Arithmétique à virgule flottante (2023)

Corrigé
*

On définit un format de nombre à virgules flottantes φ par trois entiers :

1. une précision $p \in \mathbb{N}^*$;
2. un exposant minimal $e_{\min} \in \mathbb{Z}$;
3. un exposante maximal $e_{\max} \in \mathbb{Z}$, avec $e_{\min} < 0 < e_{\max}$.

Étant donné un format $\varphi = (p, e_{\min}, e_{\max})$, l'ensemble des nombres flottants définis via ce format est :

$$\begin{aligned}\mathbb{FP}_\varphi = \{(-1)^s \cdot M \cdot 2^{e-p+1} \mid s \in \{0,1\}, e \in \mathbb{Z}, e_{\min} \leqslant e \leqslant e_{\max}, M \in \mathbb{N}, 2^{p-1} \leqslant M < 2^p\} \\ \sqcup \{(-1)^s \cdot M \cdot 2^{e_{\min}-p+1} \mid s \in \{0,1\}, M \in \mathbb{N}, 0 \leqslant M < 2^{p-1}\}\end{aligned}$$

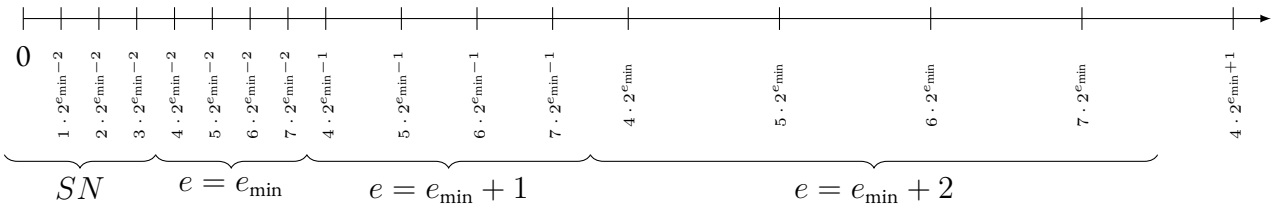
La précision p correspond ainsi au nombre maximal de chiffres dans l'écriture de M en base 2. Par exemple, le format double précision utilisé par la plupart des machines est $\varphi_{64} = (53, -1022, 1023)$. On suppose que φ est fixé.
On note Ω le plus grand nombre flottant représentable et $u = 2^{-p}$.

Question 1

On suppose dans cette question que $p = 3$ et $e_{\max} \geqslant e_{\min} + 3$. Représenter schématiquement les nombres flottants sur $[0, 2^{e_{\min}+3}]$.

Solution.

Dans la définition de $\mathbb{FP}$, on note N le premier ensemble (correspondant à $\{(-1)^s \cdot M \cdot 2^{e-p+1} \mid s \in \{0,1\}, e \in \mathbb{Z}, e_{\min} \leqslant e \leqslant e_{\max}, M \in \mathbb{N}, 2^{p-1} \leqslant M < 2^p\}$) et SN le second (correspondant à $\{(-1)^s \cdot M \cdot 2^{e_{\min}-p+1} \mid s \in \{0,1\}, M \in \mathbb{N}, 0 \leqslant M < 2^{p-1}\}$)



Les nombres de SN (appelés *nombres sous-normaux*) sont espacés de $2^{e_{\min}-2}$; puis on considère des ensembles espacés de $2^{e_{\min}-2}$, puis de $2^{e_{\min}-1}$, puis de $2^{e_{\min}}$, ...
Une interprétation utile est qu'un élément de N peut être écrit comme un nombre de la forme $x' \times 2^e$ où $1 \leqslant x' < 2$: c'est une forme d'écriture scientifique en binaire, où e est compris entre $e_{\min}$ et $e_{\max}$. Les nombres sous-normaux s'écrivent sous la forme $x' \times 2^{e_{\min}}$, où $0 \leqslant x' < 1$.

Pour tout réel x tel que $|x| < \Omega$, on note $\mathrm{RN}(x)$ le nombre flottant le plus proche de x.

Question 2

La définition de $\mathrm{RN}(x)$ est-elle bien construite ?

Solution.

Non ; il suffit de remarquer qu'il y a un nombre fini d'éléments de $\mathbb{FP}$ qui sont totalement ordonnés par l'ordre usuel. Considérons x et y deux tels flottants consécutifs : alors $\dfrac{x+y}{2}$ n'admet pas « un » nombre flottant le plus proche, mais 2.
En réalité, si x n'est pas littéralement à mi-chemin entre deux flottants consécutifs, $\mathrm{RN}(x)$ est bien défini. En général, on définit donc une règle de rupture d'égalité en « *ties to even* » : on prend la mantisse terminant par un 0.

Question 3

Montrer qu'il existe $a, b \in \mathbb{FP}$ tels que $|a+b| < \Omega$ et $a+b \notin \mathbb{FP}$.

Solution.

Soit $a = 2^{e_{\min}-p+1}$ (le plus petit réel strictement positif représentable) et $b = 2^{e_{\max}}$. Alors $|a+b| < 2b = \Omega$, mais :

$$\begin{aligned} a+b &= 2^{e_{\min}-p+1} + 2^{e_{\max}} \\ &= 2^{e_{\min}-p+1} \times (1 + 2^{e_{\max}-e_{\min}+p-1}) \end{aligned}$$

Alors, si $a+b$ était représentable, l'exposant serait $e_{\min}{}_{-p+1}$ (sans quoi, la mantisse ne serait pas entière), mais $1+2^{e_{\max}-e_{\min}+p-1} \geqslant 1+2^{p+1}$, donc cette mantisse est trop grande pour être valide dans notre représentation.

Question 4

Montrer que pour $2^{e_{\min}} \leqslant |x| \leqslant \Omega$, $\mathrm{RN}(x) = x \cdot (1+\delta)$ avec $|\delta| \leqslant u$.

Solution.

Remarque : il existe des solutions plus simples à cette question.
On étudie le cas x positif. On étudie plutôt deux nombres : $x_- = x \times (1-u)$ et $x_+ = x \times (1+u)$. On cherche à montrer qu'il existe au moins un nombre représentable dans l'intervalle $[x_-, x_+]$.
Comme $2^{e_{\min}} \leqslant x$, alors on peut écrire x comme $x = x' \times 2^e$ où $1 \leqslant x' < 2$ et $e \geqslant e_{\min}$. On pose $x' = 1 + \sum_{i=1}^{+\infty} a_i 2^{-i}$ son écriture binaire. Alors :

$$\begin{aligned} x_- &= x' \times 2^e \times (1-u) \\ &= (x' - ux') \times 2^e \\ &= (1 + \sum_{i=1}^{+\infty} a_i 2^{-i} - 2^{-p} - \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\ x_+ &= (1 + \sum_{i=1}^{+\infty} a_i 2^{-i} + 2^{-p} + \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \end{aligned}$$

Alors posons $y = 1 + \sum_{i=1}^{p-1} a_i 2^{-i} + 2^{-p}$ (ce n'est techniquement pas une écriture binaire directement, on y reviendra plus tard). Alors :

$$x_- = (1 + \sum_{i=1}^{p-1} a_i 2^{-1} + \sum_{i=p}^{+\infty} a_i 2^{-i} - 2^{-p} - \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e$$

$$
\begin{aligned}
&\leqslant (1 + \sum_{i=1}^{p-1} a_i 2^{-1} + \sum_{i=p}^{+\infty} 2^{-i} - 2^{-p} - \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\
&\leqslant (1 + \sum_{i=1}^{p-1} a_i 2^{-1} + 2^{-(p-1)} - 2^{-p} - \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\
&\leqslant (1 + \sum_{i=1}^{p-1} a_i 2^{-1} + 2^{-p} - \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\
&\leqslant (y) \times 2^e
\end{aligned}
$$

De même :

$$
\begin{aligned}
x_+ &= (1 + \sum_{i=1}^{+\infty} a_i 2^{-i} + 2^{-p} + \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\
&= (1 + \sum_{i=1}^{p-1} a_i 2^{-i} + \sum_{i=p}^{+\infty} a_i 2^{-1} + 2^{-p} + \sum_{i=p+1}^{+\infty} a_{i-p} 2^i) \times 2^e \\
&\geqslant (1 + \sum_{i=1}^{p-1} a_i 2^{-i} + 2^{-p}) \times 2^e = y \times 2^e
\end{aligned}
$$

Donc $x_- \leqslant y \times 2^e \leqslant x_+$. De plus, $y \times 2^e$ est représentable ! Donc l'intervalle $[x_-, x_+]$ contient au moins un élément représentable, donc notamment $\mathrm{RN}(x)$.

Remarque : si $|x| < 2^{e_{\min}}$, ça peut devenir faux. En effet, si $x > 0$ est ridiculement petit, alors $\mathrm{RN}(x) = 0$, mais $x \times (1 + \delta) > 0$.

On note $+_{\mathbb{FP}}$ l'addition sur les flottants, qui satisfait $x +_{\mathbb{FP}} y = \mathrm{RN}(x + y)$ lorsque $|x + y| < \Omega$. Dans la suite, on admet que l'on peut toujours utiliser l'équation de la question 4.

Question 5

Soit $(x_1, \ldots, x_n) \in \mathbb{FP}^n$. On considère la fonction SommeIterative définie par l'algorithme suivant :

```
s ← x_1
pour i = 2 à n :
    s ← s +_FP x_i
retourner s
```

On suppose que $n \cdot u < 1$. Montrer que :

$$
\left| \mathrm{SommeIterative}(x_1, \ldots, x_n) - \sum_{i=1}^{n} x_i \right| \leqslant \gamma_{n-1} \cdot \sum_{i=1}^{n} |x_i| \qquad \text{avec } \gamma_n = \frac{nu}{1 - nu}
$$

Solution.

On remarque qu'à chaque étape du calcul :

$$
\begin{aligned}
\mathrm{SommeIterative}(x_1, \ldots, x_k, x_{k+1}) &= \mathrm{RN}(\mathrm{SommeIterative}(x_1, \ldots, x_k) + x_{k+1}) \\
&= (\mathrm{SommeIterative}(x_1, \ldots, x_k) + x_{k+1}) \times (1 + \delta_k)
\end{aligned}
$$

avec un certain $|\delta_k| \leqslant u$. Donc, par récurrence, on obtient la formule :

$$\text{SommeIterative}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \left(x_i \times \prod_{k=\max(i,2)}^{n} (1 + \delta_k) \right)$$

On montre alors que chaque $\prod\prod(1 + \delta_k) \leqslant 1 + \gamma_{n-1}$, qui se prouve par récurrence sur le nombre de facteurs. On conclut alors directement.

On suppose que l'on dispose d'une primitive 2Sum : $\mathbb{FP}^2 \to \mathbb{FP}^2$ telle que pour tous flottants a et b, si $(s, t) = \text{2Sum}(a, b)$, alors :

1. $s + t = a + b$;
2. $s = \text{RN}(a + b)$;
3. $|t| \leqslant u \cdot |s|$;
4. $|t| \leqslant u \cdot |a + b|$.

Question 6

On considère la fonction SommeComposee définie par l'algorithme suivant :

$\pi_1 \leftarrow x_1$
$\sigma_1 \leftarrow 0$
pour $i = 2$ *à* n **:**
  $\pi_1, q_i \leftarrow \text{2Sum}(\pi_{i-1}, x_i)$
  $\sigma_i \leftarrow \sigma_{i-1} +_{\mathbb{FP}} q_i$
retourner $\pi_n +_{\mathbb{FP}} \sigma_n$

On suppose que $n \cdot u < 1$. Montrer que :

$$\left| \text{SommeComposee}(x_1, \ldots, x_n) - \sum_{i=1}^{n} x_i \right| \leqslant u \cdot \left| \sum_{i=1}^{n} x_i \right| + \gamma_{n-1}^2 \cdot \sum_{i=1}^{n} |x_i|$$

Justifier ensuite de l'intérêt de l'algorithme des sommes composées.

Solution.

Remarque : cette preuve est très désorganisée; si vous réussissez à trouver comment en simplifier l'écriture, n'hésitez pas à me le communiquer.

Par les propriétés de 2Sum, $\pi_i + q_i = \pi_{i-1} + x_i$.

On cherche à majorer $a_i = \sum_{k=1}^{i} x_i - \pi_i$, qui est l'erreur faite. On observe qu'on maintient σ_i, qui est censé être une approximation de a_i : en effet, à la fin, si on avait $\sigma_n = a_n$, on renverrait la somme exacte. On cherche donc à comparer a_i et σ_i : on pose $\Delta_i = \sigma_i - a_i$. Alors :

$$\begin{aligned}
\sigma_i &= (\sigma_{i-1} + q_i) \times (1 + \delta_i) \\
&= (\Delta_{i-1} + a_{i-1} + q_i) \times (1 + \delta_i) \\
&= (\Delta_{i-1} + \sum_{k=1}^{i-1} x_k - \pi_{i-1} + \pi_{i-1} + x_i - \pi_i) \times (1 + \delta_i)
\end{aligned}$$

$$= (\Delta_{i-1} + \sum_{k=1}^{i} x_k - \pi_i) \times (1 + \delta_i)$$
$$= (\Delta_{i-1} + a_i) \times (1 + \delta_i)$$

Donc :

$$\Delta_i = \sigma_i - a_i = (\Delta_{i-1} + a_i) \times (1 + \delta_i) - a_i$$
$$= \Delta_{i-1} \times (1 + \delta_i) + a_i \times \delta_i$$

On cherche à trouver une expression récursive de Δ_i. On rappelle que $\Delta_1 = a_1 - \sigma_1 = x_1 - \pi_1 - 0 = 0$. Je trouve :

$$\Delta_i = \sum_{k=1}^{i} a_k \delta_k \prod_{j=k+1}^{i} (1 + \delta_j)$$

On remarque par ailleurs que $a_i = \sum_{k=2}^{i} q_k$.
Enfin, le résultat renvoyé est $\pi_n +_{\mathbb{FP}} \sigma_n = (\pi_n + \sigma_n) \times (1 + \delta_0)$. Or :

$$\pi_n + \sigma_n = \sum x_k - a_n + \sigma_n = \sum x_k + \Delta_n$$

Donc :

$$\text{SommeComposee} - \sum x_k = (\sum x_k + \Delta_n)(1 + \delta_0) - \sum x_k = \delta_0 \sum x_k + \Delta_n(1 + \delta_0)$$

Donc :

$$\left|\text{SommeComposee} - \sum x_k\right| \leqslant |u| \cdot \left|\sum x_k\right| + |\Delta_n|(1 + u)$$

Reste à majorer $|\Delta_n|(1 + u)$. On remarque que $\Delta_n = \sigma_n - a_n$: or, ce sont tous les deux les sommes des q_j, si ce n'est que σ_n est la somme itérative, et a_n la somme réelle ! Donc par l'exo précédent :

$$|\Delta_n| \leqslant \gamma_{n-1} \sum_{i=1}^{n} |q_i|$$
$$\leqslant \gamma_{n-1} \sum_{i=1}^{n} u|\pi_i|$$
$$\leqslant \gamma_{n-1} u \sum_{i=1}^{n} (1 + \gamma_{i-1})| \sum_{j=1}^{i} |x_j|$$
$$\leqslant \gamma_{n-1} u \sum_{i=1}^{n} \frac{1}{1 - ((i-1)u)} \times \sum_{j=1}^{n} |x_j|$$

J'ai le droit de majorer $\frac{1}{1 - ((i-1)u)} \leqslant \frac{1}{1 - ((n-1)u)}$.

$$|\Delta_n| \leqslant \gamma_{n-1} u \frac{(n-1)}{1 - (n-1)u} \sum_{j=1}^{n} |x_j| = \gamma_{n-1}^2 \times \sum_{j=1}^{n} |x_j|$$

Pfiou !
L'intérêt : le premier terme d'erreur a une constante indépendante de n, et uniquement dépendante de la précision. Le second terme, lui, dépend certes de n, mais est aussi quadratique en u, donc aura tendance à très vite décroître.

Question 7

On admet les résultats suivants :

Lemme 1 : lemme de Sterbenz

Si $x, y \in \mathbb{FP}$ tels que $\frac{y}{2} \leqslant x \leqslant \frac{2}{y}$, alors $x - y$ est exactement représentable.

Lemme 2

Soient $a, b \in \mathbb{FP}$, si $s = \mathrm{RN}(a + b) < \Omega$, alors $r = (a + b) - s \in \mathbb{FP}$.

Soient a, b deux nombres flottants tels que $a \geqslant b \geqslant 0$. Comment calculer $a + b$ de manière exacte en trois opérations ?

Solution.

Si vous lisez ce corrigé, je n'ai pas le temps de résoudre cette question. Si vous avez une solution (même partielle), n'hésitez pas à me la communiquer.