UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Section de mathématiques

Analyse Numérique
Corrigé Série 5

24 octobre 2019

1. **(Arithmétique en virgule flottante et norme IEEE 754)** En cours, on a vu que sur un ordinateur les opérations élémentaires $+$, $-$, $\times$, $/$, $\sqrt{}$ sont précises jusqu'à une erreur relative, donnée par le nombre machine $\varepsilon_{\text{mach}}$. Ainsi, si on a deux nombres a et b représentés exactement dans la mémoire, on peut supposer les choses suivantes par rapport aux opérations élémentaires entre ces deux nombres :
   — $a \oplus b = \text{fl}(a+b) = (a+b)(1+\varepsilon_1)$,
   — $a \ominus b = \text{fl}(a-b) = (a-b)(1+\varepsilon_2)$,
   — $a \otimes b = \text{fl}(a \times b) = a \times b\,(1+\varepsilon_3)$,
   — $a \oslash b = \text{fl}(a/b) = \frac{a}{b}(1+\varepsilon_4)$,
   — $\text{sqrt}(a) = \sqrt{a}\,(1+\varepsilon_5)$,
   où $|\varepsilon_i| \leq \varepsilon_{\text{mach}},\ \forall i$. Ces cas sont les spécialisations aux cinq opérations de l'axiome fondamental de l'arithmétique flottante.
   Souvent les calculs impliquent plusieurs opérations élémentaires effectuées en série, de telle façon que les erreurs d'arrondi s'accumulent. Par exemple, observons ce qui se passe quand on fait l'addition de trois nombres $a, b, c > 0$ (on remarque que quand on additionne plusieurs nombres on doit spécifier l'ordre dans lequel les additions sont calculées) :

$$\begin{aligned} a \oplus (b \oplus c) &= a \oplus (b+c)(1+\varepsilon_1) \\ &= [a + (b+c)(1+\varepsilon_1)]\,(1+\varepsilon_2) \\ &= (a+b+c) + \varepsilon_2(a+b+c) + \varepsilon_1(b+c) + \varepsilon_1\varepsilon_2(b+c) \\ &= (a+b+c) + \varepsilon_2(a+b+c) + \varepsilon_1(b+c) + O(\varepsilon_{\text{mach}}^2). \end{aligned}$$

   Dans cette dernière équation on a négligé (comme on fera toujours) les termes proportionnels à ε^2 (ou $\varepsilon_{\text{mach}}^3, \varepsilon_{\text{mach}}^4, \ldots$) ; ils ont été cachés dans la notation $O(\varepsilon_{\text{mach}}^2)$. La formule nous dit que le résultat de l'addition (après les arrondis) est égal à l'addition correcte $a+b+c$, plus des autres termes. Il est utile pour borner l'erreur :

$$|a \oplus (b \oplus c) - (a+b+c)| \leq \varepsilon_{\text{mach}}\,(|a+b+c| + |b+c|) + O(\varepsilon_{\text{mach}}^2).$$

   (a) $\star$ Est-ce qu'il y a un ordre préférable dans lequel additionner a, b, c afin de réduire l'erreur ? Supposer, par exemple, d'avoir $0 < a \ll b \ll c$.
   Indice : Choisir un autre ordre pour additionner a, b, c et faire les calculs comme montré ci-dessus.
   Sol.: *On choisissant un autre ordre pour additionner a, b, c et en faisant les calculs on obtient*

$$\begin{aligned} (a \oplus b) \oplus c &= (a+b)(1+\varepsilon_1) \oplus c \\ &= [(a+b)(1+\varepsilon_1) + c]\,(1+\varepsilon_2) \\ &= (a+b+c) + \varepsilon_2(a+b+c) + \varepsilon_1(a+b) + \varepsilon_1\varepsilon_2(a+b) \\ &= (a+b+c) + \varepsilon_2(a+b+c) + \varepsilon_1(a+b) + O(\varepsilon_{\text{mach}}^2). \end{aligned}$$

   Ce qui donne l'erreur

$$|(a \oplus b) \oplus c - (a+b+c)| \leq \varepsilon_{\text{mach}}\,(|a+b+c| + |a+b|) + O(\varepsilon_{\text{mach}}^2).$$

   Comme on a supposé $0 < a \ll b \ll c$, on voit qu'en général l'erreur commise en additionnant d'abord les nombres plus grands, c.-à-d. $a \oplus (b \oplus c)$, sera plus grand de l'erreur commis en additionnant d'abord les nombres plus petits, c.-à-d. $(a \oplus b) \oplus c$.

   (b) Dans les formules précédentes, on a supposé que a, b, c sont exactement représentés. Si ce n'est pas le cas, comment est-ce que les formules changent ? Considérer les deux ordres possibles dans lesquels on peut effectuer l'addition, comme vu ci-dessus.

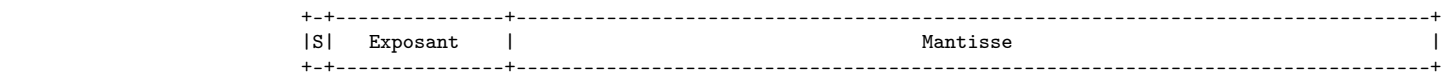
Sol.: *On aura* $\mathrm{fl}(a), \mathrm{fl}(b), \mathrm{fl}(c)$ *au lieu de* a, b, c. *Il faut écrire*

$$\begin{aligned}\mathrm{fl}(a) \oplus [\mathrm{fl}(b) \oplus \mathrm{fl}(c)] &= a\,(1+\varepsilon_1) \oplus [b\,(1+\varepsilon_2) \oplus c\,(1+\varepsilon_3)] \\ &= a\,(1+\varepsilon_1) \oplus [b\,(1+\varepsilon_2) + c\,(1+\varepsilon_3)]\,(1+\varepsilon_4) \\ &= \{a\,(1+\varepsilon_1) + [b\,(1+\varepsilon_2) + c\,(1+\varepsilon_3)]\,(1+\varepsilon_4)\}\,(1+\varepsilon_5) \\ &= (a+b+c) + \varepsilon_5(a+b+c) + \varepsilon_4(b+c) + a\,\varepsilon_1 + b\,\varepsilon_2 + c\,\varepsilon_3 + O(\varepsilon_{\mathrm{mach}}^2),\end{aligned}$$

et

$$[\mathrm{fl}(a) \oplus \mathrm{fl}(b)] \oplus \mathrm{fl}(c) = (a+b+c) + \varepsilon_5(a+b+c) + \varepsilon_4(a+b) + a\,\varepsilon_1 + b\,\varepsilon_2 + c\,\varepsilon_3 + O(\varepsilon_{\mathrm{mach}}^2).$$

(c) ⋆ (Quad précision : Quadruple précision en virgule flottante) La quadruple précision ressemble beaucoup à la double précision ; on a la base $B = 2$ sur 128 bits :



Le nombre de bits de l'exposant augmente de 11 à 15 et ces de la mantisse à 112.

i. Quelle est l'$\varepsilon_{\mathrm{mach}}$?

ii. Quel est le plus petit nombre normalisé qu'on peut représenter, en valeur absolue ?

iii. Quel est le plus grand nombre normalisé qu'on peut représenter ? *Indice : La formule* $\sum_{k=1}^{n} 2^{-k} = 1 - 2^{-n}$ *sera utile.*

Sol.:

i. Le plus petit nombre strictement plus grand que 1 *est* $(1.\underbrace{00\ldots00}_{111}1)_2$ *; la précision relative est donc* $\varepsilon_{\mathrm{mach}} = 2^{-112} \approx 1.9 \times 10^{-34}$.

ii. Le plus petit nombre normalisé qu'on peut représenter est $(1.\underbrace{00\ldots00}_{112})_2 \times 2^{-16382} \approx 3.36 \times 10^{-4932}$.

iii. Le plus grand nombre normalisé qu'on peut représenter est $(1.\underbrace{11\ldots11}_{112})_2 \times 2^{16383} = (1 + \sum_{k=1}^{112} 2^{-k}) \times 2^{16383} = (1 + (1 - 2^{-112})) \times 2^{16383} \approx 1.189 \times 10^{4932}$.

2. **(Interpolation avec une subdivision uniforme)** On peut montrer que la borne de l'erreur de l'interpolation d'une fonction $f(x)$ pour $n+1$ points équidistants dans $[a, b]$ est

$$\max_{a \le x \le b} |f(x) - p_n(x)| \le \frac{M_{n+1}}{4(n+1)} \left(\frac{b-a}{n}\right)^{n+1},$$

où $M_n = \max_{a \le \xi \le b} |f^{(n)}(\xi)|$.

(a) Soit f analytique sur $[-a, a]$, on définit R_a comme la distance entre l'intervalle $[-a, a]$ et le pôle le plus proche ($R_a = +\infty$ en l'absence de pôles). Une propriété des fonctions analytiques est que (pour les mathématiciens, voir le cours d'analyse complexe)

$$\sup_{|z| \le a} |f^{(n)}(z)| \le C_r n! r^{-n} \quad 0 < r < R_a,$$

où C_r est une constante qui dépend de r mais pas de n.

i. En utilisant la formule de Stirling qui est $(n-1)! \le n^n e^{-(n-1)}$, montrer que

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r e}{4} \left(\frac{n+1}{n}\right)^{n+1} \left(\frac{2a}{re}\right)^{n+1}$$

Sol.: *On a vu en cours que*

$$\max_{a \le x \le b} |f(x) - p_n(x)| \le \frac{M_{n+1}}{4(n+1)} \left(\frac{b-a}{n}\right)^{n+1}$$

où $M_n = \max_{a \le \xi \le b} |f^n(\xi)|$. *En utilisant la propriété des fonctions analytiques sur* $[-a, a]$ *on a*

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r (n+1)! r^{-(n+1)}}{4(n+1)} \left(\frac{2a}{n}\right)^{n+1}$$

puis en appliquant la formule de Stirling

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r (n+1)^{n+1} e}{4} \left(\frac{2a}{rne}\right)^{n+1}$$

finalement

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r e}{4} \left(\frac{n+1}{n}\right)^{n+1} \left(\frac{2a}{re}\right)^{n+1}$$

ii. En utilisant $1 + x \le e^x$ montrer que pour $n > 1$

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r e^3}{4} \left(\frac{2a}{re}\right)^{n+1}$$

Sol.: *On a que* $1 + \frac{1}{n} \le e^{\frac{1}{n}}$, *donc*

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r e}{4} (e^{\frac{1}{n}})^{n+1} \left(\frac{2a}{re}\right)^{n+1}$$

On note que $e^{1+\frac{1}{n}} \le e^2$, *donc*

$$\max_{|x| \le a} |f(x) - p_n(x)| \le \frac{C_r e^3}{4} \left(\frac{2a}{re}\right)^{n+1}$$

iii. Trouver la borne inférieure pour R_a tel qu'on obtient la convergence.
Sol.: *L'erreur tend vers zéro si* $2a < re$ *ce qui équivaut à*

$$R_a > \frac{2a}{e} \approx 0.74a.$$

(b) Pour les fonctions ci-dessous, trouvez R_a et expliquez le comportement de la convergence de leurs polynômes d'interpolation sur $[-a, a]$ par rapport aux valeurs de a.

i. $f(x) = \exp(x)$

ii. $g(x) = \log(0.1 - x)$

iii. $h(x) = \frac{1}{1+x^2}$

Sol.: $f(x) = e^x$, $R_a = \infty$ *et donc* p_n *converge vers* f *pour tout* $a > 0$ *(Aucune restriction).*
$g(x) = \log(0.1 - x)$, *si* $a \ge 0.1$, *on obtient que* $R_a = 0$ *(car* $0.1 \in [-a, a]$*) et par conséquence* $a = 0$ *ce qui est une contradiction, donc on prend* $a < 0.1$, *alors* $R_a = 0.1 - a$ *et donc* p_n *converge vers* f *pour* $0.1 - a < 0.74a$ *et donc* $0 < a < 0.1/1.74 \approx 0.06$ *(Très restrictif).*
$h(x) = \frac{1}{1+x^2}$, *les pôles sont* $\pm i$, *donc* $R_a = 1$ *et alors* p_n *converge vers* f *pour* $0 < a < 1/0.74 \approx 1.35$ *(restrictif).*