

Anithmétique Flottante

$F(B, p, e_{\min}, e_{\max})$ <ul style="list-style-type: none"> <li>- <math>B</math>: entier de la base</li> <li>- <math>p</math>: nb chiffres mantisse</li> <li>- <math>e_{\min}</math>: exp min</li> <li>- <math>e_{\max}</math>: exp max</li> </ul>	$x = s \left( \sum_{i=0}^{p-1} c_i B^{-i} \right) B^e$ $(c_0, c_1, \dots, c_{p-1})_B$	Si $s \in F$ : $\begin{cases} s = \pm 1 : \text{signe} \\ 0 \leq c_i \leq B-1 \forall i \\ e_{\min} \leq e \leq e_{\max} \end{cases}$
--	--	--

- Notat° normalisé:  $c_0 \neq 0$
- Notat° dénormalisé:  $c_0 = 0$
- M, plus grand mb positif de ff:  $M = \left( \sum_{i=0}^{p-1} (B-1) B^{-i} \right) B^{e_{\max}} = (1 - B^{-p}) B^{e_{\max}+1}$
- m, plus petit mb dénormalisé positif:  $m = (1,00\dots 0)_B B^{e_{\min}} = B^{e_{\min}}$
- N, plus petit mb dénormalisé positif:  $N = (0,00\dots 1)_B B^{e_{\min}} = B^{1-p+e_{\min}}$
- $f_f(x)$  est l'approximat° de  $x$  ds ff  $\rightarrow$  flottant le plus proche si égale distance dominé chiffre point
- Si  $|f_f(x)| > M$ : alors  $f_f(x) = \text{sign}(x) \cdot \text{inf} \Rightarrow$  Overflow

Suivi d'overflow:  $\tilde{M} = M + B^{e_{\max}+1}$

Si  $|x| \in [m; \tilde{M}]$ , alors:  $e_n = \frac{|f_f(x) - x|}{|x|} \leq U = \frac{B^{1-p}}{2}$  et  $e_a = |f_f(x) - x| \leq \frac{1}{2} B^{1-p+e}$

$f_f(x) = x(1+\epsilon)$ , avec  $|\epsilon| \leq U$

$x_c = (a \otimes b) \odot (c \otimes d)$

$a \otimes b = (a \times b)(1+\epsilon_1) \quad |\epsilon_1| \leq U$

$c \otimes d = (c \times d)(1+\epsilon_2) \quad |\epsilon_2| \leq U$

$(a \otimes b) \odot (c \otimes d) = \frac{(a \otimes b)}{(c \otimes d)} (1+\epsilon_3) \quad |\epsilon_3| \leq U$

$\Rightarrow x_c = \underbrace{\frac{a \times b}{c \times d}}_{x} (1+\epsilon_1)(1+\epsilon_3) \quad |S| \leq \frac{3U}{1-3U}$

$|S| \leq \frac{3U}{1-3U} \times \text{puiss } 2 \text{ ou } (\text{puiss } 2)^{\otimes \text{coeffs}}$

**epsilon machine**

$|S| \leq \frac{mU}{1-mU}$

$f_f(10, 3, -3, 3)$   
 $f_f(1, 015) = 1,02$   
 $f_f(1, 15 \cdot 10^{-4}) = f_f(0, 115 \cdot 10^3) = 0,12 \cdot 10^{-3}$   
 $f_f(9, 985 \cdot 10^3) = 998 \cdot 10^2$

NaN:  $\frac{0}{0}$ ,  $\text{Inf} \oplus \text{Inf}$ ,  $\text{Inf} \odot \text{Inf}$ ,  $0 \otimes \text{Inf}$

Systèmes linéaires

$\bullet$ Si A est une matrice $(m, m)$ , $A_i$ désigne la matrice $(1, m)$ et $A^T$ : $(m, 1)$	$\bullet$ $x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \dots + x_m \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \sum_{k=1}^m x_k e_k$ Base canonique ( $e_1, e_2, \dots, e_m$ )
---	---

$\bullet$  Si  $A: (m, m)$  et  $B: (m, p)$ :  $AB: (m, p)$  et  $(AB)_{i,j} = \sum_{k=1}^m a_{ik} b_{kj} \quad \forall (i,j) \in [(1,m)] \times [(1,p)]$

$\bullet$  Si  $A: (m, m)$ , alors  $A^T: (m, m)$  et  $(A^T)_{i,j} = a_{j,i}$

$\bullet$  Si  $x$  et  $y$  2 vecteurs de  $\mathbb{R}^n$ , alors:  $(x \mid y) := x_1 y_1 + \dots + x_m y_m = \sum_{k=1}^m x_k y_k$ .  $(x \mid y) = x^T y = y^T x$

$\bullet$  Résoudre un système triangulaire  $Tx = b$

Algo:  
 pour  $i = m, m-1, \dots, 1$   
 $x_i \leftarrow (b_i - \sum_{j=i+1}^m t_{i,j} x_j) / t_{i,i}$

Nombre opérations =  $\frac{m(m-1)}{2}$  multiplicat° +  $\frac{(m-1)m}{2}$  addit° + m divisions.

$\bullet$  Triangulation d'un système linéaire:

$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,m}x_m = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,m}x_m = b_2 \\ \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \dots + a_{m,m}x_m = b_m \end{cases}$

Partir des coeff sur la diag pour chaque colonne et faire en sorte d'annuler les lignes du dessous