Chapitre 1

La représentation des nombres

1.1 Introduction

Nous représentons les nombres avec une notation positionnelle. Dans les ordinateurs nous utilisons la base $\beta = 2$. Chaque nombre représentable doit pouvoir s'écrire comme suite finie de 0 et de 1.

Notation 1. On note l'ensemble des nombres représentables $\mathcal{F}(\beta, t, L, U)$, chaque nombre non nul de cet ensemble s'écrit sous la forme

$$(-1)^s \cdot (0.\alpha_1 \dots \alpha_t)_\beta \cdot \beta^e.$$

Cette représentation s'appelle représentation en *virgule flottante*, on appelle e *l'exposant*, s le *signe* et $0 \leq \alpha_i < \beta$ sont des entiers qui forment la *mantisse*.

Exemple 1.1.1. On peut considérer $\mathcal{F}(10, 3, -2, 2)$, dans cet ensemble de nombres on a

$$-48.3 = (-1)^1 \cdot (0.438) \cdot 10^2.$$

Cependant d'autres nombres n'appartiennent pas à cet ensemble comme par exemple

$$3,141 = (-1)^0 \cdot (0.3141) \cdot 10^1.$$

1.2 Représentation des nombres dans l'ordinateur

Définition 1.2.1 (64-floating point representation)**.** Dans les ordinateurs l'ensemble des nombres représentables est $\mathcal{F}(2, 53, -1021, 1024)$. Il y a un bit de signe, 52 bits pour représenter

les α_i avec $\alpha_1 \neq 0$ et 11 bits pour l'exposant, 1 pour le signe et 10 pour sa valeur. Cette représentation s'appelle *64-floating point representation.*

Nous allons étudier la structure de cet ensemble. La plus petite valeur en valeur absolue est

$$x_{\min} = (0.100\ldots0) \cdot 2^L \sim 2 \cdot 10^{-308}$$

et la plus grande est

$$x_{\max} = (0.11\ldots1) \cdot 2^U \sim 1.8 \cdot 10^{308}.$$

Remarque. Tous les nombres dans $\mathcal{F}$ sont de la forme $\frac{p}{2^n},\ n \in \mathbf{N}$ dans un ensemble borné. On en déduit que les rationnels n'appartiennent pas tous à $\mathcal{F}$. Le fait que cet ensemble soit de plus discret justifie la définition suivante.

Définition 1.2.2 (Spacing)**.** On appelle *spacing* la distance entre deux nombres consécutifs dans $\mathcal{F}$.

Pour un exposant p, le nombre le plus proche de β^p est à une distance β^{p+1-t}. Il est important de noter que la répartition est uniforme sur l'intervalle $[\beta^p, \beta^{p+1}]$ mais la distance entre chaque nombre de $\mathcal{F}$ dépend de p.

Remarque. On peut se demander pour quel p on a $\beta^{p+1-t} = 1$, c'est le cas pour $p = 52$ et donc dans l'intervalle $[2^{52}, 2^{53}]$ seuls les entiers sont représentés.

1.2.1 Approximation de R dans $\mathcal{F}(2, 53, L, U)$

La mantisse d'un nombre réel est *a priori* infinie. On pose alors une fonction d'approximation $fl : \mathbf{R} \longmapsto \mathcal{F}$ qui représente $x \in \mathbf{R}$ avec la représentation floating point.
On s'intéresse alors à estimer la valeur de la différence $\|x - fl(x)\|$. Pour $x \in [\beta^{e-1}, \beta^e]$ on a

$$\begin{aligned} \|x - fl(x)\| &\leq \frac{1}{2} \cdot \text{spacing} \\ &\leq \frac{1}{2}\beta^{e-t}. \end{aligned}$$

On peut également regarder l'erreur relative

$$\frac{\|x - fl(x)\|}{x} \leq \frac{1}{2}\beta^{e-t} \cdot \beta^{1-e} = 2^{-53} \sim 10^{-16}.$$

Définition 1.2.3 (Machine precision)**.** On appelle la valeur $2^{53} \sim 10^{-16}$ *machine precision* et on la note u.

Théorème 1.2.4. *Soit $x \in \mathbf{R}$ alors $\exists fl(x) \in \mathcal{F}(2, 53, L, U)$ tel que*

$$fl(x) = x(1+\varepsilon), \qquad \|\varepsilon\| \leq .$$

1.3 Opérations dans $\mathcal{F}$

Il est important de noter que $\mathcal{F}$ n'est pas muni d'une structure de corps, il n'est donc pas nécessairement stable par addition. Ainsi pour $x, y \in \mathbf{R}$

$$x + y \longmapsto fl(x) + fl(y)$$

peut ne pas appartenir à $\mathcal{F}$. Ainsi $x+y$ est représenté par $fl(fl(x)+fl(y))$. On veut contrôler la valeur C telle que

$$\frac{\|fl(fl(x) + fl(y)) - (x+y)\|}{\|x+y\|} \leq C \cdot u.$$

En général on peut essayer de définir la stabilité d'un problème.

Définition 1.3.1 (Stabilité d'un problème)**.** La résolution du problème $y = G(x)$ est *stable* si à petite perturbation δ_x de x correspond une petite perturbation δ_y de y.

Définition 1.3.2 (Conditionnement)**.** On appelle *conditionnement absolu* du problème la valeur

$$K_{abs} := \sup_{\delta_x} \frac{\|\delta_y\|}{\delta_x}.$$

On appelle *conditionnement relatif* du problème la valeur

$$K_{rel} := \sup_{\delta_x} \frac{\|\delta_y\|/\|y\|}{\|\delta_x\|/\|x\|}.$$

Cette dernière valeur mesure la stabilité du système.

1.3.1 L'arithmétique finie

Nous allons appliquer ces concepts de base à l'arithmétique finie.

$$\begin{aligned}\frac{\|fl(fl(x_1) + fl(x_2)) - (x_1+x_2)\|}{\|x_1+x_2\|} &= \frac{\|((x_1+x_2)(1+\varepsilon))(1+\varepsilon) - (x_1+x_2)\|}{\|x_1+x_2\|} \\ &\leq \max_{x_1,x_2}\left(\frac{\|x_1\|}{\|x_1+x_2\|} + \frac{\|x_2\|}{\|x_1+x_2\|} + 1\right) \cdot u\end{aligned}$$

.

En conclusion si x_1, x_2 sont du même signe le conditionnement relatif est faible puisque inférieur à 3, lorsque $x_1 \sim -x_2$ l'opération est instable.