Analyse Numérique

Analyse des erreurs

Laurence Lamoulie

EISTI

Ecole Internationale des Sciences du Traitement de l'Information

Exemple

Rappels : Passage du décimal au binaire

- Pour la partie entière
  $13/2 = 6$ reste $\mathbf{1}$
  $6/2 = 3$ reste $\mathbf{0}$
  $3/2 = \mathbf{1}$ reste $\mathbf{1}$
  donc on obtient : $\mathbf{13}_{10} = \mathbf{1101}_2$
- Pour la partie décimale
  $0.8125 * 2 = 1.625$ de partie entière $\mathbf{1}$
  $0.625 * 2 = 1.25$ de partie entière $\mathbf{1}$
  $0.25 * 2 = 0.5$ de partie entière $\mathbf{0}$
  $0.5 * 2 = 1$ de partie entière $\mathbf{1}$
  $0.0 * 2 = 0$ de partie entière 0
  donc obtient : $\mathbf{0.8125}_{10} = \mathbf{0.1101}_2$
- Au final on a par exemple : $\mathbf{13.8125}_{10} = \mathbf{1101.1101}_2$

Exemple (suite)

Passage en machine en simple précision

- Nombre positif : bit de signe $= 0$
- On passe à la notation avec exposant :
  $\mathbf{13.8125}_{10} = \mathbf{1101.1101}_2 = (\mathbf{1.1011101} * \mathbf{10}^{11})_2$
- Exposant : il vaut $11_2 = 3_{10}$
  Exposants disponibles : codage sur 8 bits donc variant de 0 à 255, soit 127 valeurs pour les exposants négatifs et 127 pour les exposants positifs.
  Conséquence : l'exposant est codé par $3 + 127 = 130 = \mathbf{10000010}_2$

Exemple (suite)

Passage en machine en simple précision

- Mantisse : on complète pour obtenir 23 bits et on a donc : **1(.)10111010000000000000000**
- Comme on a toujours un premier chiffre de mantisse (à gauche du . qui est omis) égal à 1, on ne l'écrit pas (bit caché) et on a la représentation définitive :
  0 10000010 10111010000000000000000

Certains nombres ne peuvent pas s'accomoder de ces conventions :

- 0 est codé par convention
  0 00000000 00000000000000000000000
  ce qui devrait correspondre à $(1.0 * 10^0)_2 = 1_{10}$

On est donc obligé de réserver l'exposant 00000000

Conséquences sur les nombres représentables

Plus grand et plus petit nombres normalisés

- Plus petit nombre normalisé :
  0 00000001 00000000000000000000000
  Il correspond à $1.0 * 2^{1-127} = 2^{-126} \approx 1.2 * 10^{-38}$ donc les exposants varient entre -126 et 127 et non pas -127 et 127.
- Plus grand nombre normalisé :
  0 11111110 11111111111111111111111
  Il correspond à
  $(1.11...1)_2 * 2^{254-127} = (2 - 2^{-23} * 2^{127} \approx 3.4 * 10^{38}$

Conséquences sur les nombres représentables

Plus grand et plus petit nombres normalisés

- Nombres spéciaux :utilisables par l'ordinateur seulement
  - Nombres sous normalisés
    Le plus petit nombre non nul stockable est :
    0 00000000 00000000000000000000001
    Il correspond $0.00..01 * 2^{-127} = 2^{-149}$
    En cas de résultat égal à un nombre sous-normalisé : *underflow*
  - Nombres sur normalisés : idem et *overflow*

Si la mantisse est nulle : on a représenté $\pm\infty$ sinon on a un *non-nombre* noté *NaN*

Arithmétique arrondie : Situations possibles

Un nombre x peut être :

- égal à un nombre normalisé
- entre le plus petit et le plus grand nombres normalisés mais pas égal à un nombre normalisé
- un nombre sous-normalisé
- une valeur infinie ($+\infty$ ou $-\infty$)
- NaN

Les normes définissent ce qui doit être fait dans chaque cas.

Définitions

Définition 1.3.1

La précision *eps* d'un ordianteur est le plus petit nombre positif normalisé tel que

$$1 + eps \neq 1$$

Fait 1.1

La précision *eps* d'une machine simple précision qui suit la norme IEEE-754 est

$$eps = 2^{-23} \approx 1.192 * 10^{-7}$$

Conséquences

Résultat

Si x est un nombre à représenter selon le standard IEEE754 alors on le remplacera soit par x_- soit par x_+. le nombre ainsi choisi est $m(x)$.
On a

$$|m(x) - x| \leq \frac{1}{2}\ eps.2^E$$

Eléments théoriques

Définitions

- **Erreur (resp. erreur absolue)**de représentation : Δx(resp. $|\Delta x|$)

$$\Delta x = m(x) - x \qquad (resp. |\Delta x| = |m(x) - x|)$$

- **Erreur relative** de représentation : $\iota(x)$

$$\iota(x) = \frac{\Delta x}{m(x)} = \frac{m(x) - x}{m(x)}$$

Eléments théoriques

Définitions - suite

- **Erreur relative (resp. erreur relative absolue)** de précision : $\eta(x)$ (resp. $|\eta(x)|$)

$$\eta(x) = \frac{\Delta x}{x} = \frac{m(x) - x}{x} = \frac{m(x)}{x} - 1$$

$$(resp. |\eta(x)| = \frac{|m(x) - x|}{|x|})$$

Conséquence

L'erreur relative de précision permet d'exprimer le nombre de précision machine *eps* par :

$$m(x) = x(1 + \eta(x))$$

Eléments théoriques

Théorème 1.4.1 : Erreur de précision relative

Sur un calculateur en base β avec une mantisse de p chiffres, l'erreur de précision relative est bornée pour tout $x \in \mathbb{R}$ par :

$$|\eta(x)| \leq \begin{cases} \beta^{1-p} & \text{si approximation par troncature} \\ \frac{\beta^{1-p}}{2} & \text{si approximation par arrondi} \end{cases}$$

Fait 1.2

L'erreur de précision relative selon le standard IEEE-754 est

$$|\eta(x)| \leq \begin{cases} 2^{1-p} & \text{si approximation par troncature} \\ 2^{-p} & \text{si approximation par arrondi} \end{cases}$$

si on utilise une représentation binaire

Références