MEMOIRE DE LICENCE MAPES **Année 2006**

Anthony Mailho
Sylvain Veloso

LES NOMBRES A VIRGULE FLOTTANTE

Sous la direction de Jean-Paul Calvi

Laboratoire Emile Picard
Bureau 241- Bâtiment 1R2
Université Paul Sabatier
31062 TOULOUSE Cédex 04
Tél : 05-61-55-76-66
calvi@picard.ups-tlse.fr

Table des Matières

INTRODUCTION :

L'évolution récente de l'informatique a necessité l'introduction de techniques pour permettre de traiter l'information numérique. Les calculateurs modernes ont la capacité d'approcher une partie assez grande de nombres réels, compris entre $-A$ et A, où A est un réel très grand. L'approximation consiste à arrondir un nombre réel appartenant à ce domaine, en fixant une précision t. Ceci permet de donner une représentation du nombre, l'arrondi, avec t chiffres significatifs (Par exemple, $t = 10$). On obtient, à partir de très peu de paramètres, un grand ensemble de nombres qui sont utilisés pour effectuer les opérations arithmétiques usuelles. Les résultats obtenus sont d'une précision très grande par rapport aux résultats exacts.

A l'aide de paramètres que nous définirons, nous allons donner une approche pour construire des sous-ensembles suffisamment grands de $\mathbb{R}$, en utilisant l'écriture d'un réel à l'aide d'une base. Cette écriture donne lieu à la représentation des nombres réels par des nombres dits à virgule flottante. Nous allons étudier les propriétés de répartition des NVF dans $\mathbb{R}$ en déterminant le domaine auquel ils appartiennent, la distance et la distance relative entre deux NVF. Nous verrons alors comment sont utilisés les NVF dans les opérations effectuées par une machine et analyserons la précision des résultats de ces opérations par rapport aux résultats mathématiques.

1. DEFINITIONS ET PROPRIETES FONDAMENTALES DES NOMBRES À VIRGULE FLOTTANTE

Notation : Soit I et J deux réels, nous noterons $\lceil I, J \rceil$ l'ensemble de tous les entiers compris entre I et J. I et J seront inclus dans cet ensemble si ce sont des entiers.

1.1 Préliminaires sur la représentation en base B

Etant donné un entier B>1, nous savons que tout entier positif $m = a_n a_{n-1} ... a_2 a_1 a_0$ peut être représenté en base B sous la forme $m = a_n.B^n + a_{n-1}.B^{n-1} + ... + a_2.B^2 + a_1.B + a_0.B^0$ avec les $a_i \in \lceil 0, B-1 \rceil$. En effet tout entier peut être décomposé par une succession de divisions par B muni de différentes puissances. Par exemple, en base $B = 10$, $9806 = 9.10^3 + 8.10^2 + 0.10^1 + 6.10^0$. Ayant fixé une base B, cette décomposition est unique car tous les $a_i \in \lceil 0, B-1 \rceil$. Nous allons étendre cette représentation au cas de tous les nombres réels.

Théorème 1.1.1. *$\forall x \in \mathbb{R}, \exists! p \in \mathbb{N}, \exists!$ (p+1)-uplet$(b_0, ..., b_p)$ avec $b_i \in \lceil 0, B-1 \rceil$ et $b_p = 0$ si et seulement si $p = 0$, $\exists!(a_n)_{n \in \mathbb{N}^*}, a_n \in \lceil 0, B-1 \rceil$ non constamment égale à $B-1$ à partir d'un certain rang, tels que :*

$$x = \epsilon b_p ... b_0, a_1 a_2 a_3 ... = \epsilon \sum_{i=0}^{p} b_i B^i + \sum_{j \in \mathbb{N}^*} a_j.B^{-j} \quad avec \quad \epsilon = 1 \ ou \ -1 \tag{1.1}$$

Définition 1.1.2. *Soit $x \in \mathbb{R}$. On appelle partie entière de x l'unique entier n$\in \mathbb{Z}$ tel que $x \in [n, n+1[$.*

Lemme 1.1.3. *Si $x = \sum_{j=1}^{\infty} a_j B^{-j}$, alors $a_1 = E(Bx)$, la suite $(a_j)_{j \geq 1}$ étant définie comme dans le théorème.*

Démonstration. $\frac{a_1}{B} \leq x = \frac{a_1}{B} + \sum_{j=2}^{\infty} a_j B^{-j} \Rightarrow$ (d'après l'hypothèse sur les a_j) $\frac{a_1}{B} \leq x < \frac{a_1}{B} + \sum_{j=2}^{\infty}(B-1)B^{-j} \Rightarrow \frac{a_1}{B} \leq x < \frac{a_1}{B} + \frac{B-1}{B} \times \frac{1}{1-\frac{1}{B}} \Rightarrow \frac{a_1}{B} \leq x < \frac{a_1+1}{B}$ $\Rightarrow a_1 \leq Bx < a_1 + 1 \Rightarrow E(Bx) = a_1$ □

Démonstration. (théorème1.1.1) Soit $x \in \mathbb{R}_+$, on commence par construire la suite $(a_n)_n$. Prenons m le plus grand entier $\leq x$. Définissons a_1 comme le plus grand entier ≥ 0 tel que $m+\frac{a_1}{B} \leq x$. On affirme que $a_1 \in [0, B-1]$. En effet, supposons que $a_1 \geq B$, alors $a_1 = qB+r$ avec $0 \leq r < B$ et $q > 0$. On a donc $m+\frac{qB+r}{B} \leq x \Leftrightarrow (m+q)+\frac{r}{B} \leq x$, d'où $m+q \leq x$; ceci contredit la définition de m donc $a_1 \in [0, B-1]$.
Construisons le reste de la suite par récurrence. Supposons $(a_n)_n$ définie pour $0 \leq n \leq k$ avec $a_n \in [0, B-1]$, alors on définit a_{k+1} comme le plus grand entier ≥ 0 tel que $m + \frac{a_1}{B} + \frac{a_2}{B^2} + ... + \frac{a_k}{B^k} + \frac{a_{k+1}}{B^{k+1}} \leq x$. On affirme que $a_{k+1} \in [0, B-1]$. En effet, supposons que $a_{k+1} \geq B$. Alors $a_{k+1} = qB + r$ avec $0 \leq r < B$ et $q > 0$. Ce qui nous donne $m + \frac{a_1}{B} + ... + \frac{a_k}{B^k} + \frac{qB+r}{B^{k+1}} = (m + \frac{a_1}{B} + ... + \frac{a_k+q}{B^k}) + \frac{r}{B^{k+1}} \leq x$, ceci contredit le fait que a_k soit le plus grand entier ≤ 0 tel que $m + \frac{a_1}{B} + ... + \frac{a_k}{B^k} \leq x$.

Montrons maintenant que $x = m + \sum_{j=1}^{\infty} \frac{a_j}{B^j}$:
On appelle $s_k = m + \sum_{j=1}^{k} \frac{a_j}{B^j}$. Montrons que $\lim_{k\to\infty} s_k = x$.
$|x - s_k| = x - s_k$ car par définition des a_i, $s_k \leq x \ \forall k \in \mathbb{N}^*$. On affirme que $x - s_k \leq \frac{1}{B^k} \Leftrightarrow x \leq s_k + \frac{1}{B^k}$ (sinon on aurait $s_k + \frac{1}{B^k} < x \Leftrightarrow m + \frac{a_1}{B} + ... + \frac{a_k+1}{B^k} \leq x$, ce qui est impossible par définition de a_k, donc on a bien $|x - s_k| = x - s_k \leq \frac{1}{B^k}$). D'où en faisant tendre $k \to \infty$, $\lim_{k\to\infty} s_k = x$.

En ce qui concerne la construction des b_i, on écrit simplement le nombre entier m suivant la base B dans laquelle on travaille. Dans la cas où $x \in \mathbb{R}_-$, il suffit de prendre $\epsilon = -1$.

Montrons maintenant que la suite (a_n) qu'on a construite n'est pas constamment égale à $B-1$ à partir d'un certain rang. Supposons que ce soit le cas : soit $s \in \mathbb{N}$ tel que $\forall n \geq s$, $a_n = B-1$. On a alors :

$$x = m + \sum_{j=1}^{s-1} \frac{a_j}{B^j} + \sum_{j=s}^{\infty} \frac{B-1}{B^j} = m + \sum_{j=1}^{s-1} \frac{a_j}{B^j} + \frac{B-1}{B^s} \sum_{j=0}^{\infty} \frac{1}{B^j} =$$

$$m + \sum_{j=1}^{s-1} \frac{a_j}{B^j} + \frac{B-1}{B^s} \times \frac{1}{1-\frac{1}{B}} = m + \sum_{j=1}^{s-1} \frac{a_j}{B^j} + \frac{1}{B^{s-1}}$$

Donc $x = m + \frac{a_1}{B} + ... + \frac{a_{s-1}+1}{B^{s-1}}$, ce qui est impossible par définition de a_{s-1}.

On a unicité des b_i à cause de l'unicité de l'entier m et de la décomposition de cet entier dans la base B abordée au début du paragraphe.

Il nous reste à démontrer l'unicité des a_i. Montrons que si $\sum_{j=1}^{\infty} a_j B^{-j} = \sum_{j=1}^{\infty} a'_j B^{-j}$ avec $a_j, a'_j \in 0, ..., B-1$ et non constamment égaux à $B-1$ à partir d'un certain rang, alors $a_j = a'_j \ \forall j \in \mathbb{N}^*$ Pour cela nous allons utiliser le lemme 1.1.3. On a $x = \sum_{j=1}^{\infty} a_j B^{-j}$ et $x = \sum_{j=1}^{\infty} a'_j B^{-j}$. D'après le lemme, $a_1 = E(Bx)$ et $a'_1 = E(Bx)$ donc $a_1 = a'_1$. On simplifie et on obtient $x = \sum_{j=2}^{\infty} a_j B^{-j}$ et $x = \sum_{j=2}^{\infty} a'_j B^{-j}$, on continue le même raisonnement, on obtient $a_2 = a'_2$, et ainsi de suite... Ce qui démontre l'unicité de la suite $(a_n)_{n\in N^*}$. □

1.2 Définition des ensembles des nombres à virgule flottante

Définition 1.2.1. *Soient $B, t \in \mathbb{N}^*$ et $L, U \in \mathbb{Z}$, avec $L \leq U$. On définit l'ensemble $\mathbb{F} = \mathbb{F}(B, t, L, U)$ des Nombres à Virgule Flottante par*

$$\mathbb{F} = \{y = \pm m \times B^{e-t}\ , (m, e) \in \mathbb{Z}^2\ /\ B^{t-1} \leq m \leq B^t - 1\ et\ L \leq e \leq U\} \qquad (1.2)$$

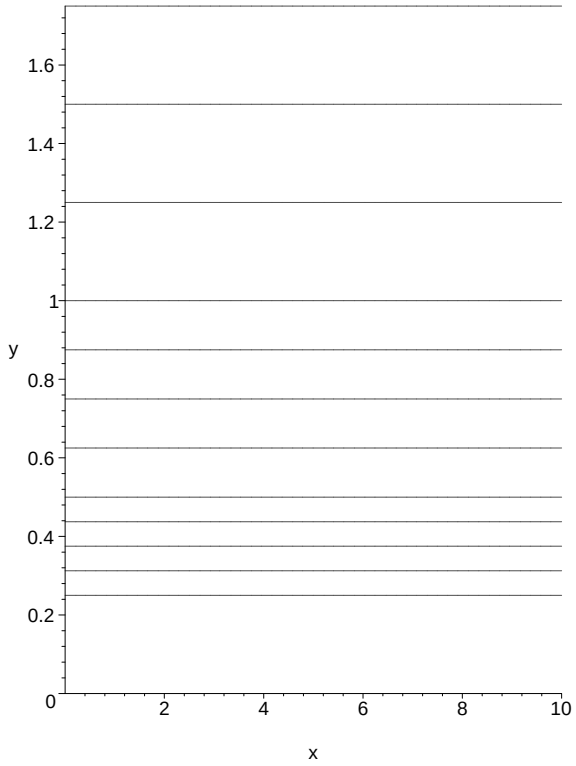
B s'appelle la Base, t la Précision, e le Rang de l'exposant et m le Signifiant. Les nombres B, t, L et U sont appelés les paramètres du systeme (B, t, L, U).

On dit que les éléments de cet ensemble sont à virgule flottante car en fixant un signifiant m entier, on peut obtenir plusieurs nombres avec le même signifiant, mais pour lesquels la virgule sera placée à différents endroits. En effet, dans $\mathbb{F}(B, t, L, U)$, soit m un entier, alors : $y_1 = m.B^{e-t}$ et $y_2 = m.B^{e'-t}$ avec $e' = e-1$ sont deux éléments distincts de $\mathbb{F}$ possédant le même signifiant (à condition que $e, e' \in [L, U]$). Par exemple, dans $\mathbb{F}\{10, 3, -2, 2\}$, avec $m = 123$, on a pour $e = 1$, $y_1 = 123.10^{1-3} = 1.23$, et pour $e' = 0$, $y_2 = 123.10^{0-3} = 0.123$.

1.3 Exemples

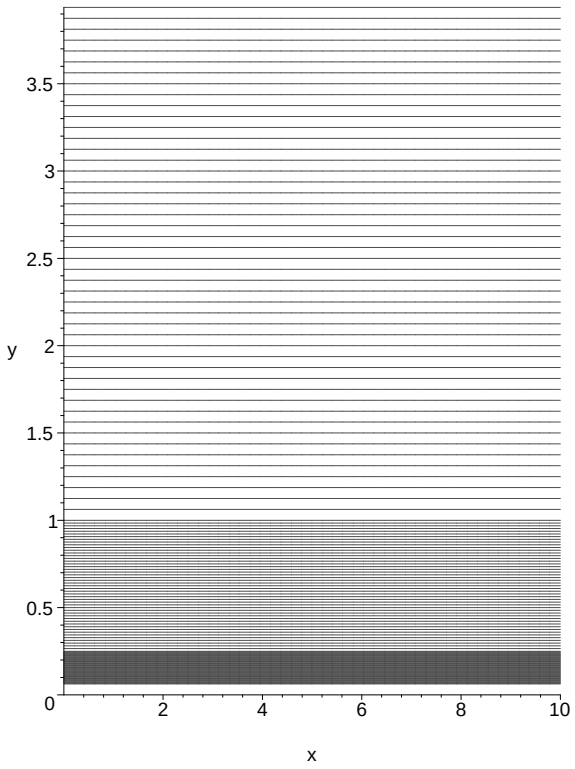
Voici quelques représentations des éléments positifs de quelques systèmes, la lecture des éléments du système se faisant sur l'axe des ordonnées.

$\mathbb{F}(2, 3, -1, 1)$:



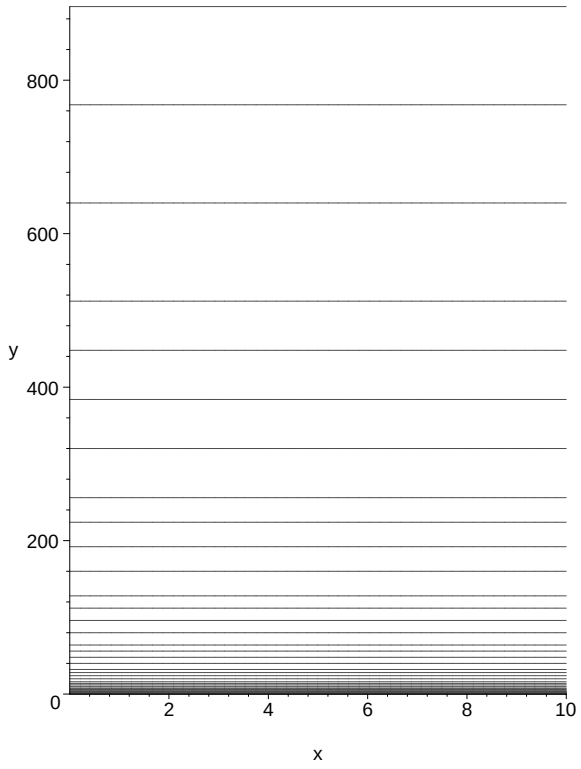
Nous avons ici un système contenant 12 élements positifs. Le plus petit élément positif est $y_{min} = 2^{-1-1} = 0.25$ et le plus grand est $y_{max} = 2^1 - 2^{1-3} = 1.75$. On remarque que plus les valeurs sont grandes, plus l'espacement entre les éléments est grand. On peut observer, d'après les exemples ci-dessus, que les nombres de $\mathbb{F}$ ne sont pas équidistribués (c'est-à-dire qu'ils ne sont pas à la même distance les uns des autres). On remarque également que ces nombres se densifient au voisinage du plus petit. Quant à l'epsilon du système ϵ_F et la plage d'entier M qui seront étudiés plus loin (cf 2.2.1 et 2.4.1), nous les indiquons dès à présent pour réutiliser ces exemples par la suite. On a ici $\epsilon_F = 0.25$ et $M = 8$.

$\mathbb{F}(4, 3, -1, 1)$:



$Card(\mathbb{F}) = 144$, $y_{min} = 0.0625$, $y_{max} = 3.9375$, $\epsilon_F = 0.0625$ et $M = 64$.

$\mathbb{F}(2, 3, -10, 10)$:



$Card(\mathbb{F}) = 84$, $y_{min} \simeq 4.9 \times 10^{-4}$, $y_{max} = 896$, $\epsilon_F = 0.25$, $M = 8$.

1.4 Unicité de la représentation

Proposition 1.4.1. *Soit $y \in \mathbb{F}(B, t, L, U)$, si $y = \epsilon m.B^{e-t}$ et $y = \epsilon' m'.B^{e'-t}$ avec $\epsilon = 1$ ou -1, $B^{t-1} \leq m, m' \leq B^t - 1$ et $L \leq e, e' \leq U$ alors $\epsilon = \epsilon'$, $m = m'$ et $e = e'$.*

Cela signifie que tout élément de $\mathbb{F}$ admet une unique représentation de la forme $x = \epsilon m.B^{e-t}$ avec m et e appartenant à leur encadrement respectif défini. Sans l'hypothèse $B^{t-1} \leq m \leq B^t - 1$, la proposition ne serait pas vraie, on n'aurait pas unicité de la représentation comme le montre l'exemple suivant.

Exemple : Dans $\mathbb{F} = \mathbb{F}(B, t, -1, 4)$, Soit $y_0 = m.B^{2-t}$ avec $B^{t-1} \leq m \leq B^t - 1$, y_0 est donc un élément de $\mathbb{F}$. Posons $m' = m.B^{-1}$, on a alors $m' \leq B^{t-1}$ (*ce qui n'est normalement pas permis d'après la définition*). On obtient alors $y_0 = m.B^{2-t}$ avec "$e = 2$" et $y_0 = m'.B^{3-t}$ avec "e=3". On remarque que les deux rangs de l'exposant e appartiennent à l'encadrement fixé qui est $[-1, 4]$. On a donc deux écritures différentes pour le même nombre. On pourrait également prendre un $m'' = m.B^{-2}$ (n'appartenant pas à l'encadrement fixé) ce qui nous donnerait une autre écriture pour y_0, etc...

Démonstration. Soient $y = m.B^{e-t}$ et $y' = m'.B^{e'-t}$ deux éléments positifs de $\mathbb{F}$. On a alors

$$y = y' \Longleftrightarrow m.B^{e-t} = m'.B^{e'-t} \Longleftrightarrow \frac{m}{m'} = B^{e'-e} \ .$$

Or

$$B^{-1} = \frac{1}{B} < \frac{\mathbf{m}}{\mathbf{m}'} \leq \frac{B^t - 1}{B^{t-1}} = B - \frac{1}{B^t - 1} < B \ .$$

Comme $y = y' \iff \frac{m}{m'} = B^{e'-e}$ et $B^{-1} < \frac{m}{m'} < B^1$, on a $e' - e = 0$ c'est-à-dire $e = e'$ et donc $\frac{m}{m'} = B^0 = 1$ (ou encore $m = m'$), ce qui démontre l'unicité de la représentation. Il en est de même pour les nombres négatifs de $\mathbb{F}$. □

La proposition précédente nous permet de calculer le cardinal d'un système $\mathbb{F}$.

Proposition 1.4.2. *Soit* $\mathbb{F} = \mathbb{F}(B, t, L, U)$,

$$Card(\mathbb{F}) = 2(U - L + 1)(B - 1)B^{t-1}.$$

Démonstration. D'après la proposition précédente, l'application

$$\begin{array}{rcl} \Phi : \{-1, 1\} \times \lceil B^{t-1}, B^t - 1\rceil \times \lceil L, U\rceil & \rightarrow & \mathbb{F} \\ (\epsilon, m, e) & \mapsto & \epsilon m.B^{e-t} \end{array}$$

est bijective (de manière précise, elle est injective par la proposition sur l'unicité 1.4.1 et elle est surjective par définition de $\mathbb{F}$). Par conséquent, les ensembles d'arrivée et de départ ont le même cardinal. Donc

$$Card(\mathbb{F}) = Card(-1, 1) \times Card\lceil B^{t-1}, B^t - 1\rceil \times Card\lceil L, U\rceil$$

$$= 2((B^t - 1) - B^{t-1} + 1)(U - L + 1) = 2((B - 1)B^{t-1})(U - L + 1).$$

□

Exemple :
Calculons le cardinal de quelques systèmes utilisés par les ordinateurs :
Cray-1 single : $\mathbb{F} = \mathbb{F}(2, 48, -8192, 8191) \rightarrow Card(\mathbb{F}) \simeq 4.6 \times 10^{18}$
Cray-1 double : $\mathbb{F} = \mathbb{F}(2, 96, -8192, 8191) \rightarrow Card(\mathbb{F}) \simeq 1.3 \times 10^{33}$
IBM 3090 extanded : $\mathbb{F} = \mathbb{F}(16, 28, -64, 63) \rightarrow Card(\mathbb{F}) \simeq 1.2 \times 10^{36}$
IEEE extanded :$\mathbb{F} = \mathbb{F}(2, 64, -16381, 16384) \rightarrow Card(\mathbb{F}) \simeq 6 \times 10^{23}$

Notation :
Nous noterons $\overline{x}^B$ la représentation de x dans la base B.
Par exemple $\overline{139,21}^{10}$ est égal à $1 * 10^2 + 3 * 10^1 + 9 * 10^0 + 2 * 10^{-1} + 1 * 10^{-2}$ dans la base 10, alors que $\overline{110,01}^2$ est égal à $1 * 2^2 + 1 * 2^1 + 0 * 2^0 + 0 * 2^{-1} + 1 * 2^{-2}$ dans la base 10.

1.5 Représentation décimale des éléments de $\mathbb{F}$:

Proposition 1.5.1. *Soit* $y \in \mathbb{F}(B, t, L, U)$, *il existe un unique t-uplet* $(d_1, d_2, ..., d_t) \in \lceil 1, B^{-1}\rceil \times \lceil 0, B^{-1}\rceil^{t-1}$ *et un unique* $e \in \lceil L, U\rceil$ *tels que*

$$y = \epsilon \overline{0, d_1 d_2 ... d_t}^B \times B^e \tag{1.3}$$

où ϵ est égal au signe de y, autrement dit

$$y = \epsilon B^e(\frac{d_1}{B} + \frac{d_2}{B^2} + ... + \frac{d_t}{B^t})$$

Réciproquement, toute écriture de la forme (1.2) définit un unique élément de $\mathbb{F}$.

Sans la restriction $d_1 \neq 0$ l'unicité de la représentation décimale ne serait plus vraie. Par exemple dans $\mathbb{F} = \mathbb{F}(B, t, -1, 4)$ on pourrait avoir $y_0 = B^3(\frac{0}{B} + \frac{d_2}{B^2} + ... + \frac{d_t}{B^t})$ élément de $\mathbb{F}$ qui pourrait également être écrit de la façon suivante, $y_0 = B^2(\frac{d_2}{B} + \frac{d_3}{B^2} + ... + \frac{d_t}{B^{t-1}})$ ou encore en posant $d'_1 = d_2$, $d'_2 = d_3$,...,$d'_{t-1} = d_t$ et $d'_t = 0$, $y_0 = B^2(\frac{d'_1}{B} + \frac{d'_2}{B^2} + ... + \frac{d'_t}{B^t})$. On remarque que dans les deux écritures, les rangs de l'exposant qui sont 2 et 3 appartiennent bien à la restriction fixée qui est $[U, L]=[-1, 4]$, donc on se retrouve avec deux écritures différentes pour le même nombre.

Démonstration. Existence :
On pose $y = m.B^{e-t}$ un élément de $\mathbb{F}(B, t, L, U)$ avec $m = \overline{d_1...d_t}^B$
on a $y = m.B^{e-t} = \overline{d_1 d_2...d_t}^B \times B^{e-t} = (\frac{d_1...d_t}{B^t})B^e = (\frac{d_1.B^{t-1}+d_2.B^{t-2}+...+d_{t-1}.B+d_t}{B^t})B^e = (\frac{d_1}{B} + \frac{d_2}{B^2} + ... + \frac{d_t}{B^t})B^e$
Unicité :
Démontrons l'unicité de la représentation d'un élément de $\mathbb{F} = \mathbb{F}(B, t, L, U)$:
Soient $y = \overline{0.d_1 d_2...d_t}^B \times B^e = 0.d_1 d_2...d_t \times B^e$ en base B et $y' = \overline{0.d'_1 d'_2...d'_t}^B \times B^{e'} = 0.d'_1 d'_2...d'_t \times B^{e'}$ en base B deux éléments positifs de $\mathbb{F}$, alors :

$$y = y' \Longleftrightarrow 0.d_1 d_2...d_t \times B^e = 0.d'_1 d'_2...d'_t \times B^{e'} \Longleftrightarrow \frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} = B^{e'-e}$$

or

$$\frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} \leq \frac{0.B-1\ B-1...B-1}{0.10...0} = \frac{\frac{B-1}{B} + \frac{B-1}{B^2} + ... + \frac{B-1}{B^t}}{B^{-1}}$$

$$= B - 1 + \frac{B-1}{B} + ... + \frac{B-1}{B^{t-1}} < B \ et \ \frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} \geq \frac{O, 10...0}{0.B-1\ B-1...B-1} > \frac{1}{B} = B^{-1}.$$

Comme $y = y' \Longleftrightarrow \frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} = B^{e'-e}$ et $B^{-1} < \frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} < B$ on a $e' - e = 0$ i.e $e = e'$ et donc $\frac{0.d_1 d_2...d_t}{0.d'_1 d'_2...d'_t} = B^0 = 1$ qui équivaut à $d_i = d'_i \ \ \forall i \in \{1, 2, ..., t\}$.
Il en est de même pour les éléments négatifs de $\mathbb{F}$. □

Cette représentation (1.2) s'appelle la représentation décimale de $\mathbb{F}$.

La condition $d_1 > 0$ de la représentation (1.2) correspond à $m \geq B^{t-1}$ dans (1.1). Quant à l'autre condition sur $m \leq B^t - 1$ dans (1.1), elle correspond à la condition t-uplet dans (1.2). Sans la condition $d_1 > 0$ on n'aurait pas unicité de la représentation.

Définition 1.5.2. *Soit $y = \overline{0, d_1 d_2...d_t}^B \times B^e$ un élément de $\mathbb{F}(B, t, L, U)$. Le chiffre d_1 est appelé le chiffre le plus significatif de y.*

2. REPARTITION DES NOMBRES A VIRGULE FLOTTANTE

2.1 Valeur minimale et maximale des Nombres à Virgule Flottante

Théorème 2.1.1. *Dans* $\mathbb{F}(B,t,L,U)$*, le plus* **petit élément positif strict** *est* $\mathbf{y_{min}} = B^{L-1}$ *et le plus* **grand élément** *est* $\mathbf{y_{max}} = B^U - B^{U-t}$

Démonstration. On utilise la représentation 1.2 ($y = m \times B^{e-t}$). Pour le plus petit élément positif strict, on choisit $m = B^{t-1}$, et $e = L$ ce qui nous donne $y_{min} = B^{t-1}.B^{L-t} = B^{L-1}$. Pour le plus grand élément, on prend $m = B^t - 1$ et $e = U$, on obtient alors $y_{max} = (B^t - 1).B^{U-t} = B^U - B^{U-t}$. □

Ces valeurs sont indiqués dans les exemples de représentation graphique.

2.2 Epsilon du systeme

Définition 2.2.1. *On note* ϵ_F *la distance entre "1" et le plus petit nombre de* $\mathbb{F}$ *qui le suit.* ϵ_F *dépend des paramètres du systeme* $\mathbb{F}$ *et est appelé "l'epsilon du système".*

Cette valeur est également indiquée dans les exemples de représentation graphique. Par exemple, dans $\mathbb{F} = \mathbb{F}(2,3,-4,4)$, appelons y_0 le premier élément de $\mathbb{F}$ supérieur à 1. En utilisant la représentation 1.3, on a $1 = B^0 = \overline{0.100}^2 \times 2^1$ et $y_0 = \overline{0.101}^2 \times 2^1 = B^0 + B^{-2}$. On obtient alors $\epsilon = y_0 - 1 = (B^{-2} + B^0) - B^0 = B^{-2}$

Proposition 2.2.2. *Si* $\mathbb{F} = \mathbb{F}(B,t,L,U)$*, alors*

$$\epsilon_F = B^{1-t}$$

On remarque donc que l'epsilon du système est indépendant de L et de U.

Démonstration. Dans $\mathbb{F} = \mathbb{F}(B,t,L,U)$, posons y_0 le premier élément de $\mathbb{F}$ supérieur à 1. D'après le modèle (1.2), 1 s'écrit $\overline{0,100}^B \times B^1$. y_0 est alors égal à $\overline{0,101}^B \times B^1$ d'où $\epsilon_F = y_0 - 1 = \overline{0,101}^B \times B^1 - \overline{0.100}^B \times B^1 = \overline{0,001}^B \times B^1 = B^{-t} \times B^1 = B^{1-t}$ □

2.3 Distance et distance relative

Définition 2.3.1. *On note Δ_y la distance entre y, un élément d'un système $\mathbb{F}$, et le premier nombre qui lui soit supérieur.*

Proposition 2.3.2. *Soit $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$ un élément de $\mathbb{F} = \mathbb{F}(B, t, L, U)$ on a alors :*

$$\Delta_y = B^{e-t} \tag{2.1}$$

Démonstration. Dans $\mathbb{F} = \mathbb{F}(B, t, L, U)$, soit $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$. Le nombre de $\mathbb{F}$ le plus proche qui lui soit supérieur est $y' = y + \overline{0, 00...1}^B \times B^e = y + B^{e-t} = y + \Delta_y$. □

On peut donner un encadrement général de Δ_y.

Proposition 2.3.3. *Dans $\mathbb{F} = \mathbb{F}(B, t, L, U)$:*

$$B^{-1}\epsilon_F|y| \leq \Delta_y \leq \epsilon_F|y| \quad \forall y \in \mathbb{F}^+ \tag{2.2}$$

Remarques :

1. Pour les éléments de $\mathbb{F}^-$, on a l'encadrement inverse.
2. L'erreur est d'autant plus importante que $|y|$ est grand.

Démonstration. Soit $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$ un élément de $\mathbb{F} = \mathbb{F}(B, t, L, U)$. On a

$$\Delta_y \leq \epsilon_F|y| \Longleftrightarrow B^{e-t} \leq B^{1-t} \times |\overline{0, d_1 d_2 ... d_t}^B \times B^e|$$

$$\Longleftrightarrow B^{e-t} \leq B^{e-t+1} \times |\overline{0, d_1 d_2 ... d_t}^B| \Longleftrightarrow B^0 = 1 \leq B \times |\overline{0, d_1 d_2 ... d_t}^B| \ .$$

Cette dernière inégalité est vraie car $d_1 \geq 1$.

On a

$$\Delta_y \geq B^{-1}\epsilon_F|y| \Longleftrightarrow B^{e-t} \geq B^{-1} \times B^{1-t} \times |\overline{0, d_1 d_2 ... d_t}^B \times B^e|$$

$$\Longleftrightarrow B^{e-t} \geq B^{e-t} \times |\overline{0, d_1 d_2 ... d_t}^B| \Longleftrightarrow 1 \geq |\overline{0, d_1 d_2 ... d_t}^B| \ ,$$

ce qui est vrai. □

Définition 2.3.4. *On définit l'erreur relative entre deux réels x_1 et x_2(par rapport à x_1) par : $\Delta R = |\frac{x_1 - x_2}{x_1}|$.*

Par définition, l'erreur relative est le rapport de l'erreur par la valeur exacte d'un nombre. Son intérêt est donc de donner une information sur l'ordre de grandeur de ce rapport. Ainsi, elle apporte plus de précision que l'erreur simple car elle fait intervenir l'erreur mais aussi le nombre que l'on veut approcher.

Exemple : Soient $x = 1, \tilde{x} = 2$ et $y = 1000, \tilde{y} = 1001$. L'erreur est de 1 dans les deux cas. Pourtant, entre 1 et 2, on passe du simple au double tandis qu'entre 1000 et 1001, l'erreur est relativement très faible. Ceci est illustré par le calcul de l'erreur relative :

- Pour 1 et 2 : $\Delta R = |\frac{2-1}{1}| = 1$.
- Pour 1000 et 1001 : $\Delta R = |\frac{1001-1000}{1000}| = \frac{1}{1000}$. L'erreur relative est 1000 fois plus faible que précédemment.

Proposition 2.3.5. *On note ΔR_y l'erreur relative entre $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$ et son successeur immédiat,*

$$\Delta R_y = (\overline{0, d_1 d_2 ... d_t}^B \times B^t)^{-1} \tag{2.3}$$

On remarque que ΔR_y est indépendant de L et de U.

Démonstration. Soit $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$ et y' son successeur immédiat, alors : $\Delta R_y = \frac{y'-y}{y} = \frac{B^{e-t}}{\overline{0,d_1 d_2 ... d_t}^B . B^e} = (\overline{0, d_1 d_2 ... d_t}^B \times B^t)^{-1}$ □

2.4 Plage d'entiers

On note **M** le plus grand entier d'un système $\mathbb{F}$ tel que tous les entiers positifs le précédant appartiennent à $\mathbb{F}$. Par exemple, dans $\mathbb{F} = \mathbb{F}(10, 3, -4, 4)$, on a $M = 10^3 = 1000$. En posant $y = 0.d1d2d3 \times 10^e$, on voit que : y prend toutes les valeurs entières comprises entre 1 et 9 avec $d_1 \in [1, 9], d_2 = d_3 = 0$ et $e = 1 \in [-4, 4]$, y prend toutes les valeurs entières comprises entre 10 et 99 avec $d_1 \in [1, 9]$, $d_2 \in [0, 9], d_3 = 0$ et $e = 2 \in [-4, 4]$, y prend toutes les valeurs entières comprises entre 100 et 999(=M-1) avec $d_1 \in [1, 9]$, $d_2, d_3 \in [0, 9]$ et $e = 3 \in [-4, 4]$ $M = 10^3 = 0,1 \times 10^4 \in \mathbb{F}$ car $4 \in [-4, 4]$. Quant à $M + 1 = 1001$, il est égal à $0,1001 \times 10^4$ qui n'appartient pas à $\mathbb{F}$ car il nécessite une précision de 4 alors que celle de $\mathbb{F}$, dans cet exemple, n'est que de 3. Le successeur de M dans $\mathbb{F}$ est $1010 = 0,101 \times 10^4$.

Proposition 2.4.1. *Dans $\mathbb{F}(B, t, L, U)$, on a*

$$M = \overline{0,1}^B \times B^{t+1} = B^t \tag{2.4}$$

à condition que $t + 1 \in [L, U]$, ce qui est le cas en pratique. Cet entier M nous donne la plus grande plage d'entiers consécutifs dans $\mathbb{F}$.

On remarque que M ne dépend que de B et de t. Cette valeur est représentée dans les exemples de représentation graphique.

Démonstration. On suppose que $t + 1 \in [L, U]$, en posant $y = \overline{0, d_1 d_2 ... d_t}^B \times B^e$ on voit que y prend toutes les valeurs entières comprises entre 1 et $B^t - 1$ en faisant varier les $d_i (1 \leq i \leq e)$ entre 0 et $B - 1$ (sauf d_1 entre 1 et $B - 1$) avec e variant entre 1 et t. Quant à B^t il est égal à $0,1 \times B^{t+1} \in \mathbb{F}$.
$B^t + 1 = \overline{0, d_1 d_2 ... d_t 1}^B \times B^{t+1}$ (avec $d_1 = 1$ *et* $d_i = 0 \; \forall i \in [2, t]$) ne peut être représenté car il nécessiterait une précision t'=t+1. □

2.5 Les Nombres à Virgule Flottante Dénormalisés

On définit cet ensemble de nombres dénormalisés afin d'avoir accès à d'autres valeurs plus petites en valeur absolue que les éléments d'un systeme $\mathbb{F}(B,t,L,U)$.

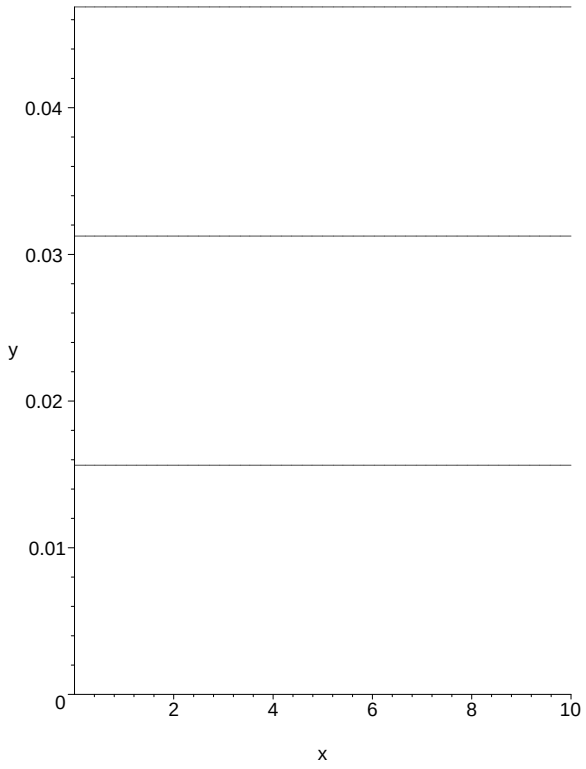
Définition 2.5.1. *Soient $B,t \in \mathbb{N}^*$ et $L \in \mathbb{Z}$, on définit l'ensemble des nombres dénormalisés $\mathbb{G} = \mathbb{G}(B,t,L)$ par*

$$\mathbb{G} = \{x = \pm m \times B^{L-t}, m \in]0, B^{t-1}[\}$$

Les nombres dénormalisés positifs, éléments de $\mathbb{G}(B,t,L)^+$, sont ainsi situés entre 0 et le plus petit élément de $\mathbb{F}(B,t,L,U)$.

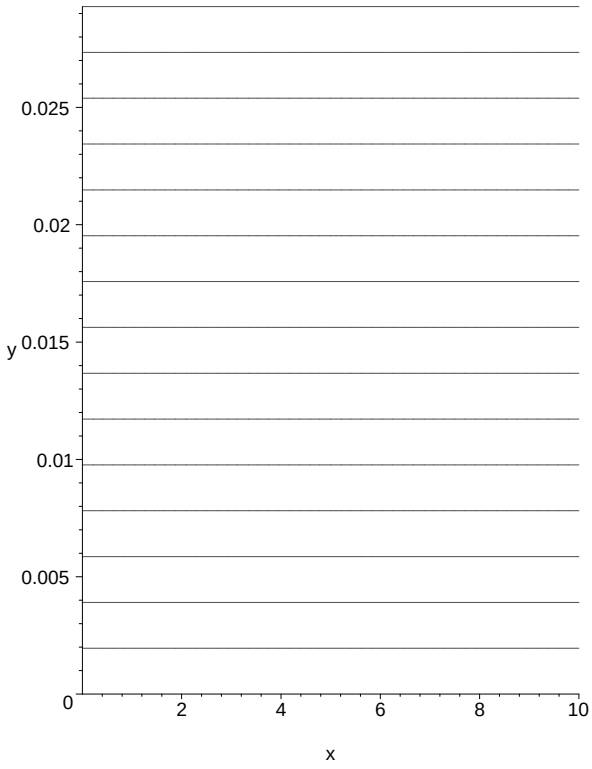
Exemple : Voici les représentations des éléments positifs de quelques systèmes $\mathbb{G}$.

$\mathbb{G}(2,3,-3)$:



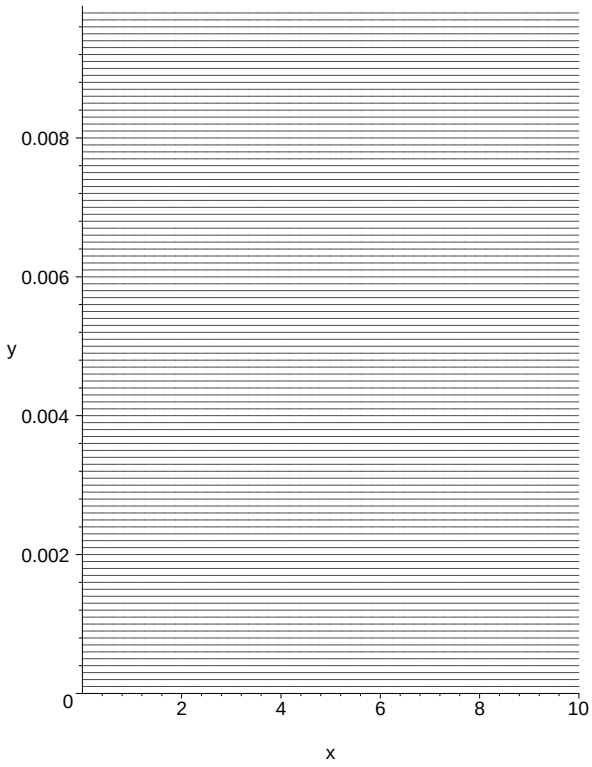
$Card(G) = 3$, le plus petit élément de $\mathbb{G}$ est $x_{min} =\simeq 0.016$, le plus grand est $x_{max} \simeq 0.047$ et l'écart entre chaque élément de $\mathbb{G}$ (cf 2.5.5) est $\Delta_G \simeq 0.016$

$\mathbb{G}(2,5,-4)$:



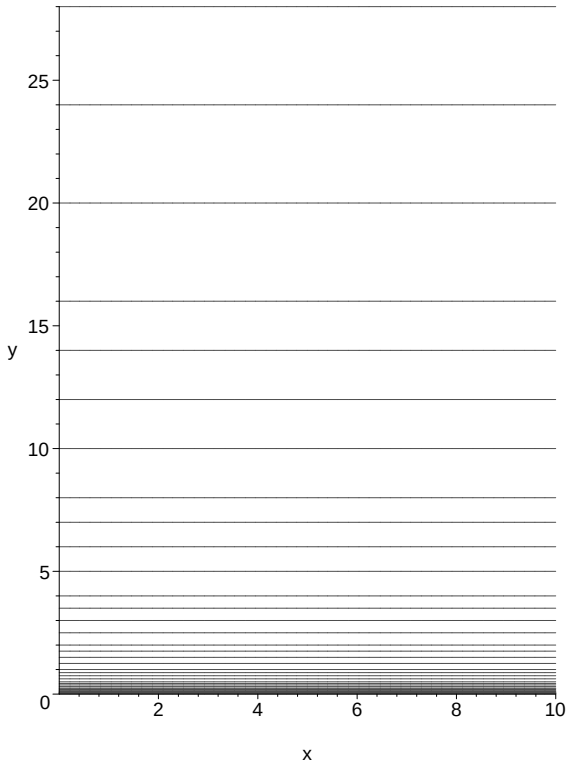
$Card(G) = 15$, $x_{min} \simeq 0.002$, $x_{max} \simeq 0.029$ et $\Delta_G \simeq 0.002$.

$\mathbb{G}(10, 3, -1)$:



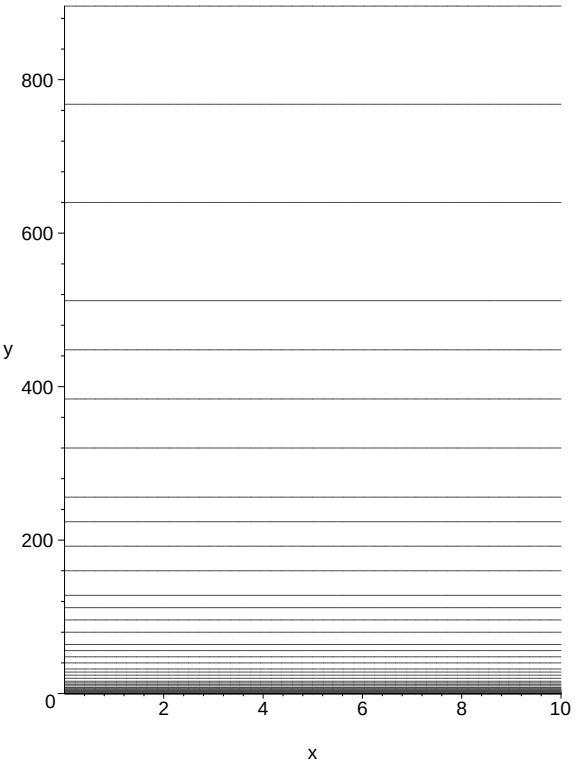
$Card(G) = 99$, $x_{min} = 10^{-4}$, $x_{max} = 0.0099$ et $\Delta_G = 10^{-4}$.

$\mathbb{F}(2,3,-5,5) + \mathbb{G}(2,3,-5)$:



$Card(F+G) = 47$, $x_{min} \simeq 0.0039$, $y_{max} = 28$, $\Delta_G \simeq 0.0039$, $\epsilon_F = 0.25$ et $M = 8$.

$\mathbb{F}(2,3,-10,10) + \mathbb{G}(2,3,-10)$:



$Card(F+G) = 87$, $x_{min} \simeq 1.22 \times 10^{-4}$, $y_{max} = 896$, $\Delta_G \simeq 1.22 \times 10^{-4}$, $\epsilon_F = 0.25$ et $M = 8$.

Proposition 2.5.2. *Soit $\mathbb{G} = \mathbb{G}(B,t,L)$, alors :*

$$Card(G) = 2(B^{t-1} - 1).$$

Démonstration. L'application

$$\begin{array}{rcl} \Phi : \{-1,1\} \times]O, B^{t-1}[& \rightarrow & \mathbb{G} \\ (\epsilon, m) & \mapsto & \epsilon m.B^{L-t} \end{array}$$

est bijective, par conséquent, les ensembles d'arrivée et de départ ont le même cardinal.
Donc $Card(\mathbb{G}) = Card(\{-1,1\}) \times Card]0, B^{t-1}[= 2 \times (B^{t-1} - 1)$. □

Proposition 2.5.3. *D'après la représentation (1.2), le chiffre le plus significatif d'un nombre dénormalisé $x \in \mathbb{G}$ est $d_1 = 0$ i.e $x = \pm 0.0 d_2 d_3 ... d_t \times B^L$.*

Démonstration. Soit $x = \epsilon m \times B^{L-t} \in \mathbb{G}(B,t,L)$, avec $\epsilon = \pm 1$. On pose $m = d_1 d_2 ... d_t = d_1 B^{t-1} + d_2 B^{t-2} + ... + d_t B^0$, comme $m < B^{t-1}$, on a $d_1 = 0$. D'où $x = \epsilon 0.0 d_2 ... d_t \times B^L$. □

Proposition 2.5.4. *Dans $\mathbb{G}(B,t,L)$, le plus petit élément positif est $x_{min} = B^{L-t}$ et le plus grand élément est $x_{max} = (B^{t-1} - 1)B^{L-t}$.*

Démonstration. Pour le plus petit élément positif $x_{min} = m \times B^{L-t}$ de $\mathbb{G}$, on prend le plus petit m possible qui est 1 et on obtient immédiatement $x_{min} = 1 \times B^{L-t} = B^{L-t}$.
Pour le plus grand élément x_{max} de $\mathbb{G}$, on prend la plus grande valeur possible de m qui est $B^{t-1} - 1$. D'où $y_{max} = (B^{t-1} - 1)B^{L-t}$. □

Proposition 2.5.5. *On note $\Delta_{\mathbb{G}}$ l'espacement entre deux éléments de $\mathbb{G}(B,t,L)$. On a alors :*

$$\Delta_{\mathbb{G}} = B^{L-t}.$$

Démonstration. Dans $\mathbb{G}^+$, soient $x_1 = m \times B^{L-t}$ avec $0 < m < B^{t-1} - 1$ et $x_2 = (m+1)B^{L-t}$ le premier élément supérieur à x_1.
On a alors $\Delta_{\mathbb{G}} = x_2 - x_1 = B^{L-t}$. □

Par la suite, pour simplifier et éviter des démonstrations trop complexes, on ne tiendra pas compte des nombres dénormalisés et on ne travaillera donc que sur des systèmes $\mathbb{F}(B,t,L,U)$.

3. LA FONCTION ARRONDISSEMENT

3.1 Définition et Exemples

Définition 3.1.1. *D'après le théorème 1.1.1, tout réel x s'écrit de manière unique (en base B) : $x = \epsilon b_m...b_0, a_1a_2...$ avec $\epsilon = 1$ ou -1, ou encore $x = \epsilon 0, b_m...b_1b_0a_1a_2... \times B^{m+1}$ avec les conventions définies dans le théorème. Autrement dit, tout réel x admet une écriture unique de la forme $x = \epsilon 0, \alpha_1\alpha_2... \times B^p$, avec $\alpha_1 \neq 0$. On définit alors la fonction arrondissement FL de manière suivante :*

Premier cas : Si $|x| > max(\mathbb{F})$ ou si $|x| < min(\mathbb{F})$, alors $FL(x) = NAN$ (not a number).

Deuxième cas : Si $min(\mathbb{F}) \leq |x| \leq max(\mathbb{F})$ (nous dirons dans ce cas que x appartient au domaine de $\mathbb{F}$, $x \in Dom(\mathbb{F})$) alors :

$FL(x) = \epsilon 0, \alpha_1\alpha_2...\alpha_t \times B^p \in F$ si $\alpha_{t+1} < \frac{B}{2}$, ou bien

$FL(x) = \epsilon 0, \alpha_1\alpha_2...\alpha_t \times B^p + B^{p-t} \in F$ si $\alpha_{t+1} \geq \frac{B}{2}$

Dans le cas où $\alpha_{t+1} \geq \frac{B}{2}$, il peut se créer un décalage dans l'arrondissement si $\alpha_t = B - 1$. Par exemple, avec un système de précision de base $B = 10$ et $t = 3$, pour un réel $x = 0,1116$, on obtient $FL(x) = 0,112 \times B^0$. Mais si on prend $x = 0,1196$, $FL(x) = 0,120 \times B^0$ ou de même, pour $x = 0,1996$, $FL(x) = 0,200 \times B^0$.

Exemples :

1. Soit $x = \overline{\frac{1}{3}}^{10} = 0.333333...$ en base 10 et calculons FL(x) dans $\mathbb{F}(2, 7, -10, 10)$. On a tout d'abord $x = \overline{0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} + 0 \times 2^{-5}...}^{10} = \overline{0,01010101...}^{2} \times 2^0$. On obtient alors $FL(x) = \overline{0,0101011}^{2} \times 2^0$ dans $\mathbb{F}$.
2. Soit $x = \overline{0.1110011001}^{2} \times 2^3 = \overline{111.0011001}^{2}$ en base 2 et calculons FL(x) dans $\mathbb{F}(10, 8, -5, 5)$. On a alors $x = \overline{7.1953125}^{10} = \overline{0.71953125}^{10} \times 10^1$
   D'où $FL(x) = \overline{0.71953125}^{10} \times 10^1 = x$ dans $\mathbb{F}$.
3. Soit $x = \overline{0.17345672}^{8} \times 8^2 = \overline{17.345672}^{8}$, calculons FL(x) dans $\mathbb{F}(10, 8, -5, 5)$. En base 10, $x = \overline{15,448951721\ldots}^{10} = \overline{0,15448951721\ldots}^{10} \times 10^2$ d'où $FL(x) = \overline{0,15448952}^{10} \times 10^2$ dans $\mathbb{F}$.

3.2 Propriétés de la fonction FL

Proposition 3.2.1. *La fonction FL(x) conserve le signe, c'est-à-dire*

$$FL(-x) = -FL(x)$$

Démonstration. Evident par définition. □

Théorème 3.2.2. *(***Monotonie***)* *Dans* $\mathbb{F}(B,t,L,U)$*, si* $x \leq y$*, avec* x *et* $y \in Dom(\mathbb{F})$*, alors* $FL(x) \leq FL(y)$.

Démonstration. D'après la proposition 3.2.1 la fonction FL conserve le signe . Considérons alors le cas où x et y sont > 0, et supposons même $y > 0$. On pose $y = \alpha_1 ... \alpha_{t-1}\alpha_t\alpha_{t+1} ... \times B^p$ et $x = \gamma_1 ... \gamma_{t-1}\gamma_t\gamma_{t+1} ... \times B^{p'}$ avec α_1 et $\gamma_1 \neq 0$
Premier cas : si $p > p'$, alors il est évident que $FL(y) > FL(x)$ car α_1 et $\gamma_1 \neq 0$
Deuxieme cas : si $p = p'$, il existe 4 cas possibles :

	$FL(x)$		$FL(y)$
$\gamma_{t+1} < B/2$ et $\alpha_{t+1} < B/2$	$0, \gamma_1 ... \gamma_t \times B^p$	$<$	$0, \alpha_1 ... \alpha_t \times B^p$
$\gamma_{t+1} < B/2$ et $\alpha_{t+1} \geq B/2$	$0, \gamma_1 ... \gamma_t \times B^p$	$<$	$0, \alpha_1 ... \alpha_t \times B^p + B^{p-t}$
$\gamma_{t+1} \geq B/2$ et $\alpha_{t+1} < B/2$	$0, \gamma_1 ... \gamma_t \times B^p + B^{p-t}$	$\leq$	$0, \alpha_1 ... \alpha_t \times B^p$
$\gamma_{t+1} \geq B/2$ $\alpha_{t+1} \geq B/2$	$0, \gamma_1 ... \gamma_t \times B^p + B^{p-t}$	$<$	$0, \alpha_1 ... \alpha_t \times B^p + B^{p-t}$

Troisieme cas : $p < p'$, ce cas est impossible car $y > x$ et $\gamma_1 \neq 0$.
On observe dans tous les cas que si $x \leq y$, alors $FL(x) \leq FL(y)$. D'où la monotonie de la fonction FL. □

Théorème 3.2.3. *Si* $x \in Dom(\mathbb{F}(B,t,L,U))$*, alors* $FL(x) = x(1+\Delta)$ *avec* $\Delta \leq u$ *et* $u = \frac{\epsilon_F}{2}$*,* $\epsilon_{\mathbb{F}}$ *étant défini précédemment dans la définition (2.2.1).*

Démonstration. Soit $x \in Dom(\mathbb{F}(B,t,L,U))$, donc x peut s'écrire sous la forme $x = r \times B^{e-t}$, avec $r \in \mathbb{R}^+$ tel que $B^{t-1} \leq r < B^t$. Soit y_1 le premier élément de $\mathbb{F}$ inférieur ou égal à x et y_2 le premier élément de $\mathbb{F}$ strictement supérieur à x. Alors $FL(x) = y_1$ ou $FL(x) = y_2$ et on a

$$|FL(x) - x| \leq \frac{|y_2 - y_1|}{2} \leq \frac{B^{e-t}}{2}$$

d'où

$$|\frac{FL(x) - x}{x}| \leq \frac{\frac{1}{2}B^{e-t}}{r \times B^{e-t}} \leq \frac{1}{2}B^{1-t} = u.$$

□

3.3 Opérations machine

A toute opération arithmétique élémentaire ($+$, $-$, $\times$, $\div$) correspond une opération effectuée par les calculateurs électroniques, uniquement avec des éléments de $\mathbb{F}$. Si o désigne une opération arithmétique, l'opération correspondante effectuée par l'opérateur électronique sera notée $\overline{o}$. Pour certaines opérations très simples, o et $\overline{o}$

coïncident. C'est par exemple le cas lorsque x et y sont des entiers inférieurs à M dans un système $\mathbb{F}$ donné (pour avoir $x+y = x \overline{+} y$, il suffit que x et y appartiennent à la plage d'entiers consécutifs de $\mathbb{F}$ inférieurs à M -cf. proposition 2.4.1.-). D'une manière générale, la loi $\overline{o}$ est définie par :

$$x \ \overline{o} \ y = FL(FL(x)oFL(y)).$$

Le calculateur doit donc faire les calculs exacts avec FL(x) et FL(y).
Exemple :

- Calcul de $\pi \overline{+} \pi$ dans $\mathbb{F}(10, 2, -10, 10)$: $FL(FL(\pi)+FL(\pi)) = FL(3.1+3.1) = 6.2$.
- Calcul de $2 \overline{\times} \pi$: $FL(FL(2)\times FL(\pi)) = FL(2\times 3.1) = 6.2$.

Propriété : (précision de l'opération machine par rapport à l'opération arithmétique) On voudrait que : $x\overline{o}y = (xoy)\times(1+\Delta)$, avec $|\Delta| \leq u = \frac{\epsilon_{\mathbb{F}}}{2}$.
En général cette propriété ne sera pas satisfaite. Le problème le plus classique vient des soustractions.

Calcul des soustractions $x - y$ par un calculateur avec $x, y \in \mathbb{F}$: On peut toujours supposer que $x \geq y$. On écrit x et y avec la même puissance mise en facteur (on dénormalise éventuellement y), puis on pose la soustraction.
Exemple : Dans $\mathbb{F}(10, 2, -20, 20)$, soient $x = 1$ et $y = 0.99$.

$x \overline{-} y = FL(FL(x) - FL(y)) = FL(0.1\times 10 - 0.99\times 10^0)$. On écrit $0.99\times 10^0 = 0.09(9) \times 10$. Posons la soustraction :

$$\begin{array}{rr} 10\times & 0.1 \\ 10\times & 0.09 \\ \hline 10\times & 0.01 \end{array}$$

(On remarque que comme la précision vaut 2, l'écriture de y en facteur de 10^1 a induit la perte du dernier chiffre de y(9)). D'où $x \overline{-} y = FL(0.01 \times 10) = 0.1$. La soustraction machine donne un résultat 10 fois plus grand que le résultat exact (l'erreur relative vaut 9).

En général, soit $x = 10^\alpha \times 0, \overbrace{\square \ldots \square}^{t \ chiffres}$ et $y = 10^\alpha \times 0, \overbrace{\square \ldots \square}^{t \ chiffres}$. Quand on écrit $y \leq x$ à la puissance α, des zéros peuvent apparaître sur les premiers carrés (si $x = m.10^\alpha, y = m'.10^{\alpha'} avec \ \alpha' < \alpha$). On perd alors le nombre correspondant de termes au niveau des derniers carrés.

Dans l'exemple précédent, l'erreur relative de la soustraction machine est élevée. Une solution consiste à utiliser temporairement une précision de t+1(le terme supplémentaire est appelé chiffre de garde). Reprenons l'exemple précédent avec un chiffre de garde : $x = 10 \times 0.1, y = 10 \times 0.099$.

$$\begin{array}{rr} 10\times & 0.1 \\ 10\times & 0.099 \\ \hline 10\times & 0.001 \end{array}$$

D'où $x \overline{-} y = FL(0.001 \times 10) = 0.01$. La soustraction machine donne le résultat exact.

Théorème 3.3.1 (Ferguson). *Si x et y sont des NVF tels que $e(x-y) \leq min(e(x), e(y))$, où $e(x)$ est l'exposant de x dans sa représentation normalisée dans $\mathbb{F}(B, t, L, U)$, alors $FL(x-y) = x-y$ (en supposant que $x-y$ appartient au domaine de $\mathbb{F}$).*

Démonstration. $e(x-y) \leq min(e(x), e(y)) \Rightarrow |e(x) - e(y)| \leq 1$. En effet : supposons par l'absurde que $|e(x) - e(y)| > 1$ soit $|e(x) - e(y)| \geq 2$. Sans perte de généralité, on peut supposer que x et y sont positifs et que $e(x) > e(y)$. On a alors : $x = 0, \alpha_1 \dots \alpha_t B^{e(x)}$ avec $\alpha_1 \neq 0$ et $y = 0, 00\beta_1 \dots B^{e(x)}$ avec éventuellement $\beta_1 = 0$. D'où $x - y > 0, \alpha_1 B^{e(x)} - 0, 0(B-1) \times B^{e(x)}$ (on majore y pour minorer $-y$) $> 0, (\alpha_1 - 1)1 \times e(x) \Rightarrow e(x-y) \geq e(x)$ si $\alpha_1 - 1 \neq 0$ ou $e(x-y) \geq e(x) - 1$ si $\alpha_1 - 1 = 0$. Dans les deux cas on a une contradiction avec $e(x-y) \leq min(e(x), e(y))$ car $e(x) - 1 > min(e(x), e(y))$. Donc on a bien $|e(x) - e(y)| \leq 1$.

On a alors deux cas : $e(x) = e(y)$ et $|e(x) - e(y)| = 1$. Si les exposants sont égaux, $FL(x-y) = x-y$. Supposons que $|e(x) - e(y)| = 1$. Alors x et y sont de même signe. En effet si x et y étaient de signes opposés, soit $x - y$ serait soit négatif et de valeur absolue supérieure à x et y, soit positif et supérieur à x et y. Ces deux possibilités conduisent à $e(x-y) = max(e(x), e(y)) > min(e(x), e(y))$, ce qui contredit la condition du théorème. Sans tenir compte de l'exposant et quitte à échanger x et y, on a donc l'encadrement : $B^{-1} \leq y < B^0 = 1 \leq x < B$. La représentation de x en base B est alors $x = x_1.x_2 \dots x_t$ et y s'écrit $y = 0.y_1 \dots y_t$. Posons la soustraction, en posant $z = x - y = z_1.z_2 \dots z_t$ ($0 < z < B$) :

$$\begin{array}{r} x_1.x_2 \dots x_{t-1}x_t \\ 0.y_1 \dots y_{t-1}y_t \\ \hline z_1.z_2 \dots z_t z_{t+1} \end{array}$$

Comme $e(x-y) \leq e(y)$ et $B^{-1} \leq y < 1$, on en déduit que $z < 1$, d'où $z_1 = 0$. La différence $z = x - y$ a t chiffres significatifs $z_2, \dots, z_{t+1}$ donc $z = FL(x-y) = x-y$. □

CONCLUSION :

A partir de la représentation d'un nombre réel en base B, il est possible d'élaborer des systèmes de représentation de NVF très étendus. Le système "IEEE extended" $\mathbb{F}(2, 64, -16381, 16384)$ permet de représenter des réels compris entre 2^{-16382} et $2^{16320}(2^{64} - 1)$, avec une erreur relative inférieure à $u = 5.10^{-20}$! Nous avons de plus montré que cette représentation de réels par des NVF est unique. Ainsi chacun des NVF caractérise un nombre réel, ce qui permet d'utiliser des nombres et d'effectuer des calculs avec une grande fiabilité au niveau des résultats mathématiques.

Les NVF sont en outre utilisés sous leur forme naturelle qui est leur représentation décimale. Ils apparaissent sous cette forme lorsque l'on utilise un calculateur électronique. Les NVF ainsi représentés ont des propriétés similaires à ceux dont la représentation est définie dans la définition 1.2.1. , ceci étant dû à l'équivalence des deux représentations que nous avons établie. La représentation décimale des NVF est donc efficace pour l'approximation et les calculs avec une partie assez grande de $\mathbb{R}$.

L'étude de la répartition des NVF d'un système donné est importante pour le choix optimal du système à utiliser en fonction des capacités de mémoire du calculateur. On a vu que si l'on veut représenter beaucoup de nombres, il faut se donner une grande précision et un domaine d'exposants large. La plage d'entiers à virgule flottante (entre 1 et $M = B^t$) sera d'autant plus étendue que la précision sera grande.

L'intérêt des NVF consiste finalement à se donner un outil accessible pour les calculs informatiques. Les NVF sont arrondis grâce à la fonction d'arrondissement, pour leur donner la taille souhaitée (en général aux alentours de 10 chiffres). Les calculs sont ensuite effectués avec les NVF. Le résultat est éventuellement corrigé en augmentant temporairement la précision pour obtenir un résultat dont l'erreur relative par rapport au résultat exact sera inférieure à $u = \frac{\epsilon_{\mathbb{F}}}{2}$ (cf. théorème 3.2.3.).

Ainsi le rôle des NVF est majeur dans la mise au point de programmes informatiques pour les opérations arithmétiques usuelles visant à une vitesse d'exécution mais aussi à une fiabilité des calculs satisfaisantes. D'où l'intérêt de l'utilisation des NVF en vue de la résolution de problèmes complexes.

A. Programmes Maple pour la représentation graphique des NVF normalisés.

A.1 $\mathbb{F}(2, 3, -1, 1)$

```
restart;
l:=[];
for m from 2^(3-1) to 2^3-1 do
   for e from -1 to 1 do l:=[op(l),evalf(m*2^(e-3))]; od; od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..1.75,color=red);
```

A.2 $\mathbb{F}(4, 3, -1, 1)$

```
restart;
l:=[]; for m from 4^(3-1) to 4^3-1 do
   for e from -1 to 1 do l:=[op(l),evalf(m*4^(e-3))];od;od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..3.9375,color=red);
```

A.3 $\mathbb{F}(2, 3, -10, 10)$

```
restart;
l:=[]; for m from 2^(3-1) to 2^3-1 do
   for e from -10 to 10 do l:=[op(l),evalf(m*2^(e-3))];od;od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..896,color=red);
```

B. Programmes Maple pour la représentation graphique des NVF dénormalisés.

B.1 $\mathbb{G}(2, 3, -3)$

```
restart;
l:=[];
for m from 1 to 2^(3-1)-1 do l:=[op(l),evalf(m*2^(-3-3))]; od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..0.046875,color=red);
```

B.2 $\mathbb{G}(2, 5, -4)$

```
restart;
l:=[];
for m from 1 to 2^(5-1)-1 do l:=[op(l),evalf(m*2^(-4-5))]; od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..0.029296875,color=red);
```

B.3 $\mathbb{G}(10, 3, -1)$

```
restart; l:=[]:
for m from 1 to 10^(3-1)-1 do l:=[op(l),evalf(m*10^(-1-3))]: od: l:
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..0.0099,color=red);
```

C. Programmes Maple pour la représentation graphique des NVF normalisés et dénormalisés.

C.1 $\mathbb{F}(2,3,-5,5) + \mathbb{G}(2,3,-5)$

```
restart;
l:=[];for m from 2^(3-1) to 2^3-1 do
   for e from -1 to 1 do l:=[op(l),evalf(m*2^(e-3))];od;od;
for m from 1 to 2^(3-1)-1 do l:=[op(l),evalf(m*2^(-5-3))]; od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..28,color=red);
```

C.2 $\mathbb{F}(2,3,-10,10) + \mathbb{G}(2,3,-10)$

```
restart;
l:=[];for m from 2^(3-1) to 2^3-1 do
   for e from -10 to 10 do l:=[op(l),evalf(m*2^(e-3))];od;od;
for m from 1 to 2^(3-1)-1 do l:=[op(l),evalf(m*2^(-10-3))]; od; l;
with(plots): plot([l[i] $i=1..nops(l)],x=0..10,y=0..896,color=red);
```

Bibliographie

[1] Alfio Quarteroni, Ricardo Sacco et Fausto Saleri. *Matematica Numerica*. Springer, Milan, 1998.

[2] Nicolas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Sian, seconde édition.