Approximation Theory and Proof Assistants: Certified Computations

Nicolas Brisebarre and Damien Pous

Master 2 Informatique Fondamentale
École Normale Supérieure de Lyon, 2024-2025

Section 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

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

Section 2.2. A little bit of quadrature: Gauss methods

First of all, if $\ell_k(x) = \prod_{\substack{j=0,\\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}$, 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 $x_0, \ldots, x_n$, 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

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

Section 2.2. A little bit of quadrature: Gauss methods

Theorem 8

There exists a unique choice of the points x_k and the weights w_k such that, whenever $f \in \mathbb{R}_{2n+1}[x]$,

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

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

Remark

The Chebyshev polynomials of the first kind satisfy

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

If $p = \sum_{k=0}^{n} c_k T_k$, *we deduce that the integral with weight* $w = 1$ *is given by*

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

Section 2.3. Lebesgue constants

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

Definition 9

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.

Section 2.3. Lebesgue constants

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

Definition 9

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

Let Λ be the Lebesgue constant for the 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

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

2.3.1. Lebesgue constants for polynomial interpolation

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

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

2.3.1. Lebesgue constants for polynomial interpolation

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

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

Theorem 10

The Lebesgue constant of degree-n Lagrange interpolation at $x_0, \ldots, x_n$ is equal to

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

2.3.1. Lebesgue constants for polynomial interpolation

Theorem 11

The Lebesgue constant Λ_n satisfies

$$\frac{2}{\pi}\left(\log(n+1)+\gamma+\log\frac{4}{\pi}\right)\leqslant\Lambda_n,\ \textit{where}\ \frac{2}{\pi}\left(\gamma+\log\frac{4}{\pi}\right)=0.52125\ldots$$

Additionally,

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

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

- *for equispaced points,*

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

2.3.1. Lebesgue constants for polynomial interpolation

Remark

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.*

2.3.2. Lebesgue constants for L_2 best approximation

When the L_2 space under consideration is $L_2\left([-1,1], \frac{1}{\sqrt{1-x^2}}\right)$, the best polynomial approximation $p_{2,n}$ is called the truncated Chebyshev series of order n.

Theorem 12

The Lebesgue constant for the $L_2\left([-1,1], \frac{1}{\sqrt{1-x^2}}\right)$ *projection onto* $\mathbb{R}_n[x]$ *is*

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

We have

$$\Lambda_n \leqslant \frac{4}{\pi^2}\log(n+1)+3 \text{ and } \Lambda_n \sim \frac{4}{\pi^2}\log n \text{ as } n \to +\infty.$$

2.3.2. Lebesgue constants for L_2 best approximation

Remark

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.*

2.3.3. Corollary: A first statement on the convergence of Chebyshev interpolants and truncated Chebyshev series

Let $f \in \mathcal{C}([a,b])$. The modulus of continuity of f is the function ω defined as

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

2.3.3. Corollary: A first statement on the convergence of Chebyshev interpolants and truncated Chebyshev series

Let $f \in \mathcal{C}([a,b])$. The modulus of continuity of f is the function ω defined as

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

Proposition

If f is a continuous function over $[0,1]$, ω its modulus of continuity, then we have

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

2.3.3. Corollary: A first statement on the convergence of Chebyshev interpolants and truncated Chebyshev series

Theorem 13

If f is Lipschitz continuous over $[a, b]$, then

1. *the sequence of interpolation polynomials at the Chebyshev nodes uniformly converges to f.*
2. *The truncated Chebyshev series of f uniformly converges to f.*

Section 2.4.2. Convergence

Remark

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.

Section 2.4.2. Convergence

Theorem 14

Let f be continuous on $[-1, 1]$. Denote by (a_k) its sequence of Chebyshev coefficients, by (f_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 $k \to \infty$.*

Section 2.4.2. Convergence

Theorem 14

Let f be continuous on $[-1, 1]$. Denote by (a_k) its sequence of Chebyshev coefficients, by (f_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 absolutely and uniformly to f and (p_n) converges uniformly to f.*

Section 2.4.2. Convergence

Theorem 14

Let f be continuous on $[-1, 1]$. Denote by (a_k) its sequence of Chebyshev coefficients, by (f_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 absolutely and uniformly to f and (p_n) converges uniformly to f.*
3. *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 $\|f - p_n\|_\infty = O(n^{-m})$.*

Bernstein Ellipse

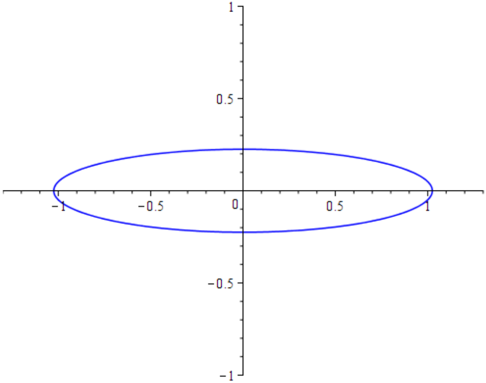
Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$

Bernstein Ellipse

Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$

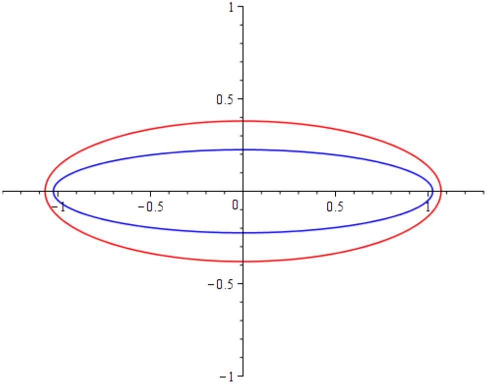


Bernstein ellipses for $\rho = 1.05$,

Bernstein Ellipse

Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$

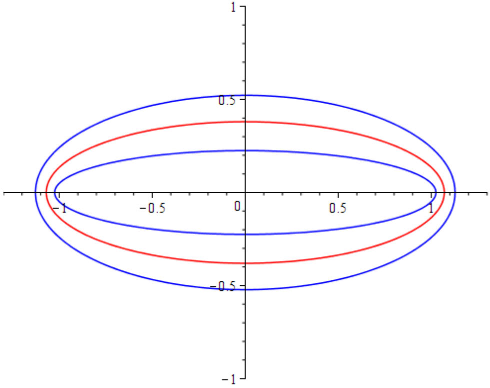


Bernstein ellipses for $\rho = 1.05$, 1.25,

Bernstein Ellipse

Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$

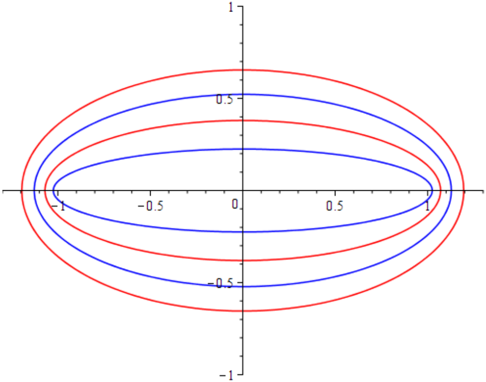


Bernstein ellipses for $\rho = 1.05$, 1.25, 1.45,

Bernstein Ellipse

Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$

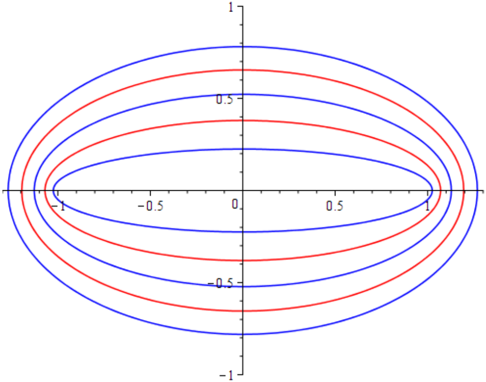


Bernstein ellipses for $\rho = 1.05, 1.25, 1.45, 1.65,$

Bernstein Ellipse

Let $\rho > 1$, let

$$\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\} = \left\{ z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho \right\}.$$



Bernstein ellipses for $\rho = 1.05,\ 1.25,\ 1.45,\ 1.65,\ 1.85$.

Section 2.4.2. Convergence

Theorem 15

Let f be continuous on $[-1, 1]$. Denote by (a_k) its sequence of Chebyshev coefficients, by (f_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. *If f is analytic inside the ellipse $\mathcal{E}_\rho :=$*
$$\left\{\frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi]\right\} = \left\{z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| \leqslant \rho\right\}$$
with $\rho > 1$, then $a_k = O(\rho^{-k})$, $\|f - f_n\|_\infty = O(\rho^{-n})$ and $\|f - p_n\|_\infty = O(\rho^{-n})$.

Section 2.4.2. Convergence

Theorem 16

Let f be continuous on $[-1, 1]$. Denote by (f_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

5. *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*

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

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

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

Approximation Theory and Proof Assistants: Certified Computations

Nicolas Brisebarre and Damien Pous

Master 2 Informatique Fondamentale
École Normale Supérieure de Lyon, 2024-2025

Chapter 3. Interval Arithmetic, Interval Analysis

Floating Point (FP) Arithmetic

Given

$$\begin{cases} \text{a radix} & \beta \geqslant 2, \\ \text{a precision} & p \geqslant 1, \\ \text{a set of exponents} & E_{\text{min}}, \cdots, E_{\text{max}}. \end{cases}$$

A finite FP number x is represented by 2 integers:

- integer mantissa : M, $\beta^{p-1} \leqslant |M| \leqslant \beta^p - 1$;
- exponent E, $E_{\text{min}} \leqslant E \leqslant E_{\text{max}}$

such that

$$x = \frac{M}{\beta^{p-1}} \times \beta^E.$$

We assume binary FP arithmetic (that is to say $\beta = 2$.)
We denote $\mathcal{F}_p$ the corresponding set of FP numbers.
Multiple-precision FP arithmetic: we let p and E vary.

IEEE Precisions

See http://en.wikipedia.org/wiki/IEEE_floating_point

	precision	minimal exponent	maximal exponent
single (binary 32)	24	-126	127
double (binary 64)	53	-1022	1023
extended double	64	-16382	16383
quadruple (binary 128)	113	-16382	16383

IEEE Rounding Modes

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): the result must be rounded.

IEEE Rounding Modes

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): the result must be rounded.

IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x) = \min\{y \in \mathcal{F}_p : y \geqslant x\}$;

IEEE Rounding Modes

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): the result must be rounded.

IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x) = \min\{y \in \mathcal{F}_p : y \geqslant x\}$;
- rounding towards $-\infty$, or downwards: $\circ_d(x) = \max\{y \in \mathcal{F}_p : y \leqslant x\}$;

IEEE Rounding Modes

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): the result must be rounded.

IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x) = \min\{y \in \mathcal{F}_p : y \geqslant x\}$;
- rounding towards $-\infty$, or downwards: $\circ_d(x) = \max\{y \in \mathcal{F}_p : y \leqslant x\}$;
- rounding towards 0: $\circ_z(x) := \circ_u(x)$ if $x < 0$, and to $\circ_d(x)$ otherwise;

IEEE Rounding Modes

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): the result must be rounded.

IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x) = \min\{y \in \mathcal{F}_p : y \geqslant x\}$;
- rounding towards $-\infty$, or downwards: $\circ_d(x) = \max\{y \in \mathcal{F}_p : y \leqslant x\}$;
- rounding towards 0: $\circ_z(x) := \circ_u(x)$ 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.

IEEE Rounding Modes

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): the result must be rounded.

IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x) = \min\{y \in \mathcal{F}_p : y \geqslant x\}$;
- rounding towards $-\infty$, or downwards: $\circ_d(x) = \max\{y \in \mathcal{F}_p : y \leqslant x\}$;
- rounding towards 0: $\circ_z(x) := \circ_u(x)$ 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.

Chapter 3. Interval Arithmetic, Interval Analysis, Rigorous Polynomial Approximations

Interval Arithmetic is "an arithmetic for inequalities".

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.

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.

In double precision, compute $x_{k+1} = (x_k)^2$ where $x_0 = 1 - 10^{-19}$.

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.

In double precision, compute $x_{k+1} = (x_k)^2$ where $x_0 = 1 - 10^{-19}$.

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 (see `https://code.google.com/p/flyspeck/`),
- W. Tucker's solution of Smale's 14th problem (see `https://www2.math.uu.se/~warwick/main/thesis.html` and also `https://paulbourke.net/fractals/lorenz/`).

Numerous additional interesting information on the website `https://www.cs.utep.edu/interval-comp/`.

Chapter 3. Interval Arithmetic, Interval Analysis

In this course, we are interested in the use of interval arithmetic for mathematical function evaluation purpose.

In this course, we are interested in the use of interval arithmetic for mathematical function evaluation purpose.
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

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

In this course, we are interested in the use of interval arithmetic for mathematical function evaluation purpose.
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 relative error $\left|1 - \frac{\widehat{f(x)}}{f(x)}\right|$ rather than 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.

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 $\eta_1 + \eta_2 \leqslant \varepsilon$.

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 course: tools that help to establish rigorous approximation error.

Regarding rounding errors, G.Melquiond has developed formal proof tools (in Coq) which address this issue (see `https://gappa.gitlabpages.inria.fr/`).

3.1. Interval arithmetic

Definition

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

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

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* $\mathbb{IR}$*.*

3.1. Interval arithmetic

Definition

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

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

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

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

It is common in the litterature to encounter the notation $(\mathrm{mid}(x), \mathrm{rad}(x)) = \{x \in \mathbb{R} : |x - \mathrm{mid}(x)| \leqslant \mathrm{rad}(x)\}$.

This mid-rad representation is the basis of the so called Ball Arithmetic, cf. the excellent software Arb, now a part of FLINT `https://flintlib.org/`.

3.1. Interval arithmetic

Remark

It is common in the litterature to encounter the notation
$(\mathrm{mid}\,(x)\,, \mathrm{rad}\,(x)) = \{x \in \mathbb{R} : |x - \mathrm{mid}\,(x)| \leqslant \mathrm{rad}\,(x)\}$.

This mid-rad representation is the basis of the so called Ball Arithmetic, cf. the excellent software Arb, now a part of FLINT `https://flintlib.org/`.

Definition

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

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

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

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

3.1.1. Operations on intervals

Proposition

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

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

which depend only on the endpoints.

Proof.

Exercise. □

3.1.1. Operations on intervals

Remark

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

Remark

In practice, multiplication (hence division) can be made more efficient (check the signs of the endpoints).

3.1.1. Operations on intervals

Proposition

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": if* $\underline{x} < 0 < \bar{x}$,

$$[\underline{x}, \bar{x}] \times [\underline{x}, \bar{x}] \neq \left[0, \max\left(\underline{x}^2, \bar{x}^2\right)\right].$$

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. □

3.1.1. Operations on intervals

A straightforward yet quite useful statement is the following.

Lemma

(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 . □

3.1.2. 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.

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

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

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).

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 "true result".

3.1.2. Floating-point interval arithmetic

For $X, Y \in \mathbb{IF}$, we set

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

where $\triangledown$ and $\vartriangle$ denote rounding to $-\infty$ and $+\infty$ respectively.

3.1.2. Floating-point interval arithmetic

Remark

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[1]*.*

[1] `http://www.mpfr.org`

3.2. Interval functions

Definition

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*.*

3.2. Interval functions

Definition

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

Finding the exact image of a (usually multivariate) function, and, in particular, a value where f *attains its minimum is a whole subdomain of Math and CS called Global Optimization.*

3.2. Interval functions

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:

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

3.2. Interval functions

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 $n \in \mathbb{Z},\ X \in \mathbb{IR}$,

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

3.2. Interval functions

Exercise

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

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

3.2. Interval functions

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.

3.2. Interval functions

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 $f(x) \in [0, 2]^2 - [0, 2] + [1] = [0, 4] + [-2, 0] + [1] = [-1, 5]$.

3.2. Interval functions

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
$f(x) \in [0,2][-1,1] + [1] = [-2,2] + [1,1] = [-1,3]$.

3.2. Interval functions

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, $R(f, [0, 2]) = [3/4, 3]$.