

Erreurs de représentation des nombres réels

Virgule fixe et virgule flottante simple (sans biais)

But

On veut estimer deux erreurs lorsqu'on représente un nombre réel x :

$$\text{erreur absolue } \Delta_x = |x - x_{\text{rep}}|, \quad \text{erreur relative } \frac{\Delta_x}{|x|} \quad (x \neq 0).$$

On suppose **arrondi au plus proche**.

1 Virgule fixe : 4 bits entiers (complément à deux) + 4 bits fraction

Idée

On place la virgule *fixe* après 4 bits de fraction. Le **pas** vaut

$$\delta = 2^{-4} = \frac{1}{16} = 0,0625.$$

Plage des valeurs possibles

- Avec 4 bits pour la partie entière en complément à deux, la plus petite valeur entière est -8 (1000_2) et la plus grande est $+7$ (0111_2).
- Chaque valeur entière est ensuite complétée de 4 bits de fraction. Cela correspond à des pas de $2^{-4} = 1/16$.
- La valeur la plus grande est donc $7 + 15/16 = 7,9375$.

$\text{Plage} = [-8, 7,9375]$

Bornes d'erreurs

$\Delta_x \leq \frac{\delta}{2} = \frac{1}{32} = 0,03125$

$\frac{\Delta_x}{|x|} \leq \frac{1}{32|x|} \quad (x \neq 0)$

Commentaires simples :

- L'erreur **absolue** est *uniforme* (même borne partout dans la plage).
- L'erreur **relative** devient grande près de $x = 0$.

Exemples rapides

- $x = 2,27 \Rightarrow x_{\text{rep}} = 2,25$. Alors $\Delta_x = 0,02 \leq 0,03125$ et $\Delta_x/|x| \approx 0,88\%$.
- $x = 0,03 \Rightarrow x_{\text{rep}} = 0$. Alors $\Delta_x = 0,03$ et $\Delta_x/|x| = 1$.

2 Virgule flottante simple : (signe / exposant / mantisse) = (1 / 3 / 4) sans biais

Format

- 1 bit de signe $s \in \{0, 1\}$.
- 3 bits d'**exposant** en complément à deux $\Rightarrow e \in \{-4, -3, -2, -1, 0, 1, 2, 3\}$.
- 4 bits de **mantisse fractionnaire** $f \in \{0, \dots, 15\}$.

$$x = (-1)^s \cdot \left(1 + \frac{f}{16}\right) \cdot 2^e$$

Plage positive (valeurs normales) :

$$x_{\min}^+ = 1,0000_2 \cdot 2^{-4} = \frac{1}{16} = 0,0625, \quad x_{\max} = 1,1111_2 \cdot 2^3 = \frac{31}{16} \cdot 8 = 15,5.$$

Espacement et erreurs

À exposant e fixé, deux nombres consécutifs diffèrent de

$$\delta = \frac{2^e}{16}.$$

Avec l'arrondi au plus proche :

$$\Delta_x \leq \frac{1}{2}\delta = \frac{2^e}{32} = \frac{|x|}{32(1+f/16)} \leq \frac{|x|}{32}$$

Donc, très simplement,

$$\frac{\Delta_x}{|x|} \leq \frac{1}{32} \approx 3,125\% \quad (x \neq 0)$$

Interprétation :

- L'erreur **relative** est (au pire) $\approx 3,1\%$, quelle que soit la taille de x dans la plage.
- L'erreur **absolue** grandit avec $|x|$ (car elle est proportionnelle à $|x|$).

Exemples rapides

- $x = 2,27$. Ici $2 \leq x < 4 \Rightarrow e = 1$. Alors $\delta = 2^1/16 = 0,125$, donc $\Delta_x \leq 0,0625$.
- $x = 0,08$. Ici $0,0625 \leq x < 0,125 \Rightarrow e = -3$. $\delta = 2^{-3}/16 = 1/128 = 0,0078125$, donc $\Delta_x \leq 0,00390625$.

Résumé en une ligne

$$\text{Fixe (4+4)} : \Delta_x \leq \frac{1}{32}, \quad \frac{\Delta_x}{|x|} \leq \frac{1}{32|x|}$$

$$\text{Flottant (1/3/4, sans biais)} : \frac{\Delta_x}{|x|} \leq \frac{1}{32}, \quad \Delta_x \leq \frac{|x|}{32}$$

Généralisation

Virgule fixe : n bits entiers (complément à deux) + m bits fraction

— Plage :

$$[-2^{n-1}, 2^{n-1} - 2^{-m}]$$

— Pas : $\delta = 2^{-m}$

— Erreurs :

$$\Delta_x \leq \frac{\delta}{2} = 2^{-(m+1)}, \quad \frac{\Delta_x}{|x|} \leq \frac{1}{2^{m+1} |x|}$$

Virgule flottante : 1 bit signe, n bits exposant, m bits mantisse (sans biais)

— Valeur représentée :

$$x = (-1)^s \cdot \left(1 + \frac{f}{2^m}\right) \cdot 2^e$$

avec e codé sur n bits en complément à deux, $f \in \{0, \dots, 2^m - 1\}$.

— Plage positive :

$$x_{\min}^+ = 1 \cdot 2^{-(2^{n-1})}, \quad x_{\max} = \left(2 - \frac{1}{2^m}\right) \cdot 2^{2^{n-1}-1}$$

(si on autorise tous les exposants de -2^{n-1} à $2^{n-1} - 1$).

— Écart entre deux nombres consécutifs δ :

$$\delta = \frac{2^e}{2^m}$$

— Erreurs :

$$\Delta_x \leq \frac{1}{2} \delta = \frac{2^e}{2^{m+1}}, \quad \frac{\Delta_x}{|x|} \leq \frac{1}{2^{m+1}}$$