EISTI - DÉPARTEMENT MATHÉMATIQUES

EXAMEN D'ANALYSE NUMÉRIQUE

février 2007 – **DURÉE 1h30**
Correction

Exercice 1 : Analyse des erreurs

1. Soit une machine décimale dont la mantisse est de 5 digits. Supposons que cette machine utilise l'arrondi au plus près. Donner, en le justifiant, l'ensemble de nombres réels x tels que :
$$m(1.0 + x) = 1.0$$
où par $m(x)$ nous avons noté le nombre machine de x.
SOL : Un nombre décimal a la représentation normalisée $x = 0.d_1d_2...d_5d_6... \times 10^e$. Donc $m(x) = 0.d_1d_2...d'_5 \times 10^e$ avec $d'_5 = d_5 + 1$ si $d_6d_7... \geq 0.000005$ et $d'_5 = d_5$ si $d_6d_7... < 0.000005$. Donc on aura $m(1.0 + x) = 1.0$ si $x < 5.10^{-6}$

2. Déterminer en le justifiant les nombres réels correspondant aux représentations binaires suivantes :

Signe s	Exposant E	Mantisse m
0	11111111	00000000000000000000000

SOL : $+\infty$ par définition

3.

Signe s	Exposant E	Mantisse m
1	10000001	01100000000000000000000

SOL : Le nombre représenté est en fait : $(1.011)_2 * 2\exp(2^7+2^0-127_{10}) = (2^0 + 2^{-2} + 2^{-3})_{10} * 2\exp(129_{10} - 127_{10}) = 1.375 * 2^2 = 1.375 * 4 = 5.5$. Le signe est négatif donc on a -5.5.

4. Soit le nombre binaire 10^m(10 et m sont des nombres binaires). En particulier on suppose que l'exposant m peut etre représenté par l'exposant de 8 bits de la norme IEEE-754.

   (a) Donner, en codage binaire selon la norme IEEE-754, la représentation de ce nombre.
   SOL :

Signe s	Exposant E	Mantisse m	Bit caché
0	$127 + m$	00000000000000000000000	1

   (on suppose que l'on a Biais=127, sinon on peut aussi répondre Biais+m pour la valeur de l'exposant)

   (b) Trouver en codage binaire selon la norme IEEE-754, le nombre binaire immédiatement supérieur à 10^m.
   SOL :

Signe s	Exposant E	Mantisse m	Bit caché
0	$127 + m$	00000000000000000000001	1

   (c) Trouver, en codage bianire selon la norme IEEE-754, le nombre binaire immédiatement supérieur à 10^m.
   SOL :

Signe s	Exposant E	Mantisse m	Bit caché
0	$127 + m - 1$	11111111111111111111111	1

(d) Soit une machine binaire qui, étant donné un nombre réel x, utilise l'arrondi vers zéro pour évaluer $m(x)$. Calculer l'erreur absolue maximale entre x et $m(x)$.
SOL : On suppose que x est un nombre positif. Sa représentation binaire est alors :

$$x = 1.b_1b_2...b_{23}b_{24}... \times 10^e$$

avec e nombre décimal et b_i le ième chiffre représentatif en binaire. La représentation machine de x est alors :

$$m(x) = 1.b_1b_2...b_{23} \times 10^e$$

Par conséquent l'erreur est :

$$x - m(x) = 0.00...0b_{24}... \times 10^e$$

On majore :

$$x - m(x) < 0.00...01000... \times 10^e = 10_2^{-23_{10}} \times 10^e = 10^{e-23}$$

5. Soit la fonction f définie par

$$f(x) = \frac{\sqrt{x^2+1}-1}{x^2}; x \in \mathbb{R}, x \neq 0$$

On cherche à faire un calcul numérique quand $x \to 0$.

(a) Indiquer les difficultés que l'on rencontre lors du calcul numérique de cette formule pour $x \to 0$.
SOL : La difficulté numérique est que $\sqrt{x^2+1}$ est proche de 1. De ce fait la soustraction $\sqrt{x^2+1}-1$ risque de provoquer une grande erreur numérique. En effet on sait que si on note $D = a - b$ la différence de deux quantités a et b, et $\eta(D)$ l'erreur de précision sera :

$$\eta(D) = \frac{a}{D}\eta(a) + \frac{b}{D}\eta(b)$$

De ce fait si D est proche de 0, la quantité $\eta(D)$ est grande.

(b) Trouver un algorithme pour dépasser ces difficultés. Donner à l'aide de cet algorithme la limite de $f(x)$ pour $x \to 0$.
SOL : On utilise la quantité conjuguée :

$$f(x) = \frac{\sqrt{x^2+1}-1}{x^2} = \frac{\left(\sqrt{x^2+1}-1\right)\left(\sqrt{x^2+1}+1\right)}{x^2\left(\sqrt{x^2+1}+1\right)} = \frac{1}{\sqrt{x^2+1}+1}$$

En utilisant la quantité $\frac{1}{\sqrt{x^2+1}+1}$ pour évaluer $f(x)$ lorsque $x \to 0$ on ne rencontre plus de problème numérique. On en déduit que si $x \to 0$ alors $f(x) = \frac{1}{\sqrt{x^2+1}+1} \to \frac{1}{2}$.

6. Considérons la formule suivante :

$$z = x^2 - y$$

Ecrire l'algorithme pour effectuer le calul et évaluer la borne supérieure de l'erreur de calcul propagée par l'utilisation de cet algorithme.
Application : $x = 1.03 \pm 0.01$ et $y = 0.45 \pm 0.01$.

SOL : On détaille le procédé comme en TD.

a) Les deux étapes de l'algorithme sont : $\begin{array}{ll} 1 & v \leftarrow x^2 \\ 2 & w \leftarrow v - y \end{array}$ donc $r = 2$

b) On a donc $\begin{array}{ll} 1 & x^{(0)} = \begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \phi^{(1)}\left(x^{(0)}\right) = \begin{pmatrix} x^2 \\ y \end{pmatrix} = x^{(1)} \\ 2 & x^{(1)} = \begin{pmatrix} v \\ y \end{pmatrix} \rightarrow \phi^{(2)}\left(x^{(1)}\right) = v - y = x^{(2)} \end{array}$ et donc $\phi = \phi^{(2)} o \phi^{(1)}$

c)On calcule les $\psi^{(k)}$. Comme ici $r = 2$, on a un seul $\psi^{(k)} = \psi^{(1)}$. $\psi^{(1)} = \phi^{(2)} \Rightarrow \psi^{(1)} \begin{pmatrix} v \\ y \end{pmatrix} = v - y$

d) On calcule les jacobiens :

$$\begin{aligned} J(\phi)(x) &= \begin{pmatrix} \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial y} \end{pmatrix} = \begin{pmatrix} 2x & -1 \end{pmatrix} \\ J\left(\psi^{(1)}\right) &= \begin{pmatrix} 1 & -1 \end{pmatrix} \end{aligned}$$

et $\Delta X = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$

e) On calcule

$$\begin{aligned} H_1 &= \begin{pmatrix} \eta_1(x^2) & 0 \\ 0 & 0 \end{pmatrix} \\ H_2 &= \eta(v - y) = \eta(x^2 - y) \end{aligned}$$

f) On applique la formule du cours (1.8.19) :

$$\begin{aligned} \Delta z &= J(\phi)(x).\Delta x + J\left(\psi^{(1)}\right)\left(x^{(1)}\right).H_1(x^{(1)}) + H_2.z \\ &= \begin{pmatrix} 2x & -1 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} + \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \eta_1(x^2) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x^2 \\ y \end{pmatrix} + \eta(x^2 - y).(x^2 - y) \\ &= 2x\Delta x - \Delta y + x^2\eta_1(x^2) + \eta(x^2 - y).(x^2 - y) \end{aligned}$$

Donc

$$|\Delta z| \leq 2x\Delta x + \Delta y + x^2\eta_1(x^2) + \left|x^2 - y\right| \eta(x^2 - y) \tag{1}$$

Application : $x = 1.03 \pm 0.01$ et $y = 0.45 \pm 0.01$.
On a

$$\eta_1(x^2) = \eta_1(x * x) = 2\frac{\Delta x}{x}$$

et

$$\eta(x^2 - y) = \frac{x^2}{x^2 - y}\eta(x^2) + \frac{y}{x^2 - y}\eta(y) = \frac{2x\Delta x}{x^2 - y} + \frac{\Delta y}{x^2 - y}$$

donc

$$\begin{aligned} |\Delta z| &\leq 2x\Delta x + \Delta y + 2x\Delta x + \left|x^2 - y\right| \left(\frac{2x\Delta x}{x^2 - y} + \frac{\Delta y}{x^2 - y}\right) \\ &\Rightarrow |\Delta z| \leq 6x\Delta x + 2\Delta y \\ &\Rightarrow |\Delta z| \leq 6 * 1.03 * 0.01 + 2 * 0.01 = 8.18 * 10^{-2} \end{aligned}$$

Les détails de l'application numérique n'étaient pas forcément demandés, il suffisait de repartir de (1) en majorant les erreurs par ε pour obtenir :

$$|\Delta z| \leq 2x\Delta x + \Delta y + x^2\varepsilon + \left|x^2 - y\right|\varepsilon \tag{2}$$

soit en majorant largement

$$|\Delta z| \leq 1.6\varepsilon + 3.2.10^{-2} \tag{3}$$

Exercice 2 : Recherche de racines

On veut calculer les solutions de l'équation

$$f(x) = \frac{x}{2} - \sin x + \frac{\pi}{6} - \frac{\sqrt{3}}{2} = 0$$

dans l'intervalle $[-\frac{\pi}{2}, \pi]$. D'après le graphe de la figure 1, la fonction admet deux zéros $\alpha_1 \in I_1 = [-\frac{\pi}{2}, 0]$ et $\alpha_2 \in I_2 = [\frac{\pi}{2}, \pi]$.

1. a) Peut-on appliquer la méthode de bissection pour calculer les deux racines ? (Justifier votre réponse)

   SOL : On peut bien appliquer la méthode de la bissection pour calculer le zéro $\alpha_2 \in I_2$, mais on ne peut pas appliquer cette méthode pour calculer α_1, car la condition $f(a).f(b) < 0$ n'est pas satisfaite pour tout $a, b \in [-\frac{\pi}{2}, \alpha_2[$

   b) Dans le(s) cas où c'est possible, estimer le nombre minimal d'itérations nécessaires pour calculer le(s) zéro(s) avec une tolérance $\alpha = 10^{-10}$, sur les intervalles I_1 et I_2.
   (On pourra utiliser $0 < \log_2 \frac{\pi}{2} < 1$ et $\log_2 10 \simeq 3$)

   SOL : On considère uniquement le zéro α_2. A la $k$ème itération de la méthode on a :

   $$\left|e^{(k)}\right| = \left|x^{(k)} - \alpha_2\right| \leq \frac{b-a}{2^{k+1}}$$

   avec $a = \frac{\pi}{2}$ et $b = \pi$ (on suppose que l'on choisit au départ les bornes de $I_2 = [\frac{\pi}{2}, \pi]$). Pour obtenir une erreur inférieure à $\alpha = 10^{-10}$, il suffit donc d'imposer

   $$\frac{b-a}{2^{k+1}} \leq 10^{-10}$$

   donc

   $$\begin{aligned} k &\geq \log_2\left(\frac{b-a}{\alpha}\right) - 1 = \log_2\left(\frac{\pi/2}{10^{-10}}\right) - 1 \\ &= \log_2(\pi/2) - \log_2\left(10^{-10}\right) - 1 = \log_2(\pi/2) + 10\log_2(10) - 1 \end{aligned}$$

   Comme $0 < \log_2 \frac{\pi}{2} < 1$ et $\log_2 10 \simeq 3$, une estimation grossière du nombre k d'itérations est 30. Avec une calculatrice on trouve que $k \geq 33$ itérations sont nécessaires.

2. On considère maintenant la méthode de point fixe

   $$x_{k+1} = \phi(x_k)$$

   avec

   $$\phi(x) = \sin x + \frac{x}{2} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right)$$

   pour calculer le zéro $\alpha_2 \in I_2$.

(a) Montrer que α_2 est un point fixe de ϕ

SOL :

$$\begin{aligned}\phi(\alpha_2) &= \sin\alpha_2 + \frac{\alpha_2}{2} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) = \sin\alpha_2 + \frac{\alpha_2}{2} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) + \frac{\alpha_2}{2} - \frac{\alpha_2}{2} \\ &= -\frac{\alpha_2}{2} + \sin\alpha_2 - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) + \alpha_2 \\ &= -f(\alpha_2) + \alpha_2 = \alpha_2\end{aligned}$$

car α_2 annule f.

(b) Montrer qu'il existe un réel $L < 1$ tel que

$$|\phi'(x)| \leq L < 1, \qquad \forall x \in I_2$$

SOL : On calcule la dérivée de ϕ

$$\phi'(x) = \cos x + \frac{1}{2}$$

Comme $-1 \leq \cos x \leq 0$ pour $x \in I_2 = [\frac{\pi}{2}, \pi]$, on a $-\frac{1}{2} \leq \cos x + \frac{1}{2} \leq \frac{1}{2}$ donc

$$-\frac{1}{2} \leq \phi'(x) \leq \frac{1}{2}$$

ce qui entraîne pour $L = \frac{3}{4}$ par exemple

$$|\phi'(x)| \leq L < 1, \qquad \forall x \in I_2$$

(c) Démontrer que

$$\phi(I_2) \subset I_2$$

SOL : La fonction ϕ est continue sur l'intervalle I_2, donc elle atteint ses bornes et on a

$$\phi(I_2) = [\min_{I_2} \phi, \max_{I_2} \phi]$$

La dérivée ϕ' s'annule pour $x_0 = \frac{2\pi}{3}$ car il correspond à $\cos x_0 = -\frac{1}{2}$. Le graphe permet de voir que la fonction y atteint son maximum et celui-ci vaut

$$\phi\left(\frac{2\pi}{3}\right) = \sin\frac{2\pi}{3} + \frac{\pi}{3} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) = \frac{\sqrt{3}}{2} + \frac{\pi}{3} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) = \sqrt{3} + \frac{\pi}{6} \simeq 2.25 \in [\frac{\pi}{2}, \pi]$$

Le minimum est atteint au point $x = \pi$ pour lequel

$$\phi(\pi) = \sin\pi + \frac{\pi}{2} - \left(\frac{\pi}{6} - \frac{\sqrt{3}}{2}\right) = \frac{\pi}{3} + \frac{\sqrt{3}}{2} \simeq 1.9 \in [\frac{\pi}{2}, \pi]$$

On obtient donc

$$\phi(I_2) = \left[\frac{\pi}{3} + \frac{\sqrt{3}}{2}, \sqrt{3} + \frac{\pi}{6}\right] \subset I_2$$

(d) Conclure, avec justifications, en ce qui concerne la convergence de la méthode de point fixe.

SOL : Les deux questions précédentes permettent d'appliquer le théorème 2.2.2 du cours, puisqu'on vérifie que ϕ est Lipschitz-bornée (ou lipschitzienne de rapport strictement inférieur à 1) et que $\phi(I_2) \subset I_2$. On en déduit l'existence d'un unique point fixe dans I_2. Du fait que l'on a montré que α_2 est un point fixe de ϕ, il est cet unique point fixe. Le théorème 2.2.5 du cours permet alors de conclure que les itérations de la forme

$$x_{k+1} = \phi(x_k)$$

engendrent une suite qui converge vers α_2.

3. a) On considère le zéro α_2 et la méthode de point fixe précédente. Montrer qu'il existe une constante positive $C < 1$ telle que

$$|x_{k+1} - \alpha_2| \leq C\,|x_k - \alpha_2|$$

SOL :

$$x_{k+1} - \alpha_2 = \phi(x_k) - \phi(\alpha_2)$$

D'après le théorème des accroissements finis, il existe une valeur $\xi_k \in [x_k, \alpha_2]$ (si on suppose $x_k < \alpha_2$) telle que

$$\phi(x_k) - \phi(\alpha_2) = \phi'(\xi_k)(x_k - \alpha_2)$$

On en déduit donc que

$$|x_{k+1} - \alpha_2| \leq \max_{\xi_k \in [x_k, \alpha_2]} |\phi'(\xi_k)|\,|x_k - \alpha_2|$$

Ayant montré à la question 2)b) que la dérivée de ϕ est inférieure ou égale à 1/2 on peut conclure que $\max_{\xi_k \in [x_k, \alpha_2]} |\phi'(\xi_k)| < 1$ et cette quantité constitue donc la constante C cherchée.

b) Montrer que $C \leq \frac{1}{2}$

SOL : On a montré à la question 2)b) que

$$-\frac{1}{2} \leq \phi'(x) \leq \frac{1}{2}$$

donc

$$\max_{\xi_k \in [x_k, \alpha_2]} |\phi'(\xi_k)| \leq \frac{1}{2}$$

soit $C \leq \frac{1}{2}$

4. On considère les itérations x_k de la méthode de point fixe du 3), initialisée avec $x_0 = \frac{\pi}{2}$.

(a) Montrer que

$$|x_k - \alpha_2| \leq C^k\,|x_0 - \alpha_2|$$

SOL : Si on applique l'inégalité de la question 3)a) en récurrence on en déduit :

$$|x_k - \alpha_2| \leq C\,|x_{k-1} - \alpha_2| \leq C^2\,|x_{k-2} - \alpha_2| \leq ... \leq C^k\,|x_0 - \alpha_2|$$

(b) En déduire le nombre d'itérations nécessaire pour que l'erreur $|x_k - \alpha_2|$ soit inférieure à 2^{-20}

SOL : Ayant montré que $C \leq \frac{1}{2}$ il suffit de trouver k tel que

$$\left(\frac{1}{2}\right)^k |x_0 - \alpha_2| \leq 2^{-20}$$

Comme $\alpha_2 \in I_2 = [\frac{\pi}{2}, \pi]$, $\frac{\pi}{2} \leq \alpha_2 \leq \pi$, on a $|x_0 - \alpha_2| = \left|\frac{\pi}{2} - \alpha_2\right| \leq \frac{\pi}{2} < 2$, il suffit donc que k soit assez grand pour que

$$2^{-k}.2 \leq 2^{-20}$$

c'est à dire que $k \geq 21$ itérations sont nécessaires.