Approximation Theory and Proof Assistants: Certified Computations

Master d'informatique fondamentale
École normale supérieure de Lyon
Fall 2020

NICOLAS BRISEBARRE   DAMIEN POUS

http://perso.ens-lyon.fr/nicolas.brisebarre/M2R/CoqApprox/

Contents

Chapter 1

Polynomial approximations

In this chapter, we present various theoretical and algorithmic results regarding polynomial approximations of functions. We will mainly deal with real-valued continuous functions over a compact interval $[a, b]$, $a, b \in \mathbb{R}$, $a \leqslant b$. We will denote $\mathcal{C}([a, b])$ the real vector space of continuous functions over $[a, b]$. In the framework of function evaluation one usually works with the following two norms over this vector space:

- the least-square norm L^2: given a weight[1] function w, if $\mathrm{d}x$ denotes the Lebesgue measure, we write
$$g \in L^2([a, b], w, \mathrm{d}x)$$
if
$$\int_a^b w(x)|g(x)|^2 \mathrm{d}x < \infty,$$
and then we define
$$\|g\|_{2,w} = \sqrt{\int_a^b w(x)|g(x)|^2 \mathrm{d}x};$$
- the supremum norm (aka Chebyshev norm, infinity norm, L^∞ norm) : if g is bounded on $[a, b]$, we set
$$\|g\|_\infty = \sup_{x \in [a,b]} |g(x)|,$$
(observe that for a continuous function g, we have $\|g\|_\infty = \max_{x \in [a,b]} |g(x)|$).

For both norms, one of the main questions we are interested in here is the following.

Problem 1.1. *(Best approximation) Given $f \in \mathcal{C}([a, b])$ and $n \in \mathbb{N}$, minimize $\|f - p\|$ where p describes the space $\mathbb{R}_n[x]$ of polynomials with real number coefficients and degree at most n.*

In the L^2 case, the answer to this question is easy. The space $\mathcal{C}([a, b])$ is a subset of $L^2([a, b], w, \mathrm{d}x)$ which is a Hilbert space, i.e. a vector space equipped with an inner product

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

and $\| \cdot \|_2$ is the associated norm, for which $L^2([a, b], w, \mathrm{d}x)$ is complete. The best polynomial approximation of degree at most n is the projection $p = \mathrm{pr}^\perp(f)$ of f onto $\mathbb{R}_n[x]$. We will give more details on the L^2 case in Chapter 2. The situation in the L^∞ case is more intricate and we will focus on it in the sequel of this chapter.

[1]Here, we will assume that it means that $w \in \mathcal{C}((a, b))$ and $w > 0$ almost everywhere.

1.1 Density of the polynomials in $(\mathcal{C}([a,b]), \|.\|_\infty)$

For all $f \in \mathcal{C}([a,b])$ and $n \in \mathbb{N}$, let

$$E_n(f) = \inf_{p \in \mathbb{R}_n[x]} \|f - p\|_\infty.$$

We first recall that $E_n(f) \to 0$ as $n \to \infty$, a result due to Weierstraß :

Theorem 1.2. *[Weierstraß, 1885] For all $f \in \mathcal{C}([a,b])$ and for all $\varepsilon > 0$, there exists $n \in \mathbb{N}$, $p \in \mathbb{R}_n[x]$ such that $\|p - f\|_\infty < \varepsilon$.*

Various proofs of this result have been published, in particular, those by Runge (1885), Picard (1891), Lerch (1892 and 1903), Volterra (1897) Lebesgue (1898), Mittag-Leffler (1900), Fejér (1900 and 1916), Landau (1908), la Vallée Poussin (1908), Jackson (1911), Sierpinski (1911), Bernstein (1912), Montel (1918). The text [Pinkus, 2000] is an interesting account on Weierstraß' contribution to Approximation Theory and, in particular, his fundamental result on the density of polynomials in $\mathcal{C}([a,b])$ stated in Theorem 1.2.

We give now one proof inspired by Bernstein's one.

Proof of Theorem 1.2. Up to a change of variable, we can assume $[a,b] = [0,1]$. Define the Bernstein polynomials as

$$B_n(g,x) = \sum_{k=0}^{n} \binom{n}{k} g(k/n) x^k (1-x)^{n-k} \text{ for } g \in \mathcal{C}([0,1]).$$

We have

$$\begin{aligned}
B_n(1,x) &= \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1, \\
B_n(x,x) &= \sum_{k=0}^{n} \binom{n}{k} \frac{k}{n} x^k (1-x)^{n-k} = x \sum_{k=1}^{n} \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k} \\
&= x \sum_{k=0}^{n-1} \binom{n-1}{k} x^k (1-x)^{n-1-k} = x, \\
B_n(x^2,x) &= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^2 x^k (1-x)^{n-k} = x \sum_{k=1}^{n} \frac{k}{n} \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k}. \\
&= x \sum_{k=1}^{n} \frac{k-1}{n} \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k} + \frac{x}{n} \\
&= \frac{n-1}{n} x^2 \sum_{k=2}^{n} \binom{n-2}{k-2} x^{k-2} (1-x)^{n-k} + \frac{x}{n} = \frac{x}{n} + x^2 \frac{n-1}{n}.
\end{aligned}$$

Now consider the sequence

$$f(x) - B_n(f,x) = \sum_{k=0}^{n} (f(x) - f(k/n)) b_{n,k}(x) \text{ where } b_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}, \forall x \in [0,1].$$

Fix $\varepsilon > 0$. The function f is continuous and hence uniformly continuous over $[0,1]$, hence there exists $\delta > 0$ such that

$$\forall x_1, x_2 \in [0,1], \quad |x_2 - x_1| < \delta \quad \Rightarrow \quad |f(x_2) - f(x_1)| < \varepsilon.$$

Let $M = \max_{x \in [0,1]} |f(x)|$. Since $b_{n,k}(x) \geqslant 0$ for all $x \in [0,1]$, we can write

$$\begin{aligned}|f(x)-B_n(f,x)| &\leqslant \left|\sum_{\substack{k=0\\|x-k/n|<\delta}}^{n}(f(x)-f(k/n))b_{n,k}(x)\right| + \left|\sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}(f(x)-f(k/n))b_{n,k}(x)\right| \\ &\leqslant \varepsilon\sum_{k=0}^{n}b_{n,k}(x) + 2M\sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}b_{n,k}(x) = \varepsilon + 2M\sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}b_{n,k}(x).\end{aligned}$$

Note that we actually have

$$\begin{aligned}\left|\sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}b_{n,k}(x)\right| &= \sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}b_{n,k}(x) \text{ since } b_{n,k}(x)\geqslant 0 \\ &\leqslant \sum_{\substack{k=0\\|x-k/n|\geqslant\delta}}^{n}\left(\frac{x-k/n}{\delta}\right)^2 b_{n,k}(x) \\ &\leqslant \sum_{k=0}^{n}\left(\frac{x-k/n}{\delta}\right)^2 b_{n,k}(x) \\ &= \frac{1}{\delta^2}\left(x^2-2x^2+x^2\frac{n-1}{n}+\frac{x}{n}\right) = \frac{x(1-x)}{n\delta^2}.\end{aligned}$$

Therefore, we obtained $|f(x)-B_n(f,x)| \leqslant \varepsilon + \frac{M}{2n\delta^2}$. The upper bound does not depend on x and can be made as small as desired. □

Remark 1.3. One of the very nice features of this proof is that it provides an explicit sequence of polynomials which converges to the function f. It is worth mentioning that Bernstein polynomials prove useful in various other domains (computer graphics, global optimization, ...). See [Farouki, 2012] for instance.

Note that, in the proof, we only used the values of the $B_n(f,x)$ for $0 \leqslant n \leqslant 2$. In fact, we have the following result.

Theorem 1.4 (Bohman and Korovkin). *Let L_n be a sequence of monotone linear operators on $\mathcal{C}(|a,b])$, that is to say: for all $f, g \in \mathcal{C}([a,b])$*

- *$L_n(\mu f + \lambda g) = \lambda L_n(f) + \mu L_n(g)$ for all $\lambda, \mu \in \mathbb{R}$,*
- *if $f(x) \geqslant g(x)$ for all $x \in [a,b]$ then $L_n f(x) \geqslant L_n g(x)$ for all $x \in [a,b]$,*

the following conditions are equivalent:

i. $L_n f \to f$ uniformly for all $f \in \mathcal{C}([a,b])$;

ii. $L_n f \to f$ uniformly for the three functions $x \mapsto 1, x, x^2$;

iii. $L_n 1 \to 1$ and $(L_n\phi_t)(t) \to 0$ uniformly in $t \in [a,b]$ where $\phi_t : x \in [a,b] \mapsto (t-x)^2$.

Proof. See [Cheney, 1998]. □

Actually, a refinement of Weierstraß's theorem yields a statement about the speed of convergence of the $B_n(f,x)$ to f. It is obtained in terms of the modulus of continuity.

Definition 1.5. *The modulus of continuity of f is the function ω defined as*

$$\text{for all } \delta > 0, \qquad \omega(\delta) = \sup_{\substack{|x-y|<\delta, \\ x,y\in[a,b]}} |f(x) - f(y)|.$$

Proposition 1.6. *If f is a continuous function over $[0,1]$, ω its modulus of continuity, then*

$$\|f - B_n(f,x)\|_\infty \leqslant \tfrac{9}{4}\omega\left(n^{-\frac{1}{2}}\right).$$

Proof. Let $\delta > 0$ and $x \in [0,1]$. Let $k \in \{0,\dots,n\}$ such that $|x - k/n| \leqslant \delta$, then $|f(x) - f(k/n)| \leqslant w(\delta)$. Since $b_{n,k}(y) \geqslant 0$ for all $y \in [0,1]$, we have

$$\left| \sum_{\substack{k=0 \\ |x-k/n|<\delta}}^{n} (f(x) - f(k/n))b_{n,k}(x) \right| \leqslant w(\delta) \sum_{k=0}^{n} b_{n,k}(x) = w(\delta).$$

Now, let $k \in \{0,\dots,n\}$ such that $|x - k/n| \geqslant \delta$. Let $M = \left\lfloor \frac{|x-k/n|}{\delta} \right\rfloor$, let $y_j = x + \frac{j}{M+1}(k/n - x)$ for $j = 0,\dots,M+1$. Note that, for all $j = 0,\dots,M$, we have $|y_{j+1} - y_j| < \delta$, from which follows

$$\begin{aligned} |f(x) - f(k/n)| &\leqslant \sum_{j=0}^{M} |f(y_{j+1}) - f(y_j)| \leqslant (M+1)w(\delta) \\ &\leqslant w(\delta)\left(1 + \frac{1}{\delta}\left|x - \frac{k}{n}\right|\right) \leqslant w(\delta)\left(1 + \frac{1}{\delta^2}\left(x - \frac{k}{n}\right)^2\right). \end{aligned}$$

For all $x \in [0,1]$, we can write

$$\begin{aligned} |f(x) - B_n(f,x)| \leqslant & \left| \sum_{\substack{k=0 \\ |x-k/n|<\delta}}^{n} (f(x) - f(k/n))b_{n,k}(x) \right| + \left| \sum_{\substack{k=0 \\ |x-k/n|\geqslant\delta}}^{n} (f(x) - f(k/n))b_{n,k}(x) \right| \\ \leqslant & w(\delta) + \sum_{\substack{k=0 \\ |x-k/n|\geqslant\delta}}^{n} w(\delta)\left(1 + \frac{1}{\delta^2}\left|x - \frac{k}{n}\right|^2\right) b_{n,k}(x) \\ \leqslant & w(\delta)\left(2 + \frac{1}{\delta^2} \sum_{\substack{k=0 \\ |x-k/n|\geqslant\delta}}^{n} \left(x - \frac{k}{n}\right)^2 b_{n,k}(x)\right) \\ \leqslant & w(\delta)\left(2 + \frac{x(1-x)}{n\delta^2}\right) \leqslant w(\delta)\left(2 + \frac{1}{4n\delta^2}\right). \end{aligned}$$

Finally, replace δ with $1/\sqrt{n}$. □

Corollary 1.7. *When f is Lipschitz continuous, $E_n(f) = O(n^{-1/2})$.*

Remark 1.8. For improvements and refinements of these results, see Section 4.6 of [Cheney, 1998] or Chapter 16 of [Powell, 1981] for a presentation of Jackson theorems. See also Theorem 2.23.

1.2 Best L^∞ (or minimax) approximation

The infimum $E_n(f)$ is reached, thanks to the following proposition.

Proposition 1.9. *Let $(E, \|\cdot\|)$ be a normed $\mathbb{R}$-vector space, let F be a finite dimensional subspace of $(E, \|\cdot\|)$. For all $f \in E$, there exists $p \in F$ such that $\|p - f\| = \min_{q\in F} \|q - f\|$. Moreover, the set of best approximations to a given $f \in E$ is convex.*

Proof. Let $f \in E$. Consider $F_0 = \{p \in F : \|p\| \leqslant 2\|f\|\}$. Then F_0 is nonempty (it contains 0), closed, bounded, and we assumed $\dim F < \infty$. Hence F_0 is compact. Let $\varphi(p) = \|f - p\|$. The function φ is 1-Lipschitz and hence continuous. It follows that $\varphi(F_0)$ is compact, which implies the existence of $p^\star \in F_0$ s.t. $\varphi(p^\star) = \min_{p\in F_0} \|f - p\|$. Moreover, if $p \in F \setminus F_0$, $\|f - p\| \geqslant \|p\| - \|f| > \|f\| \geqslant \varphi(p^\star)$ since $0 \in F_0$. Thus, $\|f - p^\star\| = \min_{p\in F} \|f - p\|$.

Now, let p and $q \in F$ be two best approximations to f. For all $\lambda \in [0, 1]$, the vector $\lambda p + (1 - \lambda)q$ is an element of the vector space F and we have, from the triangle inequality, $\|\lambda p + (1 - \lambda)q - f\| \leqslant \lambda\|p - f\| + (1 - \lambda)\|q - f\| = \min_{q\in F} \|q - f\|$: the vector $\lambda p + (1 - \lambda)q$ is also a best approximation to f. $\square$

The best L^2 approximation is unique, which is not always the case in the L^∞ setting.

Exercise 1.2.1. Consider the following simple situation : the interval is $[-1, 1]$, f is the constant function 1 and $F = \mathbb{R}g$ where $g : x \to x^2$. Determine the set of best L^∞ approximations to f.

In the case of L^∞, it is necessary to introduce an additional condition known as the Haar condition.

Definition 1.10. *Consider $n + 1$ functions $\varphi_0, \dots, \varphi_n$ defined over $[a, b]$. We say that $\varphi_0, \dots, \varphi_n$ satisfy the Haar condition iff*

a) the φ_i are continuous;

b) and the following equivalent statements hold:

- *for all $x_0, x_1, \dots, x_n \in [a, b]$,*

$$\begin{vmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix} = 0 \quad \Leftrightarrow \quad \exists i \neq j, x_i = x_j;$$

- *given pairwise distinct $x_0, \dots, x_n \in [a, b]$ and values $y_0, \dots, y_n$, there exists a unique interpolant*

$$p = \sum_{k=0}^{n} \alpha_k \varphi_k, \text{ with } \alpha_k \in \mathbb{R}, \forall k = 0, \dots, n,$$

such that $p(x_i) = y_i$, $\forall i = 0, \dots, n$;

- *any $p = \sum_{k=0}^{n} \alpha_k \varphi_k \neq 0$ has at most n distinct zeros in $[a, b]$.*

Exercise 1.2.2. Prove that the conditions above are equivalent.

A set of functions that satisfy the Haar condition is called a *Chebyshev system*. The prototype example is $\varphi_i(x) = x^i$, for which we have

$$\begin{vmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix} = \begin{vmatrix} 1 & \cdots & x_0^n \\ \vdots & & \vdots \\ 1 & \cdots & x_n^n \end{vmatrix} = V_n = \prod_{0\leqslant i<j\leqslant n} (x_j - x_i). \tag{1.1}$$

(Sketch of a proof: considering $x_n = z$ as an indeterminate and looking at the roots of the polynomial V_n, we see that $V_n = V_{n-1}(z - x_0) \cdots (z - x_{n-1})$.)

Exercise 1.2.3. Show that the following families of functions are Chebyshev systems as well:

- $\{e^{\lambda_0 x}, \dots, e^{\lambda_n x}\}$ for $\lambda_0 < \lambda_1 < \cdots < \lambda_n$;
- $\{1, \cos x, \sin x, \dots, \cos(nx), \sin(nx)\}$ over $[a, b]$ where $0 \leqslant a < b < 2\pi$;
- $\{x^{\alpha_0}, \dots, x^{\alpha_n}\}$, $\alpha_0 < \cdots < \alpha_n$, over $[a, b]$ with $a > 0$.

Let E be a real vector space, $e_1, e_2, \dots, e_m \in E$, we will denote $\mathrm{Span}_{\mathbb{R}}\{e_1, \dots, e_m\}$ the set

$$\mathrm{Span}_{\mathbb{R}}\{e_1, \dots, e_m\} = \left\{ \sum_{k=1}^{m} \alpha_k e_k ; \alpha_1, \dots, \alpha_m \in \mathbb{R} \right\}.$$

If $\{\varphi_0, \dots, \varphi_n\}$ is a Chebyshev system over $[a, b]$, any element of $\mathrm{Span}_{\mathbb{R}}\{\varphi_0, \dots, \varphi_n\}$ will be called a generalized polynomial.

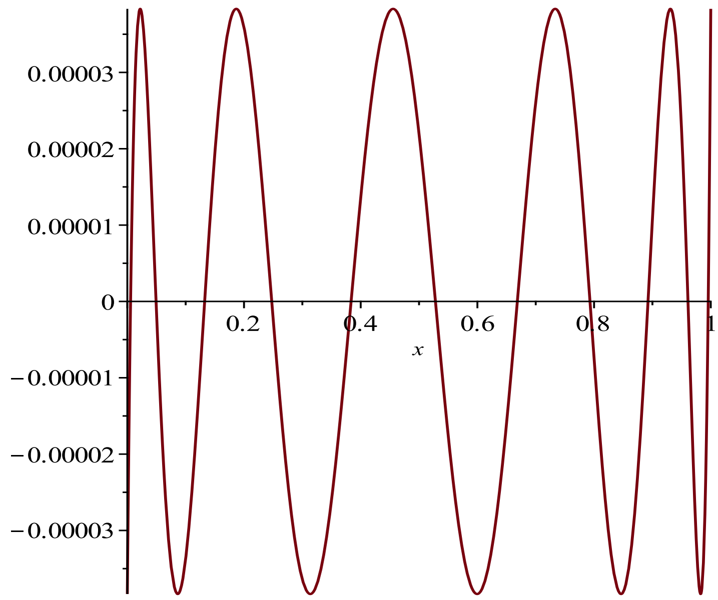
Beside its beauty, the following characterization of the minimax approximation proved crucial to the design of an algorithm for computing it.

Theorem 1.11. *[Alternation Theorem. Kirchberger (1902)] Let $\{\varphi_0, \dots, \varphi_n\}$ be a Chebyshev system over $[a, b]$. Let $f \in \mathcal{C}([a, b])$. A generalized polynomial $p = \sum_{k=0}^{n} \alpha_k \varphi_k$ is* the *best approximation (or minimax approximation) to f iff there exist $n + 2$ points $x_0, \dots, x_{n+1}$, $a \leqslant x_0 < x_1 < \cdots < x_{n+1} \leqslant b$ such that, for all k,*

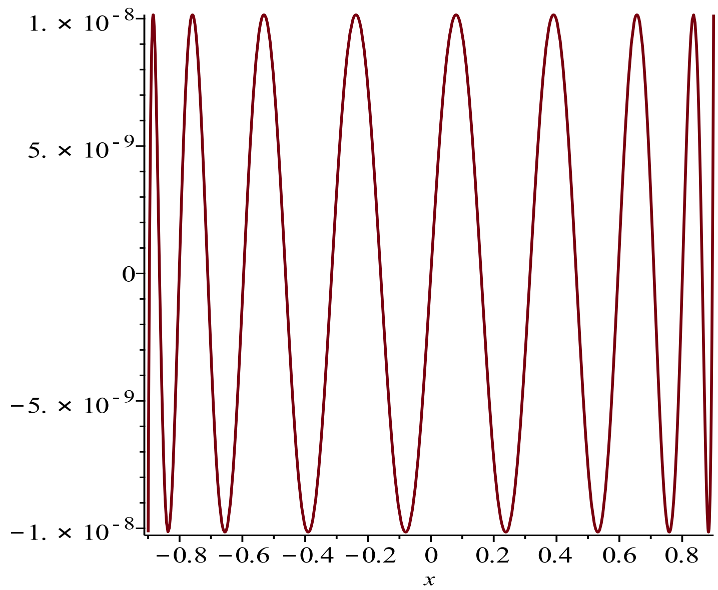
$$f(x_k) - p(x_k) = (-1)^k (f(x_0) - p(x_0)) = \pm \|f - p\|_\infty.$$

In other words, p is the best approximation if and only if the error function $f - p$ has (at least) $n + 2$ extrema, all global (of the same absolute value) and with alternating signs.

Example 1.12. Let $f : x \in [0, 1] \longmapsto e^{1/\cos(x)}$, $p = \sum_{k=0}^{10} c_k x^k$ its minimax approximation. The graph of the error function $\varepsilon = f - p$ is:



Example 1.13. Let $f : x \in [-0.9, 0.9] \longmapsto \arctan(x)$, $p = \sum_{k=0}^{15} c_k x^k$ its minimax approximation. The graph of the error function $\varepsilon = f - p$ is:



Example 1.14. The best approximation to $\cos$ over $[0, 10\pi]$ on the Chebyshev system $\{1, x, x^2\}$ is the constant function 0! Moreover, the same is true for $\{1, x, \dots, x^h\}$ up to and including $h = 9$.

Proof. We can assume that $f \notin \mathrm{Span}_{\mathbb{R}}\{\varphi_0, \dots, \varphi_n\}$.

We already proved the existence of a best approximation.

We now show that the equioscillation property implies optimality of the approximation. Let p be an approximation with equioscillating error function, and suppose that there exists $q = \sum_{j=0}^{n} \beta_j \varphi_j$ with $\|f-q\|_\infty < \|f-p\|_\infty$. Writing $p-q = (p-f)-(q-f)$, we see that $p-q$ changes sign between each pair of consecutive x_i. It follows from the intermediate value theorem that there exist $(n+1)$ points $y_0, \dots, y_n$ such that $x_0 < y_0 < x_1 < \dots < x_n < y_n < x_{n+1}$ and $p(y_i) = q(y_i)$. By definition of a Chebyshev system, this implies that $p = q$.

Conversely, optimality implies equioscillation.

For simplicity, we assume that $\{\varphi_0, \dots, \varphi_n\} = \{1, x, \dots, x^n\}$ (see [Cheney, 1998, Powell, 1981] for a proof of the general case). Let p be a best approximation. First, note that the global minimum and the global maximum of $f - p$ must have the same absolute value: otherwise, we can improve the approximation by shifting p by a constant. Now suppose that $f - p$ equioscillates at ℓ points $x_0 < x_1 < \dots < x_\ell$ at most, with $1 \leqslant \ell < n + 1$. We can choose the $\{x_i\}_{0 \leqslant i \leqslant \ell}$ as follows.

The point x_0 is the smallest number in $[a, b]$ at which $|p - f|$ reaches its maximum: the set $A = \{x \in [a, b] : |p(x) - f(x)| = \|p - f\|_\infty\}$ is nonempty, bounded and closed since $|p - f|$ is continuous and $A = (|p - f|)^{-1}(\{\|p - f\|_\infty\}) \cap [a, b]$: A is compact and let x_0 be the minimum of A. Likewise, the point x_1 is defined as the smallest number in $[x_0, b]$ at which $p - f$ is equal to $-(p - f)(x_0)$ and so on and so forth.

Assume wlog that $p(x_0) - f(x_0) = -\|p - f\|_\infty$. For $j = 0, \dots, \ell - 1$, let $B_j = \{x_j \leqslant x \leqslant x_{j+1}; p(x) = f(x)\}$, the set B_j is nonempty since $(p(x_j) - f(x_j))(p(x_{j+1}) - f(x_{j+1})) < 0$, closed ($p - f$ is continuous and $B_j = (p - f)^{-1}(\{0\}) \cap [x_j, x_{j+1}]$) and bounded: it is a compact set, which has a maximum y_{j+1} distinct from x_{j+1}. We remark that $y_1 < y_2 < \dots < y_\ell$, in particular.

We now define $Q(x) = (y_1 - x) \cdots (y_\ell - x)$. Note that $Q(a) \geqslant 0$. If we set $y_0 = a$ and $y_{\ell+1} = b$, let

$$K_1 = [a, y_1] \cup [y_2, y_3] \cup \dots = \bigcup_{k=0}^{\lfloor \ell/2 \rfloor} [y_{2k}, y_{2k+1}],$$

$$K_2 = [y_1, y_2] \cup [y_3, y_4] \cup \dots = \bigcup_{k=0}^{\lfloor (\ell-1)/2 \rfloor} [y_{2k+1}, y_{2k+2}].$$

The sets K_1 and K_2 are finite unions of compact sets, and hence compact. We have:

- $[a, b] = K_1 \cup K_2$ and $K_1 \cap K_2 = \{y_k\}_{1 \leqslant k \leqslant \ell}$,
- $-\|p - f\|_\infty \leqslant p - f < \|p - f\|_\infty$, $Q \geqslant 0$ on the compact K_1 and $Q > 0$ on $K_1 \setminus \{y_k\}_{1 \leqslant k \leqslant \ell}$,
- $-\|p - f\|_\infty < p - f \leqslant \|p - f\|_\infty$, $Q \leqslant 0$ on the compact K_2 and $Q < 0$ sur $K_2 \setminus \{y_k\}_{1 \leqslant k \leqslant \ell}$,

Hence there exists $\lambda \in (0, +\infty)$ such that

$$-\|p - f\|_\infty < p + \lambda Q - f < \|p - f\|_\infty,$$

which contradicts the optimality of p.

Finally, let us prove the uniqueness. Let p, q be two best approximations, and let

$$\mu = \|f - p\|_\infty = \|f - q\|_\infty.$$

It follows from Proposition 1.9 that $\frac{1}{2}(p+q)$ is a best approximation too. Thus there exist $t_0 < t_1 < \cdots < t_{n+1}$ such that

$$\left(\frac{p+q}{2}\right)(t_i) - f(t_i) = \pm(-1)^i \mu.$$

Thus, we have $p(t_i) - f(t_i) = q(t_i) - f(t_i) (= \pm(-1)^i \mu)$ for all $i = 0, \ldots, n+1$, and hence $p = q$ by the Haar condition.

Hint for the choice of λ: Let $M = \max_{x \in [a,b]}(p - f)(x), M_1 = \max_{x \in K_1}(p - f)(x)$ and $m_2 = \min_{x \in K_2}(p - f)(x)$. Prove that any

$$0 < \lambda < \min\left(\frac{M - M_1}{\max_{x \in K_1} Q(x)}, -\frac{M + m_2}{\min_{x \in K_2} Q(x)}\right) \text{ works.}$$

□

Next result is also a key element for the design of an algorithm for computing the minimax approximation.

Theorem 1.15. *(La Vallée Poussin) Let $f \in \mathcal{C}([a, b])$. Let $\{\varphi_0, \ldots, \varphi_n\}$ be a Chebyshev system over $[a, b]$, and let $p \in \mathrm{Span}_{\mathbb{R}}\{\varphi_0, \ldots, \varphi_n\}$. If there exist $x_0 < x_1 < \cdots < x_{n+1}$ such that $p - f$ alternates at the x_i, then*

$$\min_i |f(x_i) - p(x_i)| \leqslant E_n(f) \leqslant \|f - p\|_\infty,$$

where $E_n(f) = \inf_{q \in \mathrm{Span}_{\mathbb{R}}\{\varphi_i\}} \|f - q\|_\infty$.

Proof. The second inequality is obvious. If the first one does not hold, assume wlog that $f(x_0) > p(x_0)$. Then, if $p^*$is the best approximation of f, we have, for all $k = 0, \ldots, n+1$, $(-1)^k(f(x_k) - p(x_k)) > (-1)^k(f(x_k) - p^*(x_k))$: the generalized polynomial $p - p^*$ changes sign $n+1$ times over $[a, b]$, which is not possible. □

Remark 1.16. Let $\{\varphi_0, \ldots, \varphi_n\}$ be a Chebyshev set over $[a, b]$. The statements from Theorems 1.11 and 1.15 remain valid if $[a, b]$ is replaced with any closed subset of $[a, b]$ containing at least $n + 2$ points (see [Cheney, 1998]).

Now let's see whether the Haar condition is necessary or not to guarantee uniqueness. Before stating the result, we introduce the function $\mathrm{sgn} : \mathbb{R} \to \{-1, 0, 1\}$:

$$\mathrm{sgn}(x) = \begin{cases} 1 \text{ if } x > 0, \\ 0 \text{ if } x = 0, \\ -1 \text{ otherwise.} \end{cases}$$

Theorem 1.17 (Haar's Uniqueness Theorem)**.** *Let $\{\varphi_0, \ldots, \varphi_n\}$ be a set of continuous functions over $[a, b]$, linearly independent over $\mathbb{R}$. The minimax approximation to a continuous function f by a generalized polynomial $p = \sum_{k=0}^n \alpha_k \varphi_k$ is unique for all choices of f iff $\{\varphi_0, \ldots, \varphi_n\}$ satisfies the Haar condition.*

Proof. We already proved the "only if" direction in Theorem 1.11.

We assume that $\{\varphi_0, \dots, \varphi_n\}$ does not satisfy the Haar condition. Then, there exist $a \leqslant x_0 < \dots < x_n \leqslant b$ such that $|\varphi_j(x_i)|_{0\leqslant i,j\leqslant n} = 0$. Hence, there are $(a_0, \dots, a_n)$ and $(b_0, \dots, b_n) \in \mathbb{R}^{n+1} \setminus \{0\}$ such that $\sum_{i=0}^n a_i g_i(x_j) = 0$ for all $j = 0, \dots, n$ and $\sum_{j=0}^n b_j g_i(x_j) = 0$ for all $i = 0, \dots, n$. The latter implies

$$\sum_{j=0}^{n} b_j P(x_j) = 0 \text{ for all generalized polynomial } P. \tag{1.2}$$

Let $Q(x) = \sum_{i=0}^n a_i g_i(x)$. We have $Q(x_j) = 0$ for all $j = 0, \dots, n$ and we may assume that $||Q||_\infty < 1$. Now, we consider $f \in \mathcal{C}([a,b])$ such that $||f||_\infty = 1$ and $f(x_j) = \operatorname{sgn} b_j$ for all $j = 0, \dots, n$. Then, we introduce $F = f(1 - |Q|)$. We have $F(x_j) = f(x_j) = \operatorname{sgn} b_j$ for all $j = 0, \dots, n$.

Now, we prove that for any generalized polynomial P, we have $\|F - P\|_\infty \geqslant 1$. Suppose that there exists P_0 satisfying $||F - P_0||_\infty < 1$, then $\operatorname{sgn} P_0(x_j) = \operatorname{sgn} F(x_j) = \operatorname{sgn} b_j$ for all $j = 0, \dots, n$. And yet, we have $\sum_{j=0}^n b_j P_0(x_j) = 0$: contradiction.

Finally, we notice that for all $\lambda \in [0,1]$, the generalized polynomial λQ is a best approximation to F since, for all $x \in [a,b]$,

$$|F(x) - \lambda Q(x)| \leqslant |f(x)||1 - |Q(x)|| + \lambda|Q(x)| \leqslant 1 - |Q(x)| + \lambda|Q(x)| \leqslant 1.$$

□

Remez [Remes, 1934] published in 1934 Algorithm 1 which allows one to approximate, as close as desired, the minimax polynomial. This algorithm is used for the design of mathematical functions but it is its variant, due to Parks and McClellan [Parks and McClellan, 1972], in the framework of the design of filters for signal processing which has been extremely successful.

Algorithm 1 Remez second algorithm

Input: An interval $[a,b]$, a function $f \in \mathcal{C}([a,b])$, a natural integer n, a Chebyshev system $\{\varphi_k\}_{0\leqslant k\leqslant n}$, a tolerance Δ.

Output: An approximation of the degree n-minimax polynomial of f on the system $\{\varphi_k\}_{0\leqslant k\leqslant n}$.

Choose $n+2$ points $x_0 < x_1 < \dots < x_{n+1}$ in $[a,b]$, $\delta \leftarrow 1, \varepsilon \leftarrow 0$.

while $\delta \geqslant \Delta|\varepsilon|$ **do**

  Determine the solutions $a_0, \dots, a_n$ and ε of the linear system

$$\sum_{k=0}^{n} a_k \varphi_k(x_j) - f(x_j) = (-1)^j \varepsilon, \quad j = 0, \dots, n+1.$$

  Choose $x_{\text{new}} \in [a,b]$ such that

$$\|p - f\|_\infty = |p(x_{\text{new}}) - f(x_{\text{new}})|, \text{ with } p = \sum_{k=0}^{n} a_k \varphi_k.$$

  Replace one of the x_i with x_{new}, in such a way that the sign of $p - f$ alternates at the points of the resulting discretization $x_{0,\text{new}}, \dots, x_{n+1,\text{new}}$.

  $\delta \leftarrow |p(x_{\text{new}}) - f(x_{\text{new}})| - |\varepsilon|$.

end while

Return p.

We will not give more details concerning this algorithm. See [Cheney, 1998, Powell, 1981, Filip, 2016a, Filip, 2016b]. Regarding its speed of convergence, one can find in [Cheney, 1998] the following statement.

Theorem 1.18. *Let p_k denote the value of p after $k(n+2)$ loop turns, and let p^* be such that $E_n(f) = \|f - p^*\|_\infty$. There exists $\theta \in (0,1)$ such that $\|p_k - p^*\|_\infty = O(\theta^k)$.*

Under mild regularity assumptions, the bound $\mathrm{O}(\theta^k)$ can in fact be improved to $\mathrm{O}(\theta^{2^k})$ [Veidinger, 1960].

1.3 Polynomial interpolation

Now we restrict our study to polynomials in $\mathbb{R}_n[x]$.

At this stage, it seems natural to focus on techniques for computing polynomials that interpolate functions at a given finite family of points:

- sometimes a finite number of values is the only information we have on the function,
- Step 2.a of Remez' algorithm requires an efficient interpolation process,
- Theorem 1.11 shows that, for all n, there exists $a \leqslant z_0 < z_1 < \cdots < z_n \leqslant b$ such that $f(z_i) = p^*(z_i)$ for $i = 0, \ldots, n$, where p^* is the minimax approximation of f: the polynomial p^* is an interpolation polynomial of f.

Let A be a commutative ring (with unity). Given pairwise distinct $x_0, \ldots, x_n \in A$ and corresponding $y_0, \ldots, y_n \in A$, the interpolation problem is to find $p \in A_n[x]$ such that $p(x_i) = y_i$ for all i. Write $p = \sum_k a_k x^k$. The problem can be restated as

$$V \cdot \boldsymbol{a} = \boldsymbol{y} \tag{1.3}$$

where V is a Vandermonde matrix. If $\det V$ is invertible, there is a unique solution.

From now on we assume $A = \mathbb{R}$. The expression (1.1) of the Vandermonde determinant shows that as soon as the x_i are pairwise distinct, there is a unique solution. We now discuss several ways to compute the interpolation polynomial, and in each case, we mention the respective cost for computing and evaluating the polynomial.

Linear algebra. We could invert the system (1.3) using standard linear algebra algorithms. This takes $O(n^3)$ operations using Gaussian elimination. In theory, the best known complexity bound is currently $O(n^\theta)$ where $\theta \approx 2.3728639$ [Le Gall, 2014]. In practice, Strassen's algorithm yields a cost of $O(n^{\log_2 7})$. There are issues with this approach, though:

- the problem is ill-conditioned: a small perturbation on the y_i leads to a significant perturbation of the solution,
- we can do better from the complexity point of view: $O(n^2)$ or even $O(n \log^{O(1)} n)$ in general, $O(n \log n)$ if the x_i are so-called *Chebyshev nodes*.

Regarding the evaluation cost of p, Horner's method, which relies on the writing

$$p(x) = (\cdots(((a_n x + a_{n-1})x + a_{n-2})x + a_{n-3})\cdots)x + a_0,$$

yields a $O(n)$ complexity.

The divided-difference method. Newton's *divided-difference method* allows us to compute interpolation polynomials incrementally. The idea is as follows. Let $p_k \in \mathbb{R}_k[x]$ be such that $p_k(x_i) = y_i$ for $0 \leqslant i \leqslant k < n$, and write

$$p_{k+1}(x) = p_k(x) + a_{k+1}(x - x_0)\cdots(x - x_k).$$

Then we have

$$\begin{aligned} p_{k+1}(x_j) &= y_j, \qquad 0 \leqslant j \leqslant k, \\ p_{k+1}(x_{k+1}) &= p_k(x_{k+1}) + a_{k+1}(x_{k+1} - x_0)\cdots(x_{k+1} - x_k). \end{aligned}$$

Given $y_0, \ldots, y_k$, we denote by $[y_0, \ldots, y_k]$ the corresponding a_k. Then, we can compute a_k using the relation

$$[y_0, \ldots, y_{k+1}] = \frac{[y_1, \ldots, y_{k+1}] - [y_0, \ldots, y_k]}{x_{k+1} - x_0}.$$

Proof. Let $q_k \in \mathbb{R}_k[x]$ such that $q_k(x_i) = y_i$ for $1 \leqslant i \leqslant k+1$. Since $[y_0, \ldots, y_{k+1}]$ is the leading coefficient of p_{k+1}, we have

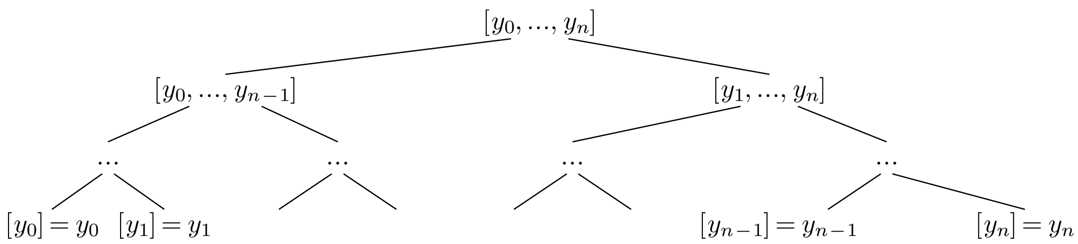
$$p_{k+1}(x) = q_k(x) + [y_0, \ldots, y_{k+1}](x - x_1) \cdots (x - x_{k+1}).$$

By definition, we also have $p_{k+1}(x) = p_k(x) + [y_0, \ldots, y_{k+1}](x - x_0) \cdots (x - x_k)$. Hence, we obtain two expressions for the coefficient of x^k of p_{k+1}:

$$[y_0, \ldots, y_{k+1}] \left(-\sum_{j=0}^{k} x_j \right) + [y_0, \ldots, y_k] = [y_0, \ldots, y_{k+1}] \left(-\sum_{j=1}^{k+1} x_j \right) + [y_1, \ldots, y_{k+1}],$$

which gives the expected result. □

This leads to a tree of the following shape.



Hence, the cost for computing the coefficients in $O(n^2)$ operations.

The evaluation cost at a given point z is in $O(n)$ operations in $\mathbb{R}$: we can adapt Horner's scheme as

$$p(z) = (\cdots(((a_n(z - x_{n-1}) + a_{n-1})(z - x_{n-2}) + a_{n-2})(z - x_{n-3}) + a_{n-3})\cdots)(z - x_0) + a_0.$$

Lagrange's Formula. For all j, let

$$\ell_j(x) = \prod_{\substack{0 \leqslant k \leqslant n, \\ k \neq j}} \frac{x - x_k}{x_j - x_k}.$$

Then we have $\deg \ell_j = n$ and $\ell_j(x_i) = \delta_{i,j}$ for all $0 \leqslant i, j \leqslant n$. The polynomials ℓ_j, $0 \leqslant j \leqslant n$, form a basis of $\mathbb{R}_n[x]$, and the interpolation polynomial p can be written

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

Thus, writing the interpolation polynomial on the Lagrange basis is straightforward.

What about the cost of evaluating the resulting polynomial at a given point z? If we do it naively, computing $\ell_j(z)$ costs (say) $2n$ subtractions, $2n+1$ multiplications and 1 division. The total cost is $O(n^2)$ operations in $\mathbb{R}$.

But we can also write

$$p(x) = W(x) \sum_{i=0}^{n} \frac{y_i}{(x - x_i) W'(x_i)}, W(x) = \prod_{i=0}^{n} (x - x_i).$$

Assuming the $W'(x_i)$ are precomputed, the cost of evaluating $p(z)$ using this formula is only $O(n)$ arithmetical operations.

The notion of "barycentric Lagrange interpolation" is particularly relevant regarding these stability issues [Trefethen, 2013].

1.4 Interpolation and approximation, Chebyshev polynomials

How useful is interpolation for our initial L^∞ approximation problem? It turns out that the choice of the points is critical. The more points, the better? Actually, with equidistant points, the error can grow with the number of points (Runge's phenomenon).

Exercise 1.4.1. Using your computer algebra system of choice, interpolate the function

$$f : x \mapsto \frac{1}{1+5x^2}$$

at the points $-1 + \frac{2k}{n}$, $0 \leqslant k \leqslant n$, for $n = 10, 15, \ldots, 30$. Compare with f on $[-1, 1]$.

In short, we should never use equidistant points when approximating a function by interpolation. Are there better choices?

Theorem 1.19. *[Faber] For each n, let a system of $n+1$ distinct nodes $\xi_0^{(n)}, \ldots, \xi_n^{(n)} \in [a, b]$. Then there exists $f \in \mathcal{C}([a,b])$ such that the sequence of errors $(\|f - p_n\|_\infty)_{n\in\mathbb{N}}$ is unbounded, where $p_n \in \mathbb{R}_n[x]$ denote the polynomial which interpolates f at the $\xi_0^{(n)}, \ldots, \xi_n^{(n)}$.*

Proof. See Remark 2.17. □

We discuss better choices below. We start with the following analogue of the Taylor-Lagrange formula.

Theorem 1.20. *Let $a \leqslant x_0 < \cdots < x_n \leqslant b$, and let $f \in \mathcal{C}^{n+1}([a, b])$. Let $p \in \mathbb{R}_n[x]$ be such that $f(x_i) = p(x_i)$ for all i. Then, for all $x \in [a, b]$, there exists $\xi_x \in (a, b)$ such that*

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} W(x), \qquad W(x) = \prod_{i=0}^{n} (x - x_i).$$

Proof. This is obvious when $x \in \{x_i, i = 0, \ldots, n\}$. Assuming $x \notin \{x_i, i = 0, \ldots, n\}$, let $\varphi = f - p - \lambda W$ where λ is chosen so that $\varphi(x) = 0$. Then, we have $\varphi(x_i) = 0$ for all i, and by Rolle's theorem there exist n+1 points $y_1 < \cdots < y_{n+1}$ with $\varphi'(y_i) = 0$. Iterating the argument, there exists $\xi \in (a, b)$ such that $\varphi^{(n+1)}(\xi) = 0$. Now recall that the polynomial W is monic and has degree $n + 1$, the polynomial p has degree at most n: this implies $W^{(n+1)}(\xi) = (n + 1)!$ and $p^{(n+1)}(\xi) = 0$, which yields the result. □

This result encourages us to search for families of x_i which make $\|W\|_\infty$ as small as possible. It's time for us to introduce Chebyshev polynomials.

Assume $[a, b] = [-1, 1]$. The n-th Chebyshev polynomial of the first kind is defined by $T_n \in \mathbb{R}_n[x]$ and

$$T_n(\cos t) = \cos(nt), \ \forall t \in [0, \pi].$$

The T_n can also be defined by

$$T_0(x) = 1, T_1(x) = x, T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x), \forall n \in \mathbb{N}. \tag{1.4}$$

Among their numerous nice features, there is the following result which suggests to consider a certain family of interpolation nodes.

Proposition 1.21. *Let $n \in \mathbb{N}, n \neq 0$. The minimum value of the set*

$$\left\{ \max_{x\in[-1,1]} |p(x)| : p \in \mathbb{R}_n[x], \mathrm{lc}(p) = 1 \right\}$$

is uniquely attained for $T_n/2^{n-1}$ and is therefore equal to 2^{-n+1}.

Proof. We have

$$A_n = \left\{ \max_{x\in[-1,1]} |p(x)| : p \in \mathbb{R}_n[x], \mathrm{lc}(p) = 1 \right\} = \left\{ \max_{x\in[-1,1]} |x^n - q(x)| : q \in \mathbb{R}_{n-1}[x] \right\}.$$

Hence, minimizing A_n is equivalent to determining $E_{n-1}(x \in [-1,1] \mapsto x^n)$. We deduce from (1.4) that the leading coefficient of T_n is 2^{n-1}. Moreover, $\|T_n\|_\infty \leqslant 1$ and $T_n\left(\cos\left(\frac{(n-k)\pi}{n}\right)\right) = (-1)^{n-k}$ for $k = 0, \ldots, n$. We can now apply Theorem 1.11 and conclude that $T_n/2^{n-1} - x^n$ is the minimax approximation of degree at most $n-1$ to x^n and $\min_{n\in\mathbb{N}, n\neq 0} A_n = E_{n-1}(x^n) = \|T_n/2^{n-1}\|_\infty = 2^{1-n}$. □

Forcing $W(x) = 2^{-n}T_{n+1}(x)$ leads to the interpolation points

$$\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n,$$

called the Chebyshev nodes of the first kind.

Another important family is that of Chebyshev polynomials of the second kind $U_n(x)$, defined by

$$U_n(\cos x) = \frac{\sin((n+1)x)}{\sin(x)}.$$

They can also be defined by

$$U_0(x) = 1, U_1(x) = 2x, U_{n+2}(x) = 2xU_{n+1}(x) - U_n(x), \forall n \in \mathbb{N}.$$

For all $n \geqslant 0$, we have $\frac{\mathrm{d}}{\mathrm{d}x}T_n = nU_{n-1}$. So the extrema of T_{n+1} are -1, 1 and the zeros of U_n, that is,

$$\nu_k = \cos\left(\frac{k\pi}{n}\right), k = 0, \ldots, n,$$

called the Chebyshev nodes of the second kind. With $W(x) = 2^{-n+1}(1-x^2)U_{n-1}(x)$, we have $\|W\|_\infty \leqslant 2^{-n+1}$.

It is obvious that $\deg T_n = \deg U_n = n$ for all $n \in \mathbb{N}$. Therefore, in particular, the family $(T_k)_{0\leqslant k\leqslant n}$ is a basis of $\mathbb{R}_n[x]$. In the sequel of the chapter, we give results that allow for the (fast) computation of the coefficients of interpolation polynomials, at the Chebyshev nodes, expressed in the basis $(T_k)_{0\leqslant k\leqslant n}$.

Let $\sum''$ denote a sum such that the first and the last terms of the sum have to be halved.

Proposition 1.22. *(Discrete orthogonality.) Let $j, \ell \in \{0, \ldots, n\}$.*

i. We have

$$\sum_{k=0}^{n} T_j(\mu_k)T_\ell(\mu_k) = \begin{cases} 0, & j \neq \ell, \\ n+1, & j = \ell = 0, \\ \frac{n+1}{2}, & j = \ell \neq 0. \end{cases}$$

ii. We have

$$\sum_{k=0}^{n}{}'' T_j(\nu_k)T_\ell(\nu_k) = \begin{cases} 0, & j \neq \ell, \\ n, & j = \ell \in \{0, n\}, \\ \frac{n}{2}, & j = \ell \notin \{0, n\}. \end{cases}$$

Exercise 1.4.2. Prove the previous proposition.

The discrete orthogonality property implies the following ($\sum'$ denotes that the first term of the sum has to be halved).

Proposition 1.23. *i. If $p_{1,n} = \sum'_{0\leqslant j\leqslant n} c_{1,j}T_j \in \mathbb{R}_n[x]$ interpolates f on the set $\{\mu_k : 0 \leqslant k \leqslant n\}$, then*

$$c_{1,j} = \frac{2}{n+1}\sum_{k=0}^{n} f(\mu_k)T_j(\mu_k) \text{ for } j = 0, \ldots, n.$$

ii. Likewise, if $p_{2,n} = \sum''_{0 \leqslant j \leqslant n} c_{2,j} T_j$ *interpolates* f *at* $\{\nu_k : 0 \leqslant k \leqslant n\}$*, then*

$$c_{2,j} = \frac{2}{n} \sum_{k=0}^{n}{}'' f(\nu_k) T_j(\nu_k) \text{ for } j = 0, \ldots, n.$$

Proof. Exercise. □

1.5 Clenshaw's method for Chebyshev sums

Given coefficients $c_0, \ldots, c_N$ and a point t, we would like to compute the sum

$$\sum_{k=0}^{N} c_k T_k(t).$$

Recall that the polynomials T_k satisfy $T_{k+2}(x) = 2x T_{k+1}(x) - T_k(x)$. A first idea would be to use this relation to compute the $T_k(t)$ that appear in the sum. Unfortunately, this method is numerically unstable. This is related to the fact that the $U_k(x)$ satisfy the same recurrence but grow faster: we have

$$\|T_k\|_\infty = 1, \qquad \|U_k\|_\infty = k + 1.$$

Clenshaw's algorithm below does better.

Algorithm 2 Clenshaw's evaluation scheme

Input: Chebyshev coefficients $c_0, \ldots, c_n$, a point $t \in [-1, 1]$
Output: $\sum_{k=0}^{n} c_k T_k(t)$

```
1: b_{n+1} ← 0, b_n ← c_n
2: for k = n − 1, n − 2, ..., 1 do
3:     b_k ← 2t b_{k+1} − b_{k+2} + c_k
4: end for
5: return c_0 + t b_1 − b_2
```

Proof. By definition of the b_k, we have

$$\sum_{k=0}^{n} c_k T_k(t) = c_0 + (b_1 - 2t b_2 + b_3) T_1(t) + \cdots + (b_{n-1} - 2t b_n + b_{n+1}) T_{n-1}(t) + c_n T_n(t).$$

The sum simplifies to $c_0 + b_1 t + b_2(T_2(t) - 2t T_1(t))$ using the recurrence relation and the values of b_n, b_{n+1}. □

This algorithm runs in $O(n)$ arithmetic operations.

1.6 Computation of the coefficients of the interpolants at Chebyshev nodes

Now, how do we compute the $c_{1,k}$ and $c_{2,k}$? First, we introduce three functions, fundamental in the field of Signal Processing [Oppenheim and Schafer, 2010]: let $M \in \mathbb{N}, M \neq 0$,

- let $\omega = e^{-2i\pi/M}$ a M-th primitive root of unity, the Discrete Fourier Transform (DFT) is the map $(x_0, \cdots, x_{M-1}) \in \mathbb{C}^M \mapsto (X_0, \cdots, X_{M-1}) \in \mathbb{C}^M$ defined by:

$$X_j = \sum_{k=0}^{M-1} x_k \omega^{jk} = \sum_{k=0}^{M-1} x_k e^{-2\pi i jk/M} \text{ for } j = 0, \ldots, M - 1.$$

The DFT sends the coefficient vector of a polynomial $P(Y) = \sum_{k=0}^{M-1} x_n Y^n$ to its values $P(1), P(\omega)$, $\dots, P(\omega^{M-1})$.

- the type I Discrete Cosine Transform (DCT-I) is the map $(x_0, \cdots, x_{M-1}) \in \mathbb{R}^M \mapsto (X_0, \cdots, X_{M-1}) \in \mathbb{R}^M$ defined by:
$$X_j = \sum_{k=0}^{M-1}{}'' x_k \cos\left(jk\frac{\pi}{M-1}\right) \text{ for } j = 0, \dots, M-1.$$
- the type II Discrete Cosine Transform (DCT-II) is the map $(x_0, \cdots, x_{M-1}) \in \mathbb{R}^M \mapsto (X_0, \cdots, X_{M-1}) \in \mathbb{R}^M$ defined by:
$$X_j = \sum_{k=0}^{M-1} x_k \cos\left(j(k+1/2)\frac{\pi}{M}\right) \text{ for } j = 0, \dots, M-1.$$

Note that DCT-I and DCT-II can be expressed in function of the DFT. For instance, a DCT-I of length M can be computed thanks to a DFT of length $2M-2$: for $j = 0, \dots, M-1$,

$$\begin{aligned}
X_j &= \sum_{k=0}^{M-1}{}'' x_k \cos\left(jk\frac{\pi}{M-1}\right) = \sum_{k=0}^{M-1}{}'' x_k \frac{e^{jk\frac{\pi}{M-1}} + e^{-jk\frac{\pi}{M-1}}}{2} \\
&= \frac{1}{2}\left(\sum_{k=0}^{M-1}{}'' x_k e^{jk\frac{\pi}{M-1}} + \sum_{k=0}^{M-1}{}'' x_k e^{-jk\frac{\pi}{M-1}}\right) \\
&= \frac{1}{2}\left(\sum_{k'=2(M-1)}^{M-1}{}'' x_{2(M-1)-k'} e^{-j(k'-2(M-1))\frac{\pi}{M-1}} + \sum_{k=0}^{M-1}{}'' x_k e^{-jk\frac{\pi}{M-1}}\right) \\
&= \frac{1}{2}\sum_{k=0}^{2M-3} x_{\min(2(M-1)-k,k)} e^{-jk\frac{\pi}{M-1}} = \frac{1}{2}\sum_{k=0}^{2M-3} x_{\min(2(M-1)-k,k)} e^{-2jk\frac{\pi}{2(M-1)}} \\
&= \frac{1}{2}\,\mathrm{DFT}\left((x_{\min(2(M-1)-k,k)})_{k=0,\dots,2M-3}\right).
\end{aligned}$$

Likewise, a DCT-II of length M can be computed thanks to a DFT of length $4M$.

The Fast Fourier Transform was introduced in 1965 by Cooley and Tukey [Cooley and Tukey, 1965, Duhamel and Vetterli, 1990, Loan, 1992], but can be traced back to Gauss [Heidemann et al., 1984]. We now briefly recall its operation [von zur Gathen and Gerhard, 2013]. Assume that $M = 2m$ is even, then, $\omega^m = -1$. Rewrite $P(Y) = \sum_{k=0}^{M-1} x_n Y^n$ as

$$P(Y) = Q_0(Y)(Y^m - 1) + R_0(Y) = Q_1(Y)(Y^m + 1) + R_1(Y)$$

with $\deg R_0, \deg R_1 < m$. More precisely,

$$R_0(Y) = \sum_{j=0}^{m-1} (x_j + x_{j+m})Y^j \text{ and } R_1(Y) = \sum_{j=0}^{m-1} (x_j - x_{j+m})Y^j.$$

Then $P(\omega^\ell) = R_0(\omega^\ell)$ if ℓ is even and $P(\omega^\ell) = R_1(\omega^\ell)$ if ℓ is odd. Therefore, if $R_1^*(Y)$ denotes the polynomial $R_1(\omega Y) = \sum_{j=0}^{m-1}(x_j - x_{j+m})\omega^j Y^j$, evaluating P at $1, \omega, \dots, \omega^{M-1}$ reduces to evaluating R_0 and R_1^* at $1, \omega^2, (\omega^2)^2, \dots, (\omega^2)^{m-1}$. If we apply this recursively, it leads to a number of operations in $O(M \log M)$, which yields a similar estimate for the computations of DCT-I and DCT-II.

Now, let's rewrite Proposition 1.23 the following way:

i. If $p_{1,n} = \sum'_{0 \leqslant j \leqslant n} c_{1,j} T_j \in \mathbb{R}_n[x]$ interpolates f on the set $\{\mu_k : 0 \leqslant k \leqslant n\}$, then, for $j = 0, \ldots, n$,

$$\begin{aligned} c_{1,j} &= \frac{2}{n+1} \sum_{k=0}^{n} f(\mu_k) T_j \left(\cos \left(\frac{(k+1/2)\pi}{n+1} \right) \right) \\ &= \frac{2}{n+1} \sum_{k=0}^{n} f(\mu_k) \cos \left(\frac{j(k+1/2)\pi}{n+1} \right) \\ &= \frac{2}{n+1} \, \text{DCT-II}(f(\mu_k)_{k=0,\ldots,n}). \end{aligned}$$

ii. Likewise, if $p_{2,n} = \sum''_{0 \leqslant j \leqslant n} c_{2,j} T_j$ interpolates f at $\{\nu_k : 0 \leqslant k \leqslant n\}$, then, for $j = 0, \ldots, n$,

$$\begin{aligned} c_{2,j} &= \frac{2}{n} \sum_{k=0}^{n}{}'' f(\nu_k) T_j \left(\cos \left(\frac{k\pi}{n} \right) \right) \\ &= \frac{2}{n} \sum_{k=0}^{n}{}'' f(\nu_k) \cos \left(\frac{jk\pi}{n} \right) \\ &= \frac{2}{n} \, \text{DCT-I}(f(\nu_k)_{k=0,\ldots,n}). \end{aligned}$$

Thus, we conclude that, if we already have the $f(\mu_k)$s, resp. the $f(\nu_k)$s, we can compute these coefficients in $O(n \log n)$ operations

Chapter 2

Orthogonal polynomials - Chebyshev series

2.1 Orthogonal polynomials

Let $(a,b) \subset \mathbb{R}$ be an open interval (note that, in this section, it does not need to be bounded), and let w be a weight function, that is to say $w : (a,b) \to (0,\infty)$ is a continuous function (this last hypothesis is not strictly necessary, we use it for ease of presentation). We assume

$$\forall n \in \mathbb{N}, \quad \int_a^b |x|^n w(x)\mathrm{d}x < \infty.$$

This is the case, for instance, if (a,b) is bounded and

$$\int_a^b w(x)\mathrm{d}x < \infty.$$

Let

$$\mathcal{E}(w) = \left\{ f \in \mathcal{C}((a,b)) : \|f\|_2 := \left(\int_a^b f(x)^2 w(x)\mathrm{d}x \right)^{1/2} < \infty \right\}.$$

Observe that $\mathbb{R}[x] \subset \mathcal{E}(w)$. The space $\mathcal{E}(w)$ is equipped with an inner product

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

and $\| \cdot \|_2$ is the norm associated to this inner product.

Definition 2.1. *A family of orthogonal polynomials associated with w is a sequence $(p_n) \in \mathbb{R}[x]^{\mathbb{N}}$ where $\deg p_k = k$ for all $k \in \mathbb{N}$, and*

$$i \neq j \quad \Rightarrow \quad \langle p_i, p_j \rangle = 0.$$

Theorem 2.2. *For any weight w, there exists a family of orthogonal polynomials associated with w. If additionally we request that the p_k are all monic, this family is unique.*

Proof. We use Gram-Schmidt orthogonalization. Starting with $p_0 = 1$, we iteratively construct polynomials p_k obeying the three conditions:

- p_k is monic;
- $\mathrm{Span}_{\mathbb{R}}\{p_0, \ldots, p_k\} = \mathbb{R}_k[x]$;

- $p_k \in \mathrm{Span}_{\mathbb{R}}\{p_0, \dots, p_{k-1}\}^{\perp}$.

In view of the first two conditions, the polynomial p_k is necessarily of the form

$$p_k = x^k + \sum_{j=0}^{k-1} \lambda_{k,j} p_j, \qquad \lambda_{k,j} \in \mathbb{R}.$$

The third condition above is equivalent to the system

$$0 = \langle p_k, p_j \rangle = \langle x^k, p_j \rangle + \lambda_{k,j} \|p_j\|_2^2, \qquad j = 0, \dots, k-1.$$

The unique solution to this system is

$$\lambda_{k,j} = -\frac{\langle x^k, p_j \rangle}{\|p_j\|_2^2}.$$

Thus uniqueness is established and the polynomial p_k thus constructed has the required properties. □

The following statement gives us a way to recursively compute a sequence of orthogonal polynomials. Note also that if you adapt Clenshaw's method (cf. Section 1.5) to this recurrence, it also yields an evaluation scheme in linear time for polynomials expressed in the corresponding basis of orthogonal polynomials.

Theorem 2.3. *The polynomials $(p_n)_{n \in \mathbb{N}}$ satisfy the recurrence relation*

$$p_n(x) = (x - \alpha_n) p_{n-1}(x) - \beta_n p_{n-2}(x) \qquad (n \geqslant 2)$$

with

$$\alpha_n = \frac{\langle x p_{n-1}, p_{n-1} \rangle}{\|p_{n-1}\|_2^2}, \qquad \beta_n = \frac{\|p_{n-1}\|_2^2}{\|p_{n-2}\|_2^2}.$$

Proof. Let $n \geqslant 2$. The polynomial $x p_{n-1}$ is monic and has degree n, hence

$$x p_{n-1} = p_n + \sum_{k=0}^{n-1} a_k p_k.$$

The orthogonality of the p_ns gives $a_k = \frac{\langle x p_{n-1}, p_k \rangle}{\|p_k\|_2^2}$ for $k = 0, \dots, n-1$. If we notice that $\langle x p_{n-1}, p_k \rangle = \langle p_{n-1}, x p_k \rangle$ for all $k = 0, \dots, n-1$, we obtain $a_k = 0$ if $k \leqslant n-3$ since $x p_k \in \mathbb{R}_{n-2}[x]$ and $p_{n-1} \in (\mathbb{R}_{n-2}[x])^{\perp}$. Hence, there are at most two nonzero coefficients:

$$\begin{aligned} a_{n-1} &= \frac{\langle p_{n-1}, x p_{n-1} \rangle}{\|p_{n-1}\|_2^2} = \alpha_n, \\ a_{n-2} &= \frac{\langle p_{n-1}, x p_{n-2} \rangle}{\|p_{n-2}\|_2^2} = \frac{\langle p_{n-1}, p_{n-1} + q \rangle}{\|p_{n-2}\|_2^2} \text{ with } q \in \mathbb{R}_{n-2}[x] \\ &= \frac{\langle p_{n-1}, p_{n-1} \rangle}{\|p_{n-2}\|_2^2} = \beta_n. \end{aligned}$$

□

Example 2.4.

$(-1, 1)$	$w(x) = (1 - x^2)^{-1/2}$	Chebyshev polynomials of the first kind (up to normalization)
$(-1, 1)$	$w(x) = 1$	Legendre polynomials
$(0, +\infty)$	$w(x) = e^{-x}$	Laguerre polynomials
$(-\infty, \infty)$	$w(x) = e^{-x^2}$	Hermite polynomials

Exercise 2.1.1. Prove that the first statement of Example 2.4 is correct.

Theorem 2.5. *For any weight w and for all n, the polynomial p_n has n distinct zeros in (a, b).*

Proof. Fix n. Let $x_1, \dots, x_k$ be the distinct zeros of p_n in (a, b), with respective multiplicities $m_1, \dots, m_k$. We introduce the polynomial

$$q(x) = \prod_{j=1}^{k} (x - x_j)^{m_j \bmod 2}.$$

If $k < n$, we have $\deg q \leqslant k < n$, and hence

$$\langle q, p_n \rangle = \int_a^b p_n(x) q(x) w(x) \mathrm{d}x = 0,$$

but the integrand is strictly positive over $(a, b) \backslash \{x_1, \dots, x_k\}$: contradiction. □

Theorem 2.6. *Let $f \in \mathcal{E}(w)$, $n \in \mathbb{N}$. There exists a unique best $L_2(w)$ polynomial approximation to f in $\mathbb{R}_n[x]$, denoted $p_{2,n}$:*

$$\|f - p_{2,n}\|_2 = \min_{p \in \mathbb{R}_n[x]} \|f - p\|_2.$$

It is characterized by

$$\forall p \in \mathbb{R}_n[x], \quad \langle f - p_{2,n}, p \rangle = 0.$$

Exercise 2.1.2. Prove this theorem.

Remark 2.7. We have $p_{2,n} = \sum_{k=0}^{n} \frac{\langle p_k, f \rangle}{\|p_k\|_2^2} p_k$.

Theorem 2.8. *If (a, b) is bounded, then for all $f \in \mathcal{E}(w)$, we have*

$$p_{2,n} \xrightarrow{\|\cdot\|_2} f$$

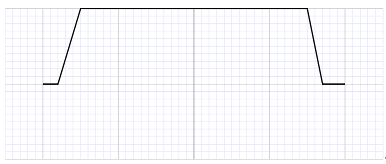
as $n \to \infty$.

Proof. First assume that $f \in \mathcal{C}([a, b])$. Let p_n^* be the minimax degree-n approximation to f: then

$$\|f - p_{2,n}\|_2 \leqslant \|f - p_n^*\|_2 = \left(\int_a^b (f - p_n^*)^2 w(x) \mathrm{d}x \right)^{1/2} \leqslant E_n(f) \left(\int_a^b w(x) \mathrm{d}x \right)^{1/2}$$

but we already know that $E_n(f) \to 0$ as $n \to \infty$.

For general f, for all $\alpha > 0$, let

$\varphi_\alpha : [a, b] \to [0, 1] =$  ,

defined more precisely by

$$\varphi_a(x) = \begin{cases} 0 \text{ if } x \in [a, a + \alpha/2] \cup [b - \alpha/2, b], \\ \frac{2}{\alpha} \left(x - a - \frac{\alpha}{2} \right) \text{ if } x \in [a + \alpha/2, a + \alpha], \\ \frac{2}{\alpha} \left(b - \frac{\alpha}{2} - x \right) \text{ if } x \in [b - \alpha, b - \alpha/2], \\ 1 \text{ if } x \in [a + \alpha, b - \alpha]. \end{cases}$$

We have $f\varphi_\alpha \in \mathcal{C}([a, b])$ if we assume $f\varphi_\alpha(a) = f\varphi_\alpha(b) = 0$. For almost all $x \in [a, b]$, we have

$$|f(x) - (f\varphi_\alpha)(x)| \leqslant |f(x)| \mathbf{1}_{[a,a+\alpha] \cup [b-\alpha,b]}(x),$$

where $\mathbf{1}_{[a,a+\alpha] \cup [b-\alpha,b]}$ denotes the indicator function of the set $[a, a + \alpha] \cup [b - \alpha, b]$. Hence, for almost all $x \in [a, b]$

- $\lim_{\alpha\to 0} |f(x) - (f\varphi_\alpha)(x)| = 0$,
- $|f(x) - (f\varphi_\alpha)(x)| \leqslant |f(x)|$, with $f \in L^2([a,b],w)$.

It follows from Lebesgue's dominated convergence theorem that

$$\int_a^b |f(x) - (f\varphi_\alpha)(x)|^2 \mathrm{d}x \xrightarrow[\alpha\to 0]{} 0.$$

Denoting by $p_{2,n}^{(\alpha)}$ the best $L_2(w)$ degree-n approximation to $f\varphi_\alpha$, we have

$$\|f - p_{2,n}\|_2 \leqslant \|f - p_{2,n}^{(\alpha)}\|_2 \leqslant \|f - f\varphi_\alpha\|_2 + \|f\varphi_\alpha - p_{2,n}^{(\alpha)}\|_2$$

for all n and α. Let $\varepsilon > 0$, there exists $\alpha > 0$ such that $\|f - f\varphi_\alpha\|_2 < \varepsilon$. For this α, there exists $n_0 \in \mathbb{N}$ such that $\|f\varphi_\alpha - p_{2,n}^{(\alpha)}\|_2 < \varepsilon$ for all $n \in \mathbb{N}, n \geqslant n_0$. □

Remark 2.9. The previous statement can be wrong if one does not assume that (a,b) is bounded. Can you give a counter-example?

Note that, from Remark 2.7, the computation of the coefficients of the best approximations in the basis of orthogonal polynomials seems to require the evaluation of several integrals. Hence, this kind of polynomials approximation is often significantly more expensive than the approach via interpolation polynomials.

2.2 A little bit of quadrature: Gauss methods

Let w be a weight function over (a,b), and let $f \in \mathcal{C}((a,b))$. We briefly study methods which approximate the integral

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

with a sum of the form

$$\sum_{k=0}^{n} w_k f(x_k), \qquad w_k \in \mathbb{R}, \quad x_k \in [a,b] \text{ pairwise distinct.} \tag{2.1}$$

First of all, if $\ell_k(x) = \prod_{\substack{0\leqslant j\leqslant n,\\ j\neq k}} \frac{x-x_j}{x_k-x_j}$, observe that if

$$p(x) = \sum_{k=0}^{n} f(x_k)\ell_k(x) \in \mathbb{R}_n[x]$$

interpolates f at the points $x_0,\dots,x_n$, then our approximation for the integral is equal to $\int_a^b p(x)w(x)\mathrm{d}x$ $=\sum_{k=0}^n w_k f(x_k)$ with

$$w_k = \int_a^b \ell_k(x)w(x)\,\mathrm{d}x \text{ for } k = 0,\dots,n.$$

Thus we obtain an approximation of the integral that is exact at least for polynomials of degree up to n. It is possible to obtain a much better result if one is allowed to choose the points $x_0,\dots,x_n$:

Theorem 2.10. *There exists a unique choice of the points x_k and the weights w_k such that, whenever $f \in \mathbb{R}_{2n+1}[x]$, the formula (2.1) is exact in the sense that*

$$\int_a^b f(x)w(x)\,\mathrm{d}x = \sum_{k=0}^{n} w_k f(x_k).$$

These points x_k belong to (a,b) and are the roots of the $(n+1)$-th orthogonal polynomial associated to w.

Proof. We start with the uniqueness. Assume that the x_js, w_js are such that the method is exact for any $f \in \mathbb{R}_m[x]$, $m \leqslant 2n+1$. Set

$$\pi_{n+1}(x) = \prod_{j=0}^{n}(x - x_j).$$

For all $p \in \mathbb{R}_n[x]$, we have $\deg(p\pi_{n+1}) \leqslant 2n+1$. Hence

$$\langle p, \pi_{n+1} \rangle = \int_a^b p(x)\pi_{n+1}(x)w(x)\mathrm{d}x = \sum_{k=0}^{n} p(x_k)\pi_{n+1}(x_k)w_k = 0.$$

The polynomial π_{n+1} is monic and belongs to $(\mathbb{R}_n[x])^\perp$: it is the $(n+1)$-th orthogonal polynomial associated to w. The x_k are its roots and, as noted above, $w_k = \sum_{k=0}^{n} w_k \ell_k(x_k) = \int_a^b \ell_k(x)w(x)\mathrm{d}x$.

As for the existence, let $x_0, \ldots, x_n$ be the distinct roots in (a,b) of the $(n+1)$-th orthogonal polynomial (cf. Proposition 2.5), and let $w_k = \int_a^b \ell_k(x)w(x)\mathrm{d}x$ where ℓ_k is the corresponding k-th Lagrange polynomial. Clearly the method is exact if $f \in \mathbb{R}_n[x]$. If now $f \in \mathbb{R}_{2n+1}[x]$, write

$$f = q\pi_{n+1} + r, \qquad \deg r \leqslant n.$$

As $\pi_{n+1} \in \mathbb{R}_n[x]^\perp$ et $\deg q \leqslant n$, we have $\int_a^b q(x)\pi_{n+1}(x)w(x)\mathrm{d}x = 0$. It follows that

$$\int_a^b f(x)w(x)\mathrm{d}x = \int_a^b r(x)w(x)\mathrm{d}x = \sum_{k=0}^{n} w_k r(x_k) = \sum_{k=0}^{n} w_k f(x_k).$$

□

See Chapter 19 of [Trefethen, 2013] for an interesting and up-to-date account on Gauss methods. Note that recent works [Hale and Townsend, 2013, Bogaert, 2014, Johansson and Mezzarobba, 2018] showed that the weights and the nodes for Gauss-Legendre or Gauss-Chebyshev quadrature, for instance, can be computed in $O(n)$ operations.

Remark 2.11. When $w = 1$, an alternative to Gauss quadrature with Legendre points is the so-called Clenshaw-Curtis quadrature, which uses Chebyshev points as interpolation nodes. The Chebyshev polynomials of the first kind satisfy

$$\int_{-1}^{1} T_k(x)\mathrm{d}x = \begin{cases} \frac{2}{1-k^2}, & k \in 2\mathbb{N}, \\ 0, & k \notin 2\mathbb{N}. \end{cases}$$

Hence, if $p = \sum_{k=0}^{n} c_k T_k$ is the interpolation polynomial of f, we deduce that the integral *with weight* $w = 1$ of f is approximated by

$$\int_{-1}^{1} p(x)\mathrm{d}x = \sum_{\substack{0 \leqslant k \leqslant n, \\ k \in 2\mathbb{N}}} \frac{2c_k}{1-k^2}.$$

Since the coefficients c_k can be computed in $O(n \log n)$ arithmetic operations using the FFT, this yields a complexity in $O(n \log n)$ for the computation of the quadrature approximant.

2.3 Lebesgue constants

For simplicity, we assume $[a,b] = [-1,1]$.

Definition 2.12. *We say that a linear mapping $L : \mathcal{C}([-1,1]) \to \mathbb{R}_n[x]$ is a projection onto $\mathbb{R}_n[x]$ if $Lp = p$ for all $p \in \mathbb{R}_n[x]$. The operator norm*

$$\Lambda = \sup_{f \in \mathcal{C}([-1,1])} \frac{\|Lf\|_\infty}{\|f\|_\infty}$$

is called the Lebesgue constant for the projection.

Proposition 2.13. *Let Λ be the Lebesgue constant for a linear projection L of $\mathcal{C}([-1,1])$ onto $\mathbb{R}_n[x]$. Let $f \in \mathcal{C}([-1,1])$ and let $p = Lf$. Let p^* denote the minimax approximation to f. Then, we have*

$$\|f - p\|_\infty \leqslant (1+\Lambda)\|f - p^*\|_\infty.$$

Proof. We have $L(f - p^*) = p - p^*$. It follows that

$$\|p - f\|_\infty - \|f - p^*\|_\infty \leqslant \|p - p^*\|_\infty = \|L(f - p^*)\|_\infty \leqslant \Lambda\|f - p^*\|_\infty.$$

□

2.3.1 Lebesgue constants for polynomial interpolation

Let $x_0, \dots, x_n$ be pairwise distinct points in $[-1,1]$. Consider the Lagrange interpolation operator

$$L_n : \mathcal{C}([-1,1]) \to \mathbb{R}_n[x], \qquad L_n f(x) = \sum_{k=0}^{n} f(x_k)\ell_k(x).$$

Clearly, L_n is a linear projection of $\mathcal{C}([-1,1])$ onto $\mathbb{R}_n[x]$. On the one hand, we have

$$|L_n f(x)| \leqslant \|f\|_\infty \sum_{k=0}^{n} |\ell_k(x)|, \text{ for all } x \in [-1,1],$$

which implies that the corresponding Lebesgue constant $\Lambda_n = |||L_n|||$ satisfies

$$\Lambda_n \leqslant A := \max_{x\in[-1,1]} \sum_{k=0}^{n} |\ell_k(x)|.$$

On the other hand, since the function $x \in [-1,1] \mapsto \sum_{k=0}^{n} |\ell_k(x)|$ is continuous, there exists $\xi \in [-1,1]$ such that $A = \sum_{k=0}^{n} |\ell_k(\xi)|$. Let $g : [-1,1] \to [-1,1]$ be a continuous piecewise affine function such that $g(x_i) = \operatorname{sgn} \ell_i(\xi)$. Then, we have

$$L_n g(\xi) = \sum_{k=0}^{n} |\ell_k(\xi)|$$

and hence $\|L_n g\|_\infty \geqslant A\|g\|_\infty$. We've just proved the following statement.

Theorem 2.14. *The Lebesgue constant of degree-n Lagrange interpolation at $x_0, \dots, x_n$ is equal to*

$$\max_{x\in[-1,1]} \sum_{k=0}^{n} |\ell_k(x)|.$$

Theorem 2.15. *The Lebesgue constant Λ_n satisfies*

$$\frac{2}{\pi}\left(\log(n+1) + \gamma + \log\frac{4}{\pi}\right) \leqslant \Lambda_n, \qquad \textit{where} \quad \frac{2}{\pi}\left(\gamma + \log\frac{4}{\pi}\right) = 0.52125\dots. \tag{2.2}$$

Additionally,

- *for Chebyshev nodes (of the first and the second kinds), we have the bound*

$$\Lambda_n \leqslant \frac{2}{\pi}\log(n+1) + 1 \textit{ and } \Lambda_n \sim \frac{2}{\pi}\log n \textit{ as } n \to +\infty \tag{2.3}$$

- *for equispaced points,*

$$\Lambda_n > \frac{2^{n-2}}{n^2} \textit{ and } \Lambda_n \sim \frac{2^{n+1}}{en\log n} \textit{ as } n \to +\infty. \tag{2.4}$$

Proof. See Chapter 15 of [Trefethen, 2013] for a commented bibliography of the proofs of these results and [Brutman, 1997] for a detailed survey on this subject.

For a proof of (2.2), see [Brutman, 1978]. See [Ehlich and Zeller, 1966] for a proof of (2.3). A proof of the left estimate of (2.4) can be found in [Trefethen and Weideman, 1991]. Two proofs of the right estimate of (2.4) were independently published in [Turetskii, 1940] and [Schönhage, 1961]. □

Remark 2.16. We deduce from this theorem that Chebyshev interpolants (i.e. interpolation polynomials at Chebyshev nodes) are "near-best" approximations:

- $\Lambda_{15} = 2.76\ldots$: one loses at most 2 bits if one uses a Chebyshev interpolant instead of the minimax polynomial;
- $\Lambda_{30} = 3.18\ldots$: one loses at most 2 bits if one uses a Chebyshev interpolant instead of the minimax polynomial;
- $\Lambda_{100} = 3.93\ldots$: one loses at most 2 bits if one uses a Chebyshev interpolant instead of the minimax polynomial;
- $\Lambda_{100000} = 8.32\ldots$: one loses at most 4 bits if one uses a Chebyshev interpolant instead of the minimax polynomial.

Remark 2.17. The estimates (2.2) imply $\sup_{n\in\mathbb{N}} \Lambda_n = +\infty$. We then deduce from Banach-Steinhaus theorem [Brezis, 2010] a proof of Theorem 1.19 (Faber's Theorem).

2.3.2 Lebesgue constants for L_2 best approximation

Definition 2.18. *When the space under consideration is $\mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right)$, the best L_2 polynomial approximation to f in $\mathbb{R}_n[x]$ is called the* truncated Chebyshev expansion *of f of order n and is denoted f_n. Its coefficients a_k are called the* Chebyshev coefficients *of f. They are given by*

$$a_k = \begin{cases} \frac{\langle f,T_k\rangle}{\langle T_k,T_k\rangle} = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x, & k \neq 0, \\ \frac{\langle f,T_0\rangle}{\langle T_0,T_0\rangle} = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x, & k = 0, \end{cases}$$

and the formal series $\sum_{k=0}^{\infty} a_k T_k(x)$ is called the Chebyshev expansion *of f.*

Remark 2.19. The Chebyshev expansion of f is the Fourier expansion of $f(\cos t)$, so that many results on the convergence of Chebyshev expansions can be deduced from corresponding results in the well-developed theory of Fourier series.

Theorem 2.20. *The Lebesgue constant for the map $f \in \mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right) \mapsto p_{2,n}$ is*

$$\Lambda_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\frac{\sin((n+1/2)t)}{\sin(t/2)}\right| \mathrm{d}t. \tag{2.5}$$

Its behaviour obeys

$$\frac{4}{\pi^2}\log(n+1) < \Lambda_n \begin{cases} = 1 \text{ if } n = 0, \\ < 2 \text{ if } n = 1, \\ < \frac{4}{\pi^2}\log(n-1) + 3 \text{ otherwise}. \end{cases} \tag{2.6}$$

Proof. For $f \in \mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right)$, $n \in \mathbb{N}$, $x \in [-1,1]$, we have

$$S_n f(x) := \sum_{k=0}^{n} a_k T_k(x) = \frac{2}{\pi}\int_{-1}^{1} \frac{f(y)}{\sqrt{1-y^2}} \left(\frac{T_0(x)T_0(y)}{2} + \sum_{k=1}^{n} T_k(x)T_k(y)\right) \mathrm{d}y.$$

If we put $x = \cos(\theta)$ and $y = \cos(u)$, it comes

$$\begin{aligned}
S_n f(x) &= \frac{2}{\pi}\int_0^\pi f(\cos u)\left(\frac{1}{2} + \sum_{k=1}^n \cos(k\theta)\cos(ku)\right)\mathrm{d}u \\
&= \frac{1}{\pi}\int_0^\pi f(\cos u)\left(1 + \sum_{k=1}^n (\cos(k(u+\theta)) + \cos(k(u-\theta)))\right)\mathrm{d}u \\
&= \frac{1}{\pi}\int_\theta^{\pi+\theta} f(\cos(v-\theta))\left(\frac{1}{2} + \sum_{k=1}^n \cos(kv)\right)\mathrm{d}v \qquad \text{(we put } v = u+\theta\text{)} \\
&\quad + \frac{1}{\pi}\int_{\theta-\pi}^{\theta} f(\cos(v-\theta))\left(\frac{1}{2} + \sum_{k=1}^n \cos(kv)\right)\mathrm{d}v \qquad \text{(we put } v = -u+\theta\text{)} \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi} f(\cos(v+\theta))\left(\frac{1}{2} + \sum_{k=1}^n \cos(kv)\right)\mathrm{d}v \qquad \text{(the integrand is even and } 2\pi\text{-periodic).}
\end{aligned}$$

Now, we use the fact that

$$\begin{aligned}
\sin((n+1/2)v) &= \sin(v/2) + \sum_{k=1}^n (\sin((k+1/2)v) - \sin((k-1/2)v)) \\
&= \sin(v/2) + \sum_{k=1}^n 2\sin(v/2)\cos(kv) = 2\sin(v/2)\left(\frac{1}{2} + \sum_{k=1}^n \cos(kv)\right).
\end{aligned}$$

This yields, for all $n \in \mathbb{N}$, $x \in [-1, 1]$,

$$S_n f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\cos(v+\theta))\,\frac{\sin((n+1/2)v)}{\sin(v/2)}\mathrm{d}v,$$

from which follows

$$\|S_n f\|_\infty \leqslant \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v\right)\|f\|_\infty.$$

hence

$$\Lambda_n \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v.$$

Let $D_n : v \in [-\pi, \pi] \mapsto \frac{\sin((n+1/2)v)}{\sin(v/2)}$ (this is the so called Dirichlet kernel). If we consider $\varphi_n : v \in [-\pi, \pi] \mapsto \mathrm{sgn}(D_n)$, we have $(\varphi_n \circ \arccos)(\cos v) = \varphi_n(v)$ for all $v \in [-\pi, \pi]$ for φ_n is even. Therefore,

$$S_n(\varphi_n \circ \arccos)(1) = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\varphi_n \circ \arccos)(\cos(v))\,\frac{\sin((n+1/2)v)}{\sin(v/2)}\mathrm{d}v = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v.$$

It follows that $\|S_n(\varphi_n \circ \arccos)\|_\infty \geqslant \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v\right)\|\varphi_n \circ \arccos\|_\infty$. For $m \in \mathbb{N}, m \geqslant 1$, let

$$\alpha_m : x \in \mathbb{R} \mapsto \begin{cases} -1 \text{ if } x \leqslant -1/m, \\ mx \text{ if } -1/m \leqslant x \leqslant 1/m, \\ 1 \text{ otherwise.} \end{cases}$$

The function $\alpha_m \circ D_n \circ \arccos$ in continuous over $[-1, 1]$, thus it belongs to $\mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right)$. For all $x \in [-1, 1]$, $\lim_{m\to+\infty} \alpha_m \circ D_n \circ \arccos(x) = \mathrm{sgn}(D_n \circ \arccos)(x)$ and $|\alpha_m \circ D_n \circ \arccos(x)| \leqslant 1 \in L_2\left(\frac{1}{\sqrt{1-x^2}}\right)$. It follows from Lebesgue's dominated convergence theorem that

$$S_n(\alpha_m \circ D_n \circ \arccos)(1) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (\alpha_m \circ D_n \circ \arccos)(\cos(v))\,\frac{\sin((n+1/2)v)}{\sin(v/2)}\mathrm{d}v$$
$$\xrightarrow[m\to+\infty]{} \frac{1}{2\pi}\int_{-\pi}^{\pi} (\varphi_n \circ \arccos)(\cos(v))\,\frac{\sin((n+1/2)v)}{\sin(v/2)}\mathrm{d}v = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v.$$

We've just proved that, for all $\varepsilon > 0$, there exists $g \in \mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right)$ such that $\|S_n g\|_\infty \geqslant |S_n g(1)| \geqslant \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v - \varepsilon\right)\|g\|_\infty$, which yields $\Lambda_n \geqslant \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\sin((n+1/2)v)}{\sin(v/2)}\right|\mathrm{d}v$.

We now prove estimates (2.6). For all $n \in \mathbb{N}$,

$$\Lambda_n = \frac{2}{\pi}\int_0^{\pi/2}\frac{|\sin((2n+1)t)|}{\sin t}\,\mathrm{d}t \qquad \text{(we have put } t = v/2\text{)},$$
$$\geqslant \frac{2}{\pi}\int_0^{\pi/2}\frac{|\sin((2n+1)t)|}{\sin t}\,\mathrm{d}t \qquad \text{since } \sin x \leqslant x \text{ for } x \in [0,\pi/2].$$

Now, we use the change of variable $v = \pi x/(2n+1)$ which gives

$$\Lambda_n \geqslant \frac{2}{\pi}\int_0^{n+1/2}\frac{|\sin(\pi v)|}{v}\,\mathrm{d}v > \frac{2}{\pi}\int_0^{n}\frac{|\sin(\pi v)|}{v}\,\mathrm{d}v.$$

Since, for $k = 0,\dots,n-1$ and $v \in [k,k+1]$, we have $\frac{|\sin(\pi v)|}{v} \geqslant \frac{|\sin(\pi v)|}{k+1}$, it comes

$$\Lambda_n > \frac{2}{\pi}\int_0^{n}\frac{|\sin(\pi v)|}{v}\,\mathrm{d}v = \frac{2}{\pi}\sum_{k=0}^{n-1}\int_k^{k+1}\frac{|\sin(\pi v)|}{v}\,\mathrm{d}v \geqslant$$
$$\frac{2}{\pi}\sum_{k=0}^{n-1}\frac{1}{k+1}\underbrace{\int_k^{k+1}|\sin(\pi v)|\,\mathrm{d}v}_{\int_0^1|\sin(\pi v)|\,\mathrm{d}v} = \frac{2}{\pi^2}[\cos(\pi v)]_0^1\sum_{k=1}^{n}\frac{1}{k} = \frac{4}{\pi^2}\sum_{k=1}^{n}\frac{1}{k}. \tag{2.7}$$

Now, for $k = 1,\dots,n$ and $v \in [k,k+1]$, $\frac{1}{k+1} \leqslant \frac{1}{v} \leqslant \frac{1}{k}$ implies $\frac{1}{k+1} \leqslant \int_k^{k+1}\frac{1}{v}\,\mathrm{d}v = \log(k+1) - \log k \leqslant \frac{1}{k}$, hence

$$\sum_{k=1}^{n}\frac{1}{k+1} \leqslant \sum_{k=1}^{n}(\log(k+1) - \log k) = \log(n+1) \leqslant \sum_{k=1}^{n}\frac{1}{k}. \tag{2.8}$$

Estimates (2.7) and (2.8) yield $\Lambda_n > \frac{4}{\pi^2}\log(n+1)$.

We now prove the second inequality of (2.6). We follow [Rivlin, 1981, Chap. 3]. The case $n = 0$ is straightforward. Then, we assume $n \geqslant 1$. First, note that

$$\Lambda_n = \frac{1}{\pi}\int_0^{\pi}\left|\frac{\sin(nv)\cos(v/2) + \cos(nv)\sin(v/2)}{\sin(v/2)}\right|\mathrm{d}v$$
$$= \frac{1}{\pi}\int_0^{\pi}\left|\frac{\sin(nv)}{\tan(v/2)} + \cos(nv)\right|\mathrm{d}v \leqslant \frac{1}{\pi}\int_0^{\pi}\frac{|\sin(nv)|}{\tan(v/2)}\mathrm{d}v + \frac{1}{\pi}\int_0^{\pi}|\cos(nv)|\mathrm{d}v. \tag{2.9}$$

Now, recall that $\tan x \geqslant x$ for all $x \in [0, \pi/2]$. Therefore,

$$\begin{aligned}
\int_0^\pi \frac{|\sin(nv)|}{\tan(v/2)}\,\mathrm{d}v &\leqslant 2\int_0^\pi \frac{|\sin(nv)|}{v}\,\mathrm{d}v \\
&= 2\int_0^{n\pi} \frac{|\sin u|}{u}\,\mathrm{d}u \qquad \text{(we have put } u = nv\text{)} \\
&= 2\sum_{k=0}^{n-1}\int_{k\pi}^{(k+1)\pi} \frac{|\sin u|}{u}\,\mathrm{d}u \\
&= 2\sum_{k=0}^{n-1}\int_0^{\pi} \frac{|\sin u|}{u+k\pi}\,\mathrm{d}u \quad \text{since } u \mapsto |\sin u| \text{ is } \pi\text{-periodic} \\
&\leqslant 2\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u + 2\left(\int_0^\pi \sin u\,\mathrm{d}u\right)\sum_{k=1}^{n-1}\frac{1}{k\pi} \\
&\leqslant 2\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u + \frac{4}{\pi}\sum_{k=1}^{n-1}\frac{1}{k} \\
&\leqslant 2\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u + \frac{4}{\pi}(1+\log(n-1)) \text{ thanks to (2.8).}
\end{aligned} \tag{2.10}$$

For all $x \geqslant 0$, $m \in \mathbb{N}$, $\left|\sin x - \sum_{k=0}^{m-1}\frac{(-1)^k}{(2k+1)!}x^{2k+1}\right| \leqslant \frac{1}{(2m+1)!}x^{2m+1}$, hence

$$\left|\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u - \sum_{k=0}^{m-1}\frac{(-1)^k}{(2k+1)!}\int_0^\pi x^{2k}\,\mathrm{d}u\right| \leqslant \frac{1}{(2m+1)!}\int_0^\pi x^{2m}\,\mathrm{d}u,$$

i.e.

$$\left|\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u - \sum_{k=0}^{m-1}\frac{(-1)^k}{(2k+1)!}\frac{\pi^{2k+1}}{2k+1}\right| \leqslant \frac{\pi^{2m+1}}{(2m+1)!(2m+1)},$$

If we set $m = 4$, we obtain

$$\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u \leqslant 1.86. \tag{2.11}$$

Finally, we have

$$\begin{aligned}
\int_0^\pi |\cos(nv)|\mathrm{d}v &= \frac{1}{n}\int_0^{n\pi} |\cos(u)|\mathrm{d}u \qquad \text{(we have put } u = nv\text{)} \\
&= \frac{1}{n}\sum_{k=0}^{n-1}\int_{k\pi}^{(k+1)\pi} |\cos(u)|\mathrm{d}u \\
&= \frac{1}{n}\sum_{k=0}^{n-1}\int_0^{\pi} |\cos(u)|\mathrm{d}u \quad \text{since } u \mapsto |\cos(u)| \text{ is } \pi\text{-periodic} \\
&= \int_0^{\pi/2}\cos(u)\mathrm{d}u - \int_{\pi/2}^{\pi}\cos(u)\mathrm{d}u = 2.
\end{aligned} \tag{2.12}$$

The estimates (2.9), (2.10), (2.11) and (2.12) yield

$$\Lambda_n \leqslant \frac{1}{\pi}\left(2\cdot 1.86 + \frac{4}{\pi} + \frac{4}{\pi}\log(n-1)\right) + \frac{2}{\pi} < 3 + \frac{4}{\pi^2}\log(n-1).$$

Note that $\Lambda_1 \leqslant \frac{2}{\pi}\int_0^\pi \frac{\sin u}{u}\,\mathrm{d}u + \frac{1}{\pi}\int_0^\pi \cos(v)\mathrm{d}v \leqslant \frac{2}{\pi}(1.86+1) < 2.$ □

Remark 2.21. We deduce from this theorem that truncated Chebyshev series are "near-best" approximations:

- $\Lambda_{15} = 4.12\ldots$: one loses at most 3 bits if one uses the truncated Chebyshev series instead of the minimax polynomial;
- $\Lambda_{30} = 4.39\ldots$: one loses at most 3 bits if one uses the truncated Chebyshev series instead of the minimax polynomial;
- $\Lambda_{100} = 4.87\ldots$: one loses at most 3 bits if one uses the truncated Chebyshev series instead of the minimax polynomial;
- $\Lambda_{100000} = 7.66\ldots$: one loses at most 3 bits if one uses the truncated Chebyshev series instead of the minimax polynomial.

Remark 2.22. The estimates (2.6) imply $\sup_{n\in\mathbb{N}} \Lambda_n = +\infty$. We then deduce from Banach-Steinhaus theorem [Brezis, 2010] that there exists $f \in \mathcal{E}\left(\frac{1}{\sqrt{1-x^2}}\right)$ such that its Chebyshev expansion that does not uniformly converge to f.

2.4 Chebyshev expansions and interpolation polynomials at Chebyshev nodes

2.4.1 Convergence results, certified estimates

Here is a summary of convergence results that we are going to rely on (see Theorems 3.1, 7.1, 7.2, 8.1, 8.2 in [Trefethen, 2013] for versions with weaker hypotheses).

Theorem 2.23. *Let f be continuous on $[-1,1]$. Denote by $(a_k)_{k\in\mathbb{N}}$ its sequence of Chebyshev coefficients, by $(f_n)_{n\in\mathbb{N}}$ its sequence of truncated Chebyshev expansions and by $(p_n)_{n\in\mathbb{N}}$ the sequence of interpolation polynomials of f at the Chebyshev nodes. Then*

1. *The coefficients a_k tend to 0 when $k \to \infty$.*
2. *If f is Lipschitz continuous on $[-1,1]$, then (f_n) converges uniformly to f and (p_n) converges uniformly to f.*
3. *If f is Lipschitz continuous on $[-1,1]$, then (f_n) converges absolutely. Consequently, it converges normally, hence uniformly, to f.*
4. *If f is $\mathcal{C}^m$ and $f^{(m)}$ is Lipschitz continuous, then $a_k = O(1/k^{m+1})$, $\|f - f_n\|_\infty = O(n^{-m})$ and $\|f - p_n\|_\infty = O(n^{-m})$.*
5. *If f is analytic inside the ellipse $\{z \in \mathbb{C} : \left|z + \sqrt{z^2-1}\right| \leqslant r\}$ with $r > 1$, then $a_k = O(r^{-k})$, $\|f - f_n\|_\infty = O(r^{-n})$ and $\|f - p_n\|_\infty = O(r^{-n})$.*
6. *Let P_n^* denote the minimax polynomial of degree at most n of f. If $f \in \mathcal{C}^{n+1}([-1,1])$, there exists $\xi_1, \xi_2, \xi_3 \in (-1,1)$ such that*

$$\|f - P_n^*\|_\infty = \frac{|f^{(n+1)}(\xi_1)|}{2^n(n+1)!}; \tag{2.13}$$

$$\|f - f_n\|_\infty = \frac{|f^{(n+1)}(\xi_2)|}{2^n(n+1)!}; \tag{2.14}$$

$$\|f - p_n\|_\infty = \frac{|f^{(n+1)}(\xi_3)|}{2^n(n+1)!}. \tag{2.15}$$

Proof. 1. This is Riemann-Lebesgue lemma [Zygmund, 2002, Chap. II].

2. This is obtained by combining Proposition 2.13, the bounds on Λ_n from Theorems 2.15 and 2.20 and Corollary 1.7 that states $E_n(f) = O(n^{-1/2})$. Thus, if $(p_n)_{n\in\mathbb{N}}$ denotes the sequence of interpolation polynomials of f at the Chebyshev nodes and $(f_n)_{n\in\mathbb{N}}$ denotes the sequence of truncated Chebyshev expansions of f, there exists K such that for large enough n,

$$\|f - p_n\|_\infty \leqslant \left(2 + \frac{2}{\pi}\log(n+1)\right)\frac{K}{\sqrt{n}} \to 0 \text{ as } n \to \infty,$$
$$\|f - f_n\|_\infty \leqslant \left(4 + \frac{4}{\pi^2}\log(n+1)\right)\frac{K}{\sqrt{n}} \to 0 \text{ as } n \to \infty.$$

3. See [Zygmund, 2002, Chap. VI] for a proof of absolute convergence. The normal convergence follows from $|a_n T_n(x)| \leqslant |a_n|$, for all $n \in \mathbb{N}$ and $x \in [-1, 1]$.

4. See Chapter 7 of [Trefethen, 2013].

5. See Chapter 8 of [Trefethen, 2013].

6. See [Bernstein, 1926, Meinardus, 1967] for a proof of (2.13) and [Elliott et al., 1987] for a proof of (2.14). The estimate (2.15) follows from Theorem 1.20 and Proposition 1.21.

□

Chapter 3

Interval Arithmetic, Interval Analysis

Interval Arithmetic is "an arithmetic for inequalities". Assume for instance that we know that $5 \leqslant a \leqslant 6$ and $10 \leqslant b \leqslant 11$: then of course $50 \leqslant ab \leqslant 66$. We will define a product of real intervals such that

$$[5, 6] \times [10, 11] = [50, 66]$$

that allows for such reasoning. Another need for interval arithmetic comes from the roundoff errors that occur when working with finite precision numbers.

Notable applications of interval arithmetic to bring rigor to numerical computations performed on a computer include T. Hales' proof of Kepler's conjecture[1] [Hales, 2005][Hales et al., 2010], and W. Tucker's solution of Smale's 14th problem[2] [Tucker, 1999][Tucker, 2002].

The interested reader will find numerous additional interesting information on the website http://www.cs.utep.edu/interval-comp/.

In this course, we are interested in the use of interval arithmetic in the evaluation of mathematical functions. Given $\varepsilon > 0$ and $f : [a, b] \to \mathbb{R}$, we would like to make sure that the evaluation $\widehat{f(x)}$ of f at any value $x \in [a, b]$ is such that

$$|\widehat{f(x)} - f(x)| \leqslant \varepsilon.$$

Note that, in practice, one commonly uses on relative error $\left|1 - \frac{\widehat{f(x)}}{f(x)}\right|$ rather than on absolute error $|\widehat{f(x)} - f(x)|$. We focus on the absolute error case for the sake of clarity. To perform the evaluation, we replace f by a polynomial p. Then we evaluate p, and $\widehat{f(x)} = \circ(p(x))$, where $\circ$ is the active rounding mode. There are two sources of error:

- *approximation error*: let η_1 be an upper bound for $\|f - p\|_\infty$,
- *rounding error*: let η_2 be an upper bound for the error $|p(x) - \circ(p(x))|$,

we have to guarantee that $\eta_1 + \eta_2 \leqslant \varepsilon$. In this chapter and in Chapter 4, we will develop tools that help to establish rigorous approximation error. Regarding rounding errors, G.Melquiond has developed formal proof tools[3] which address this issue [Melquiond, 2006] [Daumas and Melquiond, 2010] [de Dinechin et al., 2011].

3.1 Interval arithmetic

Definition 3.1. *(Real interval.) Let $\bar{x}, \underline{x} \in \mathbb{R}$, $\bar{x} \leqslant \underline{x}$. We define the interval*

$$X = [\underline{x}, \bar{x}] = \{x \in \mathbb{R} : \underline{x} \leqslant x \leqslant \bar{x}\}.$$

[1] See http://code.google.com/p/flyspeck/.

[2] See http://www2.math.uu.se/~warwick/main/thesis.html and also http://paulbourke.net/fractals/lorenz/.

[3] See http://gappa.gforge.inria.fr/.

The real numbers $\underline{x}$ and $\bar{x}$ are called the endpoints of the interval, $\underline{x}$ is its minimum, $\bar{x}$ its maximum. The set of all real intervals will be denoted $\mathbb{IR}$.

Definition 3.2. *Let $x \in \mathbb{IR}$. The width of x is denoted $w(x) = \bar{x} - \underline{x}$. We also define the center*

$$\operatorname{mid}(x) = \frac{\underline{x} + \bar{x}}{2},$$

and the radius $\operatorname{rad}(x) = \frac{1}{2}w(x)$.

Remark 3.3. It is common in the litterature to encounter the notation $(\operatorname{mid}(x), \operatorname{rad}(x)) = \{x \in \mathbb{R} : |x - \operatorname{mid}(x)| \leqslant \operatorname{rad}(x)\}$. This mid-rad representation is the basis of the so called Ball Arithmetic, cf. the excellent software Arb [4].

Definition 3.4. *A point (or degenerate, or thin) interval is one of the form $[x, x]$, also denoted $[x]$.*

3.1.1 Operations on intervals

We now define basic arithmetic operations on intervals. As you will see, monotonicity plays an essential role for obtaining sharp enclosures.

Definition 3.5. *Let $X, Y \in \mathbb{IR}$. Let $* \in \{+, -, \times, /\}$. We denote*

$$X * Y = \{x * y : x \in X, y \in Y\}$$

where we assume that $0 \notin Y$ if $* = /$.

Proposition 3.6. *We can compute the $X * Y$ above using formulae such as*

$$\begin{aligned}
[\underline{x}, \bar{x}] + [\underline{y}, \bar{y}] &= \left[\underline{x} + \underline{y}, \bar{x} + \bar{y}\right], \\
[\underline{x}, \bar{x}] - [\underline{y}, \bar{y}] &= \left[\underline{x} - \bar{y}, \bar{x} - \underline{y}\right], \\
[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] &= \left[\min\left(\underline{x} \cdot \underline{y}, \underline{x} \cdot \bar{y}, \bar{x} \cdot \underline{y}, \bar{x} \cdot \bar{y}\right), \max\left(\underline{x} \cdot \underline{y}, \underline{x} \cdot \bar{y}, \bar{x} \cdot \underline{y}, \bar{x} \cdot \bar{y}\right)\right], \\
[\underline{x}, \bar{x}] / [\underline{y}, \bar{y}] &= [\underline{x}, \bar{x}] \times \left[\frac{1}{\bar{y}}, \frac{1}{\underline{y}}\right] \quad \text{if } 0 \notin Y,
\end{aligned}$$

which depend only on the endpoints.

Proof. Exercise. □

Remark 3.7. Note that, in $\mathbb{IR}$, the operations $+$ and $\times$ are associative and commutative.

Remark 3.8. In practice, multiplication (hence division) can be made more efficient. From the formula in the previous proposition, it seems to require four real multiplications and several comparisons. And yet, if one checks the sign of the endpoints of the two intervals before starting the computation, one can reduce this amount. Note that there are nine possible cases: one of them indeed leads to four multiplications while the other eight only need two multiplications. Likewise, there are six possible cases for the division.

Remark 3.9. It can be convenient to define a result for the division even when $0 \in Y$. One can find an interesting discussion regarding this issue in Section 2.3 of W. Tucker's book [Tucker, 2011]. To do that, we work over $\overline{\mathbb{R}}$=$\mathbb{R} \cup \{+\infty, -\infty\}$, with two signed zeros $+0$ and -0 (more precisely, over $\mathbb{R} \backslash \{0\} \cup \{+\infty, -\infty, +0, -0\}$, with two signed zeros $+0$ and -0). Hence, we can take advantage of the following relations

$$1/(+\infty) = +0, \quad 1/(+0) = +\infty, \quad 1/(-\infty) = -0, \quad 1/(-0) = -\infty.$$

[4] http://arblib.org/

Assume $\underline{y} < 0 < \bar{y}$. Then we define

$$\frac{1}{[\underline{y}, 0]} = \left[-\infty, \frac{1}{\underline{y}}\right], \qquad \frac{1}{[0, \bar{y}]} = \left[\frac{1}{\bar{y}}, +\infty\right]$$

and in general

$$\frac{1}{[\underline{y}, \bar{y}]} = \left[-\infty, \frac{1}{\underline{y}}\right] \cup \left[\frac{1}{\bar{y}}, +\infty\right].$$

We will thus define the notion of extended interval by removing the condition $\underline{x} \leqslant \bar{x}$ and set

$$X = [\underline{x}, \bar{x}] = \begin{cases} \{x \in \mathbb{R} : \underline{x} \leqslant x \leqslant \bar{x}\} \text{ if } \underline{x} \leqslant \bar{x}, \\ [-\infty, \bar{x}] \cup [\underline{x}, +\infty] \text{ otherwise.} \end{cases}$$

We introduce the notation $\overline{\mathbb{IR}} = \left\{[\underline{x}, \bar{x}] : \underline{x}, \bar{x} \in \overline{\mathbb{R}}\right\}$. We then define division over $\overline{\mathbb{IR}}$ as follows

$$X/Y = \begin{cases} X \times \left[1/\bar{y}, 1/\underline{y}\right] \quad \text{if } 0 \notin Y, \\ [-\infty, +\infty] \text{ if } 0 \in X \text{ and } 0 \in Y, \\ \left[\bar{x}/\underline{y}, +\infty\right] \text{ if } \bar{x} < 0 \text{ and } \underline{y} < \bar{y} = 0, \\ \left[\bar{x}/\underline{y}, \bar{x}/\bar{y}\right] \text{ if } \bar{x} < 0 \text{ and} \underline{y} < 0 < \bar{y}, \\ [-\infty, \bar{x}/\bar{y}] \text{ if } \bar{x} < 0 \text{ and } 0 = \underline{y} < \bar{y}, \\ \left[-\infty, \underline{x}/\underline{y}\right] \text{ if } 0 < \underline{x} \text{ and } \underline{y} < \bar{y} = 0, \\ \left[\underline{x}/\bar{y}, \underline{x}/\underline{y}\right] \text{ if } 0 < \underline{x} \text{ and } \underline{y} < 0 < \bar{y}, \\ [\underline{x}/\bar{y}, +\infty] \text{ if } 0 < \underline{x} \text{ and } 0 = \underline{y} < \bar{y}, \\ \emptyset \text{ if } 0 \notin X \text{ and } Y = [0, 0]. \end{cases} \tag{3.1}$$

Proposition 3.10.

1. *Interval subtraction is not the inverse of addition.*

2. *Interval division is not the inverse of multiplication.*

3. *Interval multiplication of an interval with itself is not equivalent to "squaring the interval", i.e., in general,*

$$[\underline{x}, \bar{x}] \times [\underline{x}, \bar{x}] \neq \left[\min\left(\underline{x}^2, \bar{x}^2\right), \max\left(\underline{x}^2, \bar{x}^2\right)\right].$$

4. *Interval multiplication is sub-distributive wrt addition: for all $X, Y, Z \in \mathbb{IR}$, we have*

$$X \times (Y + Z) \subset X \times Y + X \times Z.$$

5. *For all $X \in \mathbb{IR}$, we have $X + [0] = X$ and $[0] \times X = [0]$.*

Proof. Exercise. □

A straightforward yet quite useful statement is the following.

Lemma 3.11. *(Inclusion isotonicity) If $X \subset X', Y \subset Y', * \in \{+, -, \times, /\}$, then*

$$X * Y \subset X' * Y'.$$

For division, we assume that $0 \notin Y'$.

Proof. Obvious from Definition 3.5. □

3.1.2 Floating-point interval arithmetic

When it comes to implementing interval arithmetic on a computer, we no longer work over $\mathbb{R}$, but in most cases with floating-point numbers. Let $\mathcal{F}$ be the set of machine numbers we are working with. Then we denote

$$\mathbb{I}\mathcal{F} = \{[\underline{x}, \bar{x}] : \underline{x}, \bar{x} \in \mathcal{F}\}.$$

Of course the set of floating-point numbers is not arithmetically closed (e.g., the sum of two floating-point numbers is not always a floating-point number). When we perform arithmetic operations on intervals in $\mathbb{I}\mathcal{F}$, we need to make sure to round the resulting interval outwards in order to guarantee that it contains the actual result. For $X, Y \in \mathbb{I}\mathcal{F}$, we set

$$\begin{aligned}
X + Y &= \left[\triangledown\left(\underline{x} + \underline{y}\right), \vartriangle\left(\bar{x} + \bar{y}\right)\right],\\
X - Y &= \left[\triangledown\left(\underline{x} - \bar{y}\right), \vartriangle\left(\bar{x} - \underline{y}\right)\right],\\
X \times Y &= \left[\min\left(\triangledown\left(\underline{x}\cdot\underline{y}\right), \triangledown\left(\underline{x}\cdot\bar{y}\right), \triangledown\left(\bar{x}\cdot\underline{y}\right), \triangledown\left(\bar{x}\cdot\bar{y}\right)\right),\right.\\
&\qquad \left.\max\left(\vartriangle\left(\underline{x}\cdot\underline{y}\right), \vartriangle\left(\underline{x}\cdot\bar{y}\right), \vartriangle\left(\bar{x}\cdot\underline{y}\right), \vartriangle\left(\bar{x}\cdot\bar{y}\right)\right)\right],\\
X/Y &= \left[\min\left(\triangledown\left(\underline{x}/\underline{y}\right), \triangledown\left(\underline{x}/\bar{y}\right), \triangledown\left(\bar{x}/\underline{y}\right), \triangledown\left(\bar{x}/\bar{y}\right)\right),\right.\\
&\qquad \left.\max\left(\vartriangle\left(\underline{x}/\underline{y}\right), \vartriangle\left(\underline{x}/\bar{y}\right), \vartriangle\left(\bar{x}/\underline{y}\right), \vartriangle\left(\bar{x}/\bar{y}\right)\right)\right] \quad \text{if } 0 \notin Y,
\end{aligned}$$

where $\triangledown$ and $\vartriangle$ denote rounding to $-\infty$ and $+\infty$ respectively. Using (3.1), we can extend floating-point interval arithmetic: $\overline{\mathbb{I}\mathcal{F}} = \left\{[\underline{x}, \bar{x}] : \underline{x}, \bar{x} \in \overline{\mathcal{F}}\right\}$.

Remark 3.12. Standard machine floating-point numbers are not always sufficient, e.g., to work with very small intervals. We may also use multiple-precision floating-point numbers as bounds for our intervals. An example of a library which offers support for multiple precision interval arithmetic is MPFR[5].

3.2 Interval functions

Definition 3.13. *Let $D \subset \mathbb{R}$, and let $f : D \to \mathbb{R}$. We denote*

$$R(f, D) = \{f(x) : x \in D\}$$

the range of f over D.

Remark 3.14. Finding the exact image of a (usually multivariate) function, and, in particular, a value where f attains its minimum is a whole subdomain of Mathematics and Computer Science called Optimization Theory.

Let $X = [\underline{x}, \bar{x}] \in \mathbb{IR}$. By monotonicity, interval functions defined as follows give the exact range of the corresponding real functions:

$$\begin{aligned}
e^X &= [\exp\underline{x}, \exp\bar{x}],\\
\sqrt{X} &= \left[\sqrt{\underline{x}}, \sqrt{\bar{x}}\right], \qquad \underline{x} \geqslant 0,\\
\log X &= [\log\underline{x}, \log\bar{x}], \qquad \underline{x} > 0,\\
\arctan X &= [\arctan\underline{x}, \arctan\bar{x}].
\end{aligned}$$

For some other functions like x^n, trigonometric functions..., writing down $R(f, D)$ is also possible, as long as we know their extrema. For instance, let $n \in \mathbb{Z}$, $X \in \mathbb{IR}$,

$$X^n = \mathrm{pow}(X, n) = \begin{cases}
\text{if } n \in 2\mathbb{N}+1, [\underline{x}^n, \bar{x}^n] \\
\text{if } n \in \mathbb{N}\setminus\{0\}, n \text{ even}, [\min\left(\underline{x}^n, \bar{x}^n\right), \max\left(\underline{x}^n, \bar{x}^n\right)] \text{ if } 0 \notin X, \\
\qquad\qquad\qquad\qquad [0, \max\left(\underline{x}^n, \bar{x}^n\right)] \text{ otherwise}, \\
[1, 1] \text{ if } n = 0, \\
[1/\bar{x}, 1/\underline{x}]^{-n} \text{ if } -n \in \mathbb{N} \text{ and } 0 \notin X.
\end{cases}$$

[5] http://www.mpfr.org

Exercise 3.2.1. Write the analogous formulas for sin, cos, tan. For sin and tan, consider

$$S_1^+ = \left\{2k\pi + \frac{\pi}{2}, k \in \mathbb{Z}\right\}, \quad S_1^- = \left\{2k\pi - \frac{\pi}{2}, k \in \mathbb{Z}\right\}.$$

For cos, consider

$$S_2^+ = \{2k\pi, k \in \mathbb{Z}\}, \quad S_2^- = \{2k\pi + \pi, k \in \mathbb{Z}\}.$$

The example of $f(x) = x^2 - x + 1$ over $[0, 2]$ illustrates two important issues:

- overestimation;
- dependency on the way the function is written.

We have $f(x) \in [0, 2]^2 - [0, 2] + [1] = [0, 4] + [-2, 0] + [1] = [-1, 5]$. Now write $f(x) = x(x-1)+1$. We have $f(x) \in [0, 2][-1, 1] + [1] = [-2, 2] + [1, 1] = [-1, 3]$. Actually, $R(f, [0, 2]) = [3/4, 3]$.

Definition 3.15. *(Interval extension.) Let $X \in \mathbb{IR}$, and let $f : X \to \mathbb{R}$. A function $\tilde{f} : \mathbb{IR} \to \mathbb{IR}$ is called an interval extension of f over X if:*

- *for all $x \in X$, $R(f, \{x\}) = \tilde{f}([x, x])$,*
- *for all $Y \subset \mathbb{IR}$ with $Y \subset X$, we have $R(f, Y) \subset \tilde{f}(Y)$.*

Several interval extensions are possible for the same function over the same X. Interval extensions of $\exp$ over $[-1, 1]$ include

- the function $X = [\underline{x}, \bar{x}] \mapsto [e^{\underline{x}}, e^{\bar{x}}]$.
- the function $X = [\underline{x}, \bar{x}] \mapsto [e^{\underline{x}}, e^{\bar{x}}] + X - X$.

Let us try to propose a systematic process for computing interval extensions. If $f(x)$ is a rational *expression*, one means to get an interval extension of the function it denotes is to replace each occurrence of the variable x by the interval X, and "overload" all arithmetic operations with interval operations. The resulting extension is called *the* natural interval extension.

Theorem 3.16. *Given a rational expression denoting a real-valued function f, and its natural interval extension F, which we assume to be well-defined over some interval $X \in \mathbb{IR}$, then*

1. *$Z \subset Z' \subset X$ implies $F(Z) \subset F(Z')$ (inclusion isotonicity);*
2. *$R(f, X) \subset F(X)$ (range enclosure).*

Proof. To prove assertion 1, it suffices to repeatedly use Lemma 3.11. Regarding assertion 2, assume that there exists $y \in X$ such that $f(y) \in R(f, X)$ but $f(y) \notin F(X)$. This implies that $F([y, y]) = [f(y), f(y)] \not\subset F(X)$ which contradicts assertion 1. □

We now would like to extend this notion of natural interval extension to a larger class of functions.

Definition 3.17. *We call basic (or standard) functions the elements of*

$$\mathfrak{S} = \{\sin, \cos, \exp, \tan, \log, x^{p/q} \textit{ with } p \in \mathbb{Z}, q \in \mathbb{N} \setminus \{0\}\}$$

for which we can determine the exact range over a given interval based on a simple rule.

These functions are said to have a sharp interval enclosure.

Definition 3.18. *We call elementary function a symbolic expression built from constants and basic functions using arithmetic operations and composition. The class of elementary functions will be denoted $\mathcal{E}$. A function $f \in \mathcal{E}$ is given by an expression tree (or dag, for directed acyclic graph).*

Definition 3.19. *An interval valued function $F : X \cap \mathbb{IR} \to \mathbb{IR}$ is inclusion isotonic over $X \in \mathbb{IR}$ if $Z \subset Z' \subset X$ implies $F(Z) \subset F(Z')$.*

Theorem 3.20. *Given an elementary function f and an interval X over which the natural interval extension F of f is well-defined:*

1. *F is inclusion isotonic over X;*
2. *$R(f, X) \subset F(X)$.*

Proof. The statement holds for rational functions (cf. Lemma 3.11) and, by definition, for standard functions. Let g and h be two elementary functions for which the Theorem holds. We have to prove that it holds as well for the function $g * h$, where $* \in \{+, -, \times, /, \circ\}$. Let's address the case of $\circ$ (the other cases are analogous). We assumed that $F(X)$ is well defined: this implies that neither f nor any of its sub-expressions have singularities in their domains, induced by the interval X. Then, if Z and Z' denote two sub-intervals on the domain of h such that $Z \subset Z'$, the function h is continuous over the compact sets Z and Z'. If G and H denote the natural interval extensions of g and h, the sets $H(Z)$ and $H(Z')$ are compact intervals. Using the inclusion isotonicity property satisfied by G and H, we obtain $H(Z) \subset H(Z')$ and $G \circ H(Z) = G(H(Z)) \subset G(H(Z')) = G \circ H(Z')$.

The proof of the second assertion is analogous to the proof of assertion 2 of Theorem 3.16. □

Example 3.21. Consider

$$f(x) = (\cos x - x^3 + x)(\tan x + 1/2)$$

over $[0, \pi/4]$. To show that f has no zero in this range, we compute the natural interval extension

$$f([0, \pi/4]) = \left[\frac{\sqrt{2}}{2} - \frac{\pi^3}{64}, 1 + \frac{\pi}{4}\right] \subset [0.22, 1.18].$$

Exercise 3.2.2. Show that $f(x) = x - \sin x + 2/5$ has no zero over $[0, \pi/4]$. Hint: evaluating the natural interval extension is not enough. Split the domain.

Theorem 3.22. *Let $X \in \mathbb{IR}$. Let f be an elementary function such that any subexpression of f is Lipschitz continuous. Let F be an inclusion isotonic interval extension such that $F(X)$ is well-defined. Then, there exists $\kappa > 0$, depending on F and X, such that, if $X = \bigcup_{i=1}^{k} X_i$, with $X_i \in \mathbb{IR}$ for all i,then*

$$R(f, X) \subset \bigcup_{i=1}^{k} F(X_i) \subset F(X)$$

and

$$\operatorname{rad}\left(\bigcup_{i=1}^{k} F(X_i)\right) \leqslant \operatorname{rad}(R(f, X)) + \kappa \max_{i=1,\dots,k} \operatorname{rad} X_i.$$

Proof. The first inclusion follows from the inclusion isotonicity and the range enclosure properties:

$$R(f, X) = R\left(f, \bigcup_{i=1}^{k} F(X_i)\right) = \bigcup_{i=1}^{k} R(f, X_i) \subset \bigcup_{i=1}^{k} F(X_i) \subset F\left(\bigcup_{i=1}^{k} X_i\right) = F(X).$$

Now, we are going to prove the following fact: there exists $\kappa > 0$ such that if $Z \subset X$ and $y_0 \in F(Z)$, then for all $y \in R(f, X)$, we have $|y - y_0| \leqslant \kappa \operatorname{rad}(Z)$. This implies the inequality in the theorem.

This statement is true for constants. It is also true for standard functions which are bounded (they have a sharp interval enclosure). In the same way as in the proof of Theorem 3.20, we consider two connecting branches g_1 and g_2 of the expression tree defining f and we prove that the statement is also valid for $g_1 * g_2$ where $* \in \{+, -, \times, /, \circ\}$. We focus on $\circ$, the other cases being analogous. The functions g_1 and g_2 are elementary (as sub-expressions of the elementary function f) and Lipschitz continuous. From Theorem 3.20, their natural interval extensions G_1 and G_2 are inclusion isotonic. We also know that, since $F(X)$ is well-defined, these extensions are also well-defined on their respective domains Z_{G_1} and Z_{G_2} induced by X. We assumed that the statement is true for g_1 and g_2: for $i = 1, 2$,

$$\text{if } V \subset Z_{G_i}, y_0 \in G_i(V) \text{ and } y \in R(g_i, V), \text{ then } |y - y_0| \leqslant \kappa_i \text{ rad } (V).$$

The range enclosure property of Theorem 3.20 gives: for all $V \subset Z_H$, we have

$$R(g_1 \circ g_2, V) = R(g_1, R(g_2, V)) \subset R(g_1, G_2(V)).$$

Let $z \in R(g_1 \circ g_2, V)$, there exists $u \in R(g_2, V)$ s.t. $z = g_1(u)$. The real number u also belongs to $G_2(V)$. Therefore, if $z_0 \in G_1 \circ G_2(V)$=$G_1(G_2(V))$, the inductive assumptions on g_i and G_i yield

$$|z - z_0| \leqslant \kappa_1 \operatorname{rad}(G_2(V))$$

and

$$\operatorname{rad}(G_2(V)) \leqslant \operatorname{rad}(R(g_2, V)) + \kappa_2 \operatorname{rad}(V) \leqslant (K_2 + \kappa_2) \operatorname{rad}(V),$$

where K_2 is a Lipschitz constant for g_2. If we combine these two inequalities, we obtain

$$|z - z_0| \leqslant \kappa_1(\kappa_2 + K_2) \operatorname{rad}(V).$$

Now, we use the fact that the expression tree defining f is finite by definition, which implies that the constant κ of the statement exists: it is the result of a finite accumulation of constants yielded as above. □

However, the number of subdivisions needed may be very large.

Example 3.23. Let $f(x) = e^{1/\cos x}$, and let p be a degree-10 minimax approximation of f over $[0, 1]$. Let

$$\varepsilon(x) = f(x) - p(x).$$

Using the natural interval extension of ε, we get $\|\varepsilon\|_\infty \leqslant 298$. But one can show that obtaining the true value $\|\varepsilon\|_\infty \approx 3.8325 \cdot 10^{-5}$ by subdivision would require about 10^7 subintervals.

3.3 Interval Newton method

First we recall the standard classical method.

3.3.1 Newton method

Theorem 3.24. *Let $X \in \mathbb{IR}$, let $f \in \mathcal{C}^2(X)$, s.t. $f'(x) \neq 0$ for all $x \in X$ and f has a unique, simple zero x^* in X. Then if x_0 is chosen sufficiently close to x^*, the sequence $(x_k)_{k \in \mathbb{N}}$ defined by*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \text{ for } k = 0, 1, 2, \dots$$

converges quadratically fast toward x^*: there exists a constant C such that

$$\lim_{k \to +\infty} x_k = x^* \text{ and } |x_{k+1} - x^*| \leqslant C|x_k - x^*|^2.$$

Proof. Let ε_k denote the error $x_k - x^*$. There exists ξ between x_k and x^* such that

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2} f''(\xi)(x^* - x_k)^2.$$

Since $f'(x_k) \neq 0$, the last equation is equivalent to

$$\frac{f(x_k)}{f'(x_k)} + x^* - x_k = x^* - x_{k+1} = -\frac{f''(\xi)(x^* - x_k)^2}{2f'(x_k)},$$

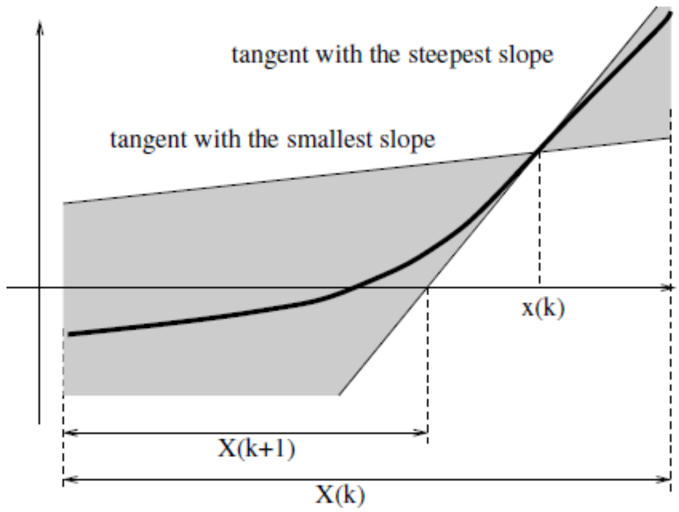


Figure 3.1: One iteration of Interval Newton method

hence

$$|\varepsilon_{k+1}| = \left|\frac{f''(\xi)}{2f'(x_k)}\right| \varepsilon_k^2 \leqslant \underbrace{\max\left\{-\frac{f''(x)}{2f'(y)}; x, y \in X\right\}}_{=:C} \varepsilon_k^2.$$

We can assume $C > 0$ (otherwise f is affine and the determination of x^* is straightforward). If $|C\varepsilon_0| < 1$ and $[x^* - \varepsilon_0, x^* + \varepsilon_0] \subset X$, then, by induction, we have, for all $k \in \mathbb{N}$,

$$x_k \in X \quad \text{and} \quad \varepsilon_k \leqslant \frac{1}{C}(C\varepsilon_0)^{2^k}.$$

□

3.3.2 Interval Newton method

Let $X \in \mathbb{IR}$, let $f \in \mathcal{C}^1(X)$, let F' an interval extension of f'. We assume $0 \notin F'(X)$ (hence f is strictly monotone). Assume that $x^* \in X$ is a zero of f. Let $x \in X$, the mean value theorem gives the existence of ξ between x and x^* such that $f(x) = f(x^*) + f(\xi)(x - x^*)$, i.e.

$$x^* = x - \frac{f(x)}{f'(\xi)}$$

since $f(x^*) = 0$ and $f'(\xi) \neq 0$. Hence, we have

$$x^* \in X \cap \left(x - \frac{f(x)}{F'(X)}\right).$$

Now, we start by giving a kind of interval version of the classical Newton method. We start with $x_0 \in X_0 = X \in \mathbb{IR}$. We obtain $x_k \in X_k$ and we define

$$X_{k+1} = \left(x_k - \frac{f(x_k)}{F'(X_k)}\right) \cap X_k.$$

This corresponds to Figure 3.1. Actually, it proves more interesting to rather set $x_k = \mathrm{mid}(X_k)$ all along the process.

We first define the *interval Newton operator*: for any $Y \in \mathbb{IR}$,

$$N(Y) = m + \frac{f(m)}{F'(Y)}, \text{ with } m = \mathrm{mid}(Y).$$

We start with $X_0 = X \in \mathbb{IR}$ and let m_k denote the middle of X_k. Let

$$X_{k+1} = N(X_k) \cap X_k, k = 0, 1, 2, \ldots.$$

Theorem 3.25. *Assume that $N(X)$ is well defined. If X contains a unique, simple zero x^*, then so do all iterates $X_k, k \in \mathbb{N}$. Moreover, the intervals X_k form a nested sequence converging to $[x^*]$.*

Proof. By construction, if $x^* \in X$, then $x^* \in X_k$ for all k and we also have $X_{k+1} \subset X_k$ for $k = 0, 1, \ldots$.

If there exists k_0 such that $X_{k_0} = [x^*, x^*]$, then $X_k = [x^*, x^*]$ for all $k \geqslant k_0$. This situation occurs if and only if $x^* = \mathrm{mid}(X_{k_0-1})$.

Now we assume that for all k, $x^* \neq \mathrm{mid}(X_k)$, hence $f(m_k) \neq 0$. Moreover, since $N(X)$ is well-defined, it implies that $0 \notin F'(X)$. Therefore, the elements of $\frac{m_k}{F'(X_k)}$ have the same sign. This implies $\mathrm{mid}(X_k) \notin X_{k+1}$, hence $\mathrm{rad}(X_{k+1}) < \mathrm{rad}(X_k)/2$. □

Theorem 3.26. *(Brouwer, 1910)*
Every continuous function f from a convex compact subset K of a Euclidean space to K itself has a fixed point.

Theorem 3.27. *Let $X \in \mathbb{IR}$, $f \in \mathcal{C}^1(X)$. Let F' an interval extension of f'. We assume $0 \notin F'(X)$.*

Let $I \in \mathbb{IR}$, $x \in I \subset X$, $N(I, x) := x - F'(I)^{-1} f(x)$

If $N(I)$ is well defined, then the following statements hold:

(1) if I contains a zero x^* of f, then so does $N(I, x) \cap I$;

(2) if $N(I, x) \cap I = \emptyset$, then I contains no zero of f;

(3) if $N(I, x) \subseteq I$, then I contains a unique zero of f.

Proof. (1) Follows from Mean Value Theorem;

(2) Contra-positive of (1);

(3) The existence is a consequence of Theorem 3.26.

The uniqueness follows from non-vanishing F', since f is then strictly monotone. □

Chapter 4

Rigorous Polynomial Approximations

Let $p(x)$ be the degree-10 minimax approximation to $e^{1/\cos x}$ over $[-1, 1]$. We observed earlier that obtaining a good enclosure of $e^{1/\cos x} - p(x)$ using the natural interval enclosure of this expression on subintervals would require about 10^7 subintervals. In this chapter, we present some tools that make it possible to get a certified enclosure, as sharp as desired, of this error function $e^{1/\cos x} - p(x)$ in an efficient way.

Definition 4.1. *Let $f \in \mathcal{C}([a, b])$. A rigorous polynomial approximation to f is a pair (p, Δ) where $p \in \mathbb{R}[x]$ and $f(x) - p(x) \in \Delta$ for all $x \in [a, b]$.*

4.1 Chebyshev Models

Let $I = [a, b]$, we define Chebyshev polynomials over I as

$$T_n^{[a,b]}(x) = T_n\left(\frac{2x - b - a}{b - a}\right), \, n \in \mathbb{N}.$$

For any $n \geqslant 0$, $T_{n+1}^{[a,b]}$ has $n + 1$ distinct real roots in $[a, b]$ (Chebyshev nodes of the first kind):

$$\mu_k = \frac{a + b}{2} + \frac{b - a}{2} \cos\left(\frac{(k + 1/2)\,\pi}{n + 1}\right), \, k = 0, \ldots, n.$$

We now adapt to the general case $[a, b]$ Proposition 1.21 and Theorem 1.20:

Proposition 4.2. *The polynomial $W_{\overline{\mu}}(x) = \prod_{k=0}^{n} (x - \mu_k)$, is the monic degree-$(n+1)$ polynomial that minimizes the supremum norm over $[a, b]$ of all monic polynomials in $\mathbb{C}[x]$ of degree at most $n + 1$. We have*

$$W_{\overline{\mu}}(x) = \frac{(b - a)^{n+1}}{2^{2n+1}} T_{n+1}^{[a,b]}(x)$$

and

$$\max_{x \in [a,b]} |W_{\overline{\mu}}(x)| = \frac{(b - a)^{n+1}}{2^{2n+1}}.$$

Theorem 4.3. *(Taylor-Lagrange-like formula.) Let $n \in \mathbb{N}$, and let $f \in \mathcal{C}^{n+1}([a, b])$. Let $P \in \mathbb{R}_n[X]$ be the interpolation polynomial of f at the Chebyshev nodes $(\mu_k)_{0 \leqslant k \leqslant n}$. For all $x \in [a, b]$, there exists $\xi_x \in (a, b)$ such that*

$$f(x) = P(x) + \frac{(b - a)^{n+1} f^{(n+1)}(\xi_x)}{2^{2n+1}} T_{n+1}^{[a,b]}(x)\,.$$

Let $n \in \mathbb{N}$, $n + 1$ times differentiable function f over $[a, b]$, we have

- $f(x) = \underbrace{\sum_{k=0}^{n} p_k T_k^{[a,b]}(x)}_{T(x)} + \underbrace{\Delta_n(x,\xi)}_{\text{remainder}}$
- $\Delta_n(x,\xi) = \frac{(b-a)^{n+1} f^{(n+1)}(\xi_x)}{2^{2n+1}} T_{n+1}^{[a,b]}(x)$, $x \in [a,b]$, ξ lies strictly between a and b

This raises two questions:

- How to compute the coefficients p_i of $T(x)$?
- How to compute an interval enclosure $\mathbf{\Delta}$ for $\Delta_n(x,\xi)$?

Computations of the coefficients $P(x) = \sum\limits_{i=0}^{n} p_i T_i^{[a,b]}(x)$, with $p_i = \sum\limits_{k=0}^{n} \frac{2}{n+1} f(\mu_k) T_i^{[a,b]}(\mu_k)$.

We replace the μ_k's and the $f(\mu_k)$'s with interval enclosures, and then perform an interval evaluation with Clenshaw's method: the coefficients p_i are intervals.

Bounding the remainder Since, for any $x \in [a,b]$, ξ_x lies strictly between a and b, we have

$$\Delta_n(x,\xi)| \leqslant \frac{(b-a)^{n+1} \left| f^{(n+1)}([a,b]) \right|}{2^{2n+1}}.$$

If f satisfies a differential equation with polynomial coefficients: it is fairly easy to retrieve an upper bound for $\left| f^{(n+1)}([a,b]) \right|$. But what about the other functions?

We will distinguish between

- *basic* or *standard* functions for which we assume to have a means to compute a tight Taylor-Lagrange-like remainder efficiently,
- and more general functions resulting from the composition of standard functions.

And, then, for bounding the remainder of a function f given as an expression tree (or a directed acyclic graph) whose nodes are basic functions:

- if f is a basic function, we use the Taylor-Lagrange-like formula,
- if f is a composite function,
  - first compute a rigorous polynomial approximation (T, Δ) for each basic function in the expression tree of the function f,
  - and then apply arithmetic rules overloading the operations.

Definition 4.4. *We call Chebyshev model a rigorous polynomial approximation computed using this scheme.*

4.1.1 Arithmetic operations on Chebyshev models

Let (P_1, Δ_1) and (P_2, Δ_2) be two Chebyshev models with $\deg P_1 \leqslant n, \deg P_2 \leqslant n$ associated respectively to f_1 and f_2.

Addition. We have for all $x \in [a,b]$

$$f_1(x) - P_1(x) \in \Delta_1, \qquad f_2(x) - P_2(x) \in \Delta_2,$$

and hence

$$(f_1(x) + f_2(x)) - (P_1(x) + P_2(x)) \in \Delta_1 + \Delta_2.$$

So we define

$$(P_1, \Delta_1) + (P_2, \Delta_2) = (P_1 + P_2, \Delta_1 + \Delta_2).$$

Multiplication. For any $m, n \in \mathbb{N}$, we have

$$T_m^{[a,b]}(x) \cdot T_n^{[a,b]}(x) = \frac{T_{m+n}^{[a,b]} + T_{|m-n|}^{[a,b]}}{2}.$$

Consider $P(x) = \sum_{i=0}^{n} p_i T_i^{[a,b]}(x)$ and $Q(x) = \sum_{i=0}^{n} q_i T_i^{[a,b]}(x)$. We have $P(x) \cdot Q(x) = \sum_{k=0}^{2n} c_k T_k^{[a,b]}(x)$, where

$$c_k = \left(\sum_{|i-j|=k} p_i q_j + \sum_{i+j=k} p_i q_j \right) /2.$$

The cost is $O(n^2)$ operations.

Given two Chebyshev Models for f_1 and f_2, over $[a,b]$, degree n: $f_1(x) - P_1(x) \in \boldsymbol{\Delta}_1$ and $f_2(x) - P_2(x) \in \boldsymbol{\Delta}_2$, $\forall x \in [a,b]$, we need an algebraic rule for $(P_1, \boldsymbol{\Delta}_1) \cdot (P_2, \boldsymbol{\Delta}_2) = (P, \boldsymbol{\Delta})$ such that $\deg P \leqslant n$ and $f_1(x) \cdot f_2(x) - P(x) \in \boldsymbol{\Delta}, \forall x \in [a,b]$.

Letting $\eta_{i,x} = f_i(x) - p_i(x)$, we have

$$f_1(x)f_2(x) = p_1(x)p_2(x) + \eta_{1,x}p_2(x) + \eta_{2,x}p_2(x) + \eta_{1,x}\eta_{2,x},$$

so

$$f_1(x)f_2(x) - p_1(x)p_2(x) \in I_1 = \tilde{p}_1([a,b])\Delta_2 + \tilde{p}_2([a,b])\Delta_1 + \Delta_1\Delta_2.$$

A difference with the previous case is that in general $\deg(p_1p_2) > n$. We write $p_1p_2 = \sum_{k=0}^{2n} a_k T_k^{[a,b]}(x) = p + q$ where p is the "order-n truncation of p_1p_2, that is to say $p(x) = \sum_{k=0}^{2n} a_k T_k^{[a,b]}(x)$. We let $I_2 = \tilde{q}([a,b])$, and we set

$$(p_1, \Delta_1)(p_2, \Delta_2) = (p, I_1 + I_2).$$

Using naive polynomial mutliplication, the cost of computing $(p_1, \Delta_1)(p_2, \Delta_2)$ is $O(n^2)$ arithmetic operations.

Composition. When one needs a Chebyshev model for $f_1 \circ f_2$, the first thing to do is to take into account the fact that the image of f_2 must be included in the definition set of f_1. Therefore we start with computing $B(P_2) + \boldsymbol{\Delta_2}$. If the inclusion is verified, then we have

$$(f_1 \circ f_2)(x) - P_1(f_2(x)) \in \boldsymbol{\Delta_1} \tag{4.1}$$

from which follows

$$(f_1 \circ f_2)(x) \in P_1(P_2(x) + \boldsymbol{\Delta_2}) + \boldsymbol{\Delta_1}.$$

In this formula, the only polynomial coefficients and the only remainders that are involved are those of the Chebyshev models of f_1 and f_2. When one uses Formula (4.1), it is not obvious to extract a polynomial and a bound for the remainder.

The idea is to use an adaptation of Clenshaw's evaluation scheme (cf. Algorithm 2). In our case, the variable x at which the sum is evaluated is a Chebyshev model and the multiplications and the additions are operations on Chebyshev models. Moreover, this algorithm requires a linear number of operations on the models.

4.1.2 Ranges of polynomials

Observe that we heavily used enclosures of ranges of polynomials. This raises (at least) two questions:

- How do we compute these enclosures?
- why would this process yield tight enclosures?

Algorithm 3 Composition of Chebyshev models : "$(P_1, \mathbf{\Delta_1}) \circ (P_2, \mathbf{\Delta_2})$"

Input: Two Chebyshev models $(P_1, \mathbf{\Delta_1})$ and $\circ(P_2, \mathbf{\Delta_2})$, respectively representing the functions f_1 and f_2

Output: A Chebyshev model for $f_1 \circ f_2$

// *Chebyshev models for 0*

$(C_{n+2}, \boldsymbol{R_{n+2}}) \leftarrow (0, [0, 0])$

$(C_{n+1}, \boldsymbol{R_{n+1}}) \leftarrow (0, [0, 0])$

// We denote $P_{1,j}$ the j-th coefficient of P_1

for $j = n, \dots, 1$ **do**

  $(C_j, \boldsymbol{R_j}) \leftarrow 2 \cdot (P_2, \mathbf{\Delta_2}) \cdot (C_{j+1}, \boldsymbol{R_{j+1}}) - (C_{j+2}, \boldsymbol{R_{j+2}}) + (P_{1,j}, [0, 0])$;

end for

$(C, \boldsymbol{R}) \leftarrow (P_2, \mathbf{\Delta_2}) \cdot (C_1, \boldsymbol{R_1}) - (C_2, \boldsymbol{R_2}) + (P_{1,0}, [0, 0])$

Return $(C, \boldsymbol{R} + \mathbf{\Delta_1})$

A first option is the following. Let $p(x) = a_0 + a_1 T_1^{[a,b]}(x) + \cdots + a_n T_n^{[a,b]}(x)$, as, $p(I)$ is bounded by $p(x) = |a_0| + |a_1| + \cdots + |a_n|$.

Another possibility is to use Bernstein's basis: indeed, one can show that if

$$p(x) = \sum_{k=0}^{n} p_k B_{n,k}(x),$$

then for all $x \in [0, 1]$, we have

$$\min_{[0,1]} p \geqslant \min_k p_k \qquad \text{and} \qquad \max_{[0,1]} p \leqslant \max_k p_k.$$

Nevertheless, in this case, there is a need for a conversion algorithm. Its cost is in $O(M(n))$ but currently, it raises problems of numerical stability.

There exist tighter methods based on Descartes' rule of signs, Sturm's theorem, sums of squares (Hilbert's 17th problem), companion matrices, etc. but they are more costful.

Second, why would this process yield tight enclosures? Our basic functions are analytic, and hence the coefficients of Chebyshev interpolants (quickly) converge to 0.

4.2 Banach fixed-point theorem

Theorem 4.5 (Banach fixed-point – global version)**.** *Let (X, d) be a complete metric space and $\mathbf{T} : X \to X$ be an operator. If $\mathbf{T}$ is contracting over X, that is, if there exists a $\mu \in [0, 1)$ such that $\mathbf{T}$ is μ-Lipschitz over X:*

$$\forall x_1, x_2 \in X, \quad d(\mathbf{T} \cdot x_1, \mathbf{T} \cdot x_2) \leqslant \mu d(x_1, x_2),$$

then $\mathbf{T}$ admits a unique fixed point x^* in X.

Moreover, for any $x^\circ \in X$, the approximation error of x° to x^* satisfies the following inequality:

$$\frac{d(x^\circ, \mathbf{T} \cdot x^\circ)}{1 + \mu} \quad \leqslant \quad d(x^\circ, x^*) \quad \leqslant \quad \frac{d(x^\circ, \mathbf{T} \cdot x^\circ)}{1 - \mu}. \tag{4.2}$$

Proof. Let $x_n := \mathbf{T}^n \cdot x^\circ$ for $n \geqslant 0$ denote the iterates of x° under $\mathbf{T}$, that is $x_0 := x^\circ$ and $x_{n+1} := \mathbf{T} \cdot x_n$. Let moreover $b := d(x^\circ, \mathbf{T} \cdot x^\circ)$.

By an easy induction, one gets $d(x_n, x_{n+1}) \leqslant \mu^n b$. Since $\mu < 1$, this yields for all n and $m \geqslant n$:

$$d(x_n, x_m) \leqslant \sum_{k=n}^{m-1} \mu^k b \leqslant \sum_{k=n}^{+\infty} \mu^k b = \frac{\mu^n b}{1 - \mu} \to 0 \quad \text{as } n \to +\infty.$$

Therefore, (x_n) is a Cauchy sequence, and hence converges to some $x^* \in X$ by completeness. Moreover, $\mathbf{T} \cdot x^* = x^*$ since $\mathbf{T}$ is Lipschitz, hence continuous.

Now, to prove the uniqueness, let $\bar{x}, x^* \in X$ be two fixed points of $\mathbf{T}$. Then we have:

$$d(\bar{x}, x^*) = d(\mathbf{T} \cdot \bar{x}, \mathbf{T} \cdot x^*) \leqslant \mu d(\bar{x}, x^*),$$

from which follows $d(\bar{x}, x^*) = 0$, hence $\bar{x} = x^*$ since $\mu < 1$.

Finally, the enclosure (4.2) is proved by using the triangle inequality:

$$\begin{aligned} d(x^\circ, x^*) &\leqslant d(x^\circ, \mathbf{T} \cdot x^\circ) + d(\mathbf{T} \cdot x^\circ, x^*) \leqslant d(x^\circ, \mathbf{T} \cdot x^\circ) + \mu d(x^\circ, x^*), \\ d(x^\circ, \mathbf{T} \cdot x^\circ) &\leqslant d(x^\circ, x^*) + d(x^*, \mathbf{T} \cdot x^\circ) \leqslant d(x^\circ, x^*) + \mu d(x^\circ, x^*). \end{aligned}$$

□

It is quite often the case that the operator $\mathbf{T}$ is contracting only over some *neighborhood* of the candidate approximation x°. In fact, one can restrict the ambient metric space X to any closed stable subset $S \subseteq X$, since S (with the induced metric d) is again a complete metric space. However, for the purpose of *effective* validation, we give the following *local* statement. By replacing the general – and difficult to check – notion of closed stable subset with that of *strongly stable ball* we obtain a formulation where rigorously verifying the preconditions only requires bounding distances and checking real inequalities.

Theorem 4.6 (Banach fixed-point – local version)**.** *Let (X, d) be a complete metric space, an operator $\mathbf{T} : X \to X$, $x^\circ \in X$, and $\mu, b, r \in \mathbb{R}_+$, satisfying the following conditions:*

1. $d(x^\circ, \mathbf{T} \cdot x^\circ) \leqslant b$;
2. $\mathbf{T}$ *is μ-Lipschitz over the closed ball $\bar{B}(x^\circ, r) := \{x \in X \mid d(x, x^\circ) \leqslant r\}$:*

$$\forall x_1, x_2 \in X, \quad x_1 \in \bar{B}(x^\circ, r) \wedge x_2 \in \bar{B}(x^\circ, r) \Rightarrow d(\mathbf{T} \cdot x_1, \mathbf{T} \cdot x_2) \leqslant \mu d(x_1, x_2);$$

3. $\mu < 1$ — $\mathbf{T}$ *is contracting over $\bar{B}(x^\circ, r)$;*
4. $b + \mu r \leqslant r$ — $\bar{B}(x^\circ, r)$ *is called a μ-*strongly stable ball with offset b.

Then $\mathbf{T}$ admits a unique fixed-point x^* in $\bar{B}(x^\circ, r)$.

Proof. The ball $\bar{B}(x^\circ, r)$ is stable under $\mathbf{T}$. Indeed, for any $x \in \bar{B}(x^\circ, r)$,

$$d(x^\circ, \mathbf{T} \cdot x) \leqslant d(x^\circ, \mathbf{T} \cdot x^\circ) + d(\mathbf{T} \cdot x^\circ, \mathbf{T} \cdot x) \leqslant b + \mu r \leqslant r.$$

Since $\mathbf{T}$ is contracting over $\bar{B}(x^\circ, r)$, Theorem 4.5 – applied to the closed subspace $\bar{B}(x^\circ, r)$ instead of X – guarantees the existence and uniqueness of a fixed point x^* of $\mathbf{T}$ in $\bar{B}(x^\circ, r)$. □

Remark 4.7. The reader may wonder why we do not provide an enclosure of $d(x^\circ, x^*)$ in Theorem 4.6, similarly to (4.2). In fact, the lower bound is not used in practice – it just tells that the enclosure is sharp for small μ – and Condition 4 is equivalent to $r \geqslant \frac{b}{1-\mu}$, so that one can directly choose $r := \frac{b}{1-\mu}$ without violating the four conditions of the theorem.

4.2.1 Newton-like Fixed-Point Methods for A Posteriori Validation

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two Banach spaces, $\mathbf{F} : X \to Y$ be a continuous and differentiable operator with respect to $\|\cdot\|_X$ and $\|\cdot\|_Y$, and $x^\circ \in X$.

Let $\mathbf{A} : X \to X$ be an injective bounded linear operator, from which we construct the Newton-like operator $\mathbf{T} : X \to X$:

$$\mathbf{T} \cdot x = x - \mathbf{A} \cdot \mathbf{F} \cdot x.$$

Theorem 4.8. *Suppose $\mathbf{A}$ "sufficiently close" to $(\mathrm{D}\mathbf{F}_{x^\circ})^{-1}$, so that $\exists r > 0$ s. t.*

(i) There exists $\lambda \in [0,1)$ *bounding* $\|\mathrm{D}\mathbf{T}_x\|_X$ *over* $\overline{B}(x^\circ, r) = \{x \in X \mid \|x - x^\circ\|_X \leqslant r\}$*:*

$$\|\mathrm{D}\mathbf{T}_x\|_X = \|\mathbf{1}_X - \mathbf{A} \cdot \mathrm{D}\mathbf{F}_x\|_X \leqslant \lambda, \qquad x \in \overline{B}(x^\circ, r).$$

(ii) $\overline{B}(x^\circ, r)$ *is stable under* $\mathbf{T}$*, which is ensured by:*

$$\|x^\circ - \mathbf{T} \cdot x^\circ\|_X + \lambda r < r.$$

Then $\mathbf{T}$ *admits a unique fixed point called* x^* *in* $\overline{B}(x^\circ, r)$*, and we have the following enclosure:*

$$\frac{\|x^\circ - \mathbf{T} \cdot x^\circ\|_X}{1 + \lambda} \leqslant \|x^\circ - x^*\|_X \leqslant \frac{\|x^\circ - \mathbf{T} \cdot x^\circ\|_X}{1 - \lambda}.$$

4.2.2 Division of two Chebyshev models

Let $I = [a, b]$, for $f, g \in \mathcal{C}(I)$ with g nonvanishing over I, the quotient f/g is the unique root of $\mathbf{F} : h \mapsto gh - f$. Let h° be a candidate approximation given by the approximation step. Constructing the Newton-like operator $\mathbf{T}$ requires an approximation $\mathbf{A}$ of $(\mathrm{D}\mathbf{F}_{h^\circ})^{-1} : k \mapsto k/g$. For that purpose, suppose $k^\circ \approx 1/g \in \mathcal{C}(I)$ is also given by an oracle, and define:

$$\mathbf{T} \cdot h = h - k^\circ(gh - f). \tag{4.3}$$

The next proposition gives an upper bound for $\|h^\circ - f/g\|$.

Proposition 4.9. *Let* $f, g, h^\circ, k^\circ \in \mathcal{C}(I)$*, and* $\mu, b \in \mathbb{R}_+$ *such that:*

(4.9 i) $\|k^\circ(gh^\circ - f)\| \leqslant b$, (4.9 ii) $\|1 - k^\circ g\| \leqslant \mu$, (4.9 iii) $\mu < 1$.

Then g *does not vanish over* I *and* $\|h^\circ - f/g\| \leqslant \frac{b}{1-\mu}$.

Proof. Conditions (4.9 ii) and (4.9 iii) imply that $\mathbf{T}$ (Equation (4.3)) is contracting over $\mathcal{C}(I)$ with ratio μ. The radius $r := \frac{b}{1-\mu}$ makes the ball $\overline{B}(h^\circ, r)$ strongly stable with offset b (4.9 i), since $b + \mu r = r$. Therefore, h^* is the (global) unique root of $\mathbf{F}$, and $\|h^\circ - h^*\| \leqslant r$.

Finally, k° and g do not vanish because $\|1 - k^\circ g\| \leqslant \mu < 1$. Hence, $h^* = f/g$ over I. □

This implies

Corollary 4.10. *Given:*

- $f, g \in \mathcal{C}(I)$ *represented by Chebyshev models* $\boldsymbol{f} = (f^\circ, \varepsilon)$ *and* $\boldsymbol{g} = (g^\circ, \eta)$*,*
- $h^\circ \in \mathbb{R}[x]$ *a polynomial approximation of* $h^* = f/g$*,*
- $k^\circ \in \mathbb{R}[x]$ *a polynomial approximation of* $1/g$*,*

we have the following rigorous upper bound on the approximation error:

$$\|h^\circ - f/g\|_\infty \leqslant \tau = \frac{\tau'}{1 - \lambda},$$

provided that we have computed τ' *and* $\lambda < 1$ *such that:*

$$\|1 - k^\circ g^\circ\|_\infty + \eta \|k^\circ\|_\infty \leqslant \lambda,$$
$$\|k^\circ(g^\circ h^\circ - f^\circ)\|_\infty + \eta \|k^\circ h^\circ\|_\infty + \varepsilon \|k^\circ\|_\infty \leqslant \tau'.$$

Hence, $\boldsymbol{h} = (h^\circ, \tau)$ *is a Chebyshev model for* $h^* = f/g$*.*

4.2.3 Square root of a Chebyshev model

Let $I = [a, b]$, let $f \in \mathcal{C}(I)$ be strictly positive over I. The square root $\sqrt{f}$ is one of the two roots of the quadratic equation $\mathbf{F} \cdot h := h^2 - f = 0$ (the other being $-\sqrt{f}$). Let h° be a candidate approximation. Since $\mathrm{D}\mathbf{F}_h : k \mapsto 2hk$, one also needs an approximation $k^\circ \approx 1/(2h^\circ) \approx 1/(2\sqrt{f}) \in \mathcal{C}(I)$ in order to define $\mathbf{A} : k \mapsto k^\circ k$, approximating $(\mathrm{D}\mathbf{F}_{h^\circ})^{-1}$. Then:

$$\mathbf{T} : \quad h \quad \mapsto \quad h - k^\circ(h^2 - f).$$

The next proposition computes an upper bound for $\|h^\circ - \sqrt{f}\|$.

Proposition 4.11. *Let $f, h^\circ, k^\circ \in \mathcal{C}(I)$, $\mu_0, \mu_1, b \in \mathbb{R}_+$ and $t_0 \in I$ such that:*

$$\text{(4.11 i)} \quad \left\| k^\circ \left(h^{\circ 2} - f \right) \right\| \leqslant b, \qquad \text{(4.11 ii)} \quad \|1 - 2k^\circ h^\circ\| \leqslant \mu_0, \qquad \text{(4.11 iii)} \quad \|k^\circ\| \leqslant \mu_1,$$

$$\text{(4.11 iv)} \quad \mu_0 < 1, \qquad \text{(4.11 v)} \quad (1 - \mu_0)^2 - 8b\mu_1 \geqslant 0, \qquad \text{(4.11 vi)} \quad k^\circ(t_0) > 0.$$

Then $f > 0$ over I and $\left\| h^\circ - \sqrt{f} \right\| \leqslant r^*$ where $r^* := \frac{1-\mu_0-\sqrt{(1-\mu_0)^2-8b\mu_1}}{4\mu_1}$.

Proof. First, since $\|1 - 2k^\circ h^\circ\| \leqslant \mu_0 < 1$ (by (4.11 ii) and (4.11 iv)) and $k^\circ(t_0) > 0$ (4.11 vi), k° and h° are strictly positive over I, by continuity. Using (4.11 iii), $\mu_1 > 0$.

If $b = 0$, then $r^* = 0$ and $h^\circ = \sqrt{f}$ over I, because $k^\circ(h^{\circ 2} - f) = 0$ (4.11 i) and $k^\circ, h^\circ > 0$. Hence the conclusion holds.

From now on, we assume $b > 0$. $\mathbf{T}$ is Lipschitz of ratio $\mu(r) := \mu_0 + 2\mu_1 r$ over $\bar{B}(h^\circ, r)$ for any $r \in \mathbb{R}_+$, because:

$$\mathbf{T} \cdot h_1 - \mathbf{T} \cdot h_2 = (h_1 - h_2) - k^\circ(h_1^2 - h_2^2) = [(1 - 2k^\circ h^\circ) + k^\circ(h^\circ - h_1) + k^\circ(h^\circ - h_2)]\,(h_1 - h_2).$$

Therefore, satisfying $b + \mu(r)r \leqslant r$ is equivalent to the quadratic inequality:

$$2\mu_1 r^2 + (\mu_0 - 1)r + b \leqslant 0. \tag{4.4}$$

Condition (4.11 v) implies that (4.4) admits solutions, and r^* is the smallest one. Moreover, since $b, \mu_1 > 0$, we get $r^* > 0$, so that $b + \mu(r^*)r^* = r^*$ also implies $\mu(r^*) < 1$.

Now, all the assumptions of Theorem 4.6 are fulfilled. Hence, $\mathbf{T}$ has a unique fixed point h^* in $\bar{B}(h^\circ, r^*)$. To obtain $h^* = \sqrt{f}$ over I, it remains to show that $h^* > 0$. This follows from $k^\circ > 0$ and:

$$\|1 - 2k^\circ h^*\| \leqslant \|1 - 2k^\circ h^\circ\| + \|2k^\circ(h^* - h^\circ)\| \leqslant \mu_0 + 2\mu_1 r^* = \mu(r^*) < 1. \qquad \square$$

Corollary 4.12. *Given:*

- *$f \in \mathcal{C}(I)$ represented by a Chebyshev model $\boldsymbol{f} = (f^\circ, \varepsilon)$,*
- *$h^\circ \in \mathbb{R}[x]$ a polynomial approximation of $h^* = \sqrt{f}$,*
- *$k^\circ \in \mathbb{R}[x]$ a polynomial approximation of $1/h^\circ$,*

we have the following rigorous upper bound on the approximation error:

$$\|h^\circ - \sqrt{f}\|_\infty \leqslant \eta = \frac{\eta'}{1 - \lambda},$$

provided that we have computed $\lambda_0, \lambda_1, \eta', \Delta, r^\circ, \lambda$ satisfying:

$$\|1 - k^\circ h^\circ\|_\infty \leqslant \lambda_0 < 1, \ \|k^\circ\|_\infty \leqslant \lambda_1, \ \|k^\circ(h^{\circ 2} - f^\circ)\|_\infty + \varepsilon\|k^\circ\|_\infty \leqslant 2\eta',$$

$$\Delta := (1 - \lambda_0)^2 - 4\lambda_1\eta' \geqslant 0, \qquad r^\circ := \frac{1 - \lambda_0 - \sqrt{\Delta}}{2\lambda_1},$$

$$\lambda := \lambda_0 + \lambda_1 r^\circ < 1.$$

Hence, $\boldsymbol{h} = (h^\circ, \eta)$ is a Chebyshev model for $h^* = \sqrt{f}$.

Appendix A

A Short Reminder on Floating-Point Arithmetic

We first recall the definition of a floating-point (FP) number.

Definition A.1. *Let $\beta, p, E_{\min}, E_{\max} \in \mathbb{Z}$, $\beta, p \geqslant 2, E_{\min} < 0 < E_{\max}$, a (normal) radix-$\beta$ FP number in precision p with exponent range $[E_{\min}, E_{\max}]$ is a number of the form*

$$x = (-1)^s \frac{M}{\beta^{p-1}} \cdot \beta^E,$$

where :

- *the* exponent *$E \in \mathbb{Z}$ is such that $E_{min} \leqslant E \leqslant E_{max}$*
- *the* integral significand *$M \in \mathbb{N}$ represented in radix β satisfies $\beta^{p-1} \leqslant |M| \leqslant \beta^p - 1$;*
- *$s \in \{0, 1\}$ is the sign bit of x.*

In these lecture notes, we leave the exponent range implicit unless it is explicitly required, and simply talk about "radix-β FP numbers in precision p".

Remark A.2. For the sake of clarity, we chose not to mention subnormal FP numbers since they will not appear in the text. One can find the complete definition in [Muller et al., 2010, Chap. 2.1].

The number zero is a special case, cf. [Muller et al., 2010, Chap. 3], that we add to the set of radix-β and precision-p FP numbers. This yields a set denoted $\mathcal{F}_{\beta,p}$.

Remark A.3. In this course, we use radix 2 for the sake of clarity but our approach remains valid for any radix, in particular radix 10, the importance of which grows at a steady pace. We will use $\mathcal{F}_p$ instead of $\mathcal{F}_{2,p}$.

In radix 2, with the required bounds on the exponent E:

- $\mathcal{F}_{24}$ is the set of the binary32 (formerly single precision) floating-point numbers,
- $\mathcal{F}_{53}$ is the set of the binary64 (formerly double precision) floating-point numbers,
- $\mathcal{F}_{113}$ is the set of the binary128 (formerly quadruple precision) floating-point numbers.

In 1985, the IEEE 754 [Cody et al., 1984, IEEE Computer Society, 1985] standard for radix-2 floating-point arithmetic was released, followed two years later by its generalization to decimal arithmetic, the IEEE 854 standard [Cody, 1985, A.N.S.I. and I.E.E.E., 1987]. This was a key step to end an era of great disorder: the improvement in terms of accuracy, reliability and portability of machine computations have been quite substantial. This standard specifies in particular various formats, the behaviour of the arithmetic operations ($+, -, \times, \div$ and $\sqrt{}$) or conversions. Currently, most systems of commercial

	precision p	minimal exponent E_{min}	maximal exponent E_{max}
binary32	24	-126	127
binary64	53	-1022	1023
binary128	113	-16382	16383

Table A.1: Main parameters of the three basic binary formats (up to 128 bits) specified by the standard [IEEE Computer Society, 2008].

significance offer compatibility 3 with IEEE 754-1985. A revision of this standard, called IEEE 754-2008 (which also include radix-10 FP arithmetic), was adopted in June 2008 [IEEE Computer Society, 2008]. Table A.1 gives the main parameters of the three basic binary formats specified by IEEE 754-2008.

The result of an arithmetic operation whose input values belong to $\mathcal{F}_p$ may not belong to $\mathcal{F}_p$ (in general it does not). Hence that result must be rounded. The IEEE standard defines 4 different rounding modes; in the sequel, x is any real number to be rounded:

- rounding towards $+\infty$, or upwards: $\circ_u(x)$ is the smallest element of $\mathcal{F}_p$ that is greater than or equal to x;
- rounding towards $-\infty$, or downwards: $\circ_d(x)$ is the largest element of $\mathcal{F}_p$ that is less than or equal to x;
- rounding towards 0: $\circ_z(x)$ is equal to $\circ_u(x)$ if $x < 0$, and to $\circ_d(x)$ otherwise;
- rounding to the nearest even: $\circ_n(x)$ is the element of $\mathcal{F}_p$ that is closest to x. If x is exactly halfway between two consecutive elements of $\mathcal{F}_p$, $\circ_n(x)$ is the one for which the integral significand j is an even number.

The first three rounding modes are called directed rounding modes.

Bibliography

[A.N.S.I. and I.E.E.E., 1987] A.N.S.I. and I.E.E.E. (1987). *IEEE Standard for Radix Independent Floating-Point Arithmetic*. ANSI/IEEE Standard 854–1987.

[Bernstein, 1926] Bernstein, S. N. (1926). *Leçons sur les propriétés extrémales et la meilleure approximation des fonctions analytiques d'une variable réelle*. Gauthier-Villars.

[Bogaert, 2014] Bogaert, I. (2014). Iteration-free computation of Gauss-Legendre quadrature nodes and weights. *SIAM J. Sci. Comput.*, 36(3):A1008–A1026.

[Brezis, 2010] Brezis, H. (2010). *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media.

[Brutman, 1978] Brutman, L. (1978). On the Lebesgue function for polynomial interpolation. *SIAM J. Numer. Anal.*, 15(4):694–704.

[Brutman, 1997] Brutman, L. (1997). Lebesgue functions for polynomial interpolation—a survey. *Ann. Numer. Math.*, 4(1-4):111–127. The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin.

[Cheney, 1998] Cheney, E. W. (1998). *Introduction to approximation theory*. AMS Chelsea Publishing, Providence, RI. Reprint of the second (1982) edition.

[Cody, 1985] Cody, W. J. (1985). A proposed radix and word length independent standard for floating-point arithmetic. *ACM SIGNUM Newsletter*, 20:37–51.

[Cody et al., 1984] Cody, W. J., Coonen, J. T., Gay, D. M., Hanson, K., Hough, D., Kahan, W., Karpinski, R., Palmer, J., Ris, F. N., and Stevenson, D. (1984). A proposed radix-and-word-length-independent standard for floating-point arithmetic. *IEEE MICRO*, 4(4):86–100.

[Cooley and Tukey, 1965] Cooley, J. and Tukey, J. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

[Daumas and Melquiond, 2010] Daumas, M. and Melquiond, G. (2010). Certification of bounds on expressions involving rounded operators. *ACM Trans. Math. Software*, 37(1):Art. 2, 20.

[de Dinechin et al., 2011] de Dinechin, F., Lauter, C., and Melquiond, G. (2011). Certifying the floating-point implementation of an elementary function using Gappa. *IEEE Trans. Comput.*, 60(2):242–253.

[Duhamel and Vetterli, 1990] Duhamel, P. and Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing*, 19:259–299.

[Ehlich and Zeller, 1966] Ehlich, H. and Zeller, K. (1966). Auswertung der Normen von Interpolationsoperatoren. *Math. Ann.*, 164:105–112.

[Elliott et al., 1987] Elliott, D., Paget, D. F., Phillips, G. M., and Taylor, P. J. (1987). Error of truncated Chebyshev series and other near minimax polynomial approximations. *J. Approx. Theory*, 50(1):49–57.

[Farouki, 2012] Farouki, R. T. (2012). The Bernstein polynomial basis: a centennial retrospective. *Comput. Aided Geom. Design*, 29(6):379–419. Available from http://mae.engr.ucdavis.edu/~farouki/bernstein.pdf.

[Filip, 2016a] Filip, S.-I. (2016a). *Robust tools for weighted Chebyshev approximation and applications to digital filter design*. PhD thesis, École Normale Supérieure de Lyon.

[Filip, 2016b] Filip, S.-I. (2016b). A robust and scalable implementation of the Parks-McClellan algorithm for designing FIR filters. *ACM Trans. Math. Software*, 43(1).

[Hale and Townsend, 2013] Hale, N. and Townsend, A. (2013). Fast and Accurate Computation of Gauss–Legendre and Gauss–Jacobi Quadrature Nodes and Weights. *SIAM Journal on Scientific Computing*, 35(2):A652–A674.

[Hales, 2005] Hales, T. C. (2005). A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185.

[Hales et al., 2010] Hales, T. C., Harrison, J., McLaughlin, S., Nipkow, T., Obua, S., and Zumkeller, R. (2010). A revision of the proof of the Kepler conjecture. *Discrete Comput. Geom.*, 44(1):1–34.

[Heidemann et al., 1984] Heidemann, M., Johnson, D., and Burrus, C. (1984). Gauss and the history of the fast Fourier transform. *IEEE ASSP Magazine*, pages 14–21.

[IEEE Computer Society, 1985] IEEE Computer Society (1985). *IEEE Standard for Floating-Point Arithmetic*. IEEE Standard 754-1985. Available at https://ieeexplore.ieee.org/document/30711.

[IEEE Computer Society, 2008] IEEE Computer Society (2008). *IEEE Standard for Floating-Point Arithmetic*. IEEE Standard 754-2008. Available at http://ieeexplore.ieee.org/servlet/opac?punumber=4610933.

[Johansson and Mezzarobba, 2018] Johansson, F. and Mezzarobba, M. (2018). Fast and rigorous arbitrary-precision computation of Gauss-Legendre quadrature nodes and weights. *SIAM J. Sci. Comput.*, 40(6):C726–C747.

[Le Gall, 2014] Le Gall, F. (2014). Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, ISSAC '14, page 296–303, New York, NY, USA. Association for Computing Machinery.

[Loan, 1992] Loan, C. V. (1992). *Computational Frameworks for the Fast Fourier Transform*. Frontiers in Applied Mathematics. SIAM.

[Meinardus, 1967] Meinardus, G. (1967). *Approximation of functions: Theory and numerical methods*. Expanded translation of the German edition. Translated by Larry L. Schumaker. Springer Tracts in Natural Philosophy, Vol. 13. Springer-Verlag New York, Inc., New York.

[Melquiond, 2006] Melquiond, G. (2006). *De l'arithmétique d'intervalles à la certification de programmes*. PhD thesis, École Normale Supérieure de Lyon, Lyon, France.

[Muller et al., 2010] Muller, J.-M., Brisebarre, N., de Dinechin, F., Jeannerod, C.-P., Lefèvre, V., Melquiond, G., Revol, N., Stehlé, D., and Torres, S. (2010). *Handbook of Floating-Point Arithmetic*. Birkhäuser Boston. ACM G.1.0; G.1.2; G.4; B.2.0; B.2.4; F.2.1., ISBN 978-0-8176-4704-9.

[Oppenheim and Schafer, 2010] Oppenheim, A. V. and Schafer, R. W. (2010). *Discrete-Time Signal Processing*. Prentice-Hall Signal Processing Series. Prentice Hall.

[Parks and McClellan, 1972] Parks, T. W. and McClellan, J. H. (1972). Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase. *IEEE Transactions on Circuit Theory*, 19(2):189–194.

[Pinkus, 2000] Pinkus, A. (2000). Weierstrass and approximation theory. *J. Approx. Theory*, 107(1):1–66. available from http://www.math.technion.ac.il/hat/fpapers/wap.pdf.

[Powell, 1981] Powell, M. J. D. (1981). *Approximation theory and methods*. Cambridge University Press.

[Remes, 1934] Remes, E. (1934). Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation (in French). *Compt. Rend. Acad. Sci.*, 198:2063–2065.

[Rivlin, 1981] Rivlin, T. J. (1981). *An introduction to the approximation of functions*. Dover Publications, Inc., New York. Corrected reprint of the 1969 original, Dover Books on Advanced Mathematics.

[Schönhage, 1961] Schönhage, A. (1961). Fehlerfortpflanzung bei Interpolation. *Numer. Math.*, 3:62–71.

[Trefethen, 2013] Trefethen, L. N. (2013). *Approximation Theory and Approximation Practice*. SIAM. See http://www.chebfun.org/ATAP/.

[Trefethen and Weideman, 1991] Trefethen, L. N. and Weideman, J. A. C. (1991). Two results on polynomial interpolation in equally spaced points. *J. Approx. Theory*, 65(3):247–260.

[Tucker, 1999] Tucker, W. (1999). The Lorenz attractor exists. *C. R. Acad. Sci. Paris Sér. I Math.*, 328(12):1197–1202.

[Tucker, 2002] Tucker, W. (2002). A rigorous ODE solver and Smale's 14th problem. *Found. Comput. Math.*, 2(1):53–117.

[Tucker, 2011] Tucker, W. (2011). *Validated numerics: a short introduction to rigorous computations*. Princeton University Press.

[Turetskii, 1940] Turetskii, A. H. (1940). The bounding of polynomials prescribed at equally distributed points. *Proc. Pedag. Inst. Vitebsk*, 3:117–127.

[Veidinger, 1960] Veidinger, L. (1960). On the numerical determination of the best approximations in the Chebyshev sense. *Numer. Math.*, 2:99–105.

[von zur Gathen and Gerhard, 2013] von zur Gathen, J. and Gerhard, J. (2013). *Modern computer algebra*. Cambridge University Press, third edition.

[Zygmund, 2002] Zygmund, A. (2002). *Trigonometric series. Vol. I, II*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition.