MT09 Chapitre 1
Introduction au calcul flottant

MT09
Vincent.Martin@utc.fr

UTC Compiègne
France

UTC, A2021

Plan

Plan

Une suite curieuse

Quelle est la limite de la suite $(u_n)_{n\in\mathbb{N}}$?

$$u_n = 111 - \frac{1130}{u_{n-1}} + \frac{3000}{u_{n-1}u_{n-2}}, \qquad u_0 = 2,\ u_1 = -4$$

3	18.5000000000000	17	7.2350211655349
4	9.37837837837838	18	22.0620784635258
5	7.80115273775217	19	78.5755748878722
6	7.15441448097533	20	98.3495031221654
7	6.80678473692481	21	99.8985692661829
8	6.59263276872179	22	99.9938709889028
9	6.44946593405393	23	99.9996303872863
10	6.34845206074662	24	99.9999777306795
11	6.27443866272812	25	99.9999986592167
12	6.21869676858216	26	99.9999999193218
13	6.17585385581539	27	99.9999999951478
14	6.14262717048101	28	99.9999999997083
15	6.12024870457016	29	99.9999999999825
16	6.16608655959810	30	99.9999999999989

Une suite curieuse

Quelle est la limite de la suite $(u_n)_{n\in\mathbb{N}}$?

$$u_n = 111 - \frac{1130}{u_{n-1}} + \frac{3000}{u_{n-1}u_{n-2}}, \qquad u_0 = 2,\ u_1 = -4$$

3	18.5000000000000	17	7.2350211655349
4	9.37837837837838	18	22.0620784635258
5	7.80115273775217	19	78.5755748878722
6	7.15441448097533	20	98.3495031221654
7	6.80678473692481	21	99.8985692661829
8	6.59263276872179	22	99.9938709889028
9	6.44946593405393	23	99.9996303872863
10	6.34845206074662	24	99.9999777306795
11	6.27443866272812	25	99.9999986592167
12	6.21869676858216	26	99.9999999193218
13	6.17585385581539	27	99.9999999951478
14	6.14262717048101	28	99.9999999997083
15	6.12024870457016	29	99.9999999999825
16	6.16608655959810	30	99.9999999999989

Pourtant :

$$u_n = \frac{3\cdot 6^{n+1} - 4\cdot 5^{n+1}}{3\cdot 6^n - 4\cdot 5^n} \Rightarrow \lim_{n\to\infty} u_n = 6$$

Plan

Écriture des entiers en base 2

En base 10

Chiffres $0, 1, \ldots, 9$

$$n = d_p 10^p + \ldots + d_1 10 + d_0, \quad 0 \leq d_i \leq 9,\ d_p \neq 0$$

$1789 = 1000 + 700 + 80 + 9 = 1 \cdot 10^3 + 7 \cdot 10^2 + 8 \cdot 10^1 + 9 \cdot 10^0$

Écriture des entiers en base 2

En base 10

Chiffres $0, 1, \ldots, 9$

$$n = d_p 10^p + \ldots + d_1 10 + d_0, \quad 0 \leq d_i \leq 9,\ d_p \neq 0$$

$$1789 = 1000 + 700 + 80 + 9 = 1 \cdot 10^3 + 7 \cdot 10^2 + 8 \cdot 10^1 + 9 \cdot 10^0$$

En base 2

Chiffres $= 0, 1$

$$n = d_p 2^p + \ldots + d_1 2 + d_0, \quad 0 \leq d_i \leq 1,\ d_p \neq 0,$$

$$\begin{aligned}(73)_{10} &= 64 + 8 + 1 \\ &= 1 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 \\ &= (1001001)_2\end{aligned}$$

Écriture scientifique des réels

$$x = \pm f \cdot 10^e, \quad 1/10 \leq f < 1 \qquad (f = \text{mantisse et } e = \text{exposant})$$

Remarque : 0 ne s'écrit pas sous cette forme.

Exemple

- $825.34 = 0.82534\,10^3$ écriture finie
- $8.2534 = 0.82534\,10^1$
- $0.0082534 = 0.82534\,10^{-2}$

- $1/2 = 0.5\,10^0$ fraction, écriture finie
- $1/3 = 0.333333333\ldots\ 10^0$ fraction, périodique
- $4/7 = 0.5714285714285\ldots$ fraction périodique

- $\pi = 0.314159265358\ldots\ 10^1$, infini, non périodique

Représentation des nombres flottants (base 2)

Taille mot mémoire limitée $\Rightarrow x \in \mathcal{F}_2$ ensemble fini (nombre flottants machine).

Definition (Nombres flottants)

Soit t, L, U trois entiers, tels que $t > 0$ et $L \leq U$.
L'ensemble des flottants en base 2 est défini par :

$$\mathcal{F}_2 = \left\{ \pm(0.1d_2 \ldots d_t)_2 \cdot 2^e \mid d_i \in \{0,1\}\ \forall i = 2, \ldots, t,\ L \leq e \leq U \right\} \cup \{0\}.$$

Définition similaire pour $\mathcal{F}_{10}$ (voir poly).

Représentation des nombres flottants (base 2)

Taille mot mémoire limitée $\Rightarrow x \in \mathcal{F}_2$ ensemble fini (nombre flottants machine).

Definition (Nombres flottants)

Soit t, L, U trois entiers, tels que $t > 0$ et $L \leq U$.
L'ensemble des flottants en base 2 est défini par :

$$\mathcal{F}_2 = \left\{ \pm(0.1d_2 \ldots d_t)_2 \cdot 2^e \mid d_i \in \{0,1\} \; \forall i = 2, \ldots, t, \; L \leq e \leq U \right\} \cup \{0\}.$$

Définition similaire pour $\mathcal{F}_{10}$ (voir poly).

Propriétés

$\mathcal{F}$ caractérisé par :

Base (ici 2 ou 10)

Nombre de chiffres t

Exposants min et max L et U

Représentation des nombres flottants (base 2)

Taille mot mémoire limitée $\Rightarrow x \in \mathcal{F}_2$ ensemble fini (nombre flottants machine).

Definition (Nombres flottants)

Soit t, L, U trois entiers, tels que $t > 0$ et $L \leq U$.
L'ensemble des flottants en base 2 est défini par :

$$\mathcal{F}_2 = \left\{ \pm(0.1d_2 \ldots d_t)_2 \cdot 2^e \mid d_i \in \{0, 1\} \ \forall i = 2, \ldots, t, \ L \leq e \leq U \right\} \cup \{0\}.$$

Définition similaire pour $\mathcal{F}_{10}$ (voir poly).

Propriétés

$\mathcal{F}$ caractérisé par :

Base (ici 2 ou 10)

Nombre de chiffres t

Exposants min et max L et U

$\mathcal{F}$: ensemble fini

$\Longrightarrow \mathcal{F}$ tout petit par rapport à $\mathbb{R}$!

$\mathrm{card}\mathcal{F}_2 = 1 + 2^t(U - L + 1)$

Représentation des nombres flottants (base 2)

Definition (Nombres flottants)

$$\mathcal{F}_2 = \left\{ \pm(0.1d_2 \dots d_t)_2 \cdot 2^e \mid d_i \in \{0,1\}\ \forall i = 2,\dots,t,\ L \leq e \leq U \right\} \cup \{0\}.$$

Propriétés de la mantisse f sur t chiffres

$$x \in \mathcal{F}_2 \setminus \{0\} \iff x = \pm f \cdot 2^e, \quad 2^{-1} \leq f < 1, \quad L \leq e \leq U$$

f mantisse, e exposant (entier, unique si $x \neq 0$)

Soit $0 \leq d_i \leq 1,\ d_1 \neq 0$, la mantisse s'écrit :

$$f = (0.d_1 d_2 \dots d_t)_2 = \frac{1}{2} + \frac{d_2}{2^2} + \dots + \frac{d_t}{2^t}.$$

Donc

$$f_2^{\min} = \frac{1}{2} = (0.10\dots0)_2 \leq f \leq (0.11\dots1)_2 = 1 - 2^{-t} = f_2^{\max}$$

Quelques exemples

- $3/2 = 1 + 1/2 = 2\,(1/2 + 1/4) = 2^1 \times (0.11)_2$

Quelques exemples

- $3/2 = 1 + 1/2 = 2\,(1/2 + 1/4) = 2^1 \times (0.11)_2$
- $5/2 = 2 + 1/2 = 4\,(1/2 + 1/8) = 2^2 \times (0.101)_2,$

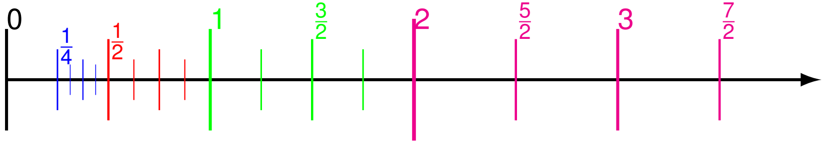
Quelques exemples

- $3/2 = 1 + 1/2 = 2\,(1/2 + 1/4) = 2^1 \times (0.11)_2$
- $5/2 = 2 + 1/2 = 4\,(1/2 + 1/8) = 2^2 \times (0.101)_2$,
- $1/10$, pas de représentation finie :

$$\begin{aligned}
1/10 &= \frac{1}{16}\frac{16}{10} = \frac{1}{16}(1 + \frac{3}{5}) = \frac{1}{16}(1 + \frac{9}{16}\,\frac{1}{1 - 1/16}) \\
&= 2^{-4}\left(1 + \frac{9}{16} + \frac{9}{16^2} + \frac{9}{16^3} + \dots\right) \\
&= 2^{-3}\left(\frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{0}{2^4} + \frac{1}{2^5} + \underbrace{\frac{1}{2^6} + \frac{0}{2^7} + \frac{0}{2^8} + \frac{1}{2^9}}_{\text{période}} + \dots\right) \\
&= 2^{-3} \times (0.1100110011001\dots)_2
\end{aligned}$$

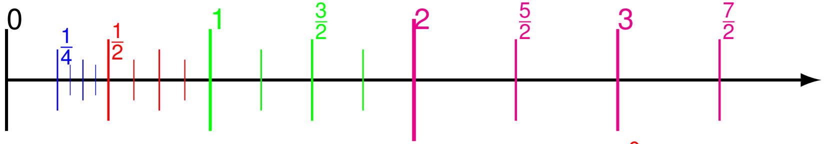
Exemple d'un ensemble $\mathcal{F}_2$

Exemple : $t = 3, L = -1, U = 2, \quad \text{card}\mathcal{F}_2 = 33$:



Exemple d'un ensemble $\mathcal{F}_2$

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \mathrm{card}\mathcal{F}_2 = 33$:

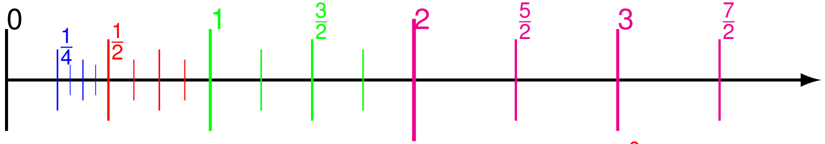


Nombres positifs de $\mathcal{F}_2$ entre 1/2 et 1 ($e = 0$) : $f = 2^0 \times (0.1d_2d_3)_2$

$$\begin{aligned} 1/2 &= & &= (0.100)_2, & 3/4 &= 1/2 + 1/4 & &= (0.110)_2 \\ 5/8 &= 1/2 + 1/8 & &= (0.101)_2, & 7/8 &= 1/2 + 1/4 + 1/8 & &= (0.111)_2 \end{aligned}$$

Exemple d'un ensemble $\mathcal{F}_2$

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \mathrm{card}\,\mathcal{F}_2 = 33$:



Nombres positifs de $\mathcal{F}_2$ entre 1/2 et 1 ($e = 0$) : $f = 2^0 \times (0.1 d_2 d_3)_2$

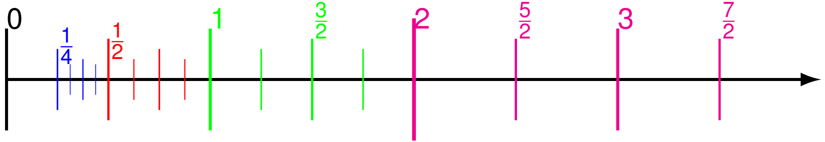
$$\begin{aligned}
1/2 &= & &= (0.100)_2, \quad & 3/4 &= 1/2 + 1/4 & &= (0.110)_2 \\
5/8 &= 1/2 + 1/8 & &= (0.101)_2, \quad & 7/8 &= 1/2 + 1/4 + 1/8 & &= (0.111)_2
\end{aligned}$$

Écart entre flottants successifs :

$$\begin{aligned}
\delta_0 &= \frac{1}{8} = 2^{-3} = 2^{-t} \quad \text{si } e = 0, \qquad & \delta_2 &= \frac{1}{2} = 2^{-1} = 2^{-t+2} \quad \text{si } e = 2, \\
\delta_1 &= \frac{1}{4} = 2^{-2} = 2^{-t+1} \quad \text{si } e = 1, \qquad & \delta_{-1} &= \frac{1}{16} = 2^{-4} = 2^{-t-1} \quad \text{si } e = -1.
\end{aligned}$$

Nombres flottants

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \mathrm{card}\,\mathcal{F}_2 = 33$

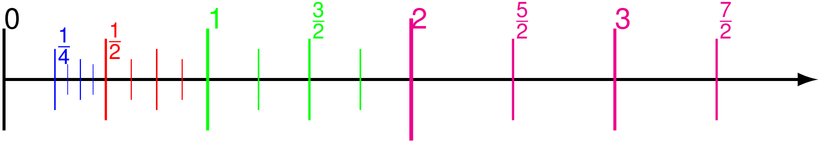


Écart absolu variable entre 2 flottants successifs
$f_1 = 2^e \times (0.1d_2d_3)_2$ et $f_2 = 2^e \times \left[(0.1d_2d_3)_2 + 2^{-3}\right]$:

$$\boxed{\delta_e = 2^{-t+e}} \qquad \text{(dépend de } e\text{).}$$

Nombres flottants

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \mathrm{card}\,\mathcal{F}_2 = 33$



Écart absolu variable entre 2 flottants successifs

$f_1 = 2^e \times (0.1d_2d_3)_2$ et $f_2 = 2^e \times \left[(0.1d_2d_3)_2 + 2^{-3}\right]$:

$$\boxed{\delta_e = 2^{-t+e}} \qquad \text{(dépend de } e\text{).}$$

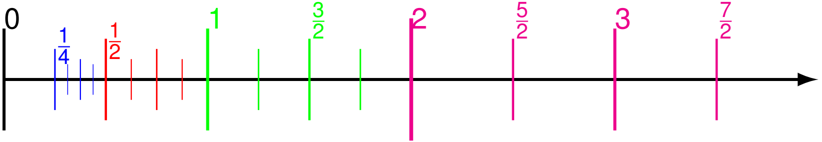
Mais écart relatif majoré par une constante

(note : $f_1 \geq 2^e \times (0.100)_2 = 2^e \times f_2^{\min} = 2^{e-1}$, donc $1/f_1 \leq 2^{-e+1}$)

$$\boxed{\delta_r = \frac{f_1+\delta_e-f_1}{f_1} \leq \frac{\delta_e}{2^{e-1}} = \frac{2^{-t}}{1/2} = 2^{-t+1} = \delta_1 = 2\varepsilon_{\mathrm{mach},2}}$$

Nombres flottants

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \text{card}\mathcal{F}_2 = 33$



Écart absolu variable entre 2 flottants successifs
$f_1 = 2^e \times (0.1d_2d_3)_2$ et $f_2 = 2^e \times \left[(0.1d_2d_3)_2 + 2^{-3}\right]$:

$$\boxed{\delta_e = 2^{-t+e}} \qquad \text{(dépend de } e\text{).}$$

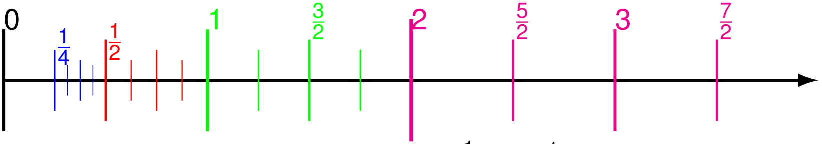
Mais écart relatif majoré par une constante
(note : $f_1 \geq 2^e \times (0.100)_2 = 2^e \times f_2^{\min} = 2^{e-1}$, donc $1/f_1 \leq 2^{-e+1}$)

$$\boxed{\delta_r = \frac{f_1+\delta_e-f_1}{f_1} \leq \frac{\delta_e}{2^{e-1}} = \frac{2^{-t}}{1/2} = 2^{-t+1} = \delta_1 = 2\varepsilon_{\text{mach},2}}$$

$$\boxed{\varepsilon_{\text{mach},2} = \frac{\delta_1}{2} = 2^{-t}}$$

Nombres flottants

Exemple : $t = 3, L = -1, U = 2, \quad \mathrm{card}\mathcal{F}_2 = 33$



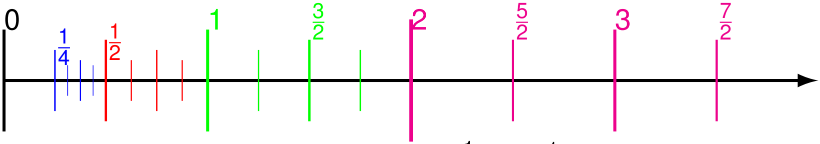
Plus petit nombre positif de $\mathcal{F}_2$: $1/4 = \frac{1}{2} \times 2^{-1}$.

Avec $f_2^{\min} = (0.10 \cdots 0)_2 = \frac{1}{2}$, il vient

$F_{\min} = 2^L \times f_2^{\min} = 2^L \times 1/2 = 2^{L-1}$.

“Trou » important autour de 0.

Exemple : $t = 3$, $L = -1$, $U = 2$, $\quad \mathrm{card}\mathcal{F}_2 = 33$



Plus petit nombre positif de $\mathcal{F}_2$: $1/4 = \frac{1}{2} \times 2^{-1}$.
Avec $f_2^{\text{min}} = (0.10\cdots0)_2 = \frac{1}{2}$, il vient
$F_{\text{min}} = 2^L \times f_2^{\text{min}} = 2^L \times 1/2 = 2^{L-1}$.
“Trou » important autour de 0.

Plus grand nombre positif de $\mathcal{F}_2$:
$7/2 = 2^2 \times (0.111)_2 = 2^2 \times ((1.00)_2 - (0.001)_2) = 4 \times (1 - 1/8) = 4 \times 7/8$.
Avec $f_2^{\text{max}} = (0.11\cdots1)_2 = (1 - 2^{-t})$, il vient
$F_{\text{max}} = 2^U \times f_2^{\text{max}} = 2^U \times (1 - 2^{-t}) \approx 2^U$.

Arrondi – epsilon machine

Approcher $x \in \mathbb{R}$ par $fl(x) \in \mathcal{F}$: $(x > 0)$

Arrondi au plus proche $fl(x)$ est l'élément de $\mathcal{F}$ le plus proche de x
($fl(x)$ éloigné au plus de $\delta_e/2$ de x,
car x est dans $[fl(x), fl(x) + \delta_e]$ ou $[fl(x) - \delta_e, fl(x)]$)

$$\boxed{\frac{|x - fl(x)|}{|x|} \leq 2^{-t} = \varepsilon_{\text{mach}}}$$

Arrondi – epsilon machine

Approcher $x \in \mathbb{R}$ par $fl(x) \in \mathcal{F}$: $(x > 0)$

Arrondi au plus proche $fl(x)$ est l'élément de $\mathcal{F}$ le plus proche de x

($fl(x)$ éloigné au plus de $\delta_e/2$ de x,

car x est dans $[fl(x), fl(x) + \delta_e]$ ou $[fl(x) - \delta_e, fl(x)]$)

$$\boxed{\frac{|x - fl(x)|}{|x|} \leq 2^{-t} = \varepsilon_{\text{mach}}}$$

Exemple (5 chiffres significatifs) : $x = \sqrt{7} \approx 2.6457513\ldots$

Arrondi $fl(x) = 2.6458$

Arrondi – epsilon machine

Approcher $x \in \mathbb{R}$ par $fl(x) \in \mathcal{F}$: $(x > 0)$

Arrondi au plus proche $fl(x)$ est l'élément de $\mathcal{F}$ le plus proche de x

($fl(x)$ éloigné au plus de $\delta_e/2$ de x,

car x est dans $[fl(x), fl(x) + \delta_e]$ ou $[fl(x) - \delta_e, fl(x)]$)

$$\boxed{\frac{|x - fl(x)|}{|x|} \leq 2^{-t} = \varepsilon_{\text{mach}}}$$

Exemple (5 chiffres significatifs) : $x = \sqrt{7} \approx 2.6457513\ldots$

Arrondi $fl(x) = 2.6458$

$\varepsilon_{\text{mach}} = 2^{-t}$ caractéristique de l'arithmétique.

$$fl(x) = x(1 + \varepsilon),\ |\varepsilon| \leq \varepsilon_{\text{mach}}$$

Calculatrice $\varepsilon_{\text{mach}} \approx 10^{-10}$

Avec Scilab $\varepsilon_{\text{mach}} = 2^{-53} \approx 1.11\,10^{-16}$ $\approx$ 16 chiffres

(Scilab : $''\%eps'' = 2\varepsilon_{\text{mach}} \approx 2.22\,10^{-16}$).

Nombre flottants : le système IEEE 754

Simple précision (`float`) :

1	8 bits	23 + 1 bits
s	e	f

32 bits, $t = 23 + 1$, $L = -126$, $U = 127$, $x_{\max} \approx 10^{38}$, $x_{\min} \approx 10^{-38}$

Nombre flottants : le système IEEE 754

Simple précision (`float`) :

1	8 bits	23 + 1 bits
s	e	f

32 bits, $t = 23 + 1$, $L = -126$, $U = 127$, $x_{\max} \approx 10^{38}$, $x_{\min} \approx 10^{-38}$

Double précision (`double`) :

1	11 bits	52 + 1 bits
s	e	f

64 bits, $t = 52 + 1$, $L = -1022$, $U = 1023$, $x_{\max} \approx 10^{308}$, $x_{\min} \approx 10^{-308}$

Propriétés de la norme IEEE 754

Utilisé par Java, processeurs Intel, PowerPC (norme internationale)

Bit caché gagne en précision

Norme précise règles d'arrondi (au plus proche, vers 0, vers $\pm\infty$)

Il existe ± 0, $\pm\infty$

Nombres dénormalisés (entre 0 et $x_{\min}$)

`NaN` = « Not a Number » pour $0/0$, ∞/∞, fonction `isnan(x)`

Plan

Calcul sur les nombres flottants

En général, le résultat exact d'une opération sur deux flottants n'est pas un flottant machine.

Exemple

En base 10 ou en base 2, avec 3 chiffres significatifs ($t = 3$) :

$0.586 + 0.728 = 1.314 =$
$10^1 \times 0.1314 \notin \mathcal{F}_{10}$

$$\begin{array}{rl} & (0.101)_2 \\ + & (0.110)_2 \\ \hline & (1.011)_2 \\ = & (0.1011)_2\, 2^1 \notin \mathcal{F}_2 \end{array}$$

Le dernier chiffre (en rouge) ne peut pas être pris en compte.

Soit une opération arithmétique notée "$*$" dans $\{+, -, \times, \backslash, \sqrt{}\}$.
Pour $(x, y) \in \mathcal{F}^2$, le résultat du calcul $x * y$ n'est pas dans $\mathcal{F}$.

Propriétés de l'arithmétique flottante

Soit une opération arithmétique notée "$*$" dans $\{+, -, \times, \backslash, \sqrt{}\}$.
Pour $(x, y) \in \mathcal{F}^2$, le résultat du calcul $x * y$ n'est pas dans $\mathcal{F}$.

Mais on peut raisonner sur les calculs flottants, car :

Axiome

Pour $(x, y) \in \mathcal{F}^2$, le résultat flottant du calcul $x * y$ est noté

$$\text{“}(x \circledast y)\text{”} \in \mathcal{F}.$$

C'est l'arrondi de la valeur exacte de $x * y$,

$$(x \circledast y) = fl(x * y) \quad \in \mathcal{F}.$$

Propriétés de l'arithmétique flottante

L'arithmétique flottante est commutative et non associative.
Pour faire les calculs (explications simplifiées) :
Soit $x_1 = f_1 \times 10^{e_1}$ et $x_2 = f_2 \times 10^{e_2}$ dans $\mathcal{F}_{10}$, tel que $f_1 > f_2 > 0$.

1. on insère des 0 dans f_2 (décalage de virgule)
   $\longrightarrow$ x_1 et x_2 écrits avec le même exposant e_1.
2. on ajoute les mantisses (calcul exact).
3. on effectue un arrondi sur le résultat $\longrightarrow$ garder t chiffres.

Propriétés de l'arithmétique flottante

L'arithmétique flottante est commutative et non associative.
Pour faire les calculs (explications simplifiées) :
Soit $x_1 = f_1 \times 10^{e_1}$ et $x_2 = f_2 \times 10^{e_2}$ dans $\mathcal{F}_{10}$, tel que $f_1 > f_2 > 0$.

1. on insère des 0 dans f_2 (décalage de virgule)
   $\longrightarrow$ x_1 et x_2 écrits avec le même exposant e_1.
2. on ajoute les mantisses (calcul exact).
3. on effectue un arrondi sur le résultat $\longrightarrow$ garder t chiffres.

Exemple 1 : (arithmétique base 10, $t = 7$ chiffres) :
$a = 0.1234567,\ b = 0.4711325 \cdot 10^4,\ c = -b$
$b \oplus c = 0, \quad (a \oplus (b \oplus c)) = a = 0.1234567$

$$\begin{array}{rl} & 0.47113250000 \quad 10^4 \\ + & 0.00001234567 \quad 10^4 \\ \hline & 0.47114484567 \quad 10^4 \end{array} \qquad \begin{array}{rl} & 0.4711448 \quad 10^4 \\ - & 0.4711325 \quad 10^4 \\ \hline & 0.0000123 \quad 10^4 \end{array}$$

$(a \oplus b) = 0.4711448\, 10^4, \quad (a \oplus b) \oplus c = 0.123$

Soustraction de deux nombres voisins

Exemple 2 : Soustraction de deux nombres voisins

$a = 0.1234567,\ b = 0.1234560$, $a \ominus b = 0.7\,10^{-6}$ (exact).
Si a et b sont connus à 6 chiffres près, $a \ominus b$ n'a qu'un chiffre significatif : révèle une perte de précision dans un calcul précédent.

Annulation destructrice

Exemple 3 : $a = 123456$, $b = 12.3456$, $c = 123450$, arithmétique (décimale) avec 6 chiffres. Calcul de $a + b - c = 18.3456$ (résultat exact).

$$\begin{array}{rr} & 123456.0000 \\ + & 12.3456 \\ \hline & 123468.\color{red}{0000} \end{array} \qquad \text{puis} \qquad \begin{array}{rr} & 123468. \\ - & 123450. \\ \hline & 18. \end{array}$$

Annulation destructrice

Exemple 3 : $a = 123456$, $b = 12.3456$, $c = 123450$, arithmétique (décimale) avec 6 chiffres. Calcul de $a + b - c = 18.3456$ (résultat exact).

$$\begin{array}{rr} & 123456.0000 \\ + & 12.3456 \\ \hline & 123468.0000 \end{array} \quad \text{puis} \quad \begin{array}{rr} & 123468. \\ - & 123450. \\ \hline & 18. \end{array}$$

Annulation destructrice : seulement deux chiffres exacts. Erreur d'arrondi dans la première opération, la seconde est exacte. L'annulation révèle une perte d'information précédente (même résultat pour $b \in [11.5, 12.5[$).

Annulation destructrice

Exemple 3 : $a = 123456$, $b = 12.3456$, $c = 123450$, arithmétique (décimale) avec 6 chiffres. Calcul de $a + b - c = 18.3456$ (résultat exact).

$$\begin{array}{r} 123456.0000 \\ +\quad 12.3456 \\ \hline 123468.\color{red}{0000} \end{array} \qquad \text{puis} \qquad \begin{array}{r} 123468. \\ -\quad 123450. \\ \hline 18. \end{array}$$

Annulation destructrice : seulement deux chiffres exacts. Erreur d'arrondi dans la première opération, la seconde est exacte. L'annulation révèle une perte d'information précédente (même résultat pour $b \in [11.5, 12.5[$).

Autre ordre

$a \ominus c = 6$, puis $b \oplus (a \ominus c) = 18.3456$, exact.

Équations du second degré

Calcul des racines de $x^2 - 2px + 1$, quand $p \gg 1$ (ex : $p = 10^7$)

Équations du second degré

Calcul des racines de $x^2 - 2px + 1$, quand $p \gg 1$ (ex : $p = 10^7$)

Algorithme 1

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = p - \sqrt{p^2 - 1}$$

Algorithme 2

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = 1/(p + \sqrt{p^2 - 1})$$

Équations du second degré

Calcul des racines de $x^2 - 2px + 1$, quand $p \gg 1$ (ex : $p = 10^7$)

Algorithme 1

$$x^+ = p + \sqrt{p^2 - 1}$$

$$x^- = p - \sqrt{p^2 - 1}$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$

$$x^- = 5.0291419029236\,10^{-8}$$

Algorithme 2

$$x^+ = p + \sqrt{p^2 - 1}$$

$$x^- = 1/(p + \sqrt{p^2 - 1})$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$

$$x^- = 5.0000000000000\,10^{-8}$$

Équations du second degré

Calcul des racines de $x^2 - 2px + 1$, quand $p \gg 1$ (ex : $p = 10^7$)

Algorithme 1

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = p - \sqrt{p^2 - 1}$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$
$$x^- = 5.0291419029236\,10^{-8}$$

Algorithme instable

Algorithme 2

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = 1/(p + \sqrt{p^2 - 1})$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$
$$x^- = 5.0000000000000\,10^{-8}$$

Algorithme stable

Équations du second degré

Calcul des racines de $x^2 - 2px + 1$, quand $p \gg 1$ (ex : $p = 10^7$)

Algorithme 1

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = p - \sqrt{p^2 - 1}$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$
$$x^- = 5.0291419029236\,10^{-8}$$

Algorithme instable

Algorithme 2

$$x^+ = p + \sqrt{p^2 - 1}$$
$$x^- = 1/(p + \sqrt{p^2 - 1})$$

Avec Scilab

$$x^+ = 2.0000000000000\,10^7$$
$$x^- = 5.0000000000000\,10^{-8}$$

Algorithme stable

Solutions exactes :
$x^+ = 1.99999999999995\,10^7$, $x^- = 5.00000000000001250\,10^{-8}$

Une suite curieuse (très simple)

$$u_{n+1} = \alpha u_n + \beta, \quad n = 0, 1, \ldots, \qquad u_0 \text{ donné}$$

Solution : $u_n = \alpha^n u_0 + \frac{\alpha^n - 1}{\alpha - 1}\beta$.
On prend : $\alpha = 4,\ \beta = -1$: $u_n = 1/3 + 4^n(u_0 - 1/3)$.
Si $u_0 = 1/3$, alors la suite est constante : $u_n = 1/3, \forall n$. Pourtant...

Une suite curieuse (très simple)

$$u_{n+1} = \alpha u_n + \beta, \quad n = 0, 1, \ldots, \qquad u_0 \text{ donné}$$

Solution : $u_n = \alpha^n u_0 + \frac{\alpha^n - 1}{\alpha - 1}\beta$.
On prend : $\alpha = 4,\ \beta = -1$: $u_n = 1/3 + 4^n(u_0 - 1/3)$.
Si $u_0 = 1/3$, alors la suite est constante : $u_n = 1/3, \forall n$. Pourtant...

```
0  0.333333333333          24   0.328125
1  0.333333333333          25   0.3125
2  0.333333333333          26   0.25
      ....                 27   0.0
11  0.333333333255         28  -1.0
    .....                  29  -5.0
23  0.33203125             30  -21.0
```

Une suite curieuse (très simple)

$$u_{n+1} = \alpha u_n + \beta, \quad n = 0, 1, \ldots, \qquad u_0 \text{ donné}$$

Solution : $u_n = \alpha^n u_0 + \frac{\alpha^n - 1}{\alpha - 1}\beta$.
On prend : $\alpha = 4,\ \beta = -1$: $u_n = 1/3 + 4^n(u_0 - 1/3)$.
Si $u_0 = 1/3$, alors la suite est constante : $u_n = 1/3, \forall n$. Pourtant...

```
0  0.333333333333          24   0.328125
1  0.333333333333          25   0.3125
2  0.333333333333          26   0.25
      ....                 27   0.0
11  0.333333333255         28  -1.0
    .....                  29  -5.0
23  0.33203125             30  -21.0
```

Si $u_0 = 1/3(1 - \delta)$ avec $\delta \approx \varepsilon_{\text{mach}}$, alors $u_n = 1/3(1 - 4^n\delta) \to -\infty$!!

Plan

À faire pour la prochaine fois : cours

- Cours
  - travailler le chapitre 1 : *Introduction au calcul flottant*.
    - Tout. Faire les exercices d'application du cours (pas les exercices de TD).
  - réviser l'algèbre. Chapitre 0 : *Algèbre linéaire*.

    0.1 Espace vectoriel,
    0.2 Applications linéaires,
    0.3 Matrices,
    0.5 Systèmes linéaires.

Analyse pour la suite curieuse

$u_n = \frac{v_{n+1}}{v_n}$, avec $v_{n+1} = 111 v_n - 1130 v_{n-1} + 3000 v_{n-2}$.

Racines caractéristiques : $100, 6, 5$, solution générale

$$u_n = \frac{\alpha\, 100^{n+1} + \beta\, 6^{n+1} + \gamma\, 5^{n+1}}{\alpha\, 100^n + \beta\, 6^n + \gamma\, 5^n}$$

CI choisies pour que $\alpha = 0$, la suite devrait converger vers 6, mais numériquement instable.

Application : dérivation par différences finies

Mission

Calculer la dérivée d'une fonction $f : I \subset \mathbb{R} \to \mathbb{R}$ par la formule

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

dans le cas où f est calculée en arithmétique flottante.

Pourquoi f trop compliquée pour dérivée analytique, ou connue seulement en des points discrets.

Application : dérivation par différences finies

Mission

Calculer la dérivée d'une fonction $f : I \subset \mathbb{R} \to \mathbb{R}$ par la formule

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

dans le cas où f est calculée en arithmétique flottante.

Pourquoi f trop compliquée pour dérivée analytique, ou connue seulement en des points discrets.

Hypothèses On calcule $\tilde{f}(x)$, avec
$\tilde{f}(x) = f(x)(1+\delta(x)),\ |\delta(x)| \leq \varepsilon_{\text{mach}},$ f dérivable 2 fois.

Application : dérivation par différences finies

Mission

Calculer la dérivée d'une fonction $f : I \subset \mathbb{R} \to \mathbb{R}$ par la formule

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

dans le cas où f est calculée en arithmétique flottante.

Pourquoi f trop compliquée pour dérivée analytique, ou connue seulement en des points discrets.

Hypothèses On calcule $\tilde{f}(x)$, avec
$\tilde{f}(x) = f(x)(1 + \delta(x)),\ |\delta(x)| \leq \varepsilon_{\text{mach}},$ f dérivable 2 fois.

Questions Comment choisir h ? comment quantifier l'approximation

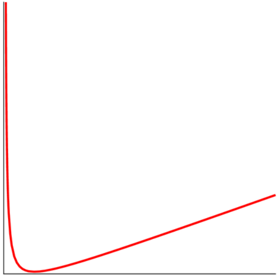
Dérivation (2)

$$E = \left|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\right| \leq \frac{2}{h} M_0\, \varepsilon_{\text{mach}} + \frac{h}{2} M_2$$

Minimiser la différence :

$$h_{\text{opt}} = 2\sqrt{\frac{M_0 \varepsilon_{\text{mach}}}{M_2}}$$

$$E_{\min} = 2\sqrt{M_0 M_2 \varepsilon_{\text{mach}}}$$



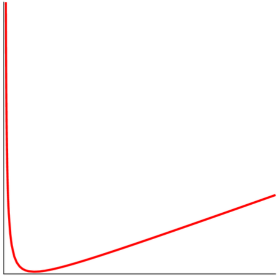
Dérivation (2)

$$E = \left|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\right| \leq \frac{2}{h} M_0\, \varepsilon_{\text{mach}} + \frac{h}{2} M_2$$

Minimiser la différence :

$$h_{\text{opt}} = 2\sqrt{\frac{M_0 \varepsilon_{\text{mach}}}{M_2}}$$

$$E_{\min} = 2\sqrt{M_0 M_2 \varepsilon_{\text{mach}}}$$



Conclusions

- On ne peut pas « faire tendre h vers 0 »

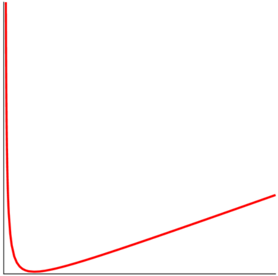
Dérivation (2)

$$E = \left|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\right| \leq \frac{2}{h} M_0 \, \varepsilon_{\text{mach}} + \frac{h}{2} M_2$$

Minimiser la différence :

$$h_{\text{opt}} = 2\sqrt{\frac{M_0 \varepsilon_{\text{mach}}}{M_2}}$$

$$E_{\min} = 2\sqrt{M_0 M_2 \varepsilon_{\text{mach}}}$$



Conclusions

- On ne peut pas « faire tendre h vers 0 »
- Valeur optimale de $h \approx \sqrt{\varepsilon_{\text{mach}}}$

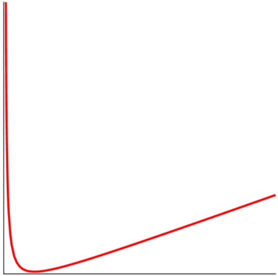
Dérivation (2)

$$E = \left|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\right| \leq \frac{2}{h} M_0\, \varepsilon_{\text{mach}} + \frac{h}{2} M_2$$

Minimiser la différence :

$$h_{\text{opt}} = 2\sqrt{\frac{M_0 \varepsilon_{\text{mach}}}{M_2}}$$

$$E_{\min} = 2\sqrt{M_0 M_2 \varepsilon_{\text{mach}}}$$



Conclusions

- On ne peut pas « faire tendre h vers 0 »
- Valeur optimale de $h \approx \sqrt{\varepsilon_{\text{mach}}}$
- f' connu avec la moitié du nbre de chiffres significatifs de f.

Dérivation (3)

Exemple

$f(x) = \exp(x), \; x = 0 \qquad (f'(0) = \exp(0) = 1).$

x	$\tilde{f}'(x)$	erreur
10^{-03}	1.000500166708385	$5.0017\,10^{-04}$
10^{-04}	1.000050001667141	$5.0002\,10^{-05}$
10^{-05}	1.000005000006965	$5.0000\,10^{-06}$
10^{-06}	1.000000499962184	$4.9996\,10^{-07}$
10^{-07}	1.000000049433680	$4.9434\,10^{-08}$
10^{-08}	0.999999993922529	$-6.0775\,10^{-09}$
10^{-09}	1.000000082740371	$8.2740\,10^{-08}$
10^{-10}	1.000000082740371	$8.2740\,10^{-08}$
10^{-11}	1.000000082740371	$8.2740\,10^{-08}$
10^{-12}	1.000088900582341	$8.8901\,10^{-05}$
10^{-13}	0.999200722162641	$-7.9928\,10^{-04}$
10^{-14}	0.999200722162641	$-7.9928\,10^{-04}$
10^{-15}	1.110223024625157	$1.1022\,10^{-01}$