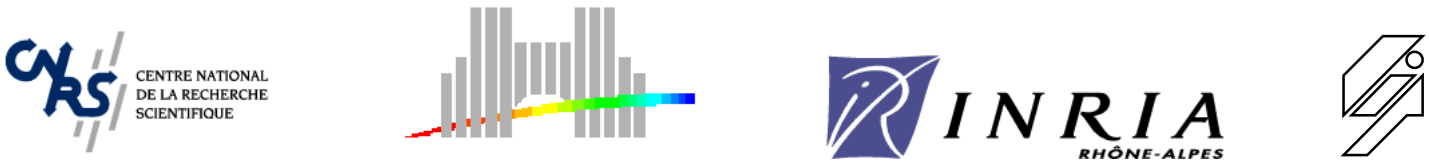
Une méthode pour produire des approximants polynomiaux efficaces en machine

J.-M. Muller et A. Tisserand
Arénaire, LIP, É.N.S. Lyon
N. Brisebarre
LArAl, Univ. St-Étienne et Arénaire, LIP, É.N.S. Lyon

Séminaire SPACES
12 juillet 2004



Évaluation de fonctions

Problème : évaluation d'une fonction φ sur $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$.

On souhaite utiliser essentiellement des additions, soustractions, multiplications (on évite les divisions) $\Rightarrow$ utilisation de polynômes.

Les algos d'évaluation des fonctions élémentaires ($\exp, \ln, \cos, \sin, \arctan, \sqrt{}, \ldots$) utilisent des approximants polynomiaux.

Évaluation des fonctions élémentaires

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

Première étape. Réduction d'argument (Payne & Hanek, Ng, Daumas *et al*) $\Rightarrow$ évaluation d'une fonction φ sur $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$ se ramène à l' évaluation d'une fonction f sur $[a, b]$.

Seconde étape. Approximation polynomiale de f :

- approximation aux moindres carrés ;
- approximation minimax.

Approximation minimax

Rappel. Soit $g : [a, b] \to \mathbb{R}$, $||g||_{[a,b]} = \sup_{a \leq x \leq b} |g(x)|$.

On note $\mathbb{R}_n[X] = \{p \in \mathbb{R}[X]; \deg p \leq n\}$.

Approximation minimax : soient $f : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, on cherche $p \in \mathbb{R}_n[X]$ t.q.

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

Algorithme de Remez donne p.

Problème : approx. minimax non utilisable en machine directement car les coefficients de p non représentables sur un nombre fini de bits.

Notre contexte : les coefficients des polynômes doivent être stockés sur un nombre imposé de bits.

Soit $m = (m_i)_{0 \leq i \leq n}$ une suite finie d'entiers naturels.

Soit $q(x) = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[x]$. Chaque q_i doit être un multiple entier de 2^{-m_i} : $q_i = a_i / 2^{m_i}$ avec $a_i \in \mathbb{Z}$.

Polynômes tronqués

Soit $m = (m_i)_{0 \leq i \leq n}$ une suite finie d'entiers naturels. On note

$$\mathcal{P}_n^m = \{q = q_0 + q_1 x + \cdots + q_n x^n \in \mathbb{R}_n[X];$$

$$q_i \text{ multiple entier de } 2^{-m_i}, \forall i\}.$$

Première idée. Remez $\rightarrow$ $p(x) = p_0 + p_1 x + \cdots + p_n x^n$. Chaque p_i arrondi vers $\hat{p}_i$, le plus proche multiple entier de 2^{-m_i} $\rightarrow$ $\hat{p}(x) = \hat{p}_0 + \hat{p}_1 x + \cdots + \hat{p}_n x^n$.

Problème : $\hat{p}$ pas nécessairement l'approx. minimax de f parmi les polynômes de $\mathcal{P}_n^m$.

Applications

Deux cibles :

- implantations matérielles spécifiques en faible précision (~ 15 bits). Réduction du coût (temps et surface de silicium) en gardant une précision correcte ;
- implantations logicielles en simple ou double précision IEEE. Très grande précision en gardant un coût (temps et mémoire) raisonnable.

Énoncé du problème

Soient $f : [a, b] \to \mathbb{R}, n \in \mathbb{N}, m = (m_i)_{0 \leq i \leq n}$ une suite finie d'entiers naturels, $p(x) = p_0 + p_1 x + \cdots + p_n x^n$ l'approx. minimax de f sur $[a, b]$ (Remez).

$$\mathcal{P}_n^m = \left\{ q(x) = \frac{a_0}{2^{m_0}} + \frac{a_1}{2^{m_1}} x + \cdots + \frac{a_n}{2^{m_n}} x^n ; a_i \in \mathbb{Z}, \forall i \right\}.$$

Chaque p_i arrondi vers $\hat{p}_i$, le plus proche multiple entier de 2^{-m_i} $\longrightarrow$ $\hat{p}(x) = \hat{p}_0 + \hat{p}_1 x + \cdots + \hat{p}_n x^n$.

On pose

$$\varepsilon = ||f - p||_{[a,b]} \text{ et } \hat{\varepsilon} = ||f - \hat{p}||_{[a,b]}.$$

On compare ε et $\hat{\varepsilon}$.

On introduit $K \in [\varepsilon, \hat{\varepsilon}]$. On cherche un polynôme tronqué $p^\star \in \mathcal{P}_n^m$ t.q.

$$||f - p^\star||_{[a,b]} = \min_{q \in \mathcal{P}_n^m} ||f - q||_{[a,b]}$$

et

$$||f - p^\star||_{[a,b]} \leq K.$$

Démarche

On pose $p^\star(x) = p_0^\star + p_1^\star x + \cdots + p_n^\star x^n$.

1. On produit des relations vérifiées par les $p_i^\star \Rightarrow$ nombre fini de polynômes candidats.
2. Si ce nombre est assez petit, on lance une recherche exhaustive : calcul des normes $||f - q||_{[a,b]}$, q parcourant les polynômes candidats.

Première approche : polynômes de Chebyshev

On se place sur $[0, a]$ (fonctionne aussi sur $[-a, a]$).

Définition . *Les polynômes de Chebyshev peuvent être définis par la récurrence*

$$\begin{cases} T_0(x) &= 1 \\ T_1(x) &= x \\ T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x); \end{cases}$$

ou par

$$T_n(x) = \begin{cases} \cos\left(n \cos^{-1} x\right) & (|x| \leq 1) \\ \cosh\left(n \cosh^{-1} x\right) & (x > 1). \end{cases}$$

T. J. Rivlin, *Chebyshev polynomials*.

P. Borwein and T. Erdélyi, *Polynomials and Polynomials Inequalities*.

Proposition . *Soient a, $b \in \mathbb{R}$, $a < b$. Le polynôme unitaire de degré n à coef. dans $\mathbb{R}$ avec la plus petite norme $||\cdot||_{[a,b]}$ est*

$$\frac{(b-a)^n}{2^{2n-1}} T_n\left(\frac{2x-b-a}{b-a}\right).$$

Soient $f : [0, a] \to \mathbb{R}$, $m_0, \ldots, m_n \in \mathbb{N}$, $p(x) = p_0 + p_1 x + \cdots + p_n x^n$ l'approx. minimax de f sur $[0, a]$ (Remez),

$$\mathcal{P}_n^m = \left\{ q(x) = \frac{a_0}{2^{m_0}} + \frac{a_1}{2^{m_1}} x + \cdots + \frac{a_n}{2^{m_n}} x^n ; a_i \in \mathbb{Z}, \forall i \right\}.$$

On détermine des bornes t. q. si les coef. de $q \in \mathcal{P}_n^m$ sont au delà de ces bornes alors

$$||f - q||_{[0,a]} > K \quad \text{i.e.} \quad q \neq p^\star.$$

Idée : introduire p. On a

$$||f - q||_{[0,a]} \geq ||p - q||_{[0,a]} - ||f - p||_{[0,a]}.$$

Si $||p - q||_{[0,a]} > \varepsilon + K$, c'est gagné.

On écrit le i-ème coef. de q sous la forme $p_i + \delta_i$, avec $\delta_i \neq 0$. On a

$$(q-p)(x) = \delta_i x^i + \sum_{\substack{0 \leq j \leq n, \\ j \neq i}} (q_j - p_j) x^j.$$

Donc, $||q-p||_{[0,a]}$ minimum implique

$$\left\| x^i + \frac{1}{\delta_i} \sum_{\substack{0 \leq j \leq n, \\ j \neq i}} (q_j - p_j) x^j \right\|_{[0,a]}$$

minimum.

Nous utilisons $T_n^*(x) = T_n(2x-1)$.

On a $T_n^*(x) = T_{2n}(x^{1/2})$.

Proposition . *Soit $a \in]0, +\infty[$, on définit*

$$\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n = T_n^* \left(\frac{x}{a}\right).$$

Soit $k \in \mathbb{N}$, $0 \leq k \leq n$, le polynôme

$$\frac{1}{\alpha_k} T_n^* \left(\frac{x}{a}\right)$$

a la plus petite norme $||.||_{[0,a]}$ parmi les polynômes de degré au plus n avec un k-ième coef. $= 1$. Cette norme vaut $|1/\alpha_k|$.

On a donc

$$\left\| x^i + \frac{1}{\delta_i} \sum_{\substack{0 \le j \le n, \\ j \neq i}} (q_j - p_j) x^j \right\|_{[0,a]} \geq \frac{1}{|\alpha_i|},$$

où α_i est le i-ème coef. de $T_n^*(x/a)$. D'où,

$$||q - p||_{[0,a]} \geq \frac{|\delta_i|}{|\alpha_i|}.$$

Rappel : si $||q - p||_{[0,a]} > \varepsilon + K$, on a $q \neq p^\star$.

Donc, s'il existe i, $0 \leq i \leq n$, t.q. $|\delta_i| > (\varepsilon + K)|\alpha_i|$ alors

$$||q - p||_{[0,a]} > \varepsilon + K.$$

Rappel : $\delta_i = q_i - p_i$. Donc, le i-ème coef. de $p^\star$ doit appartenir à

$$[p_i - (\varepsilon + K)|\alpha_i|, p_i + (\varepsilon + K)|\alpha_i|].$$

On a posé $\varepsilon = ||f - p||_{[0,a]}$, $p^\star(x) = p_0^\star + p_1^\star x + \cdots + p_n^\star x^n$. On a pour tout i

$$p_i - (\varepsilon + K)|\alpha_i| \leq p_i^\star \leq p_i + (\varepsilon + K)|\alpha_i|.$$

Rappel : $p_i^\star = a_i/2^{m_i}$ avec $a_i \in \mathbb{Z}$. On obtient pour tout i

$$\underbrace{\lceil 2^{m_i}(p_i - (\varepsilon + K)|\alpha_i|)\rceil}_{c_i} \leq 2^{m_i} p_i^\star \leq \underbrace{\lfloor 2^{m_i}(p_i + (\varepsilon + K)|\alpha_i|)\rfloor}_{d_i}.$$

On a donc $d_i - c_i + 1$ valeurs possibles pour l'entier $2^{m_i} p_i^\star$.

On a $A = \prod_{i=0}^{n}(d_i - c_i + 1)$ polynômes candidats. Si A assez petit, recherche exhaustive : calcul des normes $||f - q||_{[0,a]}$, q parcourant les polynômes candidats. Sinon, seconde approche.

Approximation de la fonction $\cos$ sur $[0, \pi/4]$ par un polynôme de degré 3

```
>m := [12,10,6,4]:polstar(cos,Pi/4,3,m);

  "minimax = ", .9998864206 +
(.00469021603 + (-.5303088665 + .06304636099 x) x) x

     "Distance between f and p =", .0001135879209

                    1  3    17  2      5
          "hatp = ", -- x  - -- x  + ---- x + 1
                    16      32       1024


      "Distance between f and hatp =", .0006939707
```

```
>Do you want to continue (y;/n;)? y;
>Enter the value of parameter lambda: 1/2;

degree 0: 4 possible values between 2047/2048 and
          4097/4096
degree 1: 22 possible values between -3/512 and
          15/1024
degree 2: 5 possible values between -9/16 and
          -1/2
degree 3: 1 possible values between 1/16 and
          1/16
440 polynomials need be checked

>Do you want to try to refine the bounds (y;/n;)?n;
```

```
                      1  3   17  2    3       4095
"pstar = ",    -- x  - -- x  + --- x + ----
               16      32      512     4096
```

```
"Distance between f and pstar =", .000244140625

        "Time elapsed (in seconds) =", 1.840
```

Dans cet exemple, on gagne $-\log_2(0.35) \approx 1.5$ bits de précision.

Approximation de la fonction exponentielle sur $[0, \log(1 + 1/2048)]$ par un polynôme de degré 3

```
>Digits:=30:
>m := [56,45,33,23]: polstar(exp,log(1.+1./2048),3,m);

"minimax = ", .99999999999999981509827946165 +
(1.00000000001212038156196482 71
 + (.49999998758606303032049391 0112
 + .16670735254986148879274879363 x) x) x
                                                    -16
  "Distance between f and p =", .1849017208895 10


                1398443  3    4294967189  2    35184372088875
"hatp =", ------- x  + ---------- x  + -------------- x
                8388608      8589934592       35184372088832
```

```
          7205759403792935
        + ----------------
          7205759403792936


    "Distance between f and hatp =",
                                          -16
                .2362422096932623522944 3 10

>Do you want to continue (y;/n;)? y;
>Enter the value of parameter lambda: 1;

degree 0: 6 possible values between
18014398509481983/18014398509481984
          and 7205759403792937/7205759403792936
```

```
degree 1: 109 possible values between
35184372088821/35184372088832
          and 35184372088929/35184372088832
degree 2: 146 possible values between
          4294967117/8589934592
          and 2147483631/4294967296
degree 3: 194 possible values between 699173/4194304
          and 1398539/8388608
18 523 896 polynomials need be checked
```

Raffinement et approche plus générale : polytopes

Première approche a plusieurs défauts :

- domaine de validité seulement $[0, a]$ ou $[-a, a]$;
- vite limitée. Pas étonnant car
  - inégalité triangulaire utilisée ;
  - coefficients traités indépendamment.

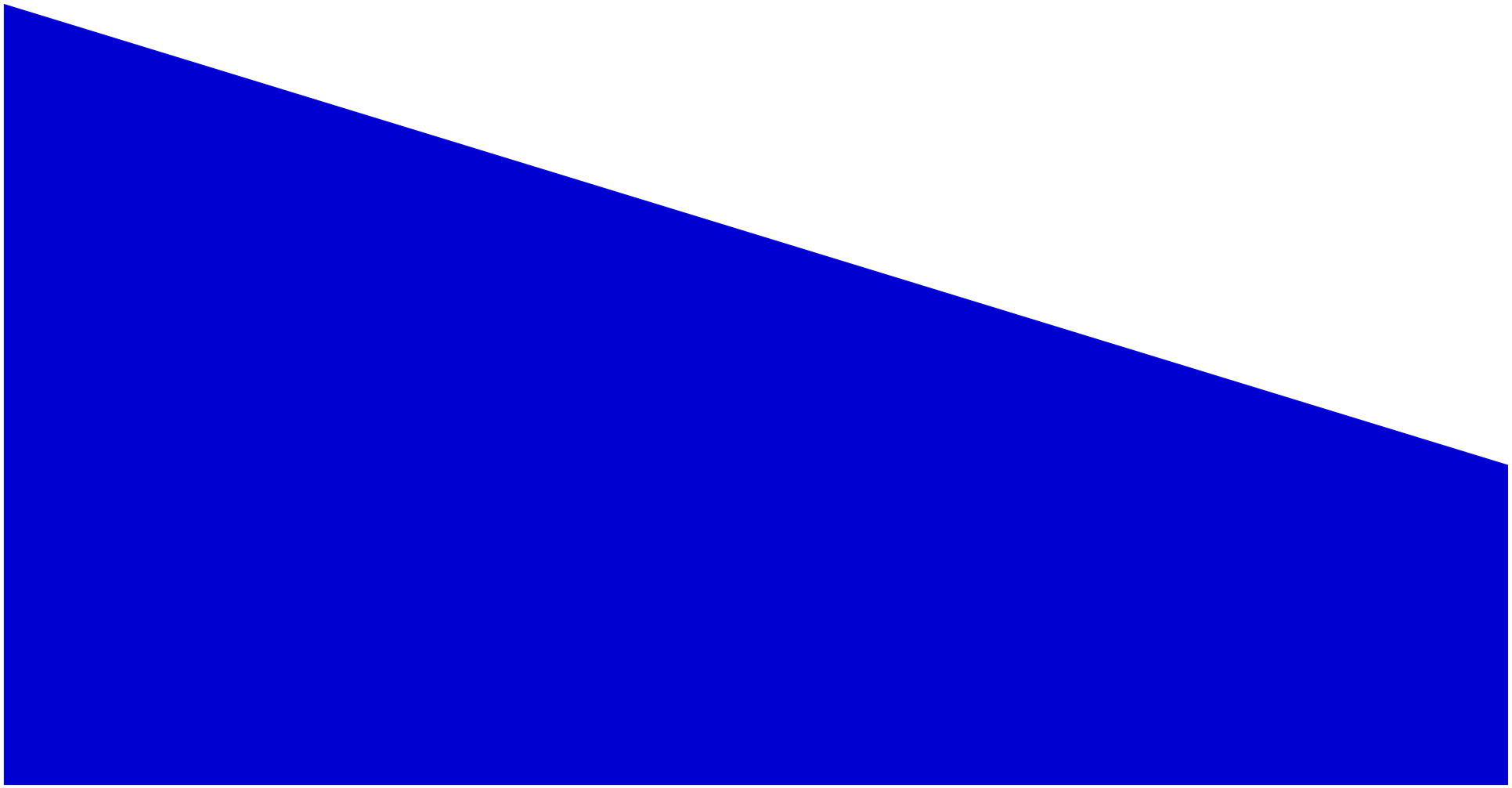
Définitions . *Soit $k \in \mathbb{N}$.*

On appelle polyèdre un sous-ensemble $\mathfrak{P}$ de $\mathbb{R}^k$ t.q. il existe une matrice $A \in \mathcal{M}_{m,k}(\mathbb{R})$ et un vecteur $b \in \mathbb{R}^m$ (avec $m \geq 0$) t. q.

$$\mathfrak{P} = \{x \in \mathbb{R}^k | Ax \leq b\}.$$

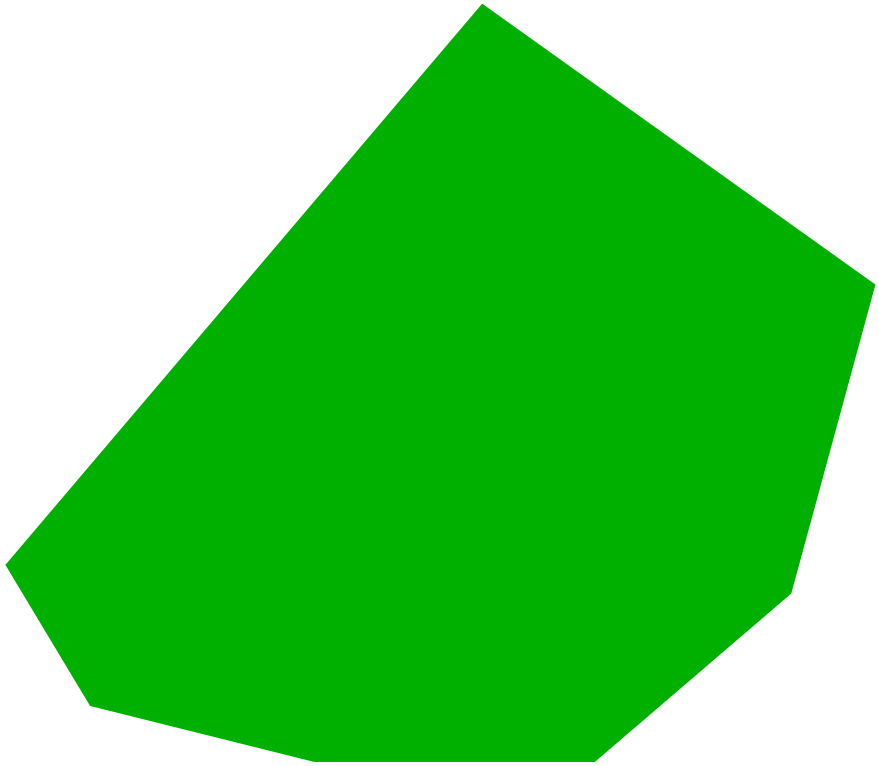
On appelle polytope un polyèdre borné.

On dit qu'un polyèdre (resp. polytope) $\mathfrak{P}$ est rationnel s'il est determiné par un système d'inégalités linéaires à coefficients rationnels.



Exemple de polyèdre : demi-plan dans $\mathbb{R}^2$.

Exemple de polyèdre : cône dans $\mathbb{R}^2$.



Exemple de polytope.

Rappel du problème

On pose

$$\varepsilon = ||f - p||_{[a,b]} \text{ et } \hat{\varepsilon} = ||f - \hat{p}||_{[a,b]}.$$

On compare ε et $\hat{\varepsilon}$.

On introduit $K \in [\varepsilon, \hat{\varepsilon}]$. On cherche un polynôme tronqué $p^\star \in \mathcal{P}_n^m$ t.q.

$$||f - p^\star||_{[a,b]} = \min_{q \in \mathcal{P}_n^m} ||f - q||_{[a,b]}$$

et

$$||f - p^\star||_{[a,b]} \leq K. \quad (1)$$

On pose $p^\star(x) = p_0^\star + p_1^\star x + \cdots + p_n^\star x^n$.

Sur $[0, a]$, on avait obtenu pour tout i

$$\underbrace{\lceil 2^{m_i}(p_i - (\varepsilon + K)|\alpha_i|)\rceil}_{c_i} \leq 2^{m_i} p_i^\star \leq \underbrace{\lfloor 2^{m_i}(p_i + (\varepsilon + K)|\alpha_i|)\rfloor}_{d_i}.$$

Elles définissent un polytope dont les entiers $2^{m_i} p_i^\star$ sont éléments.

Idée : réduire ce polytope ($\Rightarrow$ recherche exhaustive réduite).

Méthode valable sur $[a, b]$.

On doit avoir

$$f(x) - K \leq \sum_{i=0}^{n} p_i^\star x^i \leq f(x) + K \tag{2}$$

pour tout $x \in [a, b]$. On a $p_i^\star = a_i^\star / 2^{m_i}$ avec $a_i^\star \in \mathbb{Z}$.

On spécialise (2) en un certain nombre N de valeurs rationnelles de $[a, b]$. Soit $x = r/s$ avec $r \in \mathbb{Z}, s \in \mathbb{N}$. On a

$$f\left(\frac{r}{s}\right) - K \leq \sum_{i=0}^{n} \frac{a_i^\star}{2^{m_i}} \left(\frac{r}{s}\right)^i \leq f\left(\frac{r}{s}\right) + K.$$

Choisir $m(\frac{r}{s})$ et $M(\frac{r}{s}) \in \mathbb{Q}$ tels que $m(\frac{r}{s}) \leq f\left(\frac{r}{s}\right) - K$ et $f\left(\frac{r}{s}\right) + K \leq M(\frac{r}{s})$, $m(\frac{r}{s})$ “proche” de $f\left(\frac{r}{s}\right) - K$ et $M(\frac{r}{s})$ “proche” de $f\left(\frac{r}{s}\right) + K$.

Si $N \geq n+1 \Rightarrow$ on a un polytope rationnel dont les entiers $a_i^\star = 2^{m_i} p_i^\star$ sont éléments.

Si polytope assez petit, lancer recherche exhaustive en parcourant les points à coord. entières du polytope. Pour cela, on utilise des bibliothèques (comme Polylib, CLooG ou PIP) conçues pour parcourir efficacement les points entiers de polytopes.

Remarque . *Produit seulement des candidats.*

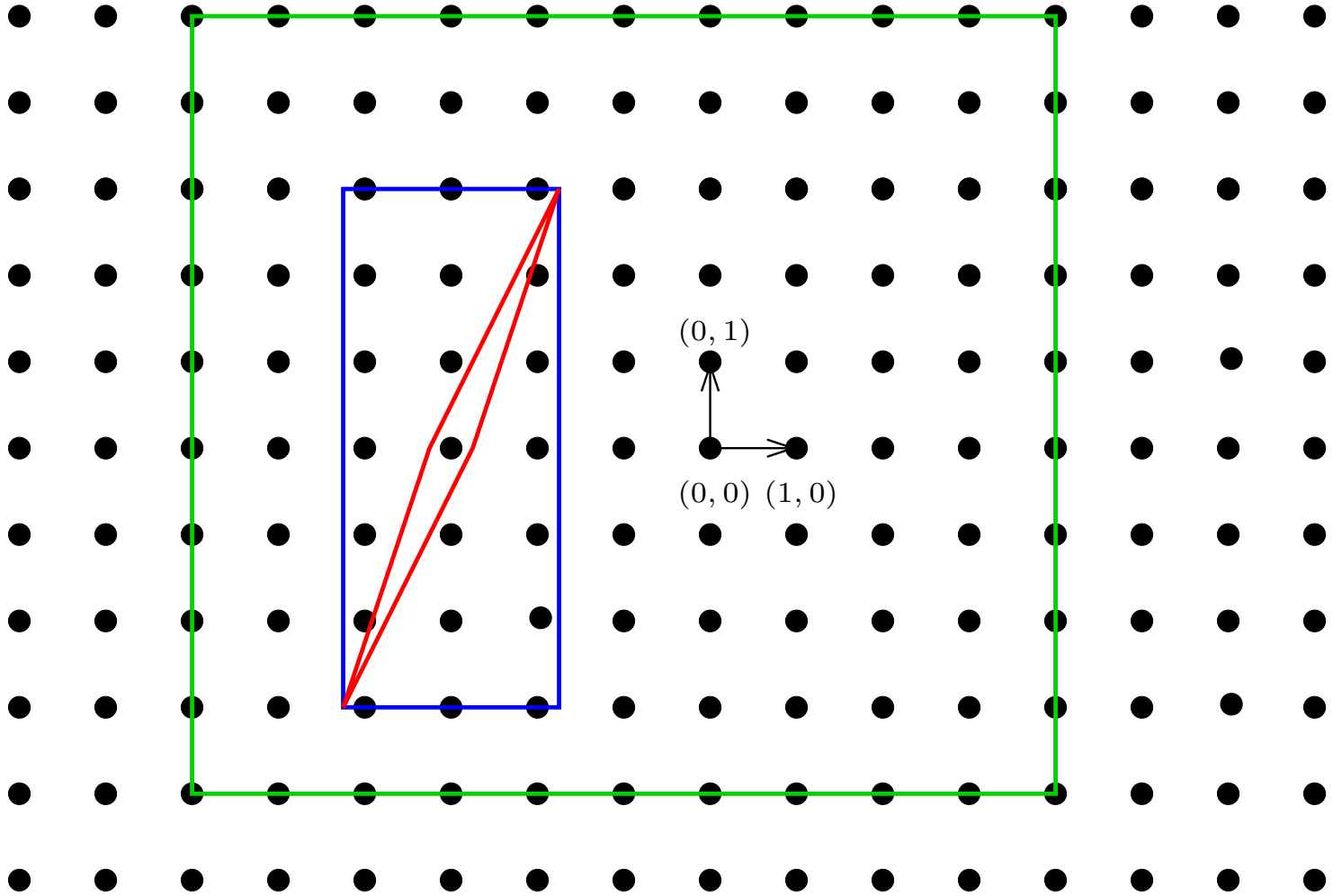
Méthode valable sur $[a, b]$.

On doit avoir

$$f(x) - K \leq \sum_{i=0}^{n} p_i^\star x^i \leq f(x) + K \tag{3}$$

pour tout $x \in [a, b]$.

1. Choisir A et $B \in \mathbb{Q}$, $a \leq A \leq B \leq b$, A “proche” de a et B “proche” de b. On définit $x_k = A + \frac{k}{d}(B - A)$ où $d \in \mathbb{N}$ choisi $\forall k$, $0 \leq k \leq d$.
2. On spécialise (3) en les x_k. On calcule approx. rationnelles des $f(x_k) - K$ et $f(x_k) + K$.
   $d \geq n \Rightarrow$ on a un polytope rationnel dont les entiers $2^{m_i} p_i^\star$ sont éléments.
3. Si polytope assez petit, lancer recherche exhaustive en parcourant les points à coord. entières du polytope. Pour cela, on utilise des bibliothèques (comme Polylib, CLooG ou PIP) conçues pour parcourir efficacement les points entiers de polytopes.



Approximation par un polynôme de degré 1. Polytope vert : 110 points de $\mathbb{Z}^2$. Polytope bleu : 21 points de $\mathbb{Z}^2$. Polytope rouge : 1 point de $\mathbb{Z}^2$.

Approximation de la fonction exponentielle sur $[0, \log(1 + 1/2048)]$ par un polynôme de degré 3

```
>Digits:=30:
>m := [56,45,33,23]: polstar(exp,log(1.+1./2048),3,m);

"minimax = ", .99999999999999981509827946165 +
(1.00000000001212038156196482 71
 + (.49999998758606303032049391 0112
 + .16670735254986148879274879363 x) x) x
                                                    -16
  "Distance between f and p =", .1849017208895 10
```

```
                1398443  3   4294967189  2   35184372088875
"hatp =", ------- x  + ---------- x  + -------------- x
                8388608     8589934592     35184372088832


                  7205759403792793 5
               + -----------------
                  7205759403792793 6

    "Distance between f and hatp =",
                                                    -16
                    .23624220969326235229443 10

>Do you want to continue (y;/n;)? y;
```

```
>Enter the value of parameter lambda: 1;

degree 0: 6 possible values between
18014398509481983/18014398509481984
          and 72057594037927937/72057594037927936
degree 1: 109 possible values between
35184372088821/35184372088832
          and 35184372088929/35184372088832
degree 2: 146 possible values between
          4294967117/8589934592
          and 2147483631/4294967296
degree 3: 194 possible values between 699173/4194304
          and 1398539/8388608
18 523 896 polynomials need be checked

>Do you want to try to refine the bounds (y;/n;)?y;
```

```
>Enter the value of parameter d: 25;

degree 0: 2 possible values between
72057594037927935/72057594037927936
          and 1
degree 1: 27 possible values between
          35184372088857/35184372088832
          and 35184372088883/35184372088832
degree 2: 32 possible values between
          536870897/1073741824
          and 4294967207/8589934592
degree 3: 44 possible values between 1398421/8388608
          and 21851/131072
76 032 polynomials need be checked

>Do you want to try to refine the bounds (y;/n;)?n;
```

```
>Do you want to change the value of Digits (y;/n;)?y;
>Enter the value of Digits: 21;
                1398443   3   2147483595   2   35184372088873
"pstar =", ------- x + ---------- x + ------------- x
                8388608     4294967296      35184372088832

                  7205759403792793 5
              + -------------------
                  7205759403792793 6

    "Distance between f and pstar =",
                                                      -16
                    .20246280367096470182285 10

          "Time elapsed (in seconds) =", 54721.961
```

Dans cet exemple, on gagne $-\log_2(0.85) \approx 0.22$ bits de précision.

La méthode à base de polytopes est souple !

On peut ajouter des contraintes (fixer des coef. par exemple) ou examiner un autre type d'erreur.

Exemples .

- *Si l'on restreint la recherche à des polynômes impairs, on considère*

$$f(x_k) - K \leq \sum_{i=0}^{I} p_i^\star x_k^{2i+1} \leq f(x_k) + K, \quad k = 0, \ldots, d$$

avec $x_k \in \mathbb{Q} \cap [a, b]$, $d \geq I$. *On calcule des approx. rationnelles* m_k *et* M_k *de* $f(x_k) - K$ *et* $f(x_k) + K$. *On obtient un polytope rationnel* $\mathfrak{P}$ *de* $\mathbb{R}^{k+1}$ *dont nous parcourons les points à coordonnées entières.*

- *Si l'on restreint la recherche à des polynômes dont le terme constant vaut* 1*, on considère*

$$f(x_k) - K \leq 1 + \sum_{i=1}^{n} p_i^\star x_k^{2i+1} \leq f(x_k) + K, \quad k = 0, \ldots, d$$

avec $x_k \in \mathbb{Q} \cap [a, b]$, $d \geq n - 1$.

- *On peut chercher le meilleur polynôme tronqué pour l'erreur relative* $||\cdot||_{\textit{rel},[a,b]}$ *définie par*

$$||f - p||_{\textit{rel},[a,b]} = \sup_{a \leq x \leq b} \frac{1}{|f(x)|} |p(x) - f(x)|.$$

Soit $K \geq 0$, on cherche un polynôme tronqué $p^\star \in \mathcal{P}_n^m$ tel que

$$||f - p^\star||_{\textbf{rel},[a,b]} = \min_{q \in \mathcal{P}_n^m} ||f - q||_{\textbf{rel},[a,b]}$$

et

$$||f - p^\star||_{\textbf{rel},[a,b]} \leq K. \tag{4}$$

On considère

$$-K|f(x)| - f(x) \leq \sum_{i=0}^{n} p_i^\star x^i \leq K|f(x)| + f(x)$$

pour au moins $n+1$ valeurs rationnelles de $x \in [a,b]$.