EISTI - DÉPARTEMENT MATHÉMATIQUES

EXAMEN D'ANALYSE NUMÉRIQUE

février 2007 Rattrapage – **DURÉE 2h**
Correction

Exercice 1 : Calcul des Erreurs

Considérons un ordinateur binaire avec

- nombre de bits des registres $p = 8$;
- nombre de bits de signe $s = 1$;
- nombre de bits de l'exposant $t = 3$;
- nombre de bits de la mantisse $q = 4$.

Le codage des nombres sur cet ordinateurs suit la norme IEEE-754.

1. Calculer la valeur du biais pour ce codage.

   SOL : Par définition $B = 2^{t-1} - 1$ donc $B = 2^{3-1} - 1 = 3$. (Attention, dans le cours le nombre de bits de l'exposant est noté q mais ici il s'agit de t)

2. Donner tous les exposants en valeurs binaire et décimale

   (a) sans tenir compte du biais;

   SOL : Les exposants sont :

en binaire	en décimal
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

   (b) en tenant compte du biais.

   SOL : Les exposants sont obtenus avec :

$$E_{10} = -biais + E_2$$

   où E_{10} représente l'exposant en décimal, et E_2 l'exposant en binaire. Il suffit donc de soustraire le chiffre 3 à tous les exposants ci-dessus :

en binaire	en décimal
000	-3
001	-2
010	-1
011	0
100	1
101	2
110	3
111	4

3. Donner le codage pour cet ordinateur de

   (a) $+0$

   SOL :

signe	exposant	mantisse
0	000	0000

   (b) -0

   SOL :

signe	exposant	mantisse
1	000	0000

   (c) $+\infty$

   SOL :

signe	exposant	mantisse
0	111	0000

   (d) $-\infty$

   SOL :

signe	exposant	mantisse
1	111	0000

   (e) +NaN

   SOL :

signe	exposant	mantisse
0	111	0001

   ou tout autre nombre de mantisse non nulle.

   (f) -NaN

   SOL :

signe	exposant	mantisse
1	111	0001

   ou tout autre nombre de mantisse non nulle.

   (g) plus petit nombre positif

   SOL :

signe	exposant	mantisse
0	001	0000

   soit en décimal $1.0 * 2^{1-3} = 2^{-2} = 0.25$. On ne tient pas compte du signe : on donne le plus petit nombre positif en valeur absolue.

   (h) plus grand nombre positif.

   SOL :

signe	exposant	mantisse
0	110	1111

   . L'exposant est égal à $110_2 = 2^1 + 2^2 = 6$. La mantisse vaut $1.1111_2 = 1 + 0.5 + 0.25 + 0.125 + 0.0625 = 1.9375$. Le nombre est donc en décimal $1.9375 * 2^{6-3} = 1.9375 * 2^3 = 15.5$.

4. Exprimer le nombre 3.25 dans cette représentation.

   SOL : On remarque que 3.25 est compris entre 0.25 et 15.5 donc il est représentable. On a : partie entière $3 = 2^1 + 2^0$; partie décimale : $0.25 = 1/4 = 2^{-2}$ donc $3.25_{10} = (2^1 + 2^0 + 2^{-2})_{10} = 11.01_2 = (1.1010 * 10^1)_2$ donc on code en tenant compte du biais $E_2 = 1 + 3 = 4 = 2^2$:

signe	exposant	mantisse
0	100	1010

5. Calculer la valeur de *eps* pour cet ordinateur.

   SOL : Par définition on a :

$$\varepsilon = 2^{-q} = 2^{-4} = \frac{1}{16} = 0.0625$$

6. Donner la relation entre l'erreur relative de représentation d'un nombre en virgule flottante et la valeur de *eps*.

   SOL : Par définition l'erreur relative de représentation est donnée par :

$$\iota(x) = \frac{\Delta x}{m(x)} = \frac{m(x) - x}{m(x)}$$

   D'autre part on a

$$m\left(x\right) = x\left(1 + \eta\left(x\right)\right)$$

   Donc

$$\iota(x) = \frac{m(x) - x}{x\left(1 + \eta\left(x\right)\right)} = \frac{m(x) - x}{x} \cdot \frac{1}{\left(1 + \eta\left(x\right)\right)}$$

   En utilisant la relation (1.4.11) du poly

$$\frac{\left|m\left(x\right) - x\right|}{|x|} \leq \beta^{1-q}$$

   avec β la base de numérotation, on obtient

$$|\iota(x)| \leq \frac{1}{|(1 + \eta\left(x\right))|} \cdot 2^{1-q} = \frac{2}{|(1 + \eta\left(x\right))|} \cdot 2^{-q} \leq 2 \cdot eps$$

Exercice 2 : Recherche de Racines

On veut calculer les racines de la fonction f définie par

$$f(x) = e^x + 3\sqrt{x} - 2$$

sur l'intervalle $[0, 1]$.

1. Montrer qu'il existe un zéro α pour la fonction f dans $[0, 1]$ et qu'il est unique.

   SOL : On voit que $f(0) = -1 < 0, f(1) = e + 1 > 0$.Puisque la fonction est continue, il existe un zéro α pour la fonction f dans $[0, 1]$. De plus $f'(x) = e^x + \frac{3}{2\sqrt{x}} > 0$ donc la fonction est strictement croissante. Par suite le zéro est unique.

   On veut calculer le zéro α de la fonction f par une méthode de point fixe convenable de la forme $\left\{ \begin{array}{c} x^{(0)} \text{ donné} \\ x^{(k+1)} = \phi(x^{(k)}) \end{array} \right.$.

   De la théorie on sait qu'il faut que ϕ soit de classe C^1 au voisinage du point fixe et que $|\phi'(x)| < 1$ au voisinage du zéro de la fonction pour que la méthode converge.

   En particulier on se donne deux méthodes de point fixe de la forme $x = \phi(x)$. Les deux méthodes se mettent sous la forme $x = \phi_1(x)$ et $x = \phi_2(x)$ avec ϕ_1 et ϕ_2 définies respectivement par :

$$\phi_1(x) = \ln\left(2 - 3\sqrt{x}\right) \quad \text{et} \quad \phi_2(x) = \frac{(2 - e^x)^2}{9}$$

2. Donner l'intervalle de définition pour chacune de ces deux fonctions.

   SOL : ϕ_1 est définie sur $[0, \frac{4}{9}[$

   ϕ_2 est définie pour tout $x \in [0, 1]$.

3. Démontrer que α est situé dans l'intervalle de définition de ces deux fonctions, et qu'il est un point fixe pour chacune d'elles.

   SOL : $\phi_1(\alpha) = \ln(2 - 3\sqrt{\alpha})$ or α est une racine de f donc $f(\alpha) = e^\alpha + 3\sqrt{\alpha} - 2 = 0$ d'où l'on tire $2 - 3\sqrt{\alpha} = e^\alpha$ donc $\phi_1(\alpha) = \ln(2 - 3\sqrt{\alpha}) = \ln e^\alpha = \alpha$. α est donc bien un point fixe de ϕ_1. De plus $f(0) = -2$ et $f(\frac{4}{9}) = e^{\frac{4}{9}} + 3\sqrt{\frac{4}{9}} - 2 = e^{\frac{4}{9}} + 3\frac{2}{3} - 2 = e^{\frac{4}{9}} > 0$ donc le zéro de f est situé entre 0 et $\frac{4}{9}$: α est donc bien dans $[0, \frac{4}{9}[$.
   $\phi_2(x) = \frac{(2-e^x)^2}{9}$ et en utilisant encore que α est une racine de f on a : $e^\alpha = 2 - 3\sqrt{\alpha}$ d'où $\phi_2(\alpha) = \frac{(2-2+3\sqrt{\alpha})^2}{9} = \alpha$. α est donc bien un point fixe de ϕ_2. De plus $\alpha \in [0, \frac{4}{9}[\subset [0, 1]$ donc $\alpha \in [0, 1]$

4. *On admet que*

   - $|\phi_1'(x)| > 4$ *pour tout* x *de l'intervalle de défintion de* $\phi_1(x)$.
   - $|\phi_2'(x)| < 0.5$ *pour tout* x *de l'intervalle de défintion de* $\phi_2(x)$.

   Pour chacune de deux méthodes ci-dessus établir, en le justifiant, si elles peuvent être utilisées pour le calcul de point fixe.

   SOL : La fonction ϕ_2 est la seule à vérifier la condition $|\phi'(x)| < 1$, qui est une condition nécessaire pour que la méthode de point fixe converge, donc seule la méthode de point fixe basée sur ϕ_2 peut être utilisée.

   On prend désormais une de deux méthodes ci-dessus à condition qu'il soit possible d'être utilisée au calcul de point fixe et l'on note la fonction correspondante $\tilde{\phi}$.

5. Montrer que $\tilde{\phi}([0, 1]) \subset [0, 1]$. Que peut-on en déduire concernant la convergence de la méthode du point fixe ?

   SOL : On prend $\tilde{\phi} = \phi_2$. ϕ_2' est croissante sur $[0, 1]$ or $\phi_2'(0) = -\frac{2}{9}(2-e)(e) = \frac{2}{9}(e-2)(e) > 0$ donc ϕ_2 est également croissante. $\phi_2(0) = \frac{1}{9}$ et $\phi_2(1) = \frac{(2-e)^2}{9} \simeq \frac{0{,}49}{9} < 1$ donc on a $\phi_2([0, 1]) \subset [0, 1]$.
   On peut donc en déduire que la méthode de point fixe est convergente quelle que soit la donnée initiale $x^{(0)} \in [0, 1]$ car les deux hypothèses nécessaires sont réunies.

6. En utilisant cette méthode avec $x^{(0)} = 0$, estimez le nombre d'itérations nécessaires pour trouver une valeur approchée de α qui soit exacte jusqu'au 20ème bit de la mantisse (il faut donc une tolérance $t = 2^{-20}$)

   SOL : On sait que si $|\phi'(x)| < C$ alors on a l'estimation d'erreur suivante : $\left|x^{(k)} - \alpha\right| < C^k \left|x^{(0)} - \alpha\right|$ En choisissant $C = 1/2$ on obtient

   $$\left|x^{(k)} - \alpha\right| < 2^{-k} |\alpha|$$

   On cherche donc k tel que $2^{-k} |\alpha| < 2^{-20}$. Comme $|\alpha| < 1$ il suffit de prendre $k = 20$ pour que la condition soit vérifiée.

On envisage maintenant le calcul de cette racine par une méthode de bissection.

7. Justifiez que la méthode est applicable à la fonction f sur l'intervalle $[0, 1]$

   SOL : La fonction f est continue et croissante sur l'intervalle. De plus on a $f(0) = e-2 > 0$ et $f(1) = e+1 > 0$ donc la fonction change de signe sur $[0, 1]$ et n'admet qu'une racine. Cela suffit pour pouvoir appliquer la méthode de bissection et être assuré de sa convergence.

8. Estimer le nombre minimal d'itérations nécessaires pour calculer le(s) zéro(s) avec une tolérance $t = 2^{-20}$.

   SOL : Du principe de la méthode on déduit que pout obtenir une précision de $t = 2^{-20}$ il faut effectuer k itérations avec

$$k > \log_2 \frac{b-a}{t} - 1$$

   où $[a, b]$ est l'intervalle d'application, soit ici

$$k > \log_2 \frac{1}{t} - 1 = -\log_2 2^{-20} - 1 = 19$$

   Le nombre d'itérations doit donc être aussi au minimum égal à 20.