Agrégation externe de mathématiques, session 2008
Épreuve de modélisation, option C : algèbre et calcul formel

(*public 2008*)

Résumé : On étudie le problème de l'arrondi correct des valeurs numériques de certaines fonctions ; on cherche en particulier à estimer la précision minimale nécessaire pour les calculs intermédiaires.

Mots clefs : représentation et manipulation de structures algébriques, $\mathbb{Z}/n\mathbb{Z}$.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Le modèle de calcul flottant

L'une des premières applications des ordinateurs est la résolution de problèmes numériques, principalement issus de l'analyse numérique (grands systèmes, équations différentielles et aux dérivées partielles). La question de la représentation informatique des nombres réels s'est alors naturellement posée. Par nature, un ordinateur ne peut représenter que des ensembles finis, et il faut donc choisir quel sous-ensemble des réels on souhaite représenter de façon exacte.

On s'est de surcroît vite aperçu que la propagation des erreurs, et, partant de là, la stabilité numérique des algorithmes dépendait fortement du choix de représentation.

Après de nombreux essais indépendants qui avaient pour effet qu'un calcul produisait des résultats potentiellement très différents sur des architectures différentes, on s'est accordé sur la nécessité de fixer une norme universelle, qui permette en particulier de disposer de résultats fiables et reproductibles sur l'ensemble des architectures.

Dans toute la suite du texte, nous fixons un intervalle de $\mathbb{Z}$, noté $[e_{\min}, e_{\max}]$, que nous appelerons la plage d'exposants. L'ensemble des nombres représentables se définit de la façon suivante :

Définition 1. *On appelle nombre représentable en précision p un réel x de la forme $\varepsilon_x m_x 2^{e_x - p}$, où $\varepsilon_x \in \{0, \pm 1\}$, $e_x \in [e_{\min}, e_{\max}]$ et m_x est un entier de l'intervalle $[2^{p-1}, 2^p[$.*

Notons que p est toujours au moins égal à 1. On notera $\mathscr{M}_p$ l'ensemble des nombres représentables en précision p. L'entier m_x est la mantisse de x, l'entier e_x son exposant. Par convention, l'exposant de 0 est $e_{\min}$ et sa mantisse est 0. *Il n'est pas inutile à ce stade de faire un dessin de $\mathscr{M}_p$ pour de petites valeurs de $p, |e_{\min}|, |e_{\max}|$, comme proposé dans les suggestions.*

2. Arrondi

Le caractère limité de l'ensemble des nombres flottants soulève une difficulté : de façon générale, le résultat mathématique d'une opération effectuée sur des nombres de $\mathscr{M}_p$ n'a aucune raison d'être dans $\mathscr{M}_p$.

Il faut alors définir le nombre représentable résultat de l'opération. Par convention, on définit une notion d'arrondi (troncature), de la façon suivante : l'arrondi d'un $x \in \mathbb{R}$ en précision p est

$$\rho_p(x) = \text{signe}(x) \max(\mathscr{M}_p \cap [0, |x|]).$$

D'un point de vue concret, cette opération revient à tronquer le développement binaire après le p-ème chiffre. Il existe d'autres possibilités pour l'arrondi qui conduisent à des situations analogues à celle étudiée dans le présent texte.

On peut alors définir des lois de composition interne sur $\mathscr{M}_p$, $\oplus_p, \otimes_p, \ominus_p$ par, pour tout $(x,y) \in \mathscr{M}_p \times \mathscr{M}_p$,

$$x \oplus_p y = \rho_p(x+y), x \ominus_p y = \rho_p(x-y), x \otimes_p y = \rho_p(xy).$$

De façon générale, si $I \subset \mathbb{R}$ est un intervalle et f une fonction de I dans $\mathbb{R}$, on peut lui associer $\tilde{f}_p : I \cap \mathscr{M}_p \to \mathscr{M}_p$, définie par

$$\tilde{f}_p(x) = \rho_p(f(x)).$$

Le problème qui se pose est alors l'évaluation des opérations et des fonctions ainsi modifiées.

Insistons sur le fait qu'il s'agit d'un problème de nature différente de l'évaluation de f à une précision arbitrairement grande ; nous supposerons d'ailleurs ce dernier problème résolu comme prérequis à l'évaluation de $\tilde{f}_p$.

3. Le dilemme du fabricant de tables

Les formules définissant les opérations et fonctions modifiées montrent que le problème est de savoir arrondir le nombre réel $f(x)$. Par définition, dans le cas où $f(x)$ est positif, cela revient à déterminer le nombre de $\mathscr{M}_p$ immédiatement inférieur à $f(x)$. Toute la difficulté réside dans le fait que $f(x)$ est un nombre réel, qui ne peut être connu exactement.

Une façon classique de procéder consiste à supposer que l'on dispose d'approximations arbitrairement bonnes de f, sous la forme suivante : pour tout $x \in \mathscr{M}_p$ et tout p', supposons que l'on sait déterminer $y \in \mathscr{M}_{p'}$ tel que $f(x) \in [y - 2^{e_y - p'}, y + 2^{e_y - p'}]$. Notons cet intervalle $[y_1, y_2]$. On a alors le lemme suivant :

Lemme 1. *On suppose $p' > p+1$ et $y \neq 0$. Les trois conditions suivantes sont équivalentes :*
- $\rho_p(y_1) = \rho_p(y_2)$
- $]m, M[\cap \mathscr{M}_p = \emptyset$, *où* $m = \min(|y_1|, |y_2|)$ *et* $M = \max(|y_1|, |y_2|)$.
- $y \notin \mathscr{M}_p$ *et* $|y| + 2^{e_y - p'} \notin \mathscr{M}_p$.

Si l'une de ces trois conditions est vérifiée, on a alors $\rho_p([y_1, y_2]) = \{\rho_p(y_1)\}$, et en particulier $\rho_p(f(x)) = \rho_p(y_1)$.

Ce lemme permet, sans connaître exactement $f(x)$, d'en calculer l'arrondi. Les conditions du lemme sont élémentaires à vérifier, puisque y_1 et y_2 sont, eux, connus exactement : leur arrondi est obtenu en tronquant leur développement binaire au p-ème chiffre. Le lemme s'applique donc lorsque y_1 et y_2 ont les mêmes premiers p chiffres. La dernière condition indique que cela arrive dès lors que le développement binaire de y ne se termine pas par $p'-p$ zéros ou uns consécutifs.

Il nous reste donc à choisir p'. Si p' est choisi grand, le coût du calcul de y, généralement quadratique en p', devient trop important. Inversement, si p' est choisi trop petit, l'hypothèse du lemme risque de ne pas être vérifiée et on ne pourra pas conclure. Le choix de p' est donc un problème délicat.

Exemple 1. *Supposons* $p = 10$*, et essayons d'évaluer* $\rho_p(\exp(937/512))$*. Le développement binaire de la valeur exacte* $\exp(937/512)$ *est*

$$110.00111011111111111111010....$$

En particulier, le calcul d'une approximation pour $p' = 15$ *indiquera que* $\exp(937/512) \in [110.001110111110, 110.001111000000]$*, et ne permet pas de décider l'arrondi. Pour décider l'arrondi, il faut choisir dans ce cas* $p' \geq 23$*. Une recherche exhaustive montre en fait qu'il s'agit du* pire cas *pour la fonction exponentielle en précision 10 pour* $x \in [1,2]$.

4. Construire des mauvais cas

Dans un premier temps, on cherche à construire, pour diverses fonctions, des cas difficiles à arrondir (en le sens que p' doit être choisi grand pour que le Lemme s'applique), que nous appelerons mauvais cas. Outre l'aspect exploratoire, cela permet de tester une implantation particulière d'une fonctions f élémentaire sur les cas les plus difficiles à arrondir.

On cherche donc par exemple $x \in \mathscr{M}_p$ tel qu'il existe $y \in \mathscr{M}_p$ avec $|f(x)-y| \leq 2^{e-p'}$, pour p' aussi grand que possible et e l'exposant de l'arrondi de $f(x)$. Le formalisme décrit ci-dessus se généralise au cas de fonctions de plusieurs variables, comme en particulier les quatre opérations élémentaires.

Il s'agit d'un cas particulier important : d'un côté, dans ce cas, on peut avoir recours à la stratégie qui consiste à calculer le résultat exact, puis à l'arrondir (sauf pour la division). D'un autre côté, cette stratégie est potentiellement très coûteuse, et il est crucial que ces opérations qui sont à la base de tous les calculs soient très efficaces.

4.1. *Multiplication et division*

Pour l'addition et la soustraction, le problème est élémentaire. La multiplication, la division et la racine carrée sont plus intéressantes. Pour la multiplication, quitte à diviser par une puissance de 2, on se ramène au cas où les exposants valent 0. Il s'agit alors de trouver x et y entiers dans l'intervalle $[2^{p-1}, 2^p-1]$ tels que $x \cdot y$ soit très proche d'un nombre de la forme $2^{2p}z$, $z \in \mathscr{M}_p$. Au vu du fait que l'exposant de xy est 0 ou -1, on obtient l'équation :

$$x \cdot y = 2^{p-i}m \pm k, \quad m \in [2^{p-1}, 2^p-1], \quad i \in \{0,1\}, \quad k \text{ petit.}$$

Pour chaque couple (y,k), y impair, on peut construire le x correspondant en calculant $\pm k \cdot y^{-1} \pmod{2^{p-i}}$.

De ces mauvais couples (x,y) pour la multiplication, on peut déduire de mauvais couples (z,x) ou (z,y) pour la division. De façon générale, quand f est une fonction monotone suffisamment régulière entre deux intervalles de $\mathbb{R}$, I et I', on peut relier les mauvais cas de f sur I aux mauvais cas de f^{-1} sur I'.

4.2. *Racine carrée*

De la même façon que le cas de la division est dual de celui de la multiplication, le cas de la racine carré est dual du carré. Le problème du carré consiste à trouver x tel que x^2 soit très voisin d'un nombre représentable. Cela revient à résoudre le problème suivant : on cherche $x \in [2^{p-1}, 2^p - 1]$ tel que $x^2 \pmod{2^{p-1}}$ ou $x^2 \pmod{2^p}$ est petit.

Du lemme suivant se déduit facilement une méthode itérative pour produire des solutions :

Lemme 2. *Soit k un entier congru à 1 modulo 8, ℓ un entier au moins égal à 3, et x un entier tel que $x^2 \equiv k \pmod{2^\ell}$. Soit y un entier tel que $xy \equiv 1 \pmod{2^{2\ell-1}}$. Alors $z = \frac{1}{2}(x+ky)$ est un entier vérifiant $z^2 \equiv k \pmod{2^{2\ell-2}}$.*

Le lemme permet en fait, pour tout $p \geq 3$, de construire itérativement 4 solutions dans $[0, 2^p - 1]$ à l'équation $x^2 = k \pmod{2^p}$, autant que de solutions de l'équation $x_0^2 = k \pmod 8$. Sur ces quatre solutions, deux sont telles que $x \geq 2^{p-1}$.

5. Bornes inférieures et méthode de Liouville

Un cas important pouvant être traité simplement tout en préservant des résultats proches des résultats attendus est le cas où f est une fonction algébrique, c'est-à-dire s'il existe un polynôme non nul $P \in \mathbb{Z}[X,Y]$ tel que $P(x, f(x)) = 0$ pour tout x. On se limite au cas de la racine carrée, où $P(X,Y) = Y^2 - X$.

On peut, dans ce cas, résoudre le dilemme du fabricant de tables :

Théorème 1. *Si $x = \ell_1 \cdot 2^{-p}$ est dans $[1/2, 1[$, on a, pour tout entier ℓ_2, soit $\sqrt{\ell_1 \cdot 2^{-p}} = \ell_2 \cdot 2^{-p}$, soit*

$$|\sqrt{\ell_1 \cdot 2^{-p}} - \ell_2 \cdot 2^{-p}| \geq 2^{-2p-1}.$$

Le cas général d'une fonction algébrique se traite de façon analogue, par un argument dû à Liouville : si X et Y sont deux rationnels tels que $P(X,Y) \neq 0$, on minore $|P(X,Y)|$ par l'inverse de son dénominateur, et on le majore en fonction de $|Y - f(X)|$, grâce à l'inégalité des accroissements finis.

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins*

fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.

- étudier le cas de l'addition et de la soustraction ;
- construire de mauvais cas pour la multiplication ;
- construire des exemples de mauvais cas pour la racine carrée.
- prouver les différents lemmes et théorèmes énoncés dans le texte ;
- étudier ce que devient l'estimation donnée pour la racine carrée si x n'est plus dans $[1/2, 1[$;
- étudier ce que donnerait la construction de mauvais cas pour la racine cubique, et la méthode de Liouville dans ce cas.
- comment peut-on obtenir une approximation du produit de deux nombres réels qui conviendrait pour mettre en œuvre la stratégie du texte ?