Integer points close to a transcendental curve and correctly-rounded evaluation of a function

Nicolas Brisebarre (C.N.R.S.) and Guillaume Hanrot (É.N.S. Lyon)

Effective Aspects in Diophantine Approximation - March 28, 2023

(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
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 $E_{\mathsf{min}} \leqslant E \leqslant E_{\mathsf{max}}$.

Rounding modes

In the IEEE 754 standard, the user defines an *active rounding mode*.

In this talk, we use:

- round to nearest (default). If $x \in \mathbb{R}$, $\mathrm{RN}(x)$: the floating-point number closest to x. In case of a tie, value whose integral significand is even.

Breakpoint: a point where the rounding function changes.

Rounding modes

In the IEEE 754 standard, the user defines an *active rounding mode*.

In this talk, we use:

- round to nearest (default). If $x \in \mathbb{R}$, $\mathrm{RN}(x)$: the floating-point number closest to x. In case of a tie, value whose integral significand is even.

Breakpoint: a point where the rounding function changes.

Here, breakpoint $=$ the middle of two consecutive FP numbers.

Correct rounding

A correctly-rounded operation whose entries are FP numbers must return what we would get by infinitely precise operation followed by rounding.

Correct rounding

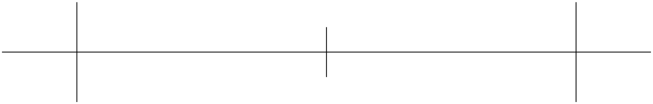
A correctly-rounded operation whose entries are FP numbers must return what we would get by infinitely precise operation followed by rounding.

IEEE-754 (1985): Correct rounding for $+$, $-$, $\times$, $\div$, $\sqrt{}$ and some conversions.

IEEE-754 (2008): suggests correct rounding for some elementary functions ($\sqrt[n]{}$, $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$...).

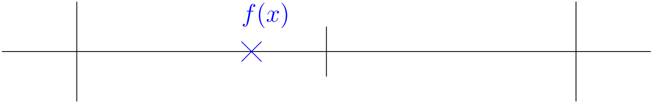
The Table Maker's Dilemma

$$x \in [1,2),\ x = \frac{j}{2^{p-1}},\ j \in \mathbb{Z},\ 2^{p-1} \leqslant j \leqslant 2^p - 1,\ f(x) \in [1,2)$$



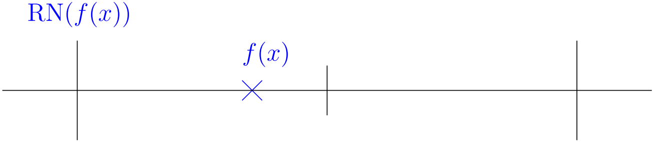
The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



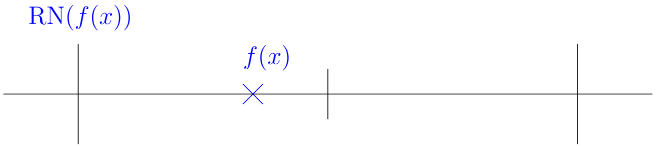
The Table Maker's Dilemma

$$x \in [1,2),\ x = \frac{j}{2^{p-1}},\ j \in \mathbb{Z},\ 2^{p-1} \leqslant j \leqslant 2^p - 1,\ f(x) \in [1,2)$$



The Table Maker's Dilemma

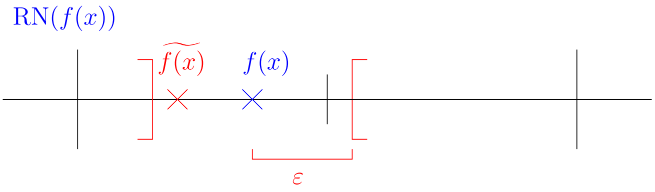
$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$

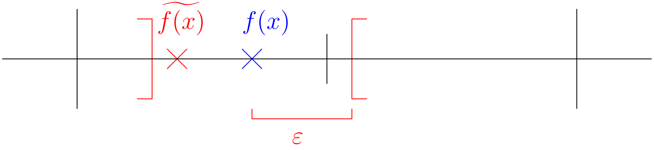


Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$

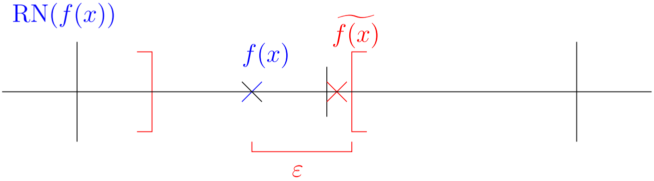
$\mathrm{RN}(f(x)) = \mathrm{RN}(\widetilde{f(x)})$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

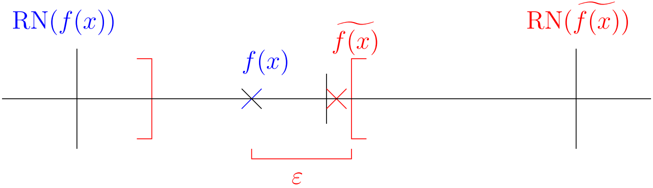
$$x \in [1,2),\ x = \frac{j}{2^{p-1}},\ j \in \mathbb{Z},\ 2^{p-1} \leqslant j \leqslant 2^p - 1,\ f(x) \in [1,2)$$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

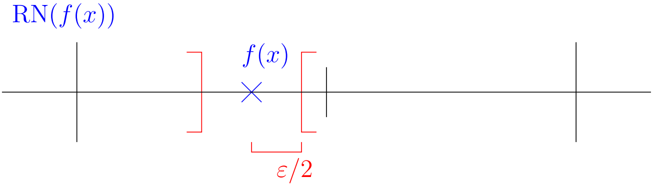
$x \in [1, 2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1, 2)$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

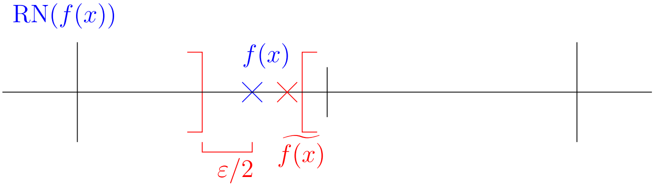
$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

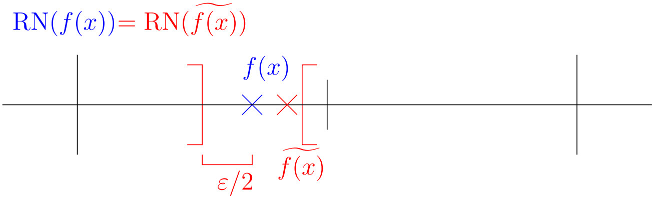
$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

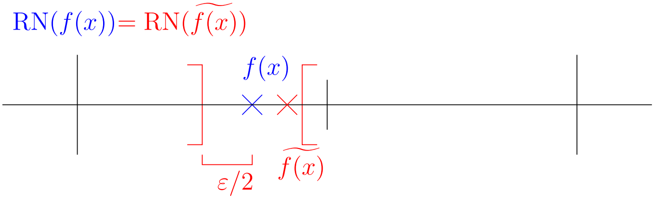
$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



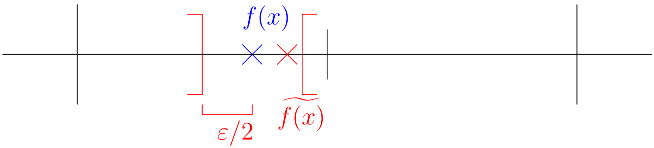
Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

Potential issues:

The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$

$\mathrm{RN}(f(x)) = \mathrm{RN}(\widetilde{f(x)})$



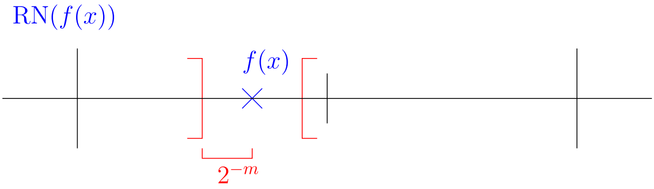
Given $\varepsilon > 0$, the computed value $\widetilde{f(x)}$ satisfies $|f(x) - \widetilde{f(x)}| < \varepsilon$.

Potential issues:

- What if $f(x)$ is a breakpoint?
- What about the number of subdivisions?
- ε should be uniform! And as large as possible!

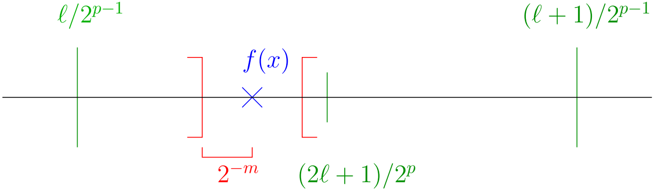
The Table Maker's Dilemma

$$x \in [1, 2),\ x = \frac{j}{2^{p-1}},\ j \in \mathbb{Z},\ 2^{p-1} \leqslant j \leqslant 2^p - 1,\ f(x) \in [1, 2)$$



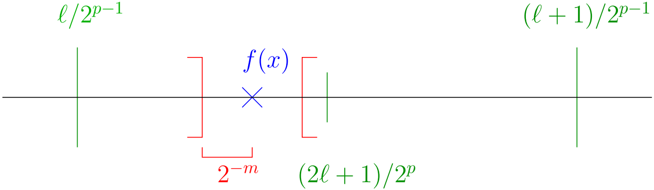
The Table Maker's Dilemma

$$x \in [1,2),\ x = \frac{j}{2^{p-1}},\ j \in \mathbb{Z},\ 2^{p-1} \leqslant j \leqslant 2^p - 1,\ f(x) \in [1,2)$$



The Table Maker's Dilemma

$x \in [1, 2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1, 2)$

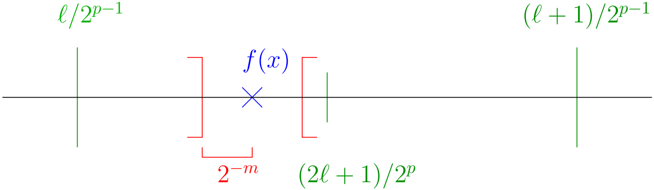


We want to find $m \in \mathbb{N}$ s.t.

- either there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $f(x) = (2\ell+1)/2^p$,

The Table Maker's Dilemma

$x \in [1,2)$, $x = \dfrac{j}{2^{p-1}}$, $j \in \mathbb{Z}$, $2^{p-1} \leqslant j \leqslant 2^p - 1$, $f(x) \in [1,2)$



We want to find $m \in \mathbb{N}$, as small as possible, s.t. for all FP x:

- either there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $f(x) = (2\ell+1)/2^p$,
- or

$$\text{for all } k \in [\![2^{p-1}, 2^p - 1]\!], \left| f(x) - \frac{2k+1}{2^p} \right| \geqslant 2^{-m}.$$

The Table Maker's Dilemma: Diophantine Problems

Assume, w.l.o.g., that x and $f(x) \in [1,2)$.

Q. (TMD) We want to determine $m \in \mathbb{N}$, as small as possible, s.t. for all $j \in [\![2^{p-1}, 2^p - 1]\!]$,

- either there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $f\left(\frac{j}{2^{p-1}}\right) = \frac{2\ell+1}{2^p}$,
- or

$$\text{for all } 2^{p-1} \leqslant k \leqslant 2^p - 1, \left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2k+1}{2^p} \right| \geqslant \frac{1}{2^m}.$$

The Table Maker's Dilemma: an Example

Consider the function 2^x and the binary64 FP number (base $2, p = 53$)

$$x = \frac{8520761231538509}{2^{62}}$$

We have

$$\begin{aligned} 2^{52+x} &= 4509371038706515.49999999999999999994026\cdots\text{(decimal)} \\ &= \underbrace{1\cdots}_{\text{53 bits}}.0\underbrace{1\cdots\cdots\cdots\cdots 1}_{\text{60 consecutive } 1's}0\cdots\text{(binary)}. \end{aligned}$$

The Table Maker's Dilemma: an Example

Consider the function 2^x and the binary64 FP number (base $2, p = 53$)

$$x = \frac{8520761231538509}{2^{62}}$$

We have

$$\begin{aligned} 2^{52+x} &= 4509371038706515.49999999999999999994026\cdots(\text{decimal}) \\ &= \underbrace{1\cdots}_{53 \text{ bits}}.0\underbrace{1\cdots\cdots\cdots\cdots 1}_{60 \text{ consecutive } 1's}0\cdots(\text{binary}). \end{aligned}$$

Hardest-to-round (HR) case for function 2^x and binary64 FP numbers.
Lefèvre et al.: the value of m is 113 ($\sim 2p, p = 53$) here.

The Table Maker's Dilemma: an Example

Consider the function 2^x and the binary64 FP number (base $2, p = 53$)

$$x = \frac{8520761231538509}{2^{62}}$$

We have

$$\begin{array}{rcl} 2^{52+x} & = & 4509371038706515.49999999999999999994026\cdots(\text{decimal}) \\ & = & \underbrace{1\cdots}_{53 \text{ bits}}.0\underbrace{1\cdots\cdots\cdots\cdots 1}_{60 \text{ consecutive } 1's}0\cdots(\text{binary}). \end{array}$$

Hardest-to-round (HR) case for function 2^x and binary64 FP numbers.
Lefèvre et al.: the value of m is 113 ($\sim 2p, p = 53$) here.

Function f: $\sqrt[n]{\ }$, $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$...

The Table Maker's Dilemma: an Example

Consider the function 2^x and the binary64 FP number (base $2, p = 53$)

$$x = \frac{8520761231538509}{2^{62}}$$

We have

$$\begin{aligned} 2^{52+x} &= 4509371038706515.49999999999999999994026 \cdots \text{(decimal)} \\ &= \underbrace{1 \cdots}_{53 \text{ bits}} .0\underbrace{1 \cdots\cdots\cdots 1}_{60 \text{ consecutive } 1's}0 \cdots \text{(binary)}. \end{aligned}$$

Hardest-to-round (HR) case for function 2^x and binary64 FP numbers.
Lefèvre et al.: the value of m is 113 ($\sim 2p, p = 53$) here.

Function f: $\sqrt[n]{\ }$, $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$...
Heuristically, $m \sim 2p$.

The Table Maker's Dilemma: Diophantine Problems

Assume, w.l.o.g., that x and $f(x) \in [1, 2)$.

Q. (TMD) We want to determine $m \in \mathbb{N}$, as small as possible, s.t. for all $j \in [\![2^{p-1}, 2^p - 1]\!]$,

- either there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $f\left(\frac{j}{2^{p-1}}\right) = \frac{2\ell + 1}{2^p}$,
- or

$$\text{for all } 2^{p-1} \leqslant \ell \leqslant 2^p - 1, \left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| \geqslant 2^{-m}.$$

The Table Maker's Dilemma: First Challenge

A breakpoint is a point where the rounding function changes. In this talk, it is the middle of two consecutive FP numbers.

First challenge:

- Determine the set BP_f of all the FP numbers $x \in [1, 2)$ such that $f(x)$ is a breakpoint.

In other words, determine all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t.

$$f\left(\frac{j}{2^{p-1}}\right) = \frac{2\ell + 1}{2^p}.$$

State of the Art

Transcendental elementary Functions $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$. Hermite-Lindemann's theorem: $\alpha \neq 0$ algebraic $\Rightarrow$ e^{α} transcendental.

State of the Art

Transcendental elementary Functions $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$. Hermite-Lindemann's theorem: $\alpha \neq 0$ algebraic $\Rightarrow e^{\alpha}$ transcendental. Therefore, let x a FP number, $f(x)$ is not a breakpoint, except for straightforward cases (e^0, $\ln(1)$, $\sin(0)$, ...).

State of the Art

Transcendental elementary Functions $\sin$, $\cos$, $\arcsin$, $\arccos$, $\tan$, $\arctan$, $\exp$, $\log$, $\sinh$, $\cosh$. Hermite-Lindemann's theorem: $\alpha \neq 0$ algebraic $\Rightarrow e^\alpha$ transcendental. Therefore, let x a FP number, $f(x)$ is not a breakpoint, except for straightforward cases (e^0, $\ln(1)$, $\sin(0)$, ...).

What about the Euler Gamma function? For $\mathrm{Re}(z) > 0$,

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} \mathrm{d}t.$$

Very little is known:
$\Gamma(1/2), \Gamma(1/3), \Gamma(1/4), \Gamma(1/6), \Gamma(2/3), \Gamma(3/4), \Gamma(5/6)$ are transcendental and there are some partial irrationality results.

Our setting

Let $f : [1, 2) \mapsto [1, 2)$, f is transcendental and analytic in the neighborhood of $[1, 2)$.

Our setting

Let $f : [1, 2) \mapsto [1, 2)$, f is transcendental and analytic in the neighborhood of $[1, 2)$.

Let $g \in \mathcal{C}([a, b])$,

$$\|g\|_{\infty,[a,b]} := \max_{x\in[a,b]} |g(x)|.$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We want to find all $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t.

$$f\left(\frac{i}{2^{p-1}}\right) = \frac{2j+1}{2^p},$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We want to find all $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t.

$$f\left(\frac{i}{2^{p-1}}\right) = \frac{2j+1}{2^p},$$

$$\text{i.e. } 2^p f\left(\frac{i}{2^{p-1}}\right) = 2j+1.$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We want to find all $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t.

$$f\left(\frac{i}{2^{p-1}}\right) = \frac{2j+1}{2^p},$$

$$\text{i.e. } 2^p f\left(\frac{i}{2^{p-1}}\right) = 2j+1.$$

We build a trap! We find a partition of $[1,2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X,Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, \quad k = 1, 2,$$

for all $u \in I_\ell$.

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, \quad k = 1, 2 \text{ for all } u \in I_\ell.$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1,2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X,Y] \setminus \{0\}$ with

$$\left| P_{\ell,k}(2^{p-1}u, 2^p f(u)) \right| < 1, \quad k = 1, 2 \text{ for all } u \in I_\ell.$$

For all ℓ, if there exist $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t. $i/2^{p-1} \in I_\ell$ and

$$f\left(\frac{i}{2^{p-1}}\right) = \frac{2j+1}{2^p},$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1,2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X,Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, \quad k = 1,2 \text{ for all } u \in I_\ell.$$

For all ℓ, if there exist $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t. $i/2^{p-1} \in I_\ell$ and

$$f\Bigg(\underbrace{\frac{i}{2^{p-1}}}_{u}\Bigg) = \underbrace{\frac{2j+1}{2^p}}_{v},$$

then we have, for $k = 1, 2$,

$$P_{\ell,k}(i, 2j+1) \in \mathbb{Z} \text{ and } |P_{\ell,k}(i, 2j+1)| < 1!$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, \quad k = 1, 2 \text{ for all } u \in I_\ell.$$

For all ℓ, if there exist $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t. $i/2^{p-1} \in I_\ell$ and

$$f\Bigg(\underbrace{\frac{i}{2^{p-1}}}_{u}\Bigg) = \underbrace{\frac{2j+1}{2^p}}_{v},$$

then we have, for $k = 1, 2$,

$$P_{\ell,k}(i, 2j+1) \in \mathbb{Z} \text{ and } |P_{\ell,k}(i, 2j+1)| < 1! \Rightarrow P_{\ell,k}(i, 2j+1) = 0.$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, \quad k = 1, 2 \text{ for all } u \in I_\ell.$$

For all ℓ, if there exist $2^{p-1} \leqslant i, j \leqslant 2^p - 1$ s.t. $i/2^{p-1} \in I_\ell$ and

$$f\left(\underbrace{\frac{i}{2^{p-1}}}_{u}\right) = \underbrace{\frac{2j+1}{2^p}}_{v},$$

then we have, for $k = 1, 2$,

$$P_{\ell,k}(i, 2j+1) \in \mathbb{Z} \text{ and } |P_{\ell,k}(i, 2j+1)| < 1! \Rightarrow P_{\ell,k}(i, 2j+1) = 0.$$

We eliminate (heuristic assumption!) one of the two variables and we get i and j, if they exist (Coppersmith; Boneh & Durfee; Stehlé).

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1,2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X,Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1,2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X,Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

1. Specify the basis that we use for these polynomials.

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left| P_{\ell,k}(2^{p-1}u, 2^p f(u)) \right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

1. Specify the basis that we use for these polynomials.
2. How do we guarantee the smallness of a function, analytic in a neighborhood of an interval $[a, b]$?

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

1. Specify the basis that we use for these polynomials.
2. How do we guarantee the smallness of a function, analytic in a neighborhood of an interval $[a, b]$? Chebyshev interpolation

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

1. Specify the basis that we use for these polynomials.
2. How do we guarantee the smallness of a function, analytic in a neighborhood of an interval $[a, b]$? Chebyshev interpolation
3. How do we compute these polynomials? Lattice basis reduction

Our Approach: Polynomial Interpolation and Lattice Basis Reduction

We build a trap! We find a partition of $[1, 2) = \cup_\ell I_\ell$ such that, for all ℓ, we can compute $P_{\ell,1}, P_{\ell,2} \in \mathbb{Z}[X, Y] \setminus \{0\}$ with

$$\left|P_{\ell,k}(2^{p-1}u, 2^p f(u))\right| < 1, k = 1, 2 \text{ for all } u \in I_\ell.$$

1. Specify the basis that we use for these polynomials.
2. How do we guarantee the smallness of a function, analytic in a neighborhood of an interval $[a, b]$? Chebyshev interpolation
3. How do we compute these polynomials? Lattice basis reduction

Actually, the lattice reduction step makes it possible to refine the choice of the basis.

Basis in Use

Let $d \in \mathbb{N}$, if $X = 2^{p-1}x$ and $Y = 2^p f(x)$ the elements of the basis that we use are:

$$\begin{array}{cccccc}
1, & & & & & \\
X, & Y, & & & & \\
X^2, & XY, & Y^2, & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
X^{d-1}, & X^{d-2}Y, & X^{d-3}Y^2, & \cdots & Y^{d-1}, & \\
X^d, & X^{d-1}Y, & X^{d-2}Y^2, & \cdots & XY^{d-1}, & Y^d,
\end{array}$$

i.e., the basis of use is $\left((2^{p-1}x)^k (2^p f(x))^\ell\right)_{\substack{0 \leqslant \ell \leqslant d \\ 0 \leqslant k \leqslant d-\ell}}$.

Dimension $N = (d+1)(d+2)/2$.

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$*, the Chebyshev nodes of the first kind of order* n *are the points* $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$*.*

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $p_n \in \mathbb{R}_n[X]$ that interpolates $\varphi \in \mathcal{C}([-1,1])$ at the $(\mu_k)_{k=0,..,n}$.

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $p_n \in \mathbb{R}_n[X]$ that interpolates $\varphi \in \mathcal{C}([-1,1])$ at the $(\mu_k)_{k=0,..,n}$.

The polynomial p_n is a quasi-optimal uniform approximation to φ:

$$\|\varphi - p_n\|_{\infty,[-1,1]} \leqslant 2\left(\frac{1}{\pi}\log(n+1)+1\right)\min_{Q\in\mathbb{R}_n[x]}\|\varphi - Q\|_{\infty,[-1,1]}.$$

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $p_n \in \mathbb{R}_n[X]$ that interpolates $\varphi \in \mathcal{C}([-1,1])$ at the $(\mu_k)_{k=0,..,n}$.

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $p_n \in \mathbb{R}_n[X]$ that interpolates $\varphi \in \mathcal{C}([-1,1])$ at the $(\mu_k)_{k=0,..,n}$.

The polynomial p_n is a quasi-optimal uniform approximation to φ.

$$\underbrace{\|\varphi\|_{\infty,[-1,1]}}_{\text{small}} \leqslant \|p_n\|_{\infty,[-1,1]} + \|\varphi - p_n\|_{\infty,[-1,1]}$$

Ensuring the Smallness of a Function: Interpolation at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $p_n \in \mathbb{R}_n[X]$ that interpolates $\varphi \in \mathcal{C}([-1,1])$ at the $(\mu_k)_{k=0,..,n}$.

The polynomial p_n is a quasi-optimal uniform approximation to φ.

$$\underbrace{\|\varphi\|_{\infty,[-1,1]}}_{\text{small}} \leqslant \underbrace{\|p_n\|_{\infty,[-1,1]}}_{\text{small}} + \underbrace{\|\varphi - p_n\|_{\infty,[-1,1]}}_{\text{small}}$$

Bounding the Interpolation Polynomial at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $P \in \mathbb{R}_n[X]$, we have

$$\max_{x \in [-1,1]} |P(x)| \leqslant \left(\frac{2}{\pi} \log(n+1) + 1\right) \max_{k=0,\ldots,n} |P(\mu_k)|.$$

Bounding the Interpolation Polynomial at Chebyshev Nodes

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \dots, n$.

Let $\varphi \in \mathcal{C}([-1,1])$, let $p_n \in \mathbb{R}_n[X]$ that interpolates φ at the $(\mu_k)_{k=0,..,n}$, we have

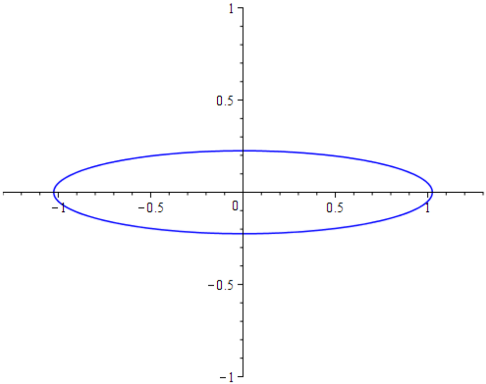
$$\begin{aligned}\|p_n\|_\infty &\leqslant \left(\frac{2}{\pi}\log(n+1)+1\right)\max_{k=0,\dots,n}|p_n(\mu_k)| \\ &= \left(\frac{2}{\pi}\log(n+1)+1\right)\max_{k=0,\dots,n}|\varphi(\mu_k)|.\end{aligned}$$

Bounding the Remainder - Bernstein Ellipse

Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.

Bounding the Remainder - Bernstein Ellipse

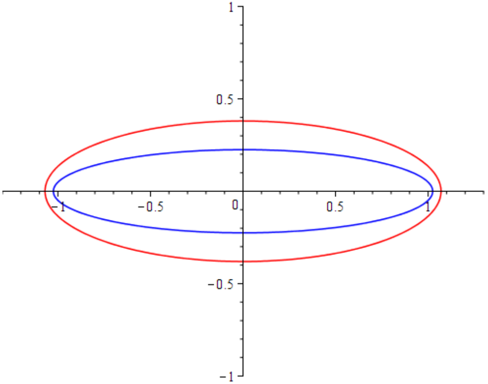
Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.



Bernstein ellipses for $\rho = 1.05$,

Bounding the Remainder - Bernstein Ellipse

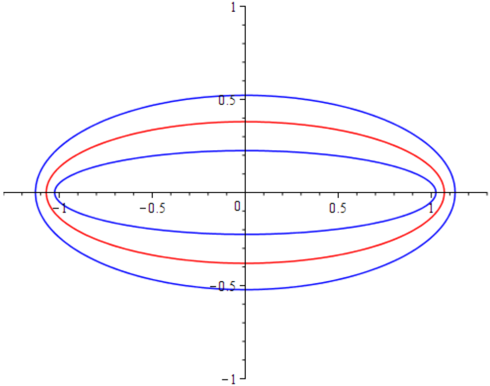
Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.



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

Bounding the Remainder - Bernstein Ellipse

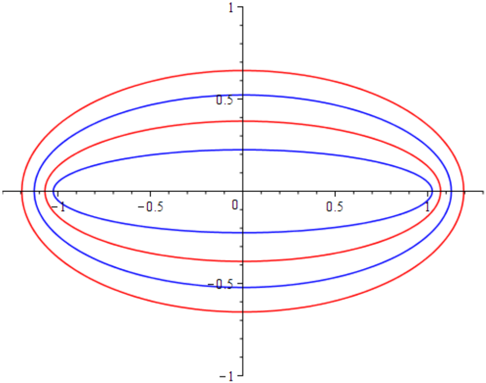
Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.



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

Bounding the Remainder - Bernstein Ellipse

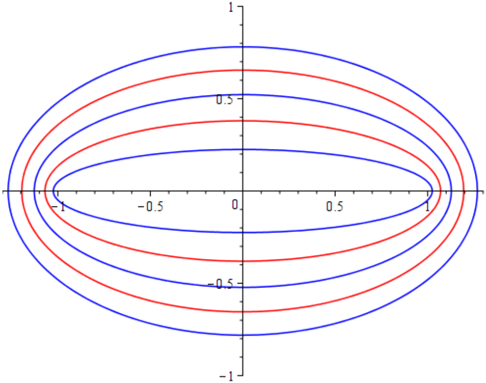
Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.



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

Bounding the Remainder - Bernstein Ellipse

Let $\rho > 1$, let $\mathcal{E}_\rho := \left\{ \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2}, \theta \in [0, 2\pi] \right\}$.



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

Bounding the Remainder

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Bounding the Remainder

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $\rho > 1$, let φ be a function analytic in a neighborhood of $\overline{\mathcal{E}_\rho}$. Let $p_n \in \mathbb{R}_n[X]$ that interpolates φ at the $(\mu_k)_{k=0,..,n}$, we have

$$\|\varphi - p_n\|_{\infty,[-1,1]} \leqslant \frac{4M_\rho(\varphi)}{\rho^n(\rho - 1)}.$$

where $M_\rho(\varphi) = \max_{z \in \mathcal{E}_\rho} |\varphi(z)|$.

Interpolation at Chebyshev Nodes and Uniform Approximation

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Interpolation at Chebyshev Nodes and Uniform Approximation

Definition

Let $n \in \mathbb{N}$, the Chebyshev nodes of the first kind of order n are the points $\mu_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), k = 0, \ldots, n$.

Let $\rho > 1$, let φ be a function analytic in a neighborhood of $\overline{\mathcal{E}_\rho}$. Let $p_n \in \mathbb{R}_n[X]$ that interpolates φ at the $(\mu_k)_{k=0,..,n}$, we have

$$\begin{aligned}\|\varphi\|_\infty &\leqslant \|p_n\|_\infty + \|\varphi - p_n\|_\infty \\ &\leqslant \left(\frac{2}{\pi}\log(n+1)+1\right)\max_{k=0,\ldots,n}|\varphi(\mu_k)| + \frac{4M_\rho(\varphi)}{\rho^n(\rho-1)}.\end{aligned}$$

where $M_\rho(\varphi) = \max_{z\in\mathcal{E}_\rho}|\varphi(z)|$.

Interpolation at Chebyshev Nodes and Uniform Approximation: The case of $[a, b]$

Let $I = [a, b]$, one defines

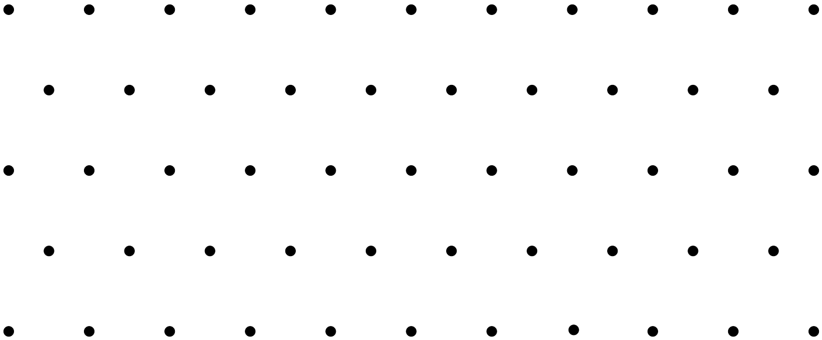
- scaled Chebyshev nodes of the first kind of order n:
  $\mu_{k,[a,b]} = \frac{b-a}{2} \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right) + \frac{a+b}{2}, k = 0, \ldots, n$,

Interpolation at Chebyshev Nodes and Uniform Approximation: The case of $[a, b]$

Let $I = [a, b]$, one defines

- scaled Chebyshev nodes of the first kind of order n:
  $\mu_{k,[a,b]} = \frac{b-a}{2} \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right) + \frac{a+b}{2}, k = 0, \ldots, n$,
- a scaled Bernstein ellipse
  $\mathcal{E}_{\rho,a,b} = \left\{ \frac{b-a}{2} \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2} + \frac{a+b}{2}, \theta \in [0, 2\pi] \right\}$.

Lattice Basis Reduction



An Approach based on Lattice Basis Reduction

Definition

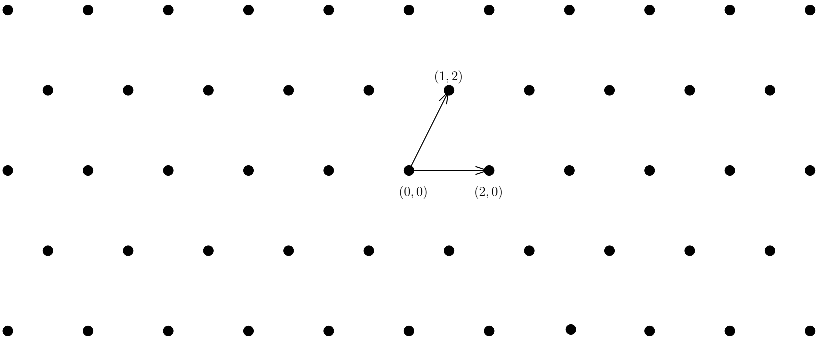
Let L be a nonempty subset of $\mathbb{R}^d$, L is a lattice iff there exists a set of vectors $b_1, \ldots, b_k$ $\mathbb{R}$-linearly independent such that

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

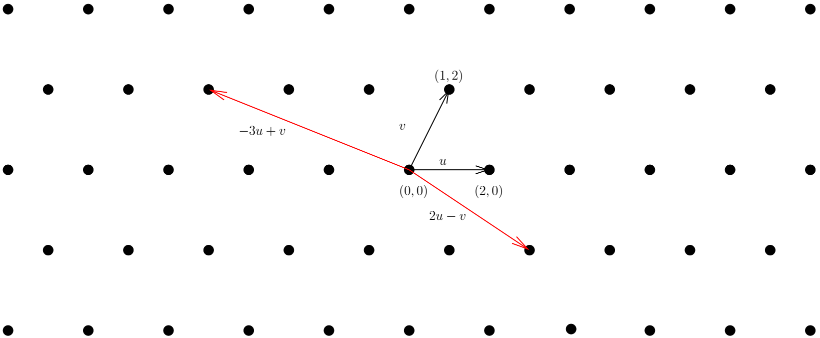
$(b_1, \ldots, b_k)$ is a basis of the lattice L.

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

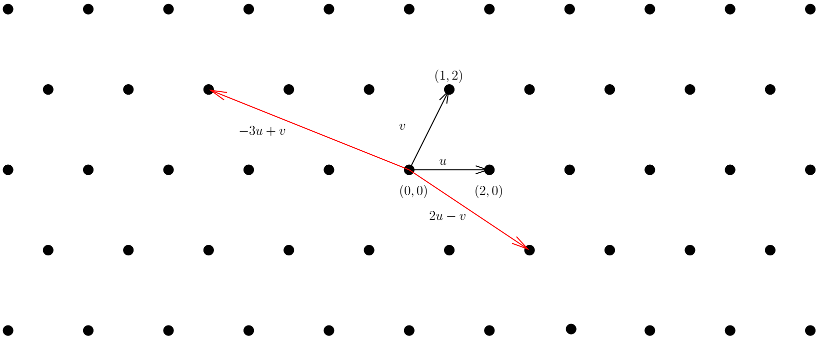
Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$



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

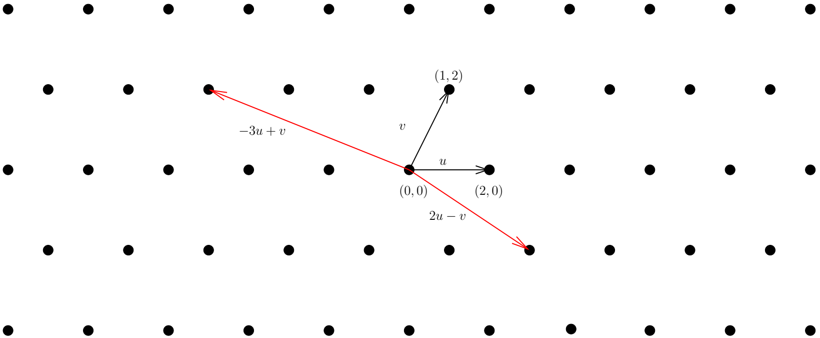


Example: The Lattice $\mathbb{Z}(2,0) \oplus \mathbb{Z}(1,2)$



SVP (Shortest Vector Problem)

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



SVP (Shortest Vector Problem) is NP-hard.

Lenstra-Lenstra-Lovász Algorithm

SVP (Shortest Vector Problem) is 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.

Lenstra-Lenstra-Lovász Algorithm

Theorem

Let L a lattice of dimension k.
LLL provides a basis $(b_1, \ldots, b_k)$ made of "pretty" short vectors. We have $||b_1|| \leqslant 2^{(k-1)/2}\lambda_1(L)$ where $\lambda_1(L)$ denotes the norm of a shortest nonzero vector of L.
It terminates in at most $O(k^6 \ln^3 B)$ operations with $B \geqslant ||b_i||^2$ for all i.

Lenstra-Lenstra-Lovász Algorithm

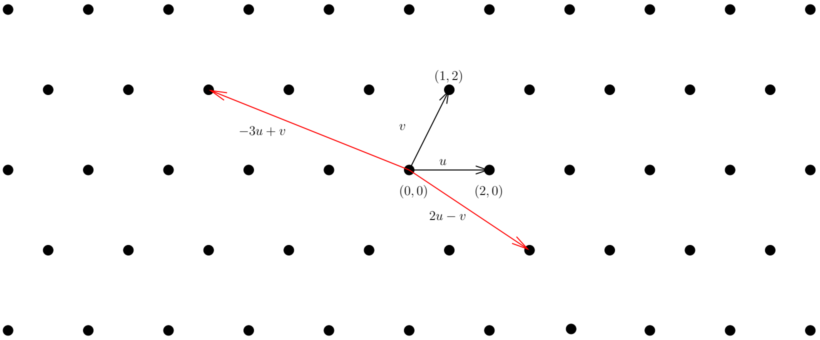
Theorem

Let L a lattice of dimension k.
LLL provides a basis $(b_1, \ldots, b_k)$ made of "pretty" short vectors. We have $||b_1|| \leqslant 2^{(k-1)/2} \lambda_1(L)$ where $\lambda_1(L)$ denotes the norm of a shortest nonzero vector of L.
It terminates in at most $O(k^6 \ln^3 B)$ operations with $B \geqslant ||b_i||^2$ for all i.

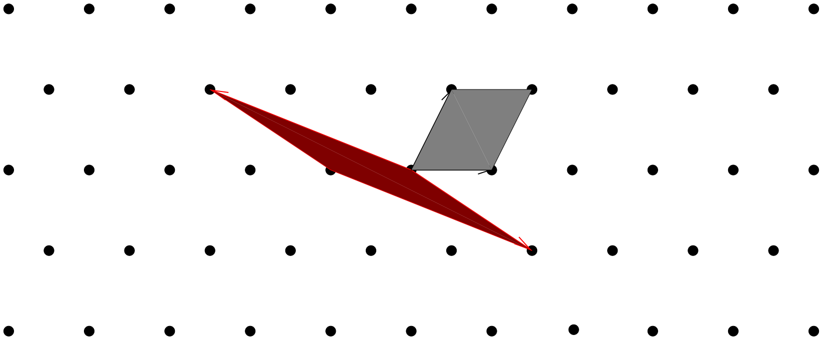
Let $(b_1, \ldots, b_k)$ be an LLL-reduced basis L then

$$||b_1|| \leqslant 2^{k/2} (\text{vol } L)^{1/k} \quad \text{and} \quad ||b_2|| \leqslant 2^{k/2} (\text{vol } L)^{\frac{1}{k-1}}.$$

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



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



Lenstra-Lenstra-Lovász Algorithm

Theorem

Let L a lattice of dimension k.
LLL provides a basis $(b_1, \ldots, b_k)$ made of "pretty" short vectors. We have $||b_1|| \leqslant 2^{(k-1)/2}\lambda_1(L)$ where $\lambda_1(L)$ denotes the norm of a shortest nonzero vector of L.
It terminates in at most $O(k^6 \ln^3 B)$ operations with $B \geqslant ||b_i||^2$ for all i.

Let $(b_1, \ldots, b_k)$ be an LLL-reduced basis L then

$$||b_1|| \leqslant 2^{k/2}(\mathrm{vol}\ L)^{1/k} \quad \text{and} \quad ||b_2|| \leqslant 2^{k/2}(\mathrm{vol}\ L)^{\frac{1}{k-1}}.$$

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P_1 = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v$ and $P_2 = \sum_{0 \leqslant u+v \leqslant d} \beta_{u,v} X^u Y^v \in \mathbb{Z}[X, Y]$. We want to have

$$\left| P_k(2^{p-1}x, 2^p f(x)) \right| < 1, \qquad k = 1, 2,$$

for all $x \in I = [a, b]$.

How do we compute P_1 and P_2? The Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. Let $(x_j)_{0 \leqslant j \leqslant N-1}$ denote Chebyshev nodes for the interval $I = [a, b]$.
We introduce, for $0 \leqslant k \leqslant d, 0 \leqslant \ell \leqslant d-k$,

$$f_{k,\ell}(x) = (2^{p-1}x)^{\ell}(2^p f(x))^k \text{ and } r_{k,\ell} = \frac{4M_\rho(f_{k,\ell})}{\rho^{N-1}(\rho-1)}.$$

How do we compute P_1 and P_2? The Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. Let $(x_j)_{0 \leqslant j \leqslant N-1}$ denote Chebyshev nodes for the interval $I = [a, b]$.
We introduce, for $0 \leqslant k \leqslant d, 0 \leqslant \ell \leqslant d-k$,

$$f_{k,\ell}(x) = (2^{p-1}x)^{\ell}(2^p f(x))^k \text{ and } r_{k,\ell} = \frac{4M_\rho(f_{k,\ell})}{\rho^{N-1}(\rho-1)}.$$

Our lattice: $\mathcal{L}$ generated by the rows of

$$\begin{pmatrix}
f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\
f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\
\vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\
f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0}
\end{pmatrix}$$

How do we compute P_1 and P_2? The Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. Let $(x_j)_{0 \leqslant j \leqslant N-1}$ denote Chebyshev nodes for the interval $I = [a, b]$.
We introduce, for $0 \leqslant k \leqslant d, 0 \leqslant \ell \leqslant d-k$,

$$f_{k,\ell}(x) = (2^{p-1}x)^\ell (2^p f(x))^k \text{ and } r_{k,\ell} = \frac{4M_\rho(f_{k,\ell})}{\rho^{N-1}(\rho - 1)}.$$

Our lattice: $\mathcal{L}$ generated by the rows of

$$\begin{array}{c} f_{0,0} \\ f_{0,1} \\ \vdots \\ \\ f_{d-1,1} \\ f_{d,0} \end{array} \begin{pmatrix} f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\ f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\ f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0} \end{pmatrix}$$

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$
\begin{array}{c}
\alpha_{0,0} \\ +\alpha_{0,1} \\ \vdots \\ \\ +\alpha_{d-1,1} \\ +\alpha_{d,0}
\end{array}
\begin{pmatrix}
f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\
f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\
\vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\
f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0}
\end{pmatrix},
$$

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$\begin{array}{c} \alpha_{0,0} \\ +\alpha_{0,1} \\ \vdots \\ +\alpha_{d-1,1} \\ +\alpha_{d,0} \end{array} \begin{pmatrix} f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\ f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\ f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0} \end{pmatrix},$$

i.e.,

$$\Bigg(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \\ \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Bigg).$$

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \Bigg(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \\ \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Bigg).$$

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X, Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \Bigg(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Bigg).$$

Let p_n the interpolation polynomial of g at the Chebyshev nodes and $r_n = \|g - p_n\|_\infty$.
If the vector V is small, then

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \Big(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Big).$$

Let p_n the interpolation polynomial of g at the Chebyshev nodes and $r_n = \|g - p_n\|_\infty$.
If the vector V is small, then

- p_n is small,

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \left(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \right.$$
$$\left. \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \right).$$

Let p_n the interpolation polynomial of g at the Chebyshev nodes and $r_n = \|g - p_n\|_\infty$.
If the vector V is small, then

- p_n is small,
- $r_n \leqslant \sum_{0 \leqslant u+v \leqslant d} |\alpha_{u,v}| r_{u,v}$ is small.

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \left(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \right. \\ \left. \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \right).$$

Let p_n the interpolation polynomial of g at the Chebyshev nodes and $r_n = \|g - p_n\|_\infty$.
If the vector V is small, then

- p_n is small,
- $r_n \leqslant \sum_{0 \leqslant u+v \leqslant d} |\alpha_{u,v}| r_{u,v}$ is small.

Hence, the function g is "small" !

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \Bigg(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Bigg).$$

If the vector V is small, the function g is "small" !

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \left(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \right.$$
$$\left. \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \right).$$

If the vector V is small, the function g is "small" !

LLL gives us two such short vectors

How do we compute P_1 and P_2?

Let $d \in \mathbb{N}$, $P = \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} X^u Y^v \in \mathbb{Z}[X,Y]$. The function $g(x) = P(2^{p-1}x, 2^p f(x))$ corresponds to the vector

$$V = \Big(\sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_0), \cdots, \sum_{0 \leqslant u+v \leqslant d} \alpha_{u,v} f_{u,v}(x_{N-1}), \alpha_{0,0} r_{0,0}, \cdots, \alpha_{d,0} r_{d,0} \Big).$$

If the vector V is small, the function g is “small” !

LLL gives us two such short vectors, as long as the volume of the lattice is small.

How do we compute P_1 and P_2? Volume of the Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. Let $(x_j)_{0 \leqslant j \leqslant N-1}$ denote Chebyshev nodes for the interval $I = [a, b]$.
We introduce, for $0 \leqslant k \leqslant d, 0 \leqslant \ell \leqslant d-k$,

$$f_{k,\ell}(x) = (2^{p-1}x)^{\ell}(2^p f(x))^k \text{ and } r_{k,\ell} = \frac{4M_\rho(f_{k,\ell})}{\rho^{N-1}(\rho-1)}.$$

Our lattice: $\mathcal{L}$ generated by the rows of

$$\begin{array}{c} f_{0,0} \\ f_{0,1} \\ \vdots \\ f_{d-1,1} \\ f_{d,0} \end{array} \begin{pmatrix} f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\ f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\ f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0} \end{pmatrix}$$

How do we compute P_1 and P_2? Volume of the Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. Let $(x_j)_{0 \leqslant j \leqslant N-1}$ denote Chebyshev nodes for the interval $I = [a, b]$.
We introduce, for $0 \leqslant k \leqslant d, 0 \leqslant \ell \leqslant d-k$,

$$f_{k,\ell}(x) = (2^{p-1}x)^\ell (2^p f(x))^k \text{ and } r_{k,\ell} = \frac{4M_\rho(f_{k,\ell})}{\rho^{N-1}(\rho-1)}.$$

Our lattice: $\mathcal{L}$ generated by the rows of

$$\begin{array}{c} f_{0,0} \\ f_{0,1} \\ \vdots \\ \\ f_{d-1,1} \\ f_{d,0} \end{array} \begin{pmatrix} f_{0,0}(x_0) & \cdots & f_{0,0}(x_{N-1}) & r_{0,0} & 0 & \cdots & \cdots & 0 \\ f_{0,1}(x_0) & \cdots & f_{0,1}(x_{N-1}) & 0 & r_{0,1} & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ & & & \vdots & & & & \\ f_{d-1,1}(x_0) & \cdots & f_{d-1,1}(x_{N-1}) & \vdots & \cdots & 0 & r_{d-1,1} & 0 \\ f_{d,0}(x_0) & \cdots & f_{d,0}(x_{N-1}) & 0 & \cdots & \cdots & 0 & r_{d,0} \end{pmatrix}$$

Its volume $\mathrm{vol}(L)$: determinant of the matrix.

How do we compute P_1 and P_2? Volume of the Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. We have, for $\rho > 1$,

$$\mathrm{vol}(\mathcal{L})^{1/N} \leqslant O(N) \frac{2^{2pd/3}}{\rho^{(N-1)/2}} \left| \frac{b-a}{2}\rho + \frac{b+a}{2} \right|^{d/3} M_{\rho,a,b}(f)^{d/3}$$

where $M_{\rho,a,b}(f) = \max_{z \in \mathcal{E}_{\rho,a,b}} |f(z)|$ and
$\mathcal{E}_{\rho,a,b} = \left\{ \frac{b-a}{2} \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2} + \frac{a+b}{2}, \theta \in [0, 2\pi] \right\}$.

How do we compute P_1 and P_2? Volume of the Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. We have, for $\rho > 1$,

$$\mathrm{vol}(\mathcal{L})^{1/N} \leqslant O(N) \frac{2^{2pd/3}}{\rho^{(N-1)/2}} \left| \frac{b-a}{2}\rho + \frac{b+a}{2} \right|^{d/3} M_{\rho,a,b}(f)^{d/3}$$

where $M_{\rho,a,b}(f) = \max_{z \in \mathcal{E}_{\rho,a,b}} |f(z)|$ and
$\mathcal{E}_{\rho,a,b} = \left\{ \frac{b-a}{2} \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2} + \frac{a+b}{2}, \theta \in [0, 2\pi] \right\}$.

Plug $\rho = 2/(b-a)$: For Euler's Gamma, $d = O(p)$ is enough to tackle the whole $[a, b]$.

How do we compute P_1 and P_2? Volume of the Lattice

Let $d \in \mathbb{N} \setminus \{0\}$ and $N = (d+1)(d+2)/2$. We have, for $\rho > 1$,

$$\mathrm{vol}(\mathcal{L})^{1/N} \leqslant O(N) \frac{2^{2pd/3}}{\rho^{(N-1)/2}} \left| \frac{b-a}{2}\rho + \frac{b+a}{2} \right|^{d/3} M_{\rho,a,b}(f)^{d/3}$$

where $M_{\rho,a,b}(f) = \max_{z \in \mathcal{E}_{\rho,a,b}} |f(z)|$ and
$\mathcal{E}_{\rho,a,b} = \left\{ \frac{b-a}{2} \frac{\rho e^{i\theta} + \rho^{-1} e^{-i\theta}}{2} + \frac{a+b}{2}, \theta \in [0, 2\pi] \right\}$.

Plug $\rho = 2/(b-a)$: For Euler's Gamma, $d = O(p)$ is enough to tackle the whole $[a,b]$.

If $[a,b] = [1,2]$, 40 CPU minutes for $p = 53$ and 46 CPU days for $p = 113$.

Computations

For Euler's Gamma, $d = O(p)$ is enough to tackle the whole $[a, b]$ ($\rho = 2/(b - a)$).

If $[a, b] = [1, 2]$, less than 40 CPU minutes for $p = 53$ and 46 CPU days for $p = 113$.

---

[1] https://www.sagemath.org/
[2] http://arblib.org/
[3] https://github.com/fplll/fplll

Computations

For Euler's Gamma, $d = O(p)$ is enough to tackle the whole $[a, b]$ ($\rho = 2/(b - a)$).

If $[a, b] = [1, 2]$, less than 40 CPU minutes for $p = 53$ and 46 CPU days for $p = 113$.

Our experiments were done in Sagemath[1] and heavily use the Arb[2] and fplll[3] libraries

[1] https://www.sagemath.org/
[2] http://arblib.org/
[3] https://github.com/fplll/fplll

The Table Maker's Dilemma

A breakpoint is a point where the rounding function changes. In this talk, it is the middle of two consecutive FP numbers.

Two-step challenge:

- Determine the set BP_f of all the FP numbers x such that $f(x)$ is a breakpoint;

The Table Maker's Dilemma

A breakpoint is a point where the rounding function changes. In this talk, it is the middle of two consecutive FP numbers.

Two-step challenge:

- Determine the set BP_f of all the FP numbers x such that $f(x)$ is a breakpoint;
- Find $m \in \mathbb{N}$, as small as possible, such that for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $j/2^{p-1} \notin BP_f$ and

$$\left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| \geqslant 2^{-m}.$$

The Table Maker's Dilemma

A breakpoint is a point where the rounding function changes. In this talk, it is the middle of two consecutive FP numbers.

Two-step challenge:

- Determine the set BP_f of all the FP numbers x such that $f(x)$ is a breakpoint;
- Find $m \in \mathbb{N}$, as small as possible, such that for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $j/2^{p-1} \notin BP_f$ and

$$\left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| \geqslant 2^{-m}.$$

  Holy Grail: $m \sim 2p$.

The Table Maker's Dilemma

A breakpoint is a point where the rounding function changes. In this talk, it is the middle of two consecutive FP numbers.

Two-step challenge:

- Determine the set BP_f of all the FP numbers x such that $f(x)$ is a breakpoint;
- Find $m \in \mathbb{N}$, as small as possible, such that for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. $j/2^{p-1} \notin BP_f$ and

$$\left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell+1}{2^p} \right| \geqslant 2^{-m}.$$

Holy Grail: $m \sim 2p$. True for $p = 53$ (V. Lefèvre et al).

Our results

Thanks to an extension of the presented ideas, we obtain for instance, for $p = 113$, for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ and

$$\left|\exp\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell+1}{2^p}\right| \geqslant \frac{1}{2^{12p}}$$

in less than 9 CPU days.

Our results

Thanks to an extension of the presented ideas, we obtain for instance, for $p = 113$, for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ and

$$\left| \exp\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| \geqslant \frac{1}{2^{12p}}$$

in less than 9 CPU days.

Not the end of the story, since $12p$ should be replaced with $\sim 2p$.

Our results

Thanks to an extension of the presented ideas, we obtain for instance, for $p = 113$, for all $j, \ell \in [\![2^{p-1}, 2^p - 1]\!]$ and

$$\left|\exp\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell+1}{2^p}\right| \geqslant \frac{1}{2^{12p}}$$

in less than 9 CPU days.

Not the end of the story, since $12p$ should be replaced with $\sim 2p$.

Still, this work should hopefully help paving the way for correctly rounded elementary functions in IEEE binary128/quadruple precision.

Additional material

Some insight (Warning: Hand-waving!)...

Assume there exist $x \in [1, 2)$, $k \in \mathbb{N} \setminus \{0\}$ and $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t.

$$\left| f(x) - \frac{2\ell + 1}{2^p} \right| < \frac{1}{2^{p+k}}.$$

Some insight (Warning: Hand-waving!)...

Assume there exist $x \in [1,2)$, $k \in \mathbb{N} \setminus \{0\}$ and $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t.

$$\left| \left(f(x) - \frac{1}{2^p} \right) - \frac{\ell}{2^{p-1}} \right| < \frac{1}{2^{p+k}}.$$

Some insight (Warning: Hand-waving!)...

Assume there exist $x \in [1,2)$, $k \in \mathbb{N} \setminus \{0\}$ and $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ s.t.

$$\left| \left(f(x) - \frac{1}{2^p} \right) - \frac{\ell}{2^{p-1}} \right| < \frac{1}{2^{p+k}}.$$

The infinitely precise significand y of $f(x)$ has the form:

$$y = y_0.y_1 y_2 \cdots y_{p-1} \underbrace{01111111\cdots 11}_{k \text{ bits}} xxxxx \cdots$$

or

$$y = y_0.y_1 y_2 \cdots y_{p-1} \overbrace{10000000\cdots 00}^{k \text{ bits}} xxxxx \cdots$$

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

$$y = y_0.y_1y_2\cdots y_{p-1}\underbrace{01111111\cdots 11}_{k \text{ bits}}xxxxx\cdots$$

or

$$y = y_0.y_1y_2\cdots y_{p-1}\overbrace{10000000\cdots 00}^{k \text{ bits}}xxxxx\cdots$$

with $k \geqslant 1$.

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

$$y = y_0.y_1 y_2 \cdots y_{p-1}\underbrace{01111111\cdots 11}_{k \text{ bits}}xxxxx\cdots$$

or

$$y = y_0.y_1 y_2 \cdots y_{p-1}\overbrace{10000000\cdots 00}^{k \text{ bits}}xxxxx\cdots$$

with $k \geqslant 1$.

- Assuming that after the k^{th} position the "1" and "0" are equally likely, the "probability" of having $k \geqslant k_0$ is 2^{1-k_0};

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

  $y = y_0.y_1y_2\cdots y_{p-1}\underbrace{01111111\cdots 11}_{k \text{ bits}}xxxxx\cdots$

  or   with $k \geqslant 1$.

  $y = y_0.y_1y_2\cdots y_{p-1}\overbrace{10000000\cdots 00}^{k \text{ bits}}xxxxx\cdots$

- Assuming that after the k^{th} position the "1" and "0" are equally likely, the "probability" of having $k \geqslant k_0$ is 2^{1-k_0};
- if we consider 2^{p-1} input FP numbers, around $2^{p-1} \times 2^{1-k_0} = 2^{p-k_0}$ values for which $k \geqslant k_0$;

Assessing the Heuristic: the Example of $\sin$

Here, $f = \sin$ over $[1, 2)$, $p = 16$.

k	Actual number of occurrences	Expected number of occurrences
1	16397	16384
2	8151	8192
3	4191	4096
4	2043	2048
5	1010	1024
6	463	512
7	255	256

Assessing the Heuristic: the Example of $\sin$

Here, $f = \sin$ over $[1, 2)$, $p = 16$.

k	Actual number of occurrences	Expected number of occurrences
8	131	128
9	62	64
10	35	32
11	16	16
12	7	8
13	6	4
14	0	2
15	1	1

Assessing the Heuristic: the Example of $\sin$

Here, $f = \sin$ over $[1, 2)$, $p = 16$.

k	Actual number of occurrences	Expected number of occurrences
8	131	128
9	62	64
10	35	32
11	16	16
12	7	8
13	6	4
14	0	2
15	1	1

Here, the heuristic seems reasonable.

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

  $$y = y_0.y_1y_2\cdots y_{p-1}\underbrace{01111111\cdots11}_{k \text{ bits}}xxxxx\cdots$$

  or with $k \geqslant 1$.

  $$y = y_0.y_1y_2\cdots y_{p-1}\overbrace{10000000\cdots00}^{k \text{ bits}}xxxxx\cdots$$

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

$$y = y_0.y_1y_2\cdots y_{p-1}\underbrace{01111111\cdots 11}_{k \text{ bits}}xxxxx\cdots$$

or

$$y = y_0.y_1y_2\cdots y_{p-1}\overbrace{10000000\cdots 00}^{k \text{ bits}}xxxxx\cdots$$

with $k \geqslant 1$.

- Assuming that after the k^{th} position the "1" and "0" are equally likely, the "probability" of having $k \geqslant k_0$ is 2^{1-k_0};

Some Insight (Warning: Hand-waving!)...

- the infinitely precise significand y of $f(x)$ has the form:

$$y = y_0.y_1y_2\cdots y_{p-1}\underbrace{01111111\cdots 11}_{k \text{ bits}}xxxxx\cdots$$

or

$$y = y_0.y_1y_2\cdots y_{p-1}\overbrace{10000000\cdots 00}^{k \text{ bits}}xxxxx\cdots$$

with $k \geqslant 1$.

- Assuming that after the k^{th} position the "1" and "0" are equally likely, the "probability" of having $k \geqslant k_0$ is 2^{1-k_0};
- if we consider 2^{p-1} input FP numbers, around $2^{p-1} \times 2^{1-k_0} = 2^{p-k_0}$ values for which $k \geqslant k_0$;

$\rightarrow$ roughly,

$$\text{"}m_{opt} \sim 2p\text{"} \qquad \text{(Q)}.$$

Proving the Heuristic

NB, G. Hanrot and O. Robert (2017)

Let $f : [1, 2) \mapsto [1, 2)$, $f \in \mathcal{C}^2$, let $k \in \mathbb{N}$.

Determine the proportion of $j \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ with

$$\left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| < \frac{1}{2^{p+k}}.$$

Proving the Heuristic

NB, G. Hanrot and O. Robert (2017)

Let $f : [1, 2) \mapsto [1, 2)$, $f \in \mathcal{C}^2$, let $k \in \mathbb{N}$.

Determine the proportion of $j \in [\![2^{p-1}, 2^p - 1]\!]$ s.t. there exists $\ell \in [\![2^{p-1}, 2^p - 1]\!]$ with

$$\left| f\left(\frac{j}{2^{p-1}}\right) - \frac{2\ell + 1}{2^p} \right| < \frac{1}{2^{p+k}}.$$

Proposition

For $\exp$ over $[1, 2)$, if $p \geqslant 24$, the heuristic is valid for $0 \leqslant k < p/3$.