Algorithmique numérique et fiabilité des calculs en arithmétique flottante
Introduction à l'arithmétique par intervalles

Cours M2 - ISFA

Nathalie Revol (INRIA)

2 février 2016

Compter sans se tromper : arithmétique par intervalles

Principe : nombres remplacés par des intervalles.
π remplacé par $[3.14159, 3.14160]$ ou $[3.14, 3.15]$ ou $[3, 4]$.
Théorème fondamental (tu ne mentiras point) : l'intervalle contient la valeur exacte.

Compter sans se tromper : arithmétique par intervalles

Principe : nombres remplacés par des intervalles.
π remplacé par $[3.14159, 3.14160]$ ou $[3.14, 3.15]$ ou $[3, 4]$.
Théorème fondamental (tu ne mentiras point) : l'intervalle contient la valeur exacte.

Exemple
Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.

Compter sans se tromper : arithmétique par intervalles

Principe : nombres remplacés par des intervalles.
π remplacé par $[3.14159, 3.14160]$ ou $[3.14, 3.15]$ ou $[3, 4]$.
Théorème fondamental (tu ne mentiras point) : l'intervalle contient la valeur exacte.

Exemple
Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.
Contenu de votre porte-monnaie : entre 5 Euros et 10 Euros,
$\in [5, 10]$ Euros.

Compter sans se tromper : arithmétique par intervalles

Principe : nombres remplacés par des intervalles.
π remplacé par $[3.14159, 3.14160]$ ou $[3.14, 3.15]$ ou $[3, 4]$.
Théorème fondamental (tu ne mentiras point) : l'intervalle contient la valeur exacte.

Exemple
Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.
Contenu de votre porte-monnaie : entre 5 Euros et 10 Euros,
$\in [5, 10]$ Euros.

Ensemble, nous avons entre 15 et 30 Euros,
$[10, 20] + [5, 10] = [15, 30]$ Euros.

Arithmétique par intervalles : premier problème

Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.

Arithmétique par intervalles : premier problème

Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.

Je vais voir vos grands-parents, qui me donnent une enveloppe pour vous, contenant entre 10 et 20 Euros.
Je range l'argent dans mon porte-monnaie, il contient maintenant entre 20 et 40 Euros : $[10, 20] + [10, 20] = [20, 40]$ Euros.

Arithmétique par intervalles : premier problème

Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.

Je vais voir vos grands-parents, qui me donnent une enveloppe pour vous, contenant entre 10 et 20 Euros.
Je range l'argent dans mon porte-monnaie, il contient maintenant entre 20 et 40 Euros : $[10, 20] + [10, 20] = [20, 40]$ Euros.

Je vous donne votre argent, entre 10 et 20 Euros.
Mon porte-monnaie contient $[20, 40] - [10, 20] = [0, 30]$ Euros.

Arithmétique par intervalles : premier problème

Contenu de mon porte-monnaie : entre 10 Euros et 20 Euros,
$\in [10, 20]$ Euros.

Je vais voir vos grands-parents, qui me donnent une enveloppe pour vous, contenant entre 10 et 20 Euros.
Je range l'argent dans mon porte-monnaie, il contient maintenant entre 20 et 40 Euros : $[10, 20] + [10, 20] = [20, 40]$ Euros.

Je vous donne votre argent, entre 10 et 20 Euros.
Mon porte-monnaie contient $[20, 40] - [10, 20] = [0, 30]$ Euros.

Autrement dit,

$$\text{porte-monnaie} + \text{enveloppe} - \text{enveloppe} \neq \text{porte-monnaie}.$$

Arithmétique par intervalles : deuxième problème

Rappel : un nombre au carré est toujours positif $\Rightarrow$ on ne peut définir la racine carrée d'un nombre que s'il est positif.

Arithmétique par intervalles : deuxième problème

Rappel : un nombre au carré est toujours positif $\Rightarrow$ on ne peut définir la racine carrée d'un nombre que s'il est positif.

Comment définir $\sqrt{[-1,2]}$?

Arithmétique par intervalles : deuxième problème

Rappel : un nombre au carré est toujours positif $\Rightarrow$ on ne peut définir la racine carrée d'un nombre que s'il est positif.

Comment définir $\sqrt{[-1,2]}$?

Par convention, $\sqrt{[-1,2]} = \sqrt{[0,2]} = [0,\sqrt{2}]$.
Il faut signaler qu'il y a eu un problème :

Arithmétique par intervalles : deuxième problème

Rappel : un nombre au carré est toujours positif $\Rightarrow$ on ne peut définir la racine carrée d'un nombre que s'il est positif.

Comment définir $\sqrt{[-1,2]}$?

Par convention, $\sqrt{[-1,2]} = \sqrt{[0,2]} = [0,\sqrt{2}]$.
Il faut signaler qu'il y a eu un problème :

- comment ? en " décorant " les intervalles 

Arithmétique par intervalles : deuxième problème

Rappel : un nombre au carré est toujours positif $\Rightarrow$ on ne peut définir la racine carrée d'un nombre que s'il est positif.

Comment définir $\sqrt{[-1,2]}$?

Par convention, $\sqrt{[-1,2]} = \sqrt{[0,2]} = [0,\sqrt{2}]$.
Il faut signaler qu'il y a eu un problème :

- comment ? en " décorant " les intervalles 
- comment faire pour ne pas pénaliser les performances des calculs sur ordinateur : temps de calcul, utilisation de la mémoire ?

Arithmétique par intervalles : troisième problème

Pour une opération entre des intervalles, il faut effectuer 2 (ou 4 ou . . .) opérations “ habituelles ”.

Pour des raisons liées à l'architecture des ordinateurs, il faut 100 fois plus de temps (ou pire) pour exécuter une série de calculs sur des intervalles qu'avec des nombres “ habituels ”.

Comment faire pour que les calculs soient moins lents ?
S'ils sont 10 à 15 fois moins lents, on est déjà satisfait.

Agenda

Agenda

Definitions : functions

Definition :
an interval extension $\boldsymbol{f}$ of a function f satisfies

$$\forall \boldsymbol{x},\ f(\boldsymbol{x}) \subset \boldsymbol{f}(\boldsymbol{x}),\ \text{and}\ \forall x,\ f(\{x\}) = \boldsymbol{f}(\{x\}).$$

Elementary functions : again, use the monotony.

$$\begin{array}{lcl} \exp \boldsymbol{x} & = & [\exp \underline{x}, \exp \overline{x}] \\ \log \boldsymbol{x} & = & [\log \underline{x}, \log \overline{x}] \ \text{if}\ \underline{x} \geq 0, [-\infty, \log \overline{x}] \ \text{if}\ \overline{x} > 0 \\ \sin[\pi/6, 2\pi/3] & = & [1/2, 1] \\ \ldots & & \end{array}$$

Agenda

Definitions : function extension

Example : $f(x) = x^2 - x + 1$ with $x \in [-2, 1]$.
$[-2, 1]^2 - [-2, 1] + 1 = [0, 4] + [-1, 2] + 1 = [0, 7]$.

Definitions : function extension

Example : $f(x) = x^2 - x + 1$ with $x \in [-2, 1]$.
$[-2,1]^2 - [-2,1] + 1 = [0,4] + [-1,2] + 1 = [0,7]$.
Since $x^2 - x + 1 = x(x-1) + 1$, we get $[-2,1] \cdot ([-2,1] - 1) + 1 = [-2,1] \cdot [-3,0] + 1 = [-3,6] + 1 = [-2,7]$.

Definitions : function extension

Example : $f(x) = x^2 - x + 1$ with $x \in [-2, 1]$.
$[-2,1]^2 - [-2,1] + 1 = [0,4] + [-1,2] + 1 = [0,7]$.
Since $x^2 - x + 1 = x(x-1) + 1$, we get $[-2,1] \cdot ([-2,1] - 1) + 1 = [-2,1] \cdot [-3,0] + 1 = [-3,6] + 1 = [-2,7]$.
Since $x^2 - x + 1 = (x - 1/2)^2 + 3/4$, we get $([-2,1] - 1/2)^2 + 3/4 = [-5/2, 1/2]^2 + 3/4 = [0, 25/4] + 3/4 = [3/4, 7] = f([-2,1])$.

Problem with this definition : infinitely many interval extensions, syntactic use (instead of semantic).

How to choose the best extension ? How to choose a good one ?

Definitions : function extension

Mean value theorem of order 1 (Taylor expansion of order 1) :

$\forall x, \forall y, \exists \xi_{x,y} \in (x, y) \ : \ f(y) = f(x) + (y - x) \cdot f'(\xi_{x,y})$

Interval interpretation :

$\forall y \in \boldsymbol{x}, \forall \tilde{x} \in \boldsymbol{x}, \ f(y) \in f(\tilde{x}) + (y - \tilde{x}) \cdot \boldsymbol{f'(x)}$

$\Rightarrow f(\boldsymbol{x}) \subset f(\tilde{x}) + (\boldsymbol{x} - \tilde{x}) \cdot \boldsymbol{f'(x)}$

Definitions : function extension

Mean value theorem of order 1 (Taylor expansion of order 1) :

$\forall x, \forall y, \exists \xi_{x,y} \in (x,y) \ : \ f(y) = f(x) + (y-x) \cdot f'(\xi_{x,y})$

Interval interpretation :

$\forall y \in \boldsymbol{x}, \forall \tilde{x} \in \boldsymbol{x}, \ f(y) \in f(\tilde{x}) + (y-\tilde{x}) \cdot \boldsymbol{f'(x)}$

$\Rightarrow f(\boldsymbol{x}) \subset f(\tilde{x}) + (\boldsymbol{x}-\tilde{x}) \cdot \boldsymbol{f'(x)}$

Mean value theorem of order 2 (Taylor expansion of order 2) :

$\forall x, \forall y, \exists \xi_{x,y} \in (x,y) : f(y) = f(x) + (y-x) \cdot f'(x) + \frac{(y-x)^2}{2} \cdot f''(\xi_{x,y})$

Interval interpretation :

$\forall y \in \boldsymbol{x}, \forall \tilde{x} \in \boldsymbol{x}, \ f(y) \in f(\tilde{x}) + (y-\tilde{x}) \cdot f'(\tilde{x}) + \frac{(y-\tilde{x})^2}{2} \cdot \boldsymbol{f''(x)}$

$\Rightarrow f(\boldsymbol{x}) \subset f(\tilde{x}) + (\boldsymbol{x}-\tilde{x}) \cdot f'(\tilde{x}) + \frac{(\boldsymbol{x}-\tilde{x})^2}{2} \cdot \boldsymbol{f''(x)}$

Definitions : function extension

No need to go further :

- it is difficult to compute (automatically) the derivatives of higher order,
  especially for multivariate functions ;
- there is no (theoretical) gain in quality.

Theorem :

- for the natural extension $\boldsymbol{f}$ of f, it holds
  $d(f(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})) \leq \mathcal{O}(w(\boldsymbol{x}))$
- for the first order Taylor extension $\boldsymbol{f_{T_1}}$ of f, it holds
  $d(f(\boldsymbol{x}), \boldsymbol{f_{T_1}}(\boldsymbol{x})) \leq \mathcal{O}(w(\boldsymbol{x})^2)$
- getting an order higher than 3 is impossible without the squaring operation, is difficult even with it...

Agenda

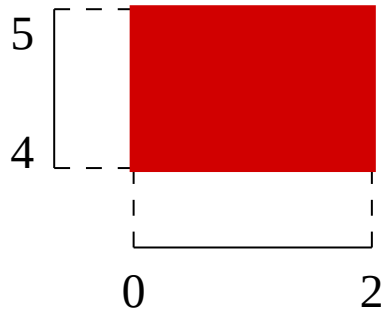
Definitions : intervals, vectors, matrices

Objects :

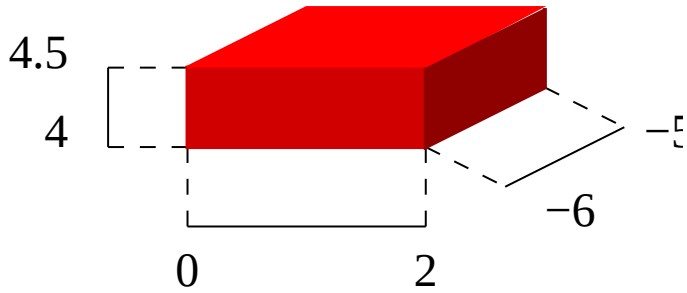
- intervals of real numbers = closed connected sets of ***R***
  - interval for π : $[3.14159, 3.14160]$
  - data d measured with an absolute error less than $\pm\varepsilon$: $[d-\varepsilon, d+\varepsilon]$
- interval vector : components = intervals ; also called *box*

$[0\,;\,2]$

$\begin{pmatrix}[0\,;\,2]\\[4\,;\,5]\end{pmatrix}$



$\begin{pmatrix}[0;2]\\[4\,;\,4.5]\\[-6\,;\,-5]\end{pmatrix}$



- interval matrix : components = intervals.

Agenda

Agenda

Cons : overestimation (1/2)

The result encloses the true result, but it is too large :
overestimation phenomenon.
Two main sources : variable dependency and wrapping effect.

(Loss of) Variable dependency :

$$\boldsymbol{x} - \boldsymbol{x} = \{x - y \ : \ x \in \boldsymbol{x}, y \in \boldsymbol{x}\} \neq \{x - x \ : \ x \in \boldsymbol{x}\} = \{0\}.$$

Cons : overestimation (2/2)

Wrapping effect

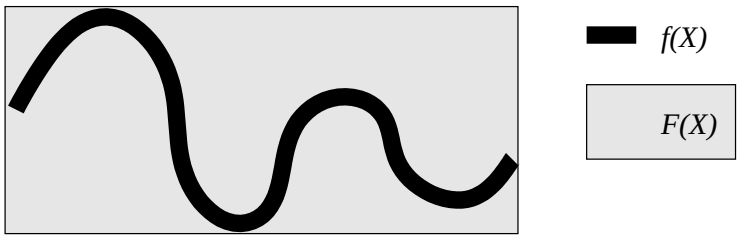
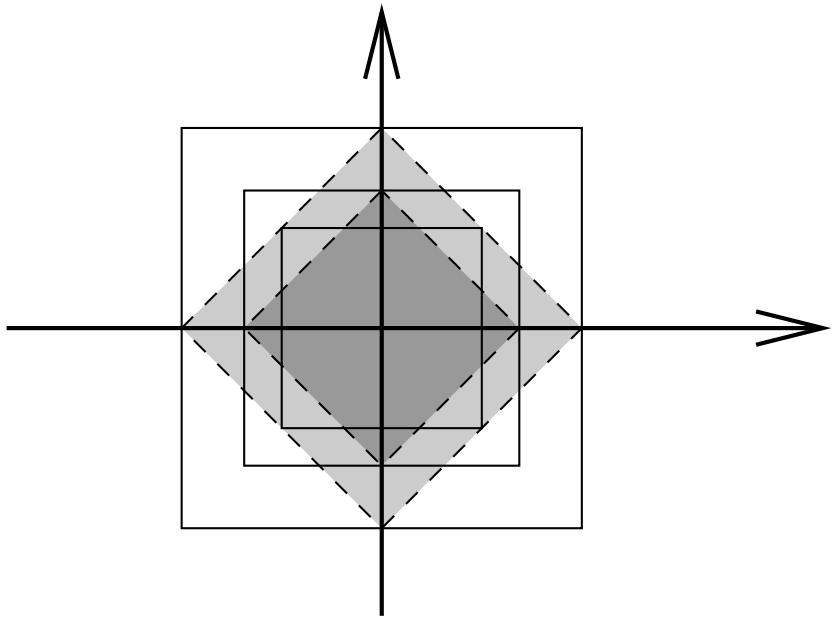


image of $f(\boldsymbol{x})$
with $f : \boldsymbol{R}^2 \rightarrow \boldsymbol{R}^2$



2 successive rotations of $\pi/4$
of the little central square

Cons : Complexity : almost every problem is NP-hard

Gaganov 1982, Rohn 1994 ff, Kreinovich...

- evaluate a function on a box (cartesian product of intervals)
- evaluate a function on a box up to ε
- solve a linear system
- solve a linear system up to $1/4n^4$ ($n =$ dim. of the system)
- determine if the solution of a linear system is bounded
- compute the matrix norm $\|\boldsymbol{A}\|_{\infty,1}$
- determine if an interval matrix ($=$ a matrix with interval coefficients) is regular, i.e. if every possible punctual matrix in it is regular
- ...

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

Idea : reduce polynomially CNF-3 to this problem.

On n boolean variables $q_1, \cdots, q_n$, a formula f in CNF-3 is defined by

$$f = \bigwedge_{i=1}^{m} f_i \text{ with } f_i = \bigvee_{j=1}^{1,2 or 3} r_{i,j}$$

with $r_{i,j} = q_{k_{i,j}}$ or $r_{i,j} = \neg q_{k_{i,j}}$.

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

To each boolean variable q_i, let us associate a real variable $x_i \in [0, 1]$.
Meaning : $x_i = 0$ if $q_i = F$ and $x_i = 1$ if $q_i = T$.

Goal : get a polynomial which takes only values in $[0, 1]$ i.e. allow only product of terms or of $(1-$ term$)$.

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

To each boolean variable q_i, let us associate a real variable $x_i \in [0, 1]$.
Meaning : $x_i = 0$ if $q_i = F$ and $x_i = 1$ if $q_i = T$.

Goal : get a polynomial which takes only values in $[0, 1]$
i.e. allow only product of terms or of $(1-$ term$)$.
A product corresponds to a conjunction and $1 - x$ to a negation

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

To each boolean variable q_i, let us associate a real variable $x_i \in [0, 1]$.
Meaning : $x_i = 0$ if $q_i = F$ and $x_i = 1$ if $q_i = T$.

Goal : get a polynomial which takes only values in $[0, 1]$
i.e. allow only product of terms or of $(1-$ term$)$.
A product corresponds to a conjunction and $1 - x$ to a negation
$\Rightarrow$ express f and the f_i using conjonctions and negations

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

To each boolean variable q_i, let us associate a real variable $x_i \in [0,1]$.
Meaning : $x_i = 0$ if $q_i = F$ and $x_i = 1$ if $q_i = T$.

Goal : get a polynomial which takes only values in $[0,1]$
i.e. allow only product of terms or of $(1-$ term$)$.
A product corresponds to a conjunction and $1-x$ to a negation
$\Rightarrow$ express f and the f_i using conjonctions and negations
$\Rightarrow$ express the f_i as $\neg \bigwedge_{j=1}^{1,2 or 3} \neg r_{i,j}$.

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

More precisely :

1. to each $r_{i,j}$ let us associate a polynomial $y_{i,j}$ (corresponding to the negation of $r_{i,j}$) defined by

$$\begin{array}{lllllll} r_{i,j} & = & q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & 1 - x_{k_{i,j}} \\ r_{i,j} & = & \neg q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & x_{k_{i,j}} \end{array}$$

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

More precisely :

1. to each $r_{i,j}$ let us associate a polynomial $y_{i,j}$ (corresponding to the negation of $r_{i,j}$) defined by
$$\begin{array}{lllllll} r_{i,j} & = & q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & 1 - x_{k_{i,j}} \\ r_{i,j} & = & \neg q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & x_{k_{i,j}} \end{array}$$
2. to each f_i, let us associate a polynomial p_i (corresponding to the negation of f_i) defined by
$f_i = \bigvee r_{i,j} = \neg \bigwedge \neg r_{i,j} \rightarrow p_i(x) = \prod y_{i,j}(x).$

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

More precisely :

1. to each $r_{i,j}$ let us associate a polynomial $y_{i,j}$ (corresponding to the negation of $r_{i,j}$) defined by
$$\begin{array}{lllllll} r_{i,j} & = & q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & 1 - x_{k_{i,j}} \\ r_{i,j} & = & \neg q_{k_{i,j}} & \rightarrow & y_{i,j}(x) & = & x_{k_{i,j}} \end{array}$$
2. to each f_i, let us associate a polynomial p_i (corresponding to the negation of f_i) defined by
$f_i = \bigvee r_{i,j} = \neg \bigwedge \neg r_{i,j} \rightarrow p_i(x) = \prod y_{i,j}(x)$.
3. to f, let us associate the polynomial p defined by
$f = \bigwedge_{i=1}^{m} f_i \rightarrow p(x) = \prod_{i=1}^{m}(1 - p_i(x))$.

Cons : Complexity : Gaganov 1982

evaluation of a multivariate polynomial with rational coeff. on a box is NP-hard

Lemma :

1. $\forall x \in [0,1]$, $p(x) \in [0,1]$.
2. if α is a boolean vector and β is the associated $0-1$ vector, then
$$\begin{array}{rcrcrcl} f(\alpha) & = & T & \Rightarrow & p(\beta) & = & 1 \\ f(\alpha) & = & F & \Rightarrow & p(\beta) & = & 0. \end{array}$$
3. if f is not feasible, then $\forall x \in [0,1]^n$, $p(x) \leq 7/8$.

Cons : Complexity : Gaganov 1982

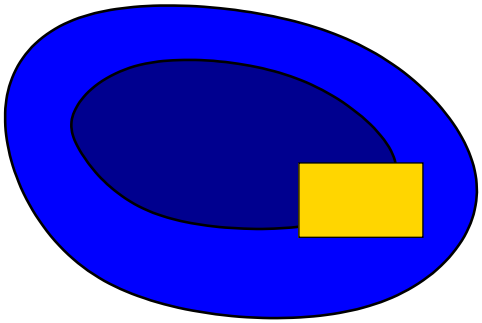
Proof of (3) : (proving (1) and (2) is easy).
$\forall x \in [0,1]^n$, let us consider β the 0-1 vector obtained by rounding x to the nearest.
Since f is not feasible, $p(\beta) = 0$.
Since $p(x) = \prod_{i=1}^{m}(1 - p_i(x))$, $\exists i_0$ such that $1 - p_{i_0}(\beta) = 0$.
One can prove that $p_{i_0}(x) \geq 1/8$, using the fact that it is the product of at most three terms, each of them $\geq 1/2$, using the fact that β is the rounding to nearest of x. Thus $1 - p_{i_0}(x) \leq 7/8$.
The remaining factors $1 - p_j(x)$ are less or equal to 1.
Thus $p(x) = \prod_{i=1}^{m}(1 - p_i(x)) \leq 7/8$.

Consequence : since checking the feasibility of a CNF-3 formula is NP-hard, evaluating a multivariate polynomial (up to a small ε) is NP-hard.

Agenda

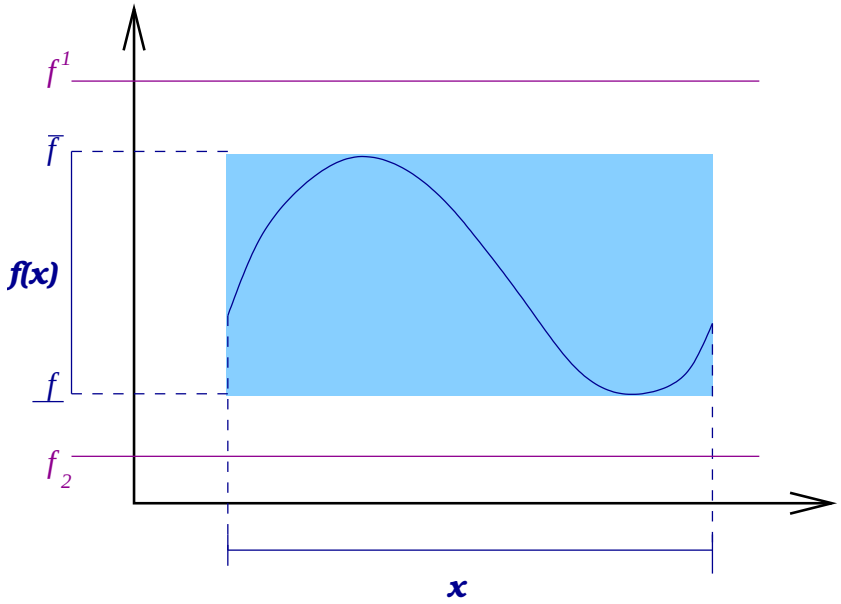
Pros : set computing

Behaviour safe ?
controllable ? dangerous ?



always controllable.

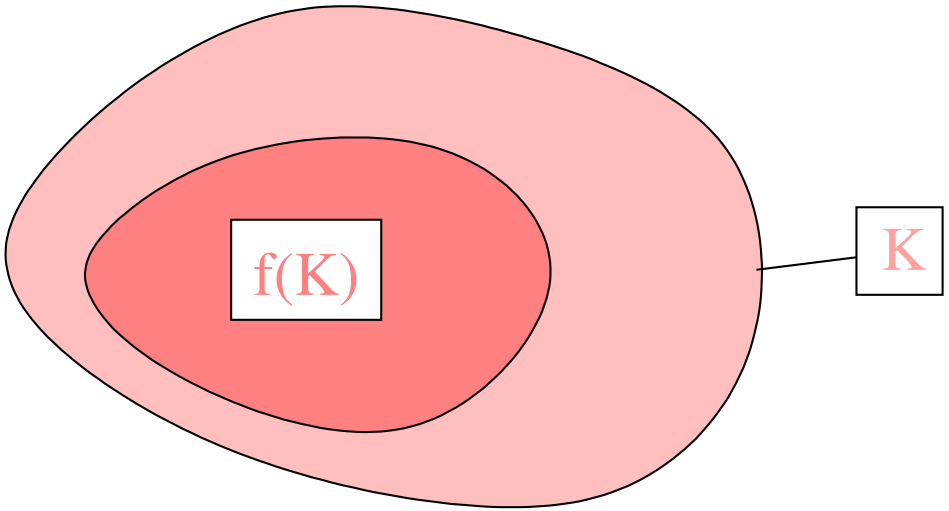
On x, are the extrema of the function f $> f^1$, $< f_2$?



No if $f(x) = [\underline{f}, \overline{f}] \subset [f_2, f^1]$.

Pros : Brouwer-Schauder theorem

A function f which is continuous on the unit ball B and which satisfies $f(B) \subset B$ has a fixed point on B.
Furthermore, if $f(B) \subset \text{int}B$ (and some other conditions) then f has a unique fixed point on B.



The theorem remains valid if B is replaced by a compact K and in particular an interval.