2. Représentat° des nombres, Erreurs d'arrondi

Sources d'erreurs : données, méthode de calcul, arrondi (précision finie)

A représentat° de nombres en base qcq : $\forall x \in \mathbb{R}$: $x = \sigma B^N \sum_{1}^{\infty} x_{\nu} B^{-\nu}$ $\sigma \in \{+1, -1\}$, $x_\nu \in \{0, 1, \ldots B-1\}$

unique (si $\forall n \in \mathbb{N}\ \exists \nu \geq n : x_{-\nu} \neq B-1$)

N : exposant
(x_ν) : mantisse

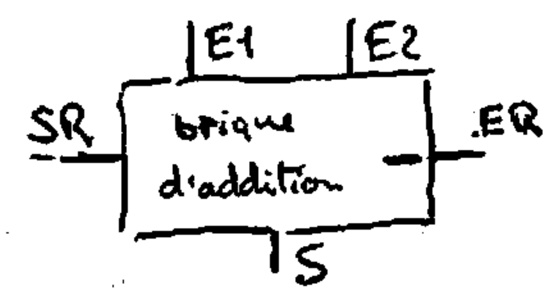
numérat°			
	binaire	$B=2$	0,1
	octale	8	0,1, .. , 7
	décimale	10	0,1, . . . , 9
	hexadécimale (sic)	16	0,1, .. 9,10, A,B, .., F

exemples $3_{10} = 11_2$
$9_{10} = 9_{16} = 11_8 = 1001_2$

cours ②

cours ③
↓ 3/10/2005

calcul binaire : $0+0 = 0$
$0+1 = 1$
$1+1 = 10$
$1+0 = 1$
suffit : on cascade



E1	E2	ER	S	SR
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

131072
262144
524288
1048576
2097152
4194304
8388608
16777216
33554432
67108864
134217728
268435456
536870912
1073741824

$2147483648 = 2^{31}$
$4294967296 = 2^{32}$

B calcul à virgule fixe

$$x = \sigma B^{N+t} \sum_{1}^{t} x_\nu B^{-\nu}$$

util courante : N ~~donné~~ fixé $\rightsquigarrow$ ne consomme pas de mémoire, on ne stocke que les chiffres (et σ)

: entiers $x = \sigma \sum_0 x_\nu B^\nu$ « int » en C : 32 bits -2^{31} ... $+2^{31} \sim 2{,}1 \cdot 10^9$ 2147483647

inconvenient : inadapté à grandeurs sur plusieurs décades

limites $\frac{x_{MAX}}{x_{MIN}} = B^t$ $\sim 2^{32}$, $x_{MAX} = B^{N}\cancel{B}^{N-T}$ $\sim 2^{32}-1$, $x_{min} = B^{N-t}$ ~ 1

nombres négatifs : $x \mapsto B^N + x$ ex: char ($N=8, B=2$) $\in \{-128 .. 127\}$ $\to 2^8 = 256$

$-1 + 1 = B^N \equiv 0 \pmod{B^N}$

$-128_{10} = 10000000_2$

$\vdots$

$-2_{10} = 256 - 2 \pmod{2^8} = 254 = 11111110_2$

$-1_{10} = 255_{10} = 11111111_2$

$0 = 0$

$1_{10} = 00000001_2$

$\vdots$

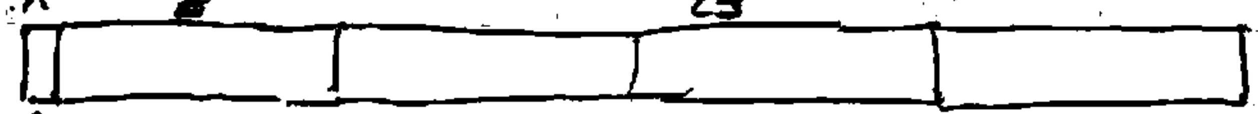
$127 = 01111111_2$

C nombres à virgule flottante

on stocke N et (x_ν)

exemple : type « float » C sur Intel Pentium : 32 bits

1 8 23



$\sigma \binom{0 : +}{1 : -}$

inconvenient : compromis de stockage, moins de place pour mantisse

avantage : $|x|_{MIN} = B^{N_{min} - T}$ $\sim 2^{-151} \sim 10^{-45}$, $|x|_{MAX} = B^{N_{max}}(1 - B^{-T})$ $\sim 2^{127} \sim 10^{40}$, $\frac{|x|_{MAX}}{|x|_{MIN}} = B^{N_{max} - N_{min}}(B^T - 1)$

float 40 $\to N_{max} - N_{min} = 255, T = 23$ $\to 2^{278}$

D Règle d'arrondi

x nombre réel , $\tilde{x}$ approx : $x - \tilde{x}$ erreur absolue

$-\varepsilon = \frac{x - \tilde{x}}{x}$ (ou $\frac{x - \tilde{x}}{\tilde{x}}$) erreur relative (p/rà $\tilde{x}$)

supposons qu'il n'y ait pas de dépassement pour l'exposant , B pair

$$x = \sigma B^N \sum_1^{\infty} x_{-\nu} B^{-\nu} \rightsquigarrow \mathrm{Arr}(x) = \begin{cases} \sigma B^N \sum_1^t x_{-\nu} B^{-\nu} & \text{si } x_{-t-1} < \frac{B}{2} \\ \sigma B^N \sum_1^t x_{-\nu} B^{-\nu} + B^{-t} & \text{si } x_{-t-1} \geq \frac{B}{2} \end{cases}$$

Arr : arrondi à t chiffres (t : longueur de mantisse)

$$\frac{\mathrm{Arr}_t(x) - x}{\mathrm{Arr}_t(x)} \leq 0.5\, B^{-t+1} \qquad \frac{\mathrm{Arr}_t(x) - x}{x} \leq \underbrace{0.5\, B^{-t+1}}_{\text{précision du calcul à virg. flottante}} \qquad \mathrm{Arr}_t(x) - x \leq 0.5\, B^{N-t}$$

ex $t = 23 \rightsquigarrow 2^{-24} \sim 10^{-7}$

$x = \sigma \cdot 10^N \cdot m$, $0.1 \leq m < 1$ $\wedge$ $\tilde{x} = \sigma 10^N \tilde{m}$ une approx.

nombre de chiffres significatifs de $\tilde{x} = \max \{ t \in \mathbb{Z} \mid |m - \tilde{m}| \leq 0.5 \cdot 10^{-t+1} \}$

E calcul en virgule flottante

pb. : si a et b sont nombres à v.f. de long. de mantisse t,

$a \square b$ ($\square \in \{+ - \cdot :\}$) n'en est pas forcément

3/10 cours ↑ ↓ cours ④ 10/10

(ex. $0.123 \cdot 10^4 + 0.456 \cdot 10^{-3}$ (donc $t=3$) $= 0.1230000456 \cdot 10^4$ ($t=10$)

⇝ le résultat doit être arrondi à la fin ⇝ $Fl_t(a \square b)$

Habituellement les ordinateurs sont construits/progr. en sorte que

$$Fl_t(a \square b) = Arr_t(a \square b)$$

$$\rightsquigarrow \quad Fl_t(a \square b) = (a \square b)(1+\varepsilon) \quad , |\varepsilon| < \tau$$

$a = \overset{Fl}{Arr_t}(x)$ ex $0.100 \cdot 10^1 = Arr_3(0.9995 \cdot 10^0)$

$b = Arr_t(y)$ $-0.998 \cdot 10^0 = Arr_3(-0.9984 \cdot 10^0)$

$$\Rightarrow |\underset{\overset{\shortparallel}{0.200 \cdot 10^{-2}}}{Fl_3(Fl_3(x) \square Fl_3(y))}| = |Fl_3(x) \square Fl_3(y)| \quad (1+\tau)$$

$$= (x(1+\varepsilon_x) \square y(1+\varepsilon_y))(1+\varepsilon)$$

$\square \in + -$ ⇝ $= \overset{0.11 \cdot 10^{-2}}{\overset{\shortparallel}{x \pm y}} + x(\varepsilon + \varepsilon_x(1+\varepsilon)) \pm y(\varepsilon + \varepsilon_y(1+\varepsilon)) =: (x \pm y)(1+\varsigma)$
$= (x \pm y) + F$

ex $F = 0.9995 \cdot 0.5003 \cdot 10^{-3} + 0.9984 \cdot 0.4006 \cdot 10^{-3} = 0.9000 \cdot 10^{-3}$

$\varsigma = \frac{F}{x \pm y} = 0.82$ (!)

$$= x \pm y + \cancel{(x \pm y)}\,\varepsilon + (\varepsilon_x x + \varepsilon_y y)(\cancel{1+\varepsilon})$$

$$|\varsigma| = \frac{|x(\varepsilon + \varepsilon_x(1+\varepsilon)) \pm y(\varepsilon + \varepsilon_y(1+\varepsilon))|}{|x \pm y|} \leq \left| \varepsilon + \frac{x\varepsilon_x (1+\varepsilon) \pm y \varepsilon_y (1+\varepsilon)}{x \pm y} \right|$$

$$\leq |\varepsilon| + |1+\varepsilon| \frac{|\varepsilon_x x| + |\varepsilon_y y|}{|x \pm y|} \leq \tau + (1+\tau)\tau \frac{|x|+|y|}{|x \pm y|}$$

$$\simeq \tau\left(1 + \frac{|x|+|y|}{|x \pm y|}\right) \quad \rightsquigarrow \quad 2 \text{ cas} \quad \varsigma \simeq 2\tau \text{ et } \varsigma \gg \tau$$

$\square \in \cdot :$ ⇝ $= x \square y + x \square y \cdot \varepsilon + (1+\varepsilon)(x\varepsilon_x \square y + \varepsilon_y x \square y + \varepsilon_x \varepsilon_y x \square y)$

$$= x \square y\,(1 + \varepsilon + (1+\varepsilon)(\varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y))$$

$$\varsigma = \varepsilon + (1+\varepsilon)(\varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y) \leq \tau + (1+\tau)(2\tau + \tau^2) = 3(\tau + \tau^2) + \tau^3 \sim 3\tau$$

F Évaluat° stable / instable

Toute formule compliquée se réduit in fine à une suite d'op. élem.

$\leadsto$ il faut s'assurer que chaque pas soit numériquement stable.

$\Rightarrow$ Importance du choix de l'algo !

ex. : $ax^2+bx+c=0$, soit $|4ac|<b^2$

$$\leadsto x_1 = \frac{1}{2a}\left(-b-\operatorname{sgn}(b)\sqrt{b^2-4ac}\right), \quad x_2 = \frac{1}{2a}\left(-b-\operatorname{sgn}(b)\sqrt{b^2-4ac}\right)$$
$$= -\frac{\operatorname{sgn}(b)}{2a}\left(|b|-\sqrt{b^2-4ac}\right)$$

$\leadsto x_2$ numériquement instable

mieux vaut utiliser $x_2 \cdot x_1 = \frac{c}{a} \leadsto x_2 = \frac{2c}{-b-\operatorname{sgn}(b)\sqrt{b^2-4ac}}$

G Condition d'un pb

$$\varphi : D \to \mathbb{R}, \quad D \subset \mathbb{R}^n, \quad \varphi \in C^2(D,\mathbb{R})$$
$$x \mapsto \varphi(x) = y$$

$$\delta y \approx \sum \frac{\partial \varphi}{\partial x_\nu}(x)\,(\hat{x}_\nu - x_\nu)$$

err. relative $\frac{\delta y}{y} = \sum \underbrace{\frac{x_\nu}{\varphi(x)}\frac{\partial \varphi}{\partial x_\nu}(x)}_{\text{coeff. de condition}} \frac{\tilde{x}_\nu - x_\nu}{x_\nu}$

≤ 1 bien conditionné
> 1 mal "

ex : $\varphi(x) = ax \leadsto \frac{x}{\varphi(x)}\left.\frac{\partial\varphi}{\partial x}\right|_x = 1$

$\varphi(x) = a+x \leadsto \frac{x}{a+x}$ $\begin{matrix} <1 \\ >1 \end{matrix}$ $x \cdot a > 0$

$\leadsto$ $\cdot$, $:$ opérat° bien conditionnées

+, − bien " si termes m̂ signe / signes différents
mal " " signes diff / m̂ signe

<u>ex</u> $x^2+2px-q=0$, $p,q>0 \wedge p \gg q$

calc. plus grand zéro $\varphi(p,q) = -p+\sqrt{p^2+q}$

sol° : • $s = p^2$
$t = s+q$
$u = \sqrt{t}$
+ 2 méthodes : ① $y = \varphi_1(u) = -p+u$
ou ② $v = -p-u$ et $y = \varphi_2(v) = \frac{-q}{v}$

$$\frac{\delta y}{y} = \frac{p}{\varphi(p,q)}\frac{\partial\varphi}{\partial p}\varepsilon_p + \frac{q}{\varphi(p,q)}\frac{\partial\varphi}{\partial q}\varepsilon_q = -\frac{p}{(p^2+q)^{1/2}}\varepsilon_p + \frac{p+(p^2+q)^{1/2}}{2(p^2+q)^{1/2}}\varepsilon_q \leadsto \text{b.c.}$$

mais méth. 1 : $\frac{\delta y}{y} = \frac{u}{-p+u}\varepsilon_u = \frac{1}{q}\left(p(p^2+q)^{1/2}+p^2+q\right)\varepsilon_u > \frac{2p^2}{q}\varepsilon_u \gg \varepsilon_u \leadsto$ m.c. !

méth. 2 : $\frac{\delta y}{y} = -\frac{(p^2+q)^{1/2}}{p+(p^2+q)^{1/2}}\varepsilon_u \leadsto$ b.c.

suite cours 4 ph332
10/10/2005

conclusion : coeffs de condit° renseignent sur la stab. num.

Il peut y avoir amplif. et atténuat^e des erreurs

coeffs < 1 : pour un bon choix des étapes de calcul $\rightsquigarrow$ stabilité
mauvais choix : quand même instab.

coeff > 1 : $\nexists$ méthode stable, tj instable

autres techniques d'analyse d'erreur

calcul direct à priori $\rightsquigarrow$ cas le pire
- analytiquement
- numériquement : calcul d'intervalles

calcul à postériori : quelles erreurs des données initiales auraient produit le résultat obtenu ?