Analyse Numerique

David Wiedemann

Table des matières

List of Theorems

Lecture 1: Representation de nombres en arithmetique finie Thu 03 Mar

1 Representation de nombres en arithmetique finie

Notons $\mathcal{F}(\beta, t, L, U)$ l'ensemble des nombres representables sous la forme $(-1)^s(0, \alpha_1 \dots \alpha_t)_\beta \beta^e$ ou e est l'exposant, $L \leq e \leq U, 0 \leq \alpha_i < \beta$, $\alpha_1, \dots, \alpha_t$ est la mantisse et s le signe.
Cette representation est la representation floating point.

1.1 Representation des nombres dans les ordinateurs

On appelle les nombres en double precision l'ensemble

$$\mathcal{F}(2, 53, -1021, 1024)$$

Bien que les valeurs maximales et minimales sont tres grandes ($2 \cdot 10^{-308}$ et $2 \cdot 10^{308}$), mais on en saute beaucoup.
Tous les nombres dans $\mathcal{F}$ sont de la forme $\frac{p}{2^n}, p \in \mathbb{N}$.
On regarde la distance entre deux nombres consecutifs de $\mathcal{F}$.
Pour un exposant fixe, $[2^p, 2^{p+1}]$, le premier nombre apres 2^p est

$$(0.10 \dots 01)2^{p+1} = 2^p + 2^{p+1-t}$$

Donc dans ce cas, on a que le spacing est donne par 2^{p-52}.

Remarque
Si on a que des entiers dans un intervalle $[\beta^p, \beta^{p+1}]$, alors $\beta^{p+1-t} = 1$.

1.2 Approximation de $\mathbb{R}$ dans $\mathcal{F}(2, 53, -1021, 1024)$

Soit $x \in \mathbb{R}$, on appelle $fl(x) \in \mathcal{F}(2, 53, -1021, 1024)$.
Notons $x = (-1)^s(0, \alpha_1 \dots \alpha_{t-1}\alpha_t\alpha_{t+1} \dots)\beta^e$, on definit alors

$$fl(x) = (-1)^s(0, \alpha_1 \dots \alpha_{t-1}\tilde{\alpha_t})\beta^e$$

on fait l'hypothese ici que au moins un des α_i est non nul.
On veut borner $|x - fl(x)| \leq \frac{1}{2}\text{spacing} = \frac{1}{2}\beta^{e-t}$.
Bien que l'erreur absolue est, en principe, grande, l'erreur relative sera bornee, on a en effet

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\beta^{e-t}\frac{1}{|x|} \leq \frac{1}{2}\beta^{1-t} (\simeq 10^{-16} \text{ dans notre systeme })$$

On appelle cette erreur la "machine precision" et on la note u

Proposition 2
On peut egalement ecrire que

$$x \in \mathbb{R} \quad fl(x) = x(1+\epsilon), |\epsilon| \leq u$$

1.3 Operations dans $\mathcal{F}$

Soit $x, y \in \mathbb{R}$, $x+y \mapsto fl[fl(x)+fl(y)]$, qu'elle est l'erreur relative commise ?

$$\frac{|fl[fl(x) + fl(y) - (x+y)|}{|x+y|}$$

En utilisant la proposition ci-dessus, notons $fl(x) = x(1+\epsilon_1), fl(y) = y(1+\epsilon_2)$, on a alors

$$|(x(1+\epsilon_1)+y(1+\epsilon_2))(1+\epsilon_3)-(x+y)|\cdot\frac{1}{|x+y|} \leq \frac{x\epsilon_1 + y\epsilon_2 + \epsilon_3(x+y) - (x+y)}{|x+y|}+\text{petit}$$

$$\leq (\frac{|x|}{|x+y|} + \frac{|y|}{|x+y|} + 1)u$$

On remarque que si $x > 0, y < 0$, il est possible de commettre une erreur tres grande.
On dit que la soustraction est une operation instable.

1.4 Parenthese sur le concept de stabilite

On veut resoudre $y = G(x)$.

Definition 1
La resolution de $y = G(x)$ est stable si une petite perturbation de x correspond a une petite perturbation de y, ie.

$$y + \delta y = G(x + \delta x)$$

On appelle alors le conditionnement absolu du probleme

$$\kappa_{abs} = \sup_{\delta x} \frac{\|\delta y\|}{\|\delta x\|}$$

Et on appelle perturbation relative du probleme

$$\kappa_{rel} = \sup_{\delta x} \frac{\|\delta y\| \,/\, \|y\|}{\|\delta x\| \,/ \delta x}$$