivative of the spline is continuous at x_1 and x_{n-1}:

$$S_3^{(3)}(x_1^+) = S_3^{(3)}(x_1^-), \qquad \text{and} \qquad S_3^{(3)}(x_{n-1}^+) = S_3^{(3)}(x_{n-1}^-). \tag{1.49}$$

In this case the nodes x_1 and x_{n-1} are interpolating points but are not partition points. In other words, the polynomial in the first two subintervals (and in the last two) are equal.

1.4. **Matlab Programming exercises.**

Task 1. Investigate on the Matlab commands *polyfit*, *polyval*, *figure*, *plot* and *subplot*. By using such commands, write an *m-file function*, named **pflagrange.m**, which compute and plot the fundamental Lagrange polynomials and the Lebesgue constants. The *m-file function* has to take the following variables as input:

1. the degree of the polynomials;
2. the interval of the interpolation $[a,b]$;
3. the variable "flag" which allows one to choose the interpolation nodes ($flag = 1$ for the equidistant nodes, $flag = 2$ for the zeros of $p_n(v^{-1/2,-1/2})$ and $flag = 3$ for the zeros of $(1-x^2)p_n(v^{1/2,1/2})$).

Then, use the *m-file function* to plot both the foundamental Lagrange polynomials of degree 6 and the Lebesgue functions based on these zeros.

Task 2. Write an *m-file function*, named **Lagrange.m**, which compute the Lagrange polynomial of degree $n-1$ defined as

$$L_n(f,x) = \sum_{k=1}^{n} l_k(x) f(x_k), \quad l_k(x) = \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{(x-x_j)}{(x_j-x_k)}.$$

The *m-file function* has to take the same input variables of the previous exercise and has to return the absolute error $E = \max_{x\in[a,b]} |f(x) - L_n(f,x)|$ and the graph of the polynomial and the function f. Then use this *m-file function* to approximate the following functions:

1 $f(x) = 1/(1+x^2), \quad x \in [-5,5]$
2 $f(x) = |x|^{9/2} e^{x+5}(1+(x+4))^3, \quad x \in [-1,1].$

with the polynomial L_n, $n = \{16, 32, 64, 128, 256\}$ by using the equidistant nodes.

Repeat the exercise by using the Chebyshev nodes of the first kind. Compare and comment the results.

Task 3. Investigate on the Matlab command *spline* to construct the cubic spline with the condition "not a knot". Hence, determine the cubic spline which interpolates the Runge function on $6, 10, 14$ equidistant nodes of the interval $[-5,5]$. Represent the relative graph and compare the results with those obtained in the previous exercise with the Lagrange polynomial.

2. Numerical Integration

Numerical integration formulae, or quadrature formulae, are methods for approximating definite integrals of the following form

$$\int_a^b f(x)dx$$

where $f \in C^0([a,b])$.

The need for numerical quadrature arises when either a definite integral cannot be evaluated analytically or when special functions involved in an analytical solution are too complicated to be of direct use. A typical example of an integral for which the antiderivative of the integrand cannot be expressed in terms of elementary functions are

$$\int_{-3}^{3} e^{-x^2}dx, \qquad \int_0^5 \frac{\sin 3x}{\sqrt{x^2+x+1}}.$$

Morover, numerical quadrature are essential when the integrand function f is known in a set of discrete data and they also provide a basic tool for other kind of problems such as the numerical solution of *integral equations* like

$$f(y) - \frac{1}{2}\int_0^1 (y+1)e^{-xy}f(x)dx = e^{-y} - \frac{1}{2} + \frac{e^{-(y+1)}}{2}, \qquad y \in [0,1]$$

or the numerical solution of *boundary value problems* such as the classical *Dirichlet problem*

$$\begin{cases} \dfrac{\partial^2 f}{\partial x_1^2} + \dfrac{\partial^2 f}{\partial x_2^2} = 0, \quad P = (x_1, x_2) \in \Omega \\ f(P) = \bar{f}(P), \quad P \in \Sigma \end{cases}$$

where Ω is a planar domain and Σ is its boundary.

In order to describe the most important quadrature formula let us consider $n+1$ points in the interval $[a,b]$

$$a \le x_0 < x_1 < ... < x_n \le b$$

and assume that the values $f(x_k)$, $\forall k = 0, 1, ..., n$ are known.

Under this assumption, we can compute the integral in the following way

$$\int_a^b f(x)dx = \sum_{k=0}^{n} a_k f(x_k) + R_n(f) := Q_n(f) + R_n(f)$$

where the sum $Q_n(f)$ is called **quadrature formula** and $R_n(f)$ represents the **remainder term**. In the quadrature formula

$$Q_n(f) = \sum_{k=0}^{n} a_k f(x_k)$$

the coefficients a_k are called **quadrature weights** or simply **weights** and the points $x_k \in [a,b]$ are called **quadrature nodes** or simply **nodes**. Moreover, Q_n is called **closed formulae** if both a and b are nodes. Conversely, it is called an **open formulae**.

When we introduce a quadrature formula we face to the following problems:

(1) The value of the sum Q_n tends to the value of the integral? In other words, the error $R_n(f)$ tends to zero?
(2) How to construct the quadrature formula? And then, if we know $(x_k, f(x_k))$ how do determine the coefficients a_k?

Let us consider the first problem.

Definition 2.1. *The quadrature formula Q_n is* **convergent** *if*

$$\lim_{n\to\infty} R_n(f) = 0.$$

Definition 2.2. *The quadrature formula Q_n has* **degree of exactness or presicion** *p if*

$$\begin{cases} Q_n(x^k) = \displaystyle\int_a^b x^k dx \quad k = 0, ..., p \\ \\ Q_n(x^k) \neq \displaystyle\int_a^b x^k dx \quad k > p \end{cases} \qquad \textit{or equivalently} \qquad \begin{cases} R_n(x^k) = 0 \quad k = 0, ..., p \\ \\ R_n(x^k) \neq 0 \quad k > p \end{cases}.$$

The following theorem states the assumption which assure that the quadrature formula $\{Q_n\}$ has to satisfy to have convergence.

Theorem 2.3. *Assume that the sequence $\{Q_n\}$ is such that*

(i) $\{Q_n\}$ has degree of exactness n;

(ii) $\sup_n \sum_{k=0}^{n} |a_k| < \infty$.

Then the sequence $\{Q_n\}$ is convergent.

Let us construct a convergent quadrature formula. To this end we fix $n+1$ distinct nodes in $[a, b]$. Taking the previous theorem into account we have to impose that the quadrature rule has degree of exactness n. In virtue of the linearity of the integrals, we can simply impose that the formula is exact on the canonical base $\{1, x, x^2, ..., x^n\}$, namely

$$\int_a^b x^p = \sum_{k=0}^{n} a_k x_k^p, \quad p = 0, 1, ..., n \tag{2.1}$$

By expliciting (2.1) we have

$$\begin{cases} a_0 + a_1 + ...a_n = (b-a) \\ a_0 x_0 + a_0 x_1 + a_1 x_2 + ... + a_n x_n = \frac{1}{2}(b^2 - a^2) \\ \vdots \qquad \vdots \\ a_0 x_0^n + a_0 x_1^n + a_1 x_2^n + ... + a_n x_n^n = \frac{1}{n+1}(b^{n+1} - a^{n+1}) \end{cases}$$

which is a linear system that can be written as

$$\begin{pmatrix} 1 & 1 & ... & 1 \\ x_0 & x_1 & ... & x_n \\ \vdots & \vdots & ... & \vdots \\ x_0^n & x_1^n & ... & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} (b-a) \\ \frac{1}{2}(b^2 - a^2) \\ \vdots \\ \frac{1}{n+1}(b^{n+1} - a^{n+1}) \end{pmatrix}.$$

By the assumptions, the nodes are distinct and then the matrix is non singular. Thus the previous system has only one solution which furnish the weights a_k of our formula.

This method is the so-called **method of undetermined coefficients**, but it is not efficiently from a computational point of view. Indeed, the matrix of coefficients is the well-known ill-conditioned Vandermonde matrix and then the results are inaccurate when n is large.

In order to overcome this problem, we have to follow an other approach which leads to the so-called **interpolatory formula** which are classified as

(i) **Newton-Cotes formula** if the quadrature points are equidistant nodes;
(ii) **Gaussian-formula** if the quadrature points are the zeros of orthogonal polynomials.

2.1. **Newton-Cotes formula.** The Newton-Cotes formula was discovered by Newton in 1711. Let $n \geq 0$ and let us introduce the following equidistant quadrature nodes

$$x_k = x_0 + kh, \quad k = 0, 1, ..., n.$$

We define **closed Newton-Cotes formula** the quadrature rules for which

$$a = x_0, \quad b = x_n, \quad h = \frac{b-a}{n}, \quad n \geq 1$$

and **open formula the quadrature rules** for which

$$a = x_0 + h, \quad b = x_n - h, \quad h = \frac{b-a}{n+2}, \quad n \geq 0.$$

In the next paragraphs we will described the most important Newton-Cotes quadrature formula.

Rectangle rule (and mid-point rule). The most simple Newton-Cotes formula is the **rectangle rule** obtained by approximating the value of the integral with the area of the rectangle we have under the curve representing the function f. Hence the integral can be approximated with the product of the value of the function at one of the end-points by the length of the interval as follows

$$\int_a^b f(x)dx = (b-a)f(a) + R_0(f) \tag{2.2}$$

or

$$\int_a^b f(x)dx = (b-a)f(b) + R_0(f). \tag{2.3}$$

A variation of the rectangle rule is the **midpoint rule**. Similarly to the rectangular rule, we approximate the value of the integral by multiplying the length of the interval by the value of the function at the mid-point

$$\int_a^b f(x)dx = (b-a)f\left(\frac{a+b}{2}\right) + R_0(f). \tag{2.4}$$

Both formulae are open Newton-Cotes formula and have degree of exactness equal to $n = 0$ that is $R_0(f) = 0$ if f is a constant.

Although the exact error R_0 is not known, an estimate for it can be derived under the assumption that f is suitably differentiable in the interval of integration $[a, b]$. More precisely if

$$\int_a^b |f'(x)|dx < \infty \tag{2.5}$$

then since

$$R_0(f) = \int_a^b f(x)dx - f(a)(b-a) = \int_a^b (f(x) - f(a))dx \int_a^b \left(\int_a^x f'(t)dt\right) dx$$

we can deduce

$$|R_0(f)| \leq \left|\int_a^b \left(\int_a^x f'(t)dt\right) dx\right| \leq (b-a)\int_a^b |f'(t)|dt. \tag{2.6}$$

Thus, the error only depends on the size of the interval. Hence, if the size is not small it is not useful to use this formula.

If $b-a>1$, it is more practical subdivide the interval integration in n sub-intervals having equal length $h=(b-a)/n$ that is

$$\int_a^b f(x)dx=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)dx$$

where

$$a=x_0<x_1<...<x_n=b$$

with $x_k=x_0+kh, \quad k=0,...,n$ and then applying a primitive rule to each of the sub-intervals. This procedure leads to the so-called **composite rectangle and mid-point quadrature formula** which read as

$$\int_a^b f(x)dx=\begin{cases}\frac{(b-a)}{n}\sum_{k=0}^{n-1} f(x_{k+1})+R_n(f)\\ \frac{(b-a)}{n}\sum_{k=0}^{n-1} f(x_k)+R_n(f)\end{cases} \tag{2.7}$$

and

$$\int_a^b f(x)dx=\frac{(b-a)}{n}\sum_{k=0}^{n-1} f\left(\frac{x_{k+1}+x_k}{2}\right)+R_n(f). \tag{2.8}$$

About the error of formula (2.7), taking (2.6) we have

$$|R_n(f)|=\sum_{k=0}^{n-1}|R_k(f)|\leq\sum_{k=0}^{n-1}(x_{k+1}-x_k)\int_{x_k}^{x_{k+1}}|f'(x)|dx=\frac{(b-a)}{n}\int_a^b|f'(x)|dx. \tag{2.9}$$

Hence, we can deduce that these formulae have order of convergence 1. Moreover, it is possible to prove that if the integrand function is smoother the error does not improve.

Trapezoidal rule. A very useful Newton-Cotes formula is the **trapezoidal rule** in which the integral is approximated with the area of the trapezoid we have under the graph of the function f:

$$\int_a^b f(x)dx=\frac{b-a}{2}\{f(a)+f(b)\}+R_1(f). \tag{2.10}$$

Let us note that this is a closed Newton-Cotes formula having degree of exactness $n=1$.

Moreover, let us remark that the previous formula could also be obtained by approximating the integrand function f by using the Lagrange polynomial based on the nodes $x_0=a$ and $x_1=b$ that is

$$\int_a^b f(x)dx=\int_a^b\left(\sum_{k=0}^{1} l_k(x)f(x_k)\right)dx=\sum_{k=0}^{1} a_k f(x_k) \tag{2.11}$$

where

$$a_0=\int_a^b l_0(x)dx=\int_a^b\frac{x-x_1}{x_0-x_1}dx=\frac{(b-a)}{2}$$
$$a_1=\int_a^b l_1(x)dx=\int_a^b\frac{x-x_0}{x_1-x_0}dx=\frac{(b-a)}{2}.$$

Hence, by replacing the above coefficients in (2.11) we get (2.10).

This is the general idea of the interpolatory formula based on $n+1$ points that is to approximate the integrand function with the Lagrange polynomialof degree n

$$\int_a^b f(x)dx = \int_a^b \sum_{k=0}^{n} \ell_k(x) f(x_k)dx = \sum_{k=0}^{n} f(x_k) \int_a^b \ell_k(x)dx = \sum_{k=0}^{n} a_k f(x_k)$$

where

$$a_k = \int_a^b \ell_k(x)dx \quad \text{with} \quad \ell_k(x) = \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}.$$

Let us also remark that taking the properties of the Lagrange polynomial into account an interpolatory formula based on $n+1$ nodes have degree of precision n.

About the error $R_1(f)$ of the trapezoidal rule if assumption (2.5) is verified then we have

$$\begin{aligned} R_1(f) &= \int_a^b f(x)dx - (b-a)\frac{f(a)+f(b)}{2} \\ &= \frac{1}{2}\int_a^b (f(x) - f(a))dx + \frac{1}{2}\int_a^b (f(x) - f(b))dx \\ &= \frac{1}{2}\int_a^b \int_a^x f'(t)dtdx - \frac{1}{2}\int_a^b \int_x^b f'(t)dtdx. \end{aligned}$$

Taking the absolute value we obtain an estimate for this error

$$|R_1(f)| \leq \frac{1}{2}\int_a^b \left| \int_a^x f'(t)dt - \int_x^b f'(t)dt \right| dx \leq \frac{(b-a)}{2}\int_a^b |f'(t)|dt. \tag{2.12}$$

Then in this case the error is exactly half of the error we could get by using formula (2.2) or (2.3).

Moreover, if the integrand function f is such that

$$\int_a^b |f''(t)|dt < \infty, \tag{2.13}$$

then

$$R_1(f) = -\frac{1}{2}\int_a^b (b-x)(x-a)f''(x)dx$$

and the following estimate holds true

$$|R_1(f)| \leq \frac{(b-a)^2}{2}\int_a^b |f''(t)|dt.$$

Summirising, if condition (2.5) is satisfied, the rectangle and trapezoidal quadrature formula have the same order of convergence but if f satisifes (2.13) then the trapezoidal rule gives better results and the order of convergence is 2.

Similarly to the rectangle rule, a better estimate can be obtained by dividing the interval $[a, b]$ into n sub-intervals of length $h = (b-a)/n$, applying the previous formula to each of the n sub-intervals and then summing the result. In this way we get the so-called **composite trapezoidal**

rule

$$\int_a^b f(x)dx = \frac{(b-a)}{n}\sum_{k=0}^{n-1}\frac{f(x_k)+f(x_{k+1})}{2} + R_1(f)$$

$$= \frac{(b-a)}{2n}\left\{f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right\} + R_1(f) \tag{2.14}$$

where $x_0 = a$, $x_k = a + kh$, $k = 1, ..., n$ and $x_n = b$. As regard the error, in this case it is possible to prove that

$$|R_1(f)| \leq \begin{cases} \dfrac{b-a}{2n}\displaystyle\int_a^b |f'(x)|dx, & if \displaystyle\int_a^b |f'(x)|dx < \infty \\ \dfrac{(b-a)^2}{2n^2}\displaystyle\int_a^b |f''(x)|dx, & if \displaystyle\int_a^b |f''(x)|dx < \infty. \end{cases} \tag{2.15}$$

The Simpson rule. The most frequently Newton-Cotes quadrature formula is the **Simpson rule** in which the integral is approximated by a parabola that passes through the two end points and a single internal point. To derive it, we approximate the function f by using the Lagrange polynomial based on three points $x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b$. Thus, we can write

$$\int_a^b f(x)dx = \sum_{k=0}^{2} a_k f(x_k) + R_2(f)$$

where

$$a_0 = \int_a^b l_0(x)dx = \int_a^b \frac{(x-x_1)}{(x_0-x_1)}\frac{(x-x_2)}{(x_0-x_2)}dx = \frac{b-a}{6}$$

$$a_1 = \int_a^b l_1(x)dx = \int_a^b \frac{(x-x_0)}{(x_1-x_0)}\frac{(x-x_2)}{(x_1-x_2)}dx = \frac{2(b-a)}{3}$$

$$a_2 = \int_a^b l_2(x)dx = \int_a^b \frac{(x-x_0)}{(x_2-x_0)}\frac{(x-x_1)}{(x_2-x_1)}dx = \frac{(b-a)}{6}$$

from which we get

$$\int_a^b f(x)dx = \frac{b-a}{6}\left\{f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right\} + R_2(f). \tag{2.16}$$

It is a closed formula and it has degree of precision $n = 2$.

As regard the error R_2, it is possible to prove that if f is such that

$$\int_a^b |f^{(IV)}(X)|dx < \infty \tag{2.17}$$

then

$$|R_2(f)| \leq \frac{(b-a)^4}{1152}.$$

Now, let us improve this error of estimate by introducing the so-called **composite Simpson rule**. To this end let us divide the interval $[a, b]$ into $2n$ sub-intervals having the same length.

Then, we can write

$$\int_a^b f(x)dx = \sum_{i=0}^{n} \int_{x_{2i}}^{x_{2i+1}} f(x)dx.$$

Now, by applying the primitive Simpson rule (2.16) to each sub-interval we get the following formula

(2.18)

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{i=1}^{n-1} f(a+2ih) + 4\sum_{i=1}^{n} f(a+(2i-1)h) + f(b)\right] + R_n(f), \quad h = \frac{b-a}{2n}$$

where under the assumption (2.17)

$$|R_n(f)| \leq \mathcal{C}\frac{(b-a)^4}{n^4}$$

with $\mathcal{C}$ a positive constant independent of f and n.

By all the theoretical estimates of the remainder term, in order to improve the error might seem natural increase the value of n and thus construct quadrature formula based on more points.

However, this can not be recommended for numerical purposes because it is possible to see that for $n \geq 8$ the weights tend to grow. As we have seen in the previous chapter the Lagrange polynomial based on equidistant nodes is not a good approximant.

2.2. **Gaussian formula.** Sometimes the integrand function is such that there exists its integral but has not some good smoothness properties because, for instance, its derivatives are singular in some points $t_k \in [a, b]$. For example,

$$g(x) = \frac{xe^{-x}}{\sqrt{1-a}}. \tag{2.19}$$

In this case we can proceed in the following way. We factorize the integrand function as

$$g(x) = w(x)f(x)$$

where w is a function having all the singularities of f and f is a smooth function; In the case of (2.19) we have

$$f(x) = xe^{-x}, \quad w(x) = \frac{1}{\sqrt{1-a}}.$$

Hence, in order to approximate the integral we approximate the function f by means of the Lagrange polynomial based on the points $\{x_k\}_{k=0}^n$. Then we have

$$\int_a^b g(x)dx = \int_a^b f(x)w(x)dx = \int_a^b \sum_{k=0}^{n} \ell_k(x)f(x_k)w(x)dx = \sum_{k=0}^{n} f(x_k)\int_a^b \ell_k(x)w(x)dx + e_n(f)$$

from which we deduce that

$$\int_a^b f(x)w(x)dx = \sum_{k=0}^{n} \lambda_k f(x_k) + e_n(f) \qquad \text{with} \qquad \lambda_k = \int_a^b \ell_k(x)w(x)dx. \tag{2.20}$$

The function w is called **weight function** and it must be such that the calculation of the coefficients λ_k should be simple. Quadrature formula (2.20), namely the quadrature rule in which the coefficients depend on w, is called **weigthed interpolatory formula**. In the case where the quadrature nodes $\{x_k\}_{k=0}^n$ of weighted interpolatory formula coincides with the zeros of orthogonal polynomials with respect to the weight function w then such quadrature is called **Gaussian quadrature formula**.

Gaussian quadrature formula is exact for polynomials of degree at most $2n-1$ and is stable. About the remainder term $e_n(f)$, there exists different type of estimates according to the smoothness of the function f. Here we only recall an estimate that is useful in the following. However for a complete discussion the reader can consult [12].

Proposizione 2.4. *Let $f \in C^0([a,b])$. Then*

$$|e_n(f)| \leq \mathcal{C} E_{2n-1}(f) \tag{2.21}$$

where $\mathcal{C} \neq \mathcal{C}(n,f)$.

Proof. Let $P \in \mathbb{P}_{2n-1}$. Then we can write

$$|e_n(f)| = |e_n(f-P)| \leq \int_{-1}^{1} |f(x)-P(x)|w(x)dx + \sum_{k=1}^{n} \lambda_k(w)|f(x_k)-P(x_k)|.$$

Hence,

$$|e_n(f)| \leq 2\mathcal{C} \int_{-1}^{1} |f(x)-P(x)|w(x)dx \leq 2\mathcal{C}\|(f-P)\|_\infty \int_{-1}^{1} w(x)dx$$

from which taking into account the infimum on P we have

$$|e_n(f)| \leq \mathcal{C} E_{2n-1}(f).$$

□

3. Integral Equations

Integral equations (i.e. equations in which the unknown function appears under the integral sign), are one of the most important classes of equations. They are classified as

1. **Fredholm integral equations** in which the limits of integrations are fixed. They have the following general form

$$p(y)f(y) - \int_a^b k(x,y)f(x)dx = g(y)$$

where $-\infty \leq a < b \leq +\infty$, f is the unknown and k, g and p are given functions. Often k and g are called kernel and right-hand side of the equation, respectively.

According to the function p, Fredholm integral equations are classified as follows:

- **Fredholm integral equations of the first kind** if $p(y) = 0$, $\forall y \in [a,b]$ that is

$$\int_a^b k(x,y)f(x)dx = g(y)$$

- **Fredholm integral equations of the second kind** if $p(y)$ is a constant for every $y \in [a,b]$. Generally, these equations read as

$$f(y) - \mu \int_a^b k(x,y)f(x)dx = g(y), \tag{3.1}$$

where $\mu := \frac{1}{p}$ and $g := \frac{g}{p}$

- **Fredholm integral equations of the third kind** if $p(y) = 0$ in some points (but not all) of the interval $[a,b]$. An example is the following

$$(1-y)f(y) - \int_{-1}^{1} k(x,y)f(x)dx = g(y).$$

Moreover, if the right-hand side g vanishes the integral equations are called **homogeneous integral equations**. Conversely, if g is non zero they are called **inhomogeneous integral equations**.

2. **Volterra integral equations** in which the limits of integrations are not fixed. Similarly to Fredholm integral equations, we refer to homogeneous or inhomogeneous integral equations according to the right-hand side. Moreover, we refer to **Volterra integral equations of the first kind** if they are of the form

$$\int_a^y k(x,y)f(x)dx = g(y), \quad x \in [a,b]$$

and **Volterra integral equations of the second kind** if they are written as

$$f(y) + \int_a^y k(x,y)f(x)dx = g(y), \quad x \in [a,b].$$

Volterra integral equation of the second kind find a natural application in the ordinary differential equation. In fact, let us consider the following ordinary differential equation of the second kind

$$x''(t) + a_1(t)x'(t) + a_2(t)x(t) = y(t), \quad t > a \tag{3.2}$$

with the initial conditions for $t = a$:

$$x(a) = \mathcal{C}_0, \quad x'(a) = \mathcal{C}_1. \tag{3.3}$$

Setting

$$x''(t) = f(t) \tag{3.4}$$

and taking into account the initial conditions we can write

$$\begin{cases} x'(t) = \displaystyle\int_a^t f(s)ds + \mathcal{C}_1 \\ \\ x(t) = \displaystyle\int_a^t (t-s)f(s)ds + \mathcal{C}_1 t + \mathcal{C}_0. \end{cases} \tag{3.5}$$

Thus by replacing (3.4) and (3.5) in (3.2) we get the following Volterra equation of second kind

$$f(t) - \int_a^t k(t,s)f(s)ds = g(t) \tag{3.6}$$

where

$$k(t,s) = a_1(t) + a_2(t)(t-s)$$

and

$$g(t) = y(t) - \mathcal{C}_1(a_1(t) - ta_2(t)) - \mathcal{C}_0 a_2(t).$$

In this way, problem (3.2) equipped with the initial conditions (3.3) has been reduced to equation (3.6). Once it is solved, by replacing the calculated solution f in (3.5) we can found the solution of our problem. We also underline that if the functions $a_i(t)$ are constants for each $i = 0, 1, 2$ the Volterra integral equation has a kernel of the type $k(t-s)$ (integral equation of convolution type).

3.1. **Fredholm integral equations of the second kind.** Let us consider equation (3.1) that is

$$f(y) - \mu \int_{-1}^{1} k(x,y) f(x) dx = g(y), \quad y \in [-1,1] \tag{3.7}$$

where k and g are two given functions.

Example 3.1. The following equation

$$f(y) + \int_{-1}^{1} e^{y-x} f(x) dx = (y+1) + (e^2 - 3) e^{y-1}, \qquad y \in [-1,1]$$

has the unique solution $f(y) = y + 1$.

In this section we will introduce some numerical methods which aim to approximate the solution of (3.7). Nevertheless before going on, in the next subsections, we will recall some basic facts on the linear functional analysis useful for our aims.

3.1.1. *Some basic facts on linear functional analysis.* In the following we assume as well-known [8,9,16] the notion of Banach space that we will denote by $(X, \|\cdot\|)$, being $\|\cdot\|$ its norm, and the notion of bounded linear operator and inverse operator. In particular we will denote by $\mathcal{B}(X,X)$ the collection of all bounded linear operators from X to X, by

$$\|K\| = \sup_{\|x\| \le 1} \|Kx\| = \sup_{x \neq 0} \frac{\|Kx\|}{\|x\|}, \tag{3.8}$$

the norm of the operator $K \in \mathcal{B}(X,X)$ and finally by K^{-1} the inverse of $K \in \mathcal{B}(X,X)$.

Now we go back to our equation (3.7). Introducing the operator

$$(Kf)(y) = \mu \int_a^b k(x,y) f(x) dx$$

(3.7) can be written in the following form

$$(I - K) f = g$$

where I is the identity operator.

Our aim is to find the necessary and sufficient conditions so that (3.7) has a unique solution for each given right-hand side g. To this end, we note that first of all we have to assure that the inverse operator $(I-K)^{-1}$ exists. In fact, in this way equation (3.7) has a unique solution, given by $f = (I-K)^{-1} g$. In this context the following theorem has an important role.

Theorem 3.2. *Let $(X, \|\cdot\|)$ be a Banach space and let $K : X \to X$ be a linear operator. If $\|K\| \le q < 1$ then $(I-K)^{-1}$ exists and*

$$\|(I-K)^{-1}\| \le \frac{1}{1 - \|K\|} \le \frac{1}{1-q}. \tag{3.9}$$

Moreover, let $K : X \to X$ be a linear operator and $\{K_m\}_m$ a sequence of linear operators with $K_m : X \to X$, such that

$$\lim_m \|K - K_m\| = 0.$$

If $(I-K)^{-1}$ exists, then for m sufficiently large (say $m > m_o$), $(I-K_m)^{-1}$ exists and the following inequality

$$\|(I-K_m)^{-1}\| \le \frac{\|(I-K)^{-1}\|}{1 - \|(I-K)^{-1}\| \, \|(K-K_m)^{-1}\|} \tag{3.10}$$

holds true.

Proof. We begin with the proof of the first statement that is the well-known Von Neuman Theorem. To this end we consider the partial sum $S_m = \sum_{i=0}^{m} K^i$ of the series $\sum_{i=0}^{\infty} K^i$.
For an arbitrary p we can write

$$\begin{aligned}\|S_{m+p} - S_m\| = \|K^{m+1} + ... + K^{m+p}\| &\leq \sum_{i=1}^{p} \|K^{m+i}\| \\ &\leq \sum_{i=1}^{p} \|K\|^{m+i} \leq \sum_{i=1}^{p} q^{i+m} = q^m \sum_{i=1}^{p} q^i \leq \left(\frac{q^{m+1}}{1-q}\right).\end{aligned}$$

Since $0 < q < 1$ we can deduce that the right-hand side tends to zero for $m \to \infty$. Moreover this relation does not depend on p and then it is true for any arbitrary p. In other words the partial sum is a Cauchy sequence and then, since X is a Banach space, is convergent.
Let us denote by S its sum. It results

$$\|S\| = \left\| \sum_{i=0}^{\infty} K^i \right\| \leq \sum_{i=0}^{\infty} q^i = \frac{1}{1-q}.$$

Now if we prove that $S = (I-K)^{-1}$ we have relation (3.9). We can write

$$S(I-K) = \lim_{m\to\infty} S_m(I-K) \tag{3.11}$$

$$= \lim_{m\to\infty} [(I + K + K^2 + ... + K^m)(I-K)] \tag{3.12}$$

$$= \lim_{m\to\infty} [I + K + K^2 + ... + K^m - K - K^2 - ... - K^{m+1}] \tag{3.13}$$

$$= \lim_{m\to\infty} [I - K^{m+1}] = I \tag{3.14}$$

since $\|I - K^{m+1} - I\|$ tends to zero for $m \to \infty$.
Then S is a left inverse operator of $(I-K)$. In the same way it is possible to prove that S is a right inverse operator and consequently (3.9) holds true. Now we prove (3.10). We first note that according to our assumptions $(I-K)$, is invertible and then it results

$$\begin{aligned} I - K_m &= (I-K) + (K - K_m) \\ &= (I-K)[I - (I-K)^{-1}(K_m - K)] \\ &:= (I-K)(I-D), \end{aligned}$$

where

$$D := (I-K)^{-1}(K_m - K).$$

Moreover since we assume that the sequence $\{K_m\}_m$ converges to K, there exists m_0 such that, for any $m > m_0$, we have

$$\|D\| \leq \|(I-K)^{-1}\| \, \|K - K_m\| < \frac{1}{2}.$$

Now by the Von Neuman Theorem the operator $(I-D)^{-1}$ exists and $(I-K_m)^{-1} = (I-D)^{-1}(I-K)^{-1}$ exists too, with

$$\begin{aligned}\|(I-K_m)^{-1}\| &\leq \|(I-K)^{-1}\|\ \|(I-D)^{-1}\| \\ &\leq \frac{\|(I-K)^{-1}\|}{1-\|(I-K)^{-1}\|\|(K-K_m)^{-1}\|}\end{aligned}$$

and (3.10) is proved. □

In the theory of the integral equation, the concept of compact operator is very important, because in general, the classical results on the uniqueness and existence of the solution of an integral equation refer to compact operators. For this reason, now we recall what means that an operator K is compact and in which way it can be characterize.

Definition 3.3. *Let X and Y be two normed spaces and let $K \in \mathcal{B}(X,Y)$. The operator K is compact (or completely continuous) if for each bounded set A of X, $K(A)$ is relatively compact* [2] *in Y. Equivalently, K is compact if every bounded sequence $\{x_n\} \subset X$ contains a subsequence $\{x_{n,k}\}$ such that $\{Kx_{n,k}\}$ converges to a point of Y.*

In the sequel we will denote by $\mathcal{K}(X,Y)$ the set of all compact operators $K : X \to Y$.

We mention that in order to characterize the set relatively compact one can use the Arzelà-Ascoli Theorem or the Riesz-Kolmogorov Theorem (see, for instance, [8,9,16]). Moreover we point out that when the space Y is complete there is a useful characterization of the set relatively compact in terms of precompactness. In fact it can be shown that a subset A of a complete metric space is relatively compact if and only if it is precompact (see, for instance, [16, p. 122-126]) i.e. if for each $\epsilon > 0$, A is contained in the union of a finite number of open balls of radius ϵ. Thus we can state the following.

Theorem 3.4. *Let Y be a Banach space and X be a normed space. Then $K \in \mathcal{K}(X,Y)$ if and only if for every bounded set $B \subset X$ the image $K(B)$ is a precompact set.*

In the following we also need the concept of conjugate operator that we recall briefly here.

Let X and Y be normed spaces and let $K : X \to Y$ be a linear and continuous operator which maps $x \in X$ in $y = Kx \in Y$. Moreover let g an element of the dual space of Y (that we will denote by Y^*) i.e. a linear continuous functional $g : Y \to \mathbb{R}$. Now we apply g to the element Kx:

$$\begin{aligned} g &: Y \to \mathbb{R} \\ Kx &\to g(Kx) := f(x). \end{aligned}$$

Note that $f \in X^*$ i.e. $g(Kx)$ is a continuous linear functional defined on X. Then, in other words, to each functional $g \in Y^*$ corresponds a functional $f \in X^*$ obtaining an operator

$$K^* : Y^* \to X^*$$

such that

$$g(Kx) = K^*(gx).$$

The operator K^* is called *conjugate operator*. The following theorem holds true (see, for instance, [16]).

[2] A set A is relatively compact if its closure $\overline{A}$ is compact.

Theorem 3.5. *Let X and Y be normed spaces and let $K \in \mathcal{B}(X,Y)$. Then $K^* \in \mathcal{B}(Y^*, X^*)$ and*

$$\|K^*\| = \|K\|.$$

Moreover if X and Y are two Banach spaces, and $K \in \mathcal{K}(X,Y)$, then $K^* \in \mathcal{K}(Y^*, X^*)$.

The last part of the theorem is the so-called Schauder Theorem.

3.1.2. *The Fredholm Alternative Theorem.* In this section we give some necessary and sufficient conditions to have the existence and the uniqueness of a large class of Fredholm integral equations of the second kind. They are contained in the so-called Fredholm Alternative Theorem that the Swedish mathematician Eric Ivan Fredholm stated in the early 1900s.

In order to establish it we consider the following

$$(I-K)f = g, \tag{3.15}$$

and the corresponding homogeneous equation

$$(I-K)f = 0 \tag{3.16}$$

whose solution are called *autosolutions*. Moreover we assume that the operator $K \in \mathcal{K}(X,X)$ with X a Banach space. We have the following.

Theorem 3.6. *Let X be a Banach space and $K \in \mathcal{K}(X,X)$. Then* [3]

$$dim(Ker(I-K)) < \infty.$$

Now, we introduce the conjugate operator $K^* : X^* \to X^*$ of K and the following equations

$$(I-K^*)f = g \quad f, g \in X^* \tag{3.17}$$

$$(I-K^*)f = 0. \tag{3.18}$$

Note that since $K \in \mathcal{K}(X,X)$, by Theorem 3.4, it results that $K^* \in \mathcal{K}(X^*, X^*)$.
The following theorems connect the equation (3.15) with (3.18) and (3.16) with (3.18).

Theorem 3.7. *Let X be a Banach space and $K \in \mathcal{K}(X,X)$. Then equation (3.15) has a unique solution for each given right-hand side g if and only if its right-hand side is orthogonal to each solutions of equation (3.18).*

Theorem 3.8. *Let X be a Banach space and $K \in \mathcal{K}(X,X)$. Then equation (3.15) has a unique solution for each given right-hand side g if and only equation (3.16) has only the trivial solution .*

Theorem 3.9. *Let X be a Banach space and $K \in \mathcal{K}(X,X)$. Then*

$$dim(Ker(I-K)) = dim(Ker(I-K^*)),$$

i.e. the homogeneous equations (3.16) and (3.18) have the same finite number of linearly independent solutions.

By means of the previous theorems we can deduce the well-known Alternative Fredholm's Theorem.

Theorem 3.10. *Let X be a Banach space and let $K \in \mathcal{K}(X,X)$. Then there are two alternatives:*

[3]The kernel or null space of a linear operator K, denoted by $Ker(K)$, is the subspace defined as

$$Ker(K) = \{x : x \in X, Kx = 0\}.$$

(i) equation (3.16) has only the trivial solution if and only if (3.15) has a unique solution for each given function g.

(ii) equation (3.16) has not only the trivial solution if and only if, according to the right-hand side g, (3.15) has not solutions or has more than one solution.

Remark 3.11. We note that the first alternative follows by Theorem 3.8. Moreover if (3.16) has not only the trivial solution, then by Theorem 3.6 it has a finite number of linear independent solutions and in virtue of Theorem 3.9, equation (3.18) has the same number of linear independent solutions. Therefore, by Theorem 3.7, (3.15) has solutions if and only if its right hand side g is orthogonal to each of these solutions. Otherwise it has not solutions.

Hence in virtue of this result only two cases are possible . In particular, it is the first case that is useful to our aims. In fact, if the operator K is compact, assuming that the homogeneous equation has only the trivial solution, our equation (3.15) has a unique solution for each right-hand side g.

Therefore, it will be important to find the necessary and sufficient conditions so that the operator K is compact. Next theorem gives these conditions.

Theorem 3.12. *Let X be a Banach space and let $K \in \mathcal{B}(X, X)$. Then K is compact if and only if*

$$\lim_{m\to+\infty} \sup_{\substack{f\in X \\ \|f\|=1}} E_m(Kf) = 0,$$

where $E_m(Kf)$ is the error of best approximation of $Kf \in X$.

3.2. **Numerical methods.** In this section we describe two numerical methods that allow us to approximate the solution of a Fredholm integral equation of the second kind. The first one is the projection method. It is based on the resolution of finite dimensional equations that are equivalent to a linear system. The second one is the Nyström method that uses a suitable quadrature rule and it also brings back to the resolution of a linear system. In both cases we give (see, for instance, [1]) the conditions on which these methods can be applied including the error estimate.

3.2.1. *The Projection Method.* We consider the operator equation (3.15) i.e.

$$(I - K)f = g.$$

Assume that $Ker\{I - K\} = \{0\}$ in X where $K \in \mathcal{K}(X, X)$ with X a Banach space equipped with the norm $\|\cdot\|_X$ and denote by f^* the unique solution of the given equation. In order to approximate it we proceed in the following way.

We choose a sequence of finite dimensional subspaces $X_m \subset X$, $m \geq 1$, and we introduce a projector $\mathcal{P}_m : X \to X_m$ i.e. a bounded operator such that $\mathcal{P}_m x = x$, $\forall x \in X_m$.

Then we consider the following finite dimensional equations

$$\mathcal{P}_m(I - K)f_m = g_m$$

i.e.

$$(I - \mathcal{P}_m K)f_m = g_m \tag{3.19}$$

where for example $g_m = \mathcal{P}_m g$.

Now if we assume that

$$\|g - g_m\|_X \to 0 \tag{3.20}$$

and

$$\|K - \mathcal{P}_m K\| \to 0 \tag{3.21}$$

where the operator norm is defined in (3.8), then, by a standard argument (see, for instance [1]), it is possible to prove the following.

Theorem 3.13. *Assume that $Ker(I-K)=\{0\}$ in X and conditions (3.21) and (3.20) are fulfilled. Then, for m sufficiently large , equation (3.19) has a unique solution $f_m^*\in X_m$. Moreover*

$$\|f^*-f_m^*\|_X \le \mathcal{C}\left(\|g-g_m\|_X+\|K-\mathcal{P}_mK\|\cdot\|g\|_X\right) \tag{3.22}$$

with $\mathcal{C}$ independent of f and m, and

$$|\mathrm{cond}\,(I-K)-\mathrm{cond}\,(I-\mathcal{P}_mK)| = O(\|K-\mathcal{P}_mK\|) \tag{3.23}$$

where $\mathrm{cond}(T)=\|T\|\cdot\left\|T^{-1}\right\|$ *denotes the condition number in infinity norm of an invertible operator T.*

Usually, the set of all algebraic polynomials of degree at most $m-1$ is chosen as subspace X_m and the Lagrange polynomial $L_m(\tilde{X},f,x)$ defined in (1.19), with $\tilde{X}$ an arbitrary triangular infinity matrix of knots belonging to $[a,b]$, is the projector $\mathcal{P}_m$.

Moreover it is possible to prove (see, for instance, [4]) that equations (3.19) are equivalent to a system of m equations in m unknowns and denoting by A_m the matrix of the obtaining system the following relation holds true:

$$cond(A_m)\le cond(I-K)\|\Lambda(\tilde{X})\|_\infty^2$$

where $\Lambda_m(\tilde{X})$ is the Lebesgue constants defined in (1.23).

Therefore if the entries of the matrix $\tilde{X}$ are the equal-spaced points, we have

$$\|\Lambda_m(\tilde{X})\|\sim\frac{2^m}{em\log m}$$

and we can have serious problems in the computations of the solution of the system. Then, the choice of matrices of knots $\tilde{X}$ for which it results $\|\Lambda_m(\tilde{X})\|_\infty\sim\log m$ is recommend.

3.2.2. *The Nyström Method.* One of the most common numerical methods used to approximate the solution of equation (3.7) is the Nyström method. It is based on the following procedure.

We consider (3.7) i.e.

$$f(y)-\mu\int_a^b k(x,y)f(x)dx=g(y)$$

and approximate the integral by means of a suitable quadrature formula

$$(G_mf)(y)=\sum_{i=1}^m w_{m,i}k(x_{m,i},y)f(x_{m,i}) \tag{3.24}$$

in which we have denoted by $w_{m,i}$ its coefficients and by $x_{m,i}$ the quadrature nodes. We assume that it is stable i.e.

$$\sup_m\|G_m\|=\sup_m\sum_{i=1}^m|w_{m,i}|<\infty$$

and is convergent i.e. if $m\to\infty$ the sum G_mf converges to the integral.

Then we consider the following

$$f(y)-\mu(G_mf)(y)=g(y) \tag{3.25}$$

and we compute it in the nodes x_i. In this way we obtain the following linear system of order m

$$f(x_{m,j}) - \mu \sum_{i=1}^{m} w_{m,i} k(x_{m,i}, x_{m,j}) f(x_{m,i}) = g(x_{m,j}) \quad j = 1, ..., m \tag{3.26}$$

that is equivalent to (3.25).

In fact, each solution $f^*(y)$ of equation furnishes a solution $f^* \equiv [f^*(x_1), ..., f^*(x_m)]$ for the system (3.26) and vice versa for each solution of system (3.26) there is a unique solution for the equation (3.25) given by

$$f_m^*(y) = g(y) + \mu \sum_{i=1}^{m} w_i k(x_i, y) f^*(x_i).$$

The formula just written is called *Nyström interpolation formula.*

Of course we have to prove that (3.25) or (3.26) admit a unique solution. To this end the following theorem is useful (see, for instance, [1, Chapter 4]).

Theorem 3.14. *Let $(X, \|\cdot\|)$ a Banach space of continuous functions, $K : X \to X$ a compact operator, $Ker\{I - K\} = \{0\}$ and denote with $f^* \in X$ the unique solution of the given equation. Then if*

- $\sup_m \|G_m\| < \infty$
- $\|Kf - G_m f\|$ *tends to zero for any* $f \in X$
- $\|(K - G_m)G_m\|$ *tends to zero*

for m sufficiently large, the operator $(I - G_m)^{-1}$ exists and it is uniformly bounded since

$$\|(I - G_m)^{-1}\| \leq \frac{1 + \|(I - K)^{-1}\|\|G_m\|}{1 - \|(I - K)^{-1}\|\|(K - G_m)G_m\|}.$$

Moreover it results

$$\|f^* - f_m^*\| \leq \|(I - G_m)^{-1}\| \quad \|(K - G_m)f^*\| \tag{3.27}$$

and

$$cond(A_m) \leq cond(I - G_m) = \|I - G_m\| \ \|(I - G_m)^{-1}\| \tag{3.28}$$

where $cond(A_m)$ denotes the condition number of the matrix of coefficients of system (3.26).

Remark 3.15. Note that $\sup_m \|G_m\| < \infty$ follows by the statement $\|Kf - G_m f\| \to 0$, $\forall f \in X$ in virtue of the principle of uniform boundedness.

References

[1] K. E. Atkinson, *The Numerical Solution of Integral Equations of the second kind*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 1997. 30, 31, 32

[2] S. Bernstein, *Quelques remarques sur l'interpolation*, Zap. Kharkov Mat. Ob-va (Comm. Kharkov Math. Soc.), **15** (2) (1916), 49-61. 9

[3] S. N. Bernstein, *Quelques remarques sur l'interpolation*, Math. Ann., **79** (1918), 1-12. 9

[4] M. C. De Bonis, G. Mastroianni, *Projection Methods and condition numbers in uniform norm for Fredholm and Cauchy singular integral equations*, SIAM Journal on Numerical Analysis, **44** (2006), no 4, 1351-1374. 31

[5] G. Faber, *Über die interpolatorische Darstellung stetiger Funktionen*, Jahresber. der dentschen Math. Verein, **23**(1914), 190-210. 8

[6] G. Grünwald, *Über Divergenzerscheinungen der agrangeschen Interpolationspolynome*, Acta.Sci. Math (Szeged), **7** (1935), 207-221.

[7] G. Grünwald, *Über Divergenzerscheinungen der agrangeschen Interpolationspolynome stetiger Funktionen*, Ann. Math., **37** (1936), 908-918.

[8] L. V. Kantarovic, G.P. Akilov, *Analisi funzionale*, Edizioni Mir- Editori Riuniti, 1980. 26, 28
[9] A.N. Kolmogorov and S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Graylock Press, Rochester, N.Y., 1957. 26, 28
[10] J. Marczinkiewicz, *Sur la divergence des polynomes d'interpolation*, Acta Sci. Math (Szeged), **8** (1937), 131-135.
[11] G. Mastroianni, *Uniform convergence of derivatives of Lagrange interpolation*, Journal of Computational and Applied Mathematics, **43** (1992), 37-51. 10
[12] G. Mastroianni and G. V. Milovanovic, *Interpolation Processes Basic Theory and Applications*, Series Springer Monographs in Mathematics 2009. 24
[13] P. Nevai, *Orthogonal polynomials*, Memoirs Amer. Math. Soc., vol. 213 (Amer. MAthematical Soc., Providence, RI, 1979).
[14] P.O. Runck, P. Vèrtesi, *Some good point systems for derivatives of Lagrange interpolatory operators*, Acta Math. Hungarica, **56** (1990), 16-26. 10
[15] G. Szëgo, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, 4th ed., Amer. Math Soc, Providence R.I., 1975. 9
[16] A.E. Taylor and D. C. Lay, *Introduction to functional analysis*, Robert E. Krieger Publishing Company, Malabar, Florida, 1986. 26, 28
[17] A.F. Timan, *Theory of approximation of functions of a real variable*, The Macmillan Co., New York, 1963. 9

Department of Mathematics and Computer Science
University of Cagliari
Viale L. Merello 92, 09123 Cagliari, Italy
Email address: fermo@unica.it