Machine-efficient polynomial approximation

Nicolas Brisebarre Sylvain Chevillard Guillaume Hanrot
Tom Hubrecht Serge Torres

SMAI 2025

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

(Binary) Floating Point (FP) Arithmetic

Given

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

A finite FP number x is represented by 2 integers:

- integer significand M, $2^{p-1} \leqslant |M| \leqslant 2^p - 1$,
- exponent E, $E_{\mathsf{min}} \leqslant E \leqslant E_{\mathsf{max}}$

such that

$$x = \frac{M}{2^{p-1}} \times 2^E.$$

IEEE Precisions

IEEE 754 standard (1984 then 2008).

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

	precision p	min. exponent E_{min}	maximal exponent E_{max}
binary32 (single)	24	-126	127
binary64 (double)	53	-1022	1023
extended double	64	-16382	16383
binary128 (quadruple)	113	-16382	16383

We have $x = \frac{M}{2^{p-1}} \times 2^E$ with $2^{p-1} \leqslant |M| \leqslant 2^p - 1$

and $E_{\mathrm{min}} \leqslant E \leqslant E_{\mathrm{max}}$.

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

- Step 1. Argument reduction (Payne & Hanek, Ng, Daumas *et al*): evaluation of a function φ over $\mathbb{R}$ or a subset of $\mathbb{R}$ is reduced to the evaluation of a function f over $[a, b]$.

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

- Step 1. Argument reduction (Payne & Hanek, Ng, Daumas *et al*): evaluation of a function φ over $\mathbb{R}$ or a subset of $\mathbb{R}$ is reduced to the evaluation of a function f over $[a, b]$.
- Step 2. Computation of $p^\star$, a "machine-efficient" polynomial approximation of f.

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

- Step 1. Argument reduction (Payne & Hanek, Ng, Daumas *et al*): evaluation of a function φ over $\mathbb{R}$ or a subset of $\mathbb{R}$ is reduced to the evaluation of a function f over $[a, b]$.
- Step 2. Computation of $p^\star$, a "machine-efficient" polynomial approximation of f.
- Step 3. Computation of a rigorous approximation error $||f - p^\star||$.

Evaluation of Elementary Functions

$\exp, \ln, \cos, \sin, \arctan, \sqrt{\ }, \ldots$

Goal: evaluation of φ to a given accuracy η.

- Step 1. Argument reduction (Payne & Hanek, Ng, Daumas *et al*): evaluation of a function φ over $\mathbb{R}$ or a subset of $\mathbb{R}$ is reduced to the evaluation of a function f over $[a, b]$.
- Step 2. Computation of $p^\star$, a "machine-efficient" polynomial approximation of f.
- Step 3. Computation of a rigorous approximation error $||f - p^\star||$.
- Step 4. Computation of a certified evalutation error of $p^\star$: GAPPA (G. Melquiond).

Why Use Polynomial Approximation?

The floating-point operations $+, -, \times, /, \sqrt{}$ are as accurate as possible (correct rounding).

Why Use Polynomial Approximation?

The floating-point operations $+, -, \times, /, \sqrt{}$ are as accurate as possible (correct rounding).

The floating-point operations $+, -, \times$ are very fast.

Why Use Polynomial Approximation?

The floating-point operations $+, -, \times, /, \sqrt{}$ are as accurate as possible (correct rounding).

The floating-point operations $+, -, \times$ are very fast.

We have very fine approximation and evaluation schemes for polynomials.

Why Use Polynomial Approximation?

The floating-point operations $+, -, \times, /, \sqrt{}$ are as accurate as possible (correct rounding).

The floating-point operations $+, -, \times$ are very fast.

We have very fine approximation and evaluation schemes for polynomials.

$\Longrightarrow$ Let's use polynomials!

Minimax Approximation

Reminder. Let $g : [a, b] \to \mathbb{R}$, $||g||_{\infty,[a,b]} = \sup_{a \leqslant x \leqslant b} |g(x)|$.

We denote $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leqslant n\}$.

Minimax Approximation

Reminder. Let $g : [a,b] \to \mathbb{R}$, $||g||_{\infty,[a,b]} = \sup_{a \leqslant x \leqslant b} |g(x)|$.

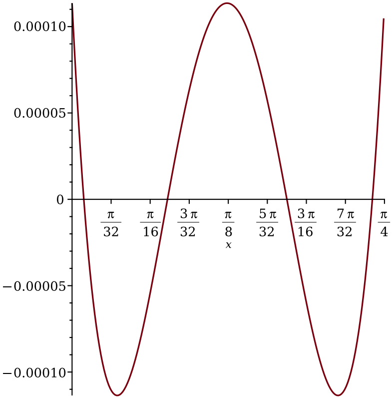
We denote $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leqslant n\}$.

Minimax approximation: let $f : [a,b] \to \mathbb{R}$, $n \in \mathbb{N}$, we search for $p \in \mathbb{R}_n[X]$ s.t.

$$||p - f||_{\infty,[a,b]} = \inf_{q \in \mathbb{R}_n[X]} ||q - f||_{\infty,[a,b]}.$$

Minimax Approximation

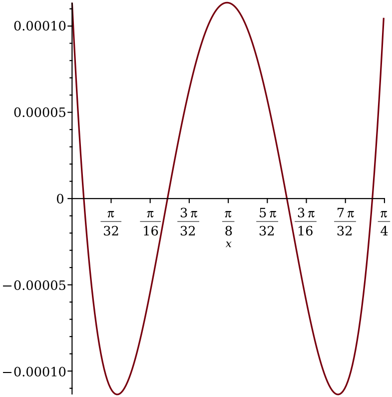
Degree-3 minimax approximation to $\cos$



Chebyshev's theorem (1902).

Minimax Approximation

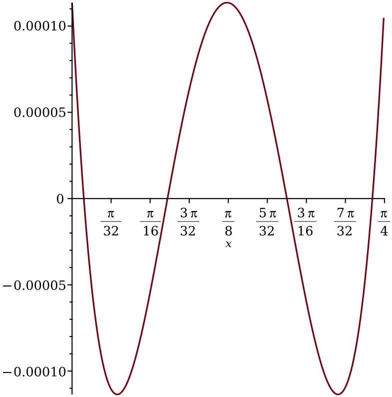
Degree-3 minimax approximation to $\cos$



~~Chebyshev~~/Borel's theorem (1902).

Minimax Approximation

Degree-3 minimax approximation to $\cos$



~~Chebyshev~~/~~Borel~~/Kirchberger's theorem (1902).

Minimax Approximation

Reminder. Let $g : [a, b] \to \mathbb{R}$, $||g||_{\infty,[a,b]} = \sup_{a \leqslant x \leqslant b} |g(x)|$.

We denote $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leqslant n\}$.

Minimax approximation: let $f : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, we search for $p \in \mathbb{R}_n[X]$ s.t.

$$||p - f||_{\infty,[a,b]} = \inf_{q \in \mathbb{R}_n[X]} ||q - f||_{\infty,[a,b]}.$$

Minimax Approximation

Reminder. Let $g : [a, b] \to \mathbb{R}$, $||g||_{\infty,[a,b]} = \sup_{a \leqslant x \leqslant b} |g(x)|$.

We denote $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leqslant n\}$.

Minimax approximation: let $f : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, we search for $p \in \mathbb{R}_n[X]$ s.t.

$$||p - f||_{\infty,[a,b]} = \inf_{q \in \mathbb{R}_n[X]} ||q - f||_{\infty,[a,b]}.$$

An algorithm due to Remez (1934) gives p (`minimax` function in Maple, `remez` function in Sollya `http://sollya.gforge.inria.fr/`).

Minimax Approximation

Reminder. Let $g : [a, b] \to \mathbb{R}$, $||g||_{\infty,[a,b]} = \sup_{a \leqslant x \leqslant b} |g(x)|$.

We denote $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leqslant n\}$.

Minimax approximation: let $f : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, we search for $p \in \mathbb{R}_n[X]$ s.t.

$$||p - f||_{\infty,[a,b]} = \inf_{q \in \mathbb{R}_n[X]} ||q - f||_{\infty,[a,b]}.$$

An algorithm due to Remez (1934) gives p (`minimax` function in Maple, `remez` function in Sollya `http://sollya.gforge.inria.fr/`).

Problem: we can't directly use minimax approx. in a computer since the coefficients of p can't be represented on a finite number of bits.

Machine-efficient Polynomials

Our context: the coefficients of the polynomials must be written on a finite (imposed) number of bits.

Machine-efficient Polynomials

Our context: the coefficients of the polynomials must be written on a finite (imposed) number of bits.

Let $m = (m_i)_{0 \leqslant i \leqslant n}$ a finite sequence of rational integers. Let

$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X]; q_i$ integer multiple of $2^{-m_i}, \forall i\}$.

Machine-efficient Polynomials

Our context: the coefficients of the polynomials must be written on a finite (imposed) number of bits.

Let $m = (m_i)_{0 \leqslant i \leqslant n}$ a finite sequence of rational integers. Let

$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X]; q_i$ integer multiple of $2^{-m_i}, \forall i\}$.

Question: find $p^\star \in \mathcal{P}_n^m$ which minimizes $\|f - q\|_\infty$, $q \in \mathcal{P}_n^m$.

Machine-efficient Polynomials

Our context: the coefficients of the polynomials must be written on a finite (imposed) number of bits.

Let $m = (m_i)_{0 \leqslant i \leqslant n}$ a finite sequence of rational integers. Let

$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X]; q_i$ integer multiple of $2^{-m_i}, \forall i\}$.

Question: find $p^\star \in \mathcal{P}_n^m$ which minimizes $\|f - q\|_\infty$, $q \in \mathcal{P}_n^m$.

First idea. Remez $\to p(x) = p_0 + p_1 x + \cdots + p_n x^n$. Every p_i rounded to $\hat{a}_i / 2^{m_i}$, the nearest integer multiple of $2^{-m_i} \to$

$\hat{p}(x) = \dfrac{\hat{a}_0}{2^{m_0}} + \dfrac{\hat{a}_1}{2^{m_1}} x + \cdots + \dfrac{\hat{a}_n}{2^{m_n}} x^n$.

Machine-efficient Polynomials

Our context: the coefficients of the polynomials must be written on a finite (imposed) number of bits.

Let $m = (m_i)_{0 \leqslant i \leqslant n}$ a finite sequence of rational integers. Let

$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X]; q_i$ integer multiple of $2^{-m_i}, \forall i\}$.

Question: find $p^\star \in \mathcal{P}_n^m$ which minimizes $\|f - q\|_\infty$, $q \in \mathcal{P}_n^m$.

First idea. Remez $\rightarrow p(x) = p_0 + p_1 x + \cdots + p_n x^n$. Every p_i rounded to $\hat{a}_i / 2^{m_i}$, the nearest integer multiple of $2^{-m_i} \rightarrow$
$\hat{p}(x) = \dfrac{\hat{a}_0}{2^{m_0}} + \dfrac{\hat{a}_1}{2^{m_1}} x + \cdots + \dfrac{\hat{a}_n}{2^{m_n}} x^n$.

Problem: $\hat{p}$ not necessarily a minimax approx. of f among the polynomials of $\mathcal{P}_n^m$.

Approximation of the Function $\cos$ over $[0, \pi/4]$ by a Degree-3 Polynomial

Maple or Sollya tell us that the polynomial

$p = 0.9998864206 + 0.00469021603x - 0.5303088665x^2 + 0.06304636099x^3$

is $\sim$ the best approximant to $\cos$. We have
$\varepsilon = || \cos - p||_{[0,\pi/4]} = 0.0001135879...$.

We look for $a_0, a_1, a_2, a_3 \in \mathbb{Z}$ such that

$$\max_{0 \leqslant x \leqslant \pi/4} \left| \cos x - \left(\frac{a_0}{2^{12}} + \frac{a_1}{2^{10}} x + \frac{a_2}{2^6} x^2 + \frac{a_3}{2^4} x^3 \right) \right|$$

is minimal.

Approximation of the Function $\cos$ over $[0, \pi/4]$ by a Degree-3 Polynomial

Maple or Sollya tell us that the polynomial

$p = 0.9998864206 + 0.00469021603x - 0.5303088665x^2 + 0.06304636099x^3$

is $\sim$ the best approximant to $\cos$. We have
$\varepsilon = || \cos - p||_{[0,\pi/4]} = 0.0001135879...$.

We look for $a_0, a_1, a_2, a_3 \in \mathbb{Z}$ such that

$$\max_{0 \leqslant x \leqslant \pi/4} \left| \cos x - \left(\frac{a_0}{2^{12}} + \frac{a_1}{2^{10}} x + \frac{a_2}{2^6} x^2 + \frac{a_3}{2^4} x^3 \right) \right|$$

is minimal.

The naive approach gives the polynomial

$$\hat{p} = \frac{2^{12}}{2^{12}} + \frac{5}{2^{10}} x - \frac{34}{2^6} x^2 + \frac{1}{2^4} x^3.$$

We have $\hat{\varepsilon} = || \cos - \hat{p}||_{[0,\pi/4]} = 0.00069397...$.

Approximation of the Function $\cos$ over $[0, \pi/4]$ by a Degree-3 Polynomial

Maple or Sollya computes a polynomial p which is $\sim$ the best approximant to $\cos$. We have $\varepsilon = ||\cos - p||_{[0,\pi/4]} = 0.0001135879...$.
We look for $a_0, a_1, a_2, a_3 \in \mathbb{Z}$ such that

$$\max_{0 \leqslant x \leqslant \pi/4} \left| \cos x - \left(\frac{a_0}{2^{12}} + \frac{a_1}{2^{10}} x + \frac{a_2}{2^6} x^2 + \frac{a_3}{2^4} x^3 \right) \right|$$

is minimal.
The naive approach gives the polynomial $\hat{p}$ and
$\hat{\varepsilon} = ||\cos - \hat{p}||_{[0,\pi/4]} = 0.00069397...$

Approximation of the Function $\cos$ over $[0, \pi/4]$ by a Degree-3 Polynomial

Maple or Sollya computes a polynomial p which is $\sim$ the best approximant to $\cos$. We have $\varepsilon = ||\cos - p||_{[0,\pi/4]} = 0.0001135879...$.
We look for $a_0, a_1, a_2, a_3 \in \mathbb{Z}$ such that

$$\max_{0 \leqslant x \leqslant \pi/4} \left| \cos x - \left(\frac{a_0}{2^{12}} + \frac{a_1}{2^{10}} x + \frac{a_2}{2^6} x^2 + \frac{a_3}{2^4} x^3 \right) \right|$$

is minimal.
The naive approach gives the polynomial $\hat{p}$ and $\hat{\varepsilon} = ||\cos - \hat{p}||_{[0,\pi/4]} = 0.00069397...$ But the best "constrained" approximant:

$$p^\star = \frac{4095}{2^{12}} + \frac{6}{2^{10}}\, x - \frac{34}{2^6}\, x^2 + \frac{1}{2^4}\, x^3$$

which gives $||\cos - p^\star||_{[0,\pi/4]} = 0.0002441406250$.

In this example, we gain $-\log_2(0.35) \approx 1.5$ bits of accuracy.

Back to our Problem

We're given $[a,b]$, $f : [a,b] \mapsto \mathbb{R}$ and $m = (m_i)_{0 \leqslant i \leqslant n}$ a finite sequence of rational integers.
Let

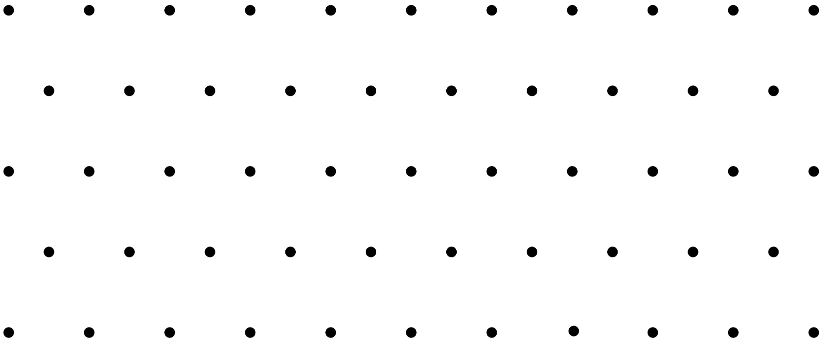
$$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X]; q_i \text{ integer multiple of } 2^{-m_i}, \forall i\}.$$

Question: find $a_0^\star, \ldots, a_n^\star \in \mathbb{Z}$ s.t.

$$p^\star = a_0^\star \underbrace{\frac{1}{2^{m_0}}}_{e_0(x)} + a_1^\star \underbrace{\frac{x}{2^{m_1}}}_{e_1(x)} + \cdots + a_n^\star \underbrace{\frac{x^n}{2^{m_n}}}_{e_n(x)}$$

minimizes $\|f - q\|_\infty$, $q \in \mathcal{P}_n^m$.

An Approach via Lattice Basis Reduction



An Approach via Lattice Basis Reduction

Definition

Let L be a nonempty subset of $\mathbb{R}^d$, L is a lattice iff there exists a set of vectors $b_1, \ldots, b_k$ $\mathbb{R}$-linearly independent such that

$$L = \mathbb{Z}.b_1 \oplus \cdots \oplus \mathbb{Z}.b_k.$$

$(b_1, \ldots, b_k)$ is a basis of the lattice L.

Examples. $\mathbb{Z}^d$, every subgroup of $\mathbb{Z}^d$.

An Approach via Lattice Basis Reduction

Definition

Let L be a nonempty subset of $\mathbb{R}^d$, L is a lattice iff there exists a set of vectors $b_1, \ldots, b_k$ $\mathbb{R}$-linearly independent such that

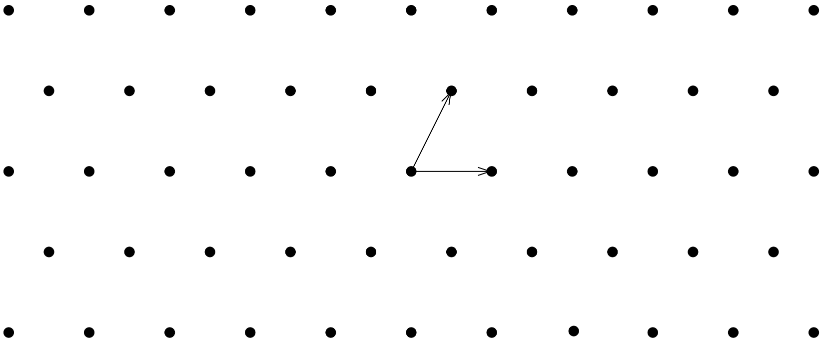
$$L = \mathbb{Z}.b_1 \oplus \cdots \oplus \mathbb{Z}.b_k.$$

$(b_1, \ldots, b_k)$ is a basis of the lattice L.

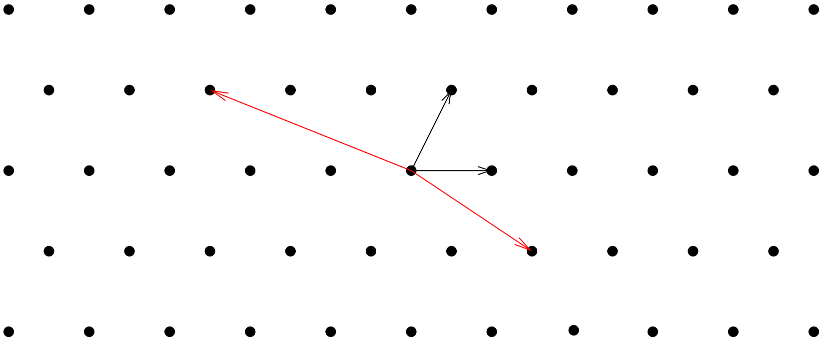
Examples. $\mathbb{Z}^d$, every subgroup of $\mathbb{Z}^d$.

L is equipped with $|(y_i)_{1 \leqslant i \leqslant d}|_2 = (\sum_{i=1}^{d} y_i^2)^{1/2}$ for all $(y_i)_{1 \leqslant i \leqslant d} \in \mathbb{R}^d$.

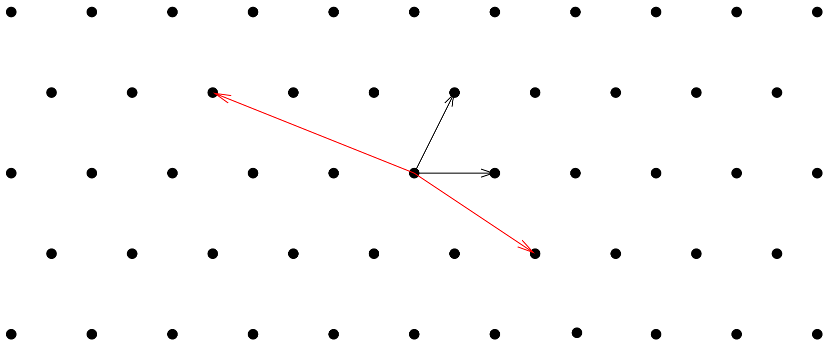
Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$



Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$

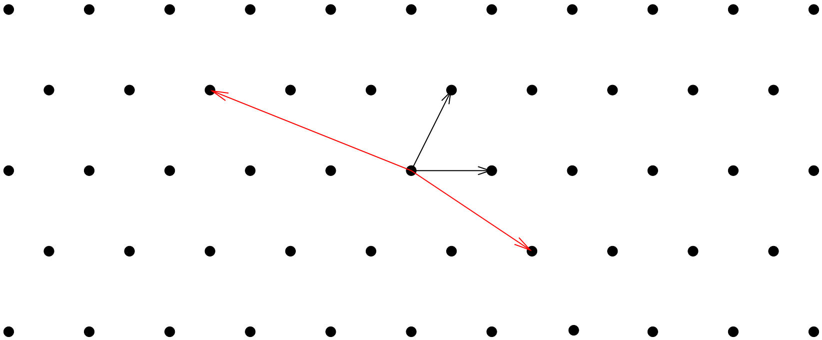


Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$



SVP (Shortest Vector Problem) and CVP (Closest Vector Problem)

Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$



SVP (Shortest Vector Problem) and CVP (Closest Vector Problem) are NP-hard.

Lenstra-Lenstra-Lovász Algorithm

SVP (Shortest Vector Problem) and CVP (Closest Vector Problem) are NP-hard.

Lenstra-Lenstra-Lovász Algorithm

SVP (Shortest Vector Problem) and CVP (Closest Vector Problem) are NP-hard.

Factoring Polynomials with Rational Coefficients, A. K. Lenstra, H. W. Lenstra and L. Lovász, Math. Annalen **261**, 515-534, 1982.

The LLL algorithm gives an approximate solution to SVP in polynomial time.

Babai's algorithm (based on LLL) gives an approximate solution to CVP in polynomial time.

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \ldots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = \frac{a_0^\star}{2^{m_0}} + \frac{a_1^\star}{2^{m_1}} X + \cdots + \frac{a_n^\star}{2^{m_n}} X^n$$

where $a_i^\star \in \mathbb{Z}$ that minimizes $\|f - p\|_\infty$.

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \ldots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = \frac{a_0^\star}{2^{m_0}} + \frac{a_1^\star}{2^{m_1}} X + \cdots + \frac{a_n^\star}{2^{m_n}} X^n$$

where $a_i^\star \in \mathbb{Z}$ that minimizes $\|f - p\|_\infty$.

First step: replace $\|g\|_\infty$ with $\|g\|_2 = \left(\int_{-1}^{1} g(x)^2 \frac{\mathrm{d}x}{\sqrt{1 - x^2}} \right)^{1/2}$.

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \ldots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = \frac{a_0^\star}{2^{m_0}} + \frac{a_1^\star}{2^{m_1}} x + \cdots + \frac{a_n^\star}{2^{m_n}} x^n$$

where $a_i^\star \in \mathbb{Z}$ which minimizes

$$\|f - p\|_2 = \left(\int_{-1}^{1} (f(x) - p(x))^2 \frac{\mathrm{d}x}{\sqrt{1 - x^2}} \right)^{1/2}.$$

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \ldots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = a_0^\star \underbrace{\frac{1}{2^{m_0}}}_{e_0(x)} + a_1^\star \underbrace{\frac{x}{2^{m_1}}}_{e_1(x)} + \cdots + a_n^\star \underbrace{\frac{x^n}{2^{m_n}}}_{e_n(x)}$$

where $a_i^\star \in \mathbb{Z}$ which minimizes

$$\|f - p\|_2 = \left(\int_{-1}^{1} (f(x) - p(x))^2 \frac{\mathrm{d}x}{\sqrt{1 - x^2}} \right)^{1/2}.$$

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \ldots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = a_0^\star \underbrace{\frac{1}{2^{m_0}}}_{e_0(x)} + a_1^\star \underbrace{\frac{x}{2^{m_1}}}_{e_1(x)} + \cdots + a_n^\star \underbrace{\frac{x^n}{2^{m_n}}}_{e_n(x)}$$

where $a_i^\star \in \mathbb{Z}$ which minimizes

$$\|f - p\|_2 = \|f - p_{\mathbb{R}_n[x]}(f)\|_2 + \|p_{\mathbb{R}_n[x]}(f) - p\|_2$$

where $p_{\mathbb{R}_n[x]}(f)$: orthogonal projection of f onto $\mathbb{R}_n[x]$.

Let $p_{\mathbb{R}_n[x]}(f) = \sum_{i=0}^{n} f_i e_i(x)$.

Absolute Error Problem

Given $f : [-1, 1] \mapsto \mathbb{R}$ and $m_0, \dots, m_n \in \mathbb{Z}$, we search for (one of the) best(s) polynomial of the form

$$p^\star = a_0^\star \underbrace{\frac{1}{2^{m_0}}}_{e_0(x)} + a_1^\star \underbrace{\frac{x}{2^{m_1}}}_{e_1(x)} + \cdots + a_n^\star \underbrace{\frac{x^n}{2^{m_n}}}_{e_n(x)}$$

where $a_i^\star \in \mathbb{Z}$ which minimizes

$$\|p_{\mathbb{R}_n[x]}(f) - p\|_2 = \left(\int_{-1}^{1} (p_{\mathbb{R}_n[x]}(f) - p(x))^2 \frac{\mathrm{d}x}{\sqrt{1-x^2}} \right)^{1/2}.$$

where $p_{\mathbb{R}_n[x]}(f)$: orthogonal projection of f onto $\mathbb{R}_n[x]$.

Let $p_{\mathbb{R}_n[x]}(f) = \sum_{i=0}^{n} f_i e_i(x)$.

Absolute Error Problem

We wish to minimize

$$\|p_{\mathbb{R}_n[x]}(f) - p^\star\|_2 = \|\sum_{i=0}^{n}(f_i - a_i^\star)e_i(x)\|_2$$

where $p_{\mathbb{R}_n[x]}(f) = \sum_{i=0}^{n} f_i e_i(x)$ and $a_0^\star, \ldots, a_n^\star \in \mathbb{Z}$.

Absolute Error Problem

We wish to minimize

$$\|p_{\mathbb{R}_n[x]}(f) - p^\star\|_2 = \|\sum_{i=0}^{n}(f_i - a_i^\star)e_i(x)\|_2$$

where $p_{\mathbb{R}_n[x]}(f) = \sum_{i=0}^{n} f_i e_i(x)$ and $a_0^\star, \ldots, a_n^\star \in \mathbb{Z}$.

Let $G = ((e_i|e_j))_{0 \leqslant i,j \leqslant n}$, let M such that $G = M^t M$ (Cholesky decomposition). Actually,

$$\|p_{\mathbb{R}_n[x]}(f) - p^\star\|_2 = |Ma^\star - v|_2$$

where $a^\star = (a_j^\star)_{0 \leqslant j \leqslant n} \in \mathbb{Z}^{n+1}$ and $v = (M^t)^{-1}(f_j)_{0 \leqslant j \leqslant n}$.

Absolute Error Problem

We wish to minimize

$$\|p_{\mathbb{R}_n[x]}(f) - p^\star\|_2 = \|\sum_{i=0}^{n}(f_i - a_i^\star)e_i(x)\|_2$$

where $p_{\mathbb{R}_n[x]}(f) = \sum_{i=0}^{n} f_i e_i(x)$ and $a_0^\star, \ldots, a_n^\star \in \mathbb{Z}$.

Let $G = ((e_i|e_j))_{0 \leqslant i,j \leqslant n}$, let M such that $G = M^t M$ (Cholesky decomposition). Actually,

$$\|p_{\mathbb{R}_n[x]}(f) - p^\star\|_2 = |Ma^\star - v|_2$$

where $a^\star = (a_j^\star)_{0 \leqslant j \leqslant n} \in \mathbb{Z}^{n+1}$ and $v = (M^t)^{-1}(f_j)_{0 \leqslant j \leqslant n}$.

This is a closest vector problem in a lattice!

It is NP-hard: LLL algorithm gives an approximate solution.

An Example: Arctan Function on [-1,1]

A question from the numerics group of Intel Portland.

Each approximant is of the form

$$x + x^3(\underbrace{p_0}_{d.} + \underbrace{p_1}_{d.}x^2 + \underbrace{p_2}_{d.}x^4 + \cdots + \underbrace{p_{21}}_{d.}x^{42} + \underbrace{p_{22}}_{d.}x^{44})$$

where the p_i are all double precision numbers (d.).

An Example: Arctan Function on [-1,1]

A question from the numerics group of Intel Portland.

Each approximant is of the form

$$x + x^3(\underbrace{p_0}_{d.} + \underbrace{p_1}_{d.} x^2 + \underbrace{p_2}_{d.} x^4 + \cdots + \underbrace{p_{21}}_{d.} x^{42} + \underbrace{p_{22}}_{d.} x^{44})$$

where the p_i are all double precision numbers (d.).

Here, we minimize the relative error

$$\left\| 1 - \frac{x + x^3(\underbrace{p_0}_{d.} + \underbrace{p_1}_{d.} x^2 + \cdots + \underbrace{p_{21}}_{d.} x^{42} + \underbrace{p_{22}}_{d.} x^{44})}{f} \right\|_\infty$$

An Example: Arctan Function on [-1,1]

A question from the numerics group of Intel Portland.

Each approximant is of the form

$$x + x^3(\underbrace{p_0}_{d.} + \underbrace{p_1}_{d.} x^2 + \underbrace{p_2}_{d.} x^4 + \cdots + \underbrace{p_{21}}_{d.} x^{42} + \underbrace{p_{22}}_{d.} x^{44})$$

where the p_i are all double precision numbers (d.).

binary logarithm of the relative error of several approximants

Minimax	-65.41
Rounded minimax	-56.59
Our polynomial	-65.08

An Example: Arctan Function on [-1,1]

A question from the numerics group of Intel Portland.

Each approximant is of the form

$$x + x^3(\underbrace{p_0}_{d.} + \underbrace{p_1}_{d.}x^2 + \underbrace{p_2}_{d.}x^4 + \cdots + \underbrace{p_{21}}_{d.}x^{42} + \underbrace{p_{22}}_{d.}x^{44})$$

where the p_i are all double precision numbers (d.).

binary logarithm of the relative error of several approximants

Minimax	-65.41
Rounded minimax	-56.59
Our polynomial	-65.08

We save 8 bits with our method.

Extension

Very nice work by D. Arzelier, F. Bréhard, T. Hubrecht and M. Joldes who minimize both the approximation and evaluation errors.