Chapitre 1

Les nombres flottants

Dans ce chapitre, on s'intéresse à la représentation des nombres réels. Le système décimal originaire de l'inde nous a été transmis par l'intermediaire des mathématiciens arabes. On peut représenter un nombre dans d'autres bases que le système décimal. Ce chapitre est divisé en deux parties. Dans la première partie, on traite de la représentation des entiers et des réels dans une base β quelconque. Dans la deuxième partie, on s'intéresse à la représentation des réels sur ordinateur. Un nombre réel peut nécessiter un nombre infini de chiffres pour sa représentation. Mais un ordinateur n'a qu'une capacité finie de stockage. Cette simple remarque implique que l'on travaille toujours avec des approximations. Il faut donc veiller à ce que les approximations soient assez précises, en dépit de quoi on peut aboutir à des erreurs conséquentes.

1.1 Numérotation en base β

Exercice 1. Soit β un nombre entier strictement supérieur à 1. Montrer que pour tout entier n supérieur ou égal à 1, il existe un unique entier p et des entiers d_i, $0 \leq i \leq p$ compris entre 0 et $\beta - 1$, avec $d_p \neq 0$ tels que

$$n = \sum_{i=0}^{p} \beta^i d_i \tag{1.1}$$

Le membre de droite de l'équation (1.1) donne la représentation de n en base β, également notée

$$n = (d_p d_{p-1} ... d_1 d_0)_\beta.$$

Correction

Soit n un entier supérieur ou égal à 1. On sait alors qu'il existe un unique entier

$p \in \mathbb{N}$ tel que

$$\beta^p \leq n < \beta^{p+1}$$

On effectue alors la division euclidienne de n par β^p. Il existe un unique couple (d_p, r_p) tel que

$$n = d_p\beta^p + r_p$$

où $d_p \in \{1, ..., \beta - 1\}$ et $r_p \in \{0, ..., \beta^p - 1\}$. Si $r_p = 0$ alors, c'est terminé, sinon, on écrit :

$$r_p = d_{p-1}\beta^{p-1} + r_{p-1}$$

où $d_{p-1} \in \{0, ...\beta - 1\}$ et $r_{p-1} \in \{1, \beta^{p-1} - 1\}$.
Alors,

$$n = d_p\beta^p + d_{p-1}\beta^{p-1} + r_{p-1}$$

En poursuivant ainsi, on trouve finalement que

$$n = d_p\beta^p + d_{p-1}\beta^{p-1} + ... + d_1\beta^1 + r_1 = d_p\beta^p + d_{p-1}\beta^{p-1} + ... + d_1\beta^1 + d_0$$

.
Cette écriture est unique. En effet, supposons qu'il y en ait deux, alors :

$$0 = n - n = \sum_i \tilde{d}_i\beta^i \neq 0$$

Ce qui est une contradiction car si I est le grand entier tel que $\tilde{d}_I = 0$ alors

$$\beta^I > \sum_{i=0}^{I-1} |\tilde{d}_i|\beta^i$$

.

Nous utilisons d'ordinaire la base 10. Cet exercice montre que l'on pourrait utiliser toute autre base. Sur ordinateur, les calculs se font en base 2 (numérotation binaire, avec les chiffres 0 et 1), en base 8 (numérotation octale, avec les chiffres 0 à 7) ou en base 16 (avec les chiffres 0 à 9 et les lettres A à F).

Exercice 2. Soit $b = (142)_8$ en numérotation octale, donner la représentation de b en base 10 et en base 16.
Correction

$$\begin{aligned} (142)_8 &= 2 \times 1 + 4 \times 8 + \times 8^2 \\ &= 64 + 31 + 2 \\ &= 98 \text{ en base } 10 \end{aligned}$$

En base 16, on a :

$$16 < 98 < 16^2$$

et,

$$98 = 6 \times 16 + 2$$

D'où

$$b = (62)_{16} \text{ en base } 16.$$

Par extension, on peut représenter tout nombre réel x par la somme suivante :

$$x = \sum_{i=-\infty}^{p} \beta^i d_i \tag{1.2}$$

Cependant avec cette notation un nombre entier peut avoir deux représentations distinctes en base β, comme le montre l'exercice suivant.

Exercice 3. Soit $b = \beta - 1$. Montrer que dans toute base β, on a :

$$1 = (0.bbbbb...)_\beta = \sum_{i=-\infty}^{-1} b$$

Par exemple en base dix, on a :

$$36 = 35.9999....$$

En base 10, lorsque le quotient d'un nombre rationnel ne tombe pas juste, sa représentation est toujours périodique à partir d'un certain rang. Le résultat est vrai dans n'importe quelles base comme le montre l'exercice suivant.

Exercice 4.

1. Calculer le développement décimal de $\frac{1}{7}$.
2. Soient m et n deux entiers premiers entre eux tels que $m < n$. On pose $r_0 = m$, et on définit par récurrence $d_{-(j+1)}$ et $r_{-(j+1)}$ comme étant respectivement le quotient et le reste de la division euclidienne de βr_{-j} par n :
   $$\beta r_{-j} = n d_{-(j+1)} + r_{-(j+1)} \text{ avec } 0 \leq r_{-(j+1)} < n.$$
   Montrer que pour tout $j \geq 1$, $0 \leq d_{-j} < \beta$.
3. Montrer que $\overline{0.d_{-1}d_{-2}...}$ est le développement de $\frac{m}{n}$ en base β.
4. Montrer qu'il existe deux entiers k et l tels que r_{-k} et r_{-l} sont égaux. En déduire que le développement de $\frac{m}{n}$ est périodique à partir d'un certain rang.

Correction

1. On a que
$$\frac{1}{7} = 0,142857142857...$$
2. Puisque $\beta m = d_{-1}n + r_{-1}$, on a que $d_{-1} < \beta$. Sinon on aurait $d_{-1}n \geq \beta n \geq \beta m$. Par ailleurs $0 \leq r_{-1} < n$ (reste de la division euclidienne). De même, on a que $0 \leq d_{-j} \leq \beta$ et $0 \leq r_{-j} < n$ pour tout $j \in \mathbb{N}$.
3. On montre par récurrence sur i que,
$$\frac{m}{n} = \sum_{j=1}^{i} d_{-j}\beta^{-j} + \frac{r_{-i}}{n}\beta^{-i}.$$
Ceci est vrai à l'ordre 1 puisque,
$$\beta m = d_{-1}n + r_{-1}.$$
Donc,
$$\frac{m}{n} = d_{-1}\beta^{-1} + \frac{r_{-1}}{n}\beta_{-1}.$$
Supposons le résultat vrai jusqu'à l'ordre i. Montrons qu'il est vrai à l'ordre $i+1$. On sait que
$$\beta r_{-j} = nd_{-(j+1)} + r_{-(j+1)},$$
donc,
$$\frac{r_{-j}}{n} = d_{-(j+1)}\beta^{-1} + \frac{r_{-(j+1)}}{n}\beta^{-1}.$$
Or d'après l'hypothèse de récurrence, on a :
$$\frac{m}{n} = \sum_{j=1}^{i} d_{-j}\beta^{-j} + \frac{r_{-i}}{n}\beta^{-i}.$$
Donc,
$$\begin{aligned}\frac{m}{n} &= \sum_{j=1}^{i} d_{-j}\beta^{-j} + (d_{-(i+1)}\beta^{-1} + \frac{r_{-(i+1)}}{n}\beta^{-1})\beta^{-i} \\ &= \sum_{j=1}^{i+1} d_{-j}\beta^{-j} + \frac{r_{-(i+1)}}{n}\beta^{-(i+1)}.\end{aligned}$$
Ce qui termine la preuve par récurrence. Par ailleurs,
$$|\frac{r_{-i}}{n}\beta^{-i}| \leq \beta^{-i} \to 0 \text{ quand } i \to +\infty.$$

Finalement,

$$\frac{m}{n} = \sum_{j=1}^{+\infty} d_{-j}\beta^{-j}.$$

4. On sait que pour tout $j \in \mathbb{N}^*$, $r_{-j} \in \{0, ..., n-1\}$. Donc les r_{-j} ne peuvent prendre qu'un nombre fini de valeurs distinctes (au plus n). Donc il existe un couple (k, l) tel que $r_{-k} = r_{-l}$. D'après l'unicité de la division euclidienne, cela conduit à la périodicité du développement à partir d'un certain rang. On peut noter que dans le développement périodique, la séquence sera d'une longeur d'au plus $n - 1$.

Exercice 5.

1. En base 10, quel nombre rationnel est égal à 0,123123123...(le développement est périodique de période 123)
2. Soit $\beta \in \mathbb{N}$, $\beta > 1$. En généralisant le raisonnement de la question précédente, montrer que tout nombre dont l'écriture est périodique à partir d'un certain rang, est égal à un nombre rationnel.

Correction

1.

$$\begin{aligned} 0,123123123... &= 0,123 \times \sum_{j=0}^{+\infty} (10^{-3})^j \\ &= 0,123 \times (\frac{1}{1-10^{-3}}) \\ &= 0,123 \times \frac{10^3}{10^3-1} \\ &= \frac{123}{999} \\ &= \frac{41}{333} \end{aligned}$$

2. Soit x un nombre compris entre 0 et 1 dont le développement en base β est périodique à partir d'un certain rang. Sans perte de généralité, on suppose que

$$x = (0, d_{-1}d_{-2}...d_{-q}d_{-1}d_{-2}...d_{-q}...)_\beta.$$

Soit $v = \sum_{i=1}^{q} d_{-i}\beta^{-i}$. Alors,

$$\begin{aligned} x &= v \sum_{j=0}^{+\infty} (\beta^{-q})^{-j} \\ &= v \frac{1}{1-\beta^{-q}} \\ &= \frac{\beta^q v}{\beta^q - 1} \end{aligned}$$

1.2 Représentation des nombres en machine

Sur machine on représente l'ensemble des nombres réels par les flottants (float) ou nombres à virgule flottante. L'ensemble de ces nombres est décrit par une base β, un nombre r de chiffres significatifs, et deux entiers e^- et e^+. Tout nombre à virgule flottante est de la forme :

$$s(m)_\beta \beta^j$$

où j est compris entre e^- et e^+, m est la mantisse codée sur r chiffres significatifs et s désigne le signe du nombre. On adjoint à ce système le nombre 0. Il convient de remarquer qu'avec cette représentation un nombre peut être écrit de plusieurs manières. Ainsi pour $\beta = 10$ et $r = 5$, on a , $34562 \times 10^1 = 3,4562 = 34562 \times 10^{-4}$. Pour avoir l'unicité de la notation, on fixe par convention la place de la virgule dans la mantisse. Généralement la virgule peut se situer juste après le premier chiffre non nul de la mantisse, juste avant, ou ne pas avoir de virgule. On définit ainsi un ensemble fini de nombres, noté $\mathcal{F}(\beta, r, e^-, e^+)$ qui vont représenter l'ensemble des réels. La norme IEEE754 (1985) a définit 4 standards de nombres flottants (simple précision, simple précision étendue devenu obsolète, double précision, double précision étendue). La représentation simple précision est codée sur 32 bits tandis que la double précision est codée sur 64 bits. Pour donner un exemple, le type float du langage C est de type simple précision tandis que le type double est de type double précision.

Le format simple précision

Dans le format simple précision, on a $\beta = 2, e^- = -126, e^+ = 127$, l'exposant est codé sur 8 bits, la mantisse m sur 23 bits et le signe sur un bit. Le bit de signe est 1 si le chiffre est négatif, 0 si le chiffre est positif. La virgule dans la mantisse est placée juste après le premier chiffre non nul. Comme on est en base 2, il n'y a pas nécessité de coder ce chiffre qui est nécessairement égal à 1. Ceci est l'écriture dite normalisée. Ainsi le plus petit nombre normalisé est égal à :

$signe$	$exposant$	$mantisse$
$1bit$	$8bits$	$23bits$

FIGURE 1.1 – Le format simple précision

Format	taille	r	e^-	e^+	Nombre normalisé minimum	Nombre normalisé maximum
Simple	32	23	-126	127	$1.175... \times 10^{-38}$	$3.048... \times 10^{38}$
Double	64	52	-1022	1023	$1.112... \times 10^{-308}$	$1.797... \times 10^{308}$

FIGURE 1.2 – Les formats simple et double précicions

$$2^{-126} \simeq 1.17549449095 \times 10^{-38},$$

et le plus grand nombre normalisé est égal à :

$$2^{127}(2 - 2^{-23}) \simeq 3.40282326356 \times 10^{38}.$$

En plus des nombres normalisés, la norme définit également les valeurs $+\infty$, $-\infty$, NaN(Not-a-Number) et des nombres dénormalisés. Les nombres dénormalisés sont utilisés pour représenter des nombres très petits. Dans ce cas, le nombre situé avant la virgule est nul. Ces informations sont codées grâce à l'exposant. Ainsi comme on l'a vu, huit bits sont utilisés pour l'exposant. Pour huit bits, il y a $2^8 = 256$ nombres différents. Pourtant, on a vu que l'exposant variait de -126 à 127 soit 254 nombres. En fait, la valeur d'exposant -127 est utilisée pour désigner un nombre dénormalisé. La valeur 128 est utilisée pour représenter les valeurs $+\infty$, $-\infty$, NaN. Enfin pour représenter les valeurs négatives en binaire, l'exposant est représenté décalé de $+127$. C'est à dire que l'exposant -127 sera représenté par la valeur 0 en binaire, -126 sera représenté par la valeur 1 en binaire,..., 128 sera représenté par la valeur 255 en binaire. Si l'exposant vaut -127 et si la mantisse est nulle, c'est le nombre 0 qui est représenté. Il y a donc un 0^+ et un 0^- selon le bit de signe. Si l'exposant vaut -127 et si la mantisse est non nulle, c'est le nombre dénormalisé qui est représenté. Ainsi le plus petit nombre dénormalisé vaut :

$$2^{-126} \times 2^{-23} = 2^{-149} \simeq 1.40129846432 \times 10^{-45}.$$

Remarquez que bien que la valeur de l'exposant soit -127 on conserve -126 pour le calcul pour assurer la continuité avec les nombres normalisés. Si l'exposant vaut $+128$, et si la mantisse est nulle, c'est les valeurs $+\infty$ et $-\infty$ qui sont représentées. Enfin si l'exposant vaut $+128$, et si la mantisse est non nulle c'est le NaN qui est représenté.

Quelques erreurs en arithmétique flottante célèbres

Le crash d'Ariane 5

Le quatre juin 1996, c'est une erreur de programmation qui a causé le crash de la fusée européenne Ariane 5. Plus précisément, un réel codé sur 64 bits donnant la vitesse horizontale de la fusée était converti en un entier signé sur 16 bits. Or l'entier obtenu était plus grand que $32767 = 2^{15} - 1$, le plus grand entier représentable sur 16 bits. La convertion échouait donc. Tous les tests logiciels avaient pourtant réussi, mais ils avaient été effectués avec les données d'Ariane 4 pour laquelle la vitesse horizontale restait inférieure au maximum de 32767.

L'anti-missile manque sa cible

Pendant la guerre du Golfe, un anti-missile Patriot tiré de Dahran (Arabie Saoudite) a manqué l'interception d'un missile irakien Scud. Ce dernier a tué 28 soldats et blessé quelque cent autres personnes. L'erreur a découlé d'une imprécision dans le calcul de la date de l'anti-missile Patriot. Celui-ci dispose en effet d'un processeur interne, qui calcule l'heure en multiples de dixièmes de secondes. Le nombre de dixièmes de secondes depuis le démarrage du processeur est stocké dans un registre entier, puis multiplié par une approximation sur 24 bits de $\frac{1}{10}$ pour obtenir le temps en secondes. Or l'écriture de $\frac{1}{10}$ en base 2 est infinie puisque l'on a $(\frac{1}{10})_{10} = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + ...$ L'approximation de 0.1 sur 24 bits stockée valait $\frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + ... + \frac{1}{2^{-21}}$. Cela donne une erreur d'environ 9.5×10^{-8}. Le processeur du missile ayant été démarré une centaine d'heures auparavant, l'erreur totale était donc de 0,34 secondes ($100 \times 36000 \times 0.95 \times 10^{-}7$). Sachant qu'un missile scud parcourt 1676 mètres par seconde, le missile Scud parcourt plus de 500 mètres en 0,34 secondes.

Exercice 6. On choisit $\beta = 2$, $r = 3$, $e^- = -1$ et $e^+ = 2$. On suppose que la virgule est placée juste après le premier chiffre non nul. Donner la valeur en base de 10 de tous les nombres flottants normalisés de ce système, et les dessiner sur un segment de droite centré en zéro.

Même question pour les nombres dénormalisés (avec $j = -1$).

Correction

Le différents nombres normalisés sont représentés ci-dessous :

Mantisse	$j=-1$	$j=0$	$j=1$	$j=2$
000	$1/2$	1	2	4
001	$9/16$	$9/8$	$9/4$	$9/2$
010	$10/16$	$10/8$	$10/4$	$10/2$
011	$11/16$	$11/8$	$11/4$	$11/2$
100	$12/16$	$12/8$	$12/4$	$12/2$
101	$13/16$	$13/8$	$13/4$	$13/2$
110	$14/16$	$14/8$	$14/4$	$14/2$
111	$15/16$	$15/8$	$15/4$	$15/2$

Pour $j=-1$, les nombres dénormalisés sont :

Mantisse	$j=-1$
000	0
001	$1/16$
010	$2/16$
011	$3/16$
100	$4/16$
101	$5/16$
110	$6/16$
111	$7/16$

On définit maintenant l'application arrondi A qui est l'approximation dans $\mathcal{F}(\beta, r, e^-, e^+)$ d'un réel quelconque : pour tout $x \in \mathbb{R}$, soit $[f, f']$ le plus petit intervalle contenant x tel que f et f' appartiennent à $\mathcal{F}(\beta, r, e^-, e^+)$. Alors $A(x)$ est égal à celui qui est le plus proche de x. Si f et f' sont équidistants de x, alors $A(x)$ sera déterminé de diverses manières qui dépend de la machine utilisée. Par exemple en base 2, on peut choisir, le nombre pair (dernier bit à 0) entre f et f'. On définit également les opérations mathématiques sur les flotants en posant :

$$
\begin{aligned}
f \oplus f' &= A(f + f') \\
f \ominus f' &= A(f - f') \\
f \otimes f' &= A(f * f') \\
f \oslash f' &= A(f / f')
\end{aligned}
$$

Exercice 7.

1. Calculer $A(\frac{1}{3})$ dans le système $\mathcal{F}(2, 3, -1, 2)$
2. calculer :
$$1 \oplus \frac{1}{16} \ominus \frac{1}{8} \text{ et } 1 \ominus \frac{1}{8} \oplus \frac{1}{16}.$$
Que remarquez vous ?

Correction

1. On a :
$$\frac{15}{48} < \frac{1}{3} < \frac{18}{48},$$
donc $\mathcal{A}(\frac{1}{3}) = \frac{5}{16}$ (si l'on inclut les nombres dénormalisés.
2. $$1 \oplus \frac{1}{16} \ominus \frac{1}{8} = \frac{7}{8} \text{ et } 1 \ominus \frac{1}{8} \oplus \frac{1}{16} = \frac{15}{16}.$$
On en déduit que l'addition n'est pas commutative.

L2, 2012-2013. Méthodes numériques Université du Havre

Feuille de TP 1

Exercice 1. Déterminer le système $\mathcal{F}(\beta, r, e^-, e^+)$ utilisé par Scilab pour la représentation des nombres flottants.

Exercice 2. Soit l'équation du second degré :

$$p(x) = x^2 - 160x + 1 = 0 \tag{1.3}$$

1. Calculer numériquement les valeurs x_1 et x_2 ($x_1 < x_2$) des deux solutions de l'équation en utilisant les formules usuelles.
2. Calculer $p(x_1)$ et $p(x_2)$. Que constatez vous ?
3. Déterminer la nouvelle valeur numérique $\tilde{x}_1$ en utilisant

$$\tilde{x}_1 = \frac{1}{x_2}.$$

   Calculer $p(\tilde{x}_1)$. Que constatez vous ?
4. Pouvez-vous proposer une explication de ce phénomène ?

Exercice 3. Un calcul approché de π

1. Montrer que :

$$\lim_{n \to +\infty} n \sin(\frac{\pi}{n}) = \pi.$$

2. On pose

$$x_k = 2^k \sin(\frac{\pi}{2^k}).$$

   Montrer que pour tout nombre réel α compris entre 0 et $\frac{\pi}{2}$ on a

$$\sin(\frac{\alpha}{2}) = \sqrt{\frac{1}{2}(1 - \sqrt{1 - \sin^2 \alpha})}$$

   et en déduire que

$$x_{k+1} = 2^k \sqrt{2(1 - \sqrt{1 - (x_k/2^k)^2})}.$$