Arithmétique à virgule flottante (2023)

Sujet
*

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 représentant 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 $[0, 2^{e_{\min}+3}]$.

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 ?

Question 3

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

Question 4

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

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, dans l'hypothèse où, à toute étape de l'algorithme, $2^{e_{\min}} \leqslant |s| \leqslant \Omega$:

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

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 **:**
  $\pi_i, q_i \leftarrow \text{2Sum}(\pi_{i-1}, x_i)$
  $\sigma_i \leftarrow \sigma_{i-1} +_{\mathbb{FP}} q_i$
retourner $\pi_n +_{\mathbb{FP}} \sigma_n$

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

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

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

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 $s = \text{RN}(a+b) < \Omega$, 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 ?