EISTI - DÉPARTEMENT MATHÉMATIQUES

EXAMEN D'ANALYSE NUMÉRIQUE – CORRIGÉ

7 juin 2010 – **DURÉE 3h00**

Exercice 1 Considérons un ordinateur qui utilise, pour la représentation machine des réels, un système selon le standard IEEE-754, avec base binaire, 6 bits pour la mantisse et 4 bits pour l'exposant.

On notera dans la suite ce système par $\mathcal{S}$.

1. Évaluer le plus grand $\mathrm{x}_{\max}$ et le plus petit $\mathrm{x}_{\min}$ nombres normalisés et positifs que nous pouvons représenter avec avec le système $\mathcal{S}$.
   SOL.- Le biais est égal à $B = 2^{q-1} - 1 = 2^{4-1} - 1 = 7_{10} = 111_2$.
   Plus grand nombre normalisé positif :
   $\mathrm{x}_{\max}$ = 0 1110 111111 = $\left(1.111111 \times 10^{111}\right)_2 = 254_{10}$.
   Plus petit nombre normalisé positif :
   $\mathrm{x}_{\min}$ =0 0001 000000 = $\left(1.000000 \times 10^{-110}\right)_2 = 0.015625_{10}$.

2. Donner la valeur de la plus grande erreur $\varepsilon_{\max}$ de conversion du système $\mathcal{S}$ lorsque pour la conversion on utilise la troncature.
   SOL. $|\eta(x)| = \beta^{1-p} = 2^{1-6} = 0.03125$, où $\beta = 2$ la base du système, $p = 6$ le nombre de bits de la mantisse.

3. Donner, en la justifiant, la valeur de eps pour le système $\mathcal{S}$.
   SOL.- eps est le plus petit positif pour lequel on $1 + eps \neq eps$. On a
   $1 \rightarrow$ 0 0111 000000 = 1.000000×10^0 = $1.000000 \times 10^{-111}$ et 1+eps $\rightarrow$ 0 0111 000001 = 1.000001×10^0= $1.000001 \times 10^{-111}$. Donc $eps = 0.0000001_2 = 0.0078125_{10}$.

4. Expliquer pour quelle raison le réel 0.1_{10} ne peut pas être représenté sans erreur avec le système $\mathcal{S}$ et donner la représentation du nombre-machine correspondant à 0.1_{10} selon ce même système.
   SOL.- On a $0.1_{10} = 0.0001\ 1001\ 1001...{}_2 = \left[(1.1001\ 1001...) \times 10^{-100}\right]_2$. Par conséquent si on utilise 6 bits pour la mantisse, on a
   $\left[1.100110 \times 10^{-100}\right]_2 = 0.099609375_{10} \neq 0.1$ c'est-à-dire une troncature de la valeur initiale qui induit une erreur de représentation.
   Représentation de 0.1_{10} : On a pour l'exposant $E = -100_2 = -4_{10}$, et sa représentation selon le standard IEEE 754 s'écrit $e = E + B = 11_2$. Donc nombre machine : 0 100110 11.

5. On sait que la série infinie $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k}$ est divergente.

   (a) Expliquer pour quelle raison le calcul par ordinateur de cette série permet la convergence à partir d'une valeur de $k > K_0$.
   SOL.- Le calcul par ordinateur a une précision finie. Donc, dès que le terme $\frac{1}{k}$ depasse la capacité de la représentation de la machine, il est converti à 0 et donc son ajout à la somme ne change pas la valeur de cette dernière. Il y a donc convergence numérique.

   (b) Calculer le nombre K_0 pour le système de $\mathcal{S}$ ci-dessus dans le cas de la représentation par troncature.
   SOL.- Notons par $S_n = \displaystyle\sum_{k=1}^{n} \frac{1}{k}$ la somme partielle de n premiers termes de cette série. Nous avons la convergence dès que $S_{n+1} = S_n$ ce qui donne $fl\left(S_n + \frac{1}{n+1}\right) = S_n$. Il faut donc que $\frac{1}{n+1} < \varepsilon_{\max}$, où $\varepsilon_{\max}$ l'erreur maximale de représentation par arrondi. On a finalement $\frac{1}{n+1} < 2^{1-6} \Rightarrow n + 1 > 2^5 = 32 \Rightarrow K_0 = 32$

6. (Histoire de montrer que les financiers avaient tort de delaisser l'analyse numérique)
   Un logiciel de calcul des "options exotiques" permet aux traders de choisir le moment propice pour vendre ces options. Il comporte la ligne de code suivante (écrite en pseudocode) `if (xNew - xOld) < 1.0e-6 then Attendre, sinon Vendre;`

   où `xNew, xOld` sont deux prix successifs d'une option.
   À votre avis, faut-il se servir ou non de ce logiciel, si l'ordinateur utilise comme système de représentation le système $\mathcal{S}$? Justifier votre réponse .
   (On suppose que `xNew, xOld` $\in [x_{\max}\ , x_{\min}]$).
   SOL.- Non, car l'erreur maximale étant $\varepsilon_{\max} = 10^{-5}$, la représentation machine de la différence de deux nombres ne pourra jamais être inférieure à 10^{-5} si on veut que l'ordinateur détecte une valeur non nulle. Donc il faut corriger ce test à `if (xNew - xOld) < 1.0 e-4 then Attendre, sinon Vendre;`

Exercice 2 Soit la matrice $\mathbf{A}$ définie par

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

de taille n qui intervient lors de la discrétisation de l'opérateur $\frac{\partial^2}{\partial x^2}$ sur l'intervalle $[-1, 1]$ avec les conditions aux bords $u(-1) = u(1) = 0$.

Partie A : On s'intéresse à la décomposition $\mathbf{LU}$ de $\mathbf{A}$.

1. Expliciter la structure (la forme) des matrices $\mathbf{L}$ et $\mathbf{U}$ et justifier votre réponse.
   SOL.- $\mathbf{L}$ est triangulaire inférieure avec des 1 sur la diagonale et ne comporte qu'une sous diagonale non nulle car elle "hérite" de la structure bande de $\mathbf{A}$.
   $\mathbf{U}$ est triangulaire supérieure et ne comporte qu'une sur diagonale non nulle car elle "hérite" de la structure bande de $\mathbf{A}$. On a donc

$$\mathbf{U} = \begin{bmatrix} \times & \times & & & \\ 0 & \times & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \times \\ & & & 0 & \times \end{bmatrix} \quad et \ \mathbf{L} = \begin{bmatrix} 1 & 0 & & & 0 \\ \times & 1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 0 \\ 0 & & & \times & 1 \end{bmatrix}$$

2. On note u_{ij} les éléments de la matrice $\mathbf{U}$, et l_{ij} ceux de $\mathbf{L}$. Calculer les valeurs de
   (a) u_{11}
   SOL.- : $u_{11} = 2$
   (b) u_{12}
   SOL.- $u_{12} = -1$
   (c) l_{21}
   SOL.- - $l_{21} = -1/2$

3. Pour tout i compris entre 2 et n, exprimer :
   (a) u_{ii} en fonction de $l_{i,i-1}$ et $u_{i-1,i}$
   SOL.- $u_{ii} = 2 - l_{i,i-1}u_{i-1,i}$
   (b) $u_{i,i+1}$
   SOL.- .- $u_{i,i+1} = a_{i,i+1}$ =-1

(c) $l_{i,i-1}$en fonction de $u_{i-1,i-1}$
SOL.- $l_{i,i-1} = -1/u_{i-1,i-1}$

4. En déduire l'expression de u_{ii} en fonction de $u_{i-1,i-1}$
SOL.- En reportant $l_{i,i-1}$ dans u_{ii} on trouve $u_{ii} = 2 + u_{i-1,i}/u_{i-1,i-1} = 2 - 1/u_{i-1,i-1}$
5. Montrer que
$$u_{ii} = \frac{i+1}{i}$$
SOL.- $u_{11} = 2$ donc $u_{22} = 2 - 1/2 = 3/2$. Par suite $u_{33} = 2 - 2/3 = 4/3$. On démontre par récurrence la formule.
6. En déduire le déterminant de la matrice A.
SOL.- $\det(\mathbf{A}) = \det(\mathbf{LU}) = \det(\mathbf{L}).\det(\mathbf{U}) = \det(\mathbf{U}) = \prod_{i=1}^{n} \frac{i+1}{i} = n+1$

Partie B : On s'intéresse à la méthode de Jacobi et à la méthode de Gauss-Seidel pour résoudre le système $\mathbf{Ax} = \mathbf{b}$

1. Déterminer la matrice d'itération de Jacobi, notée $\mathbf{J}$.
SOL.-
$$\mathbf{J} = \frac{1}{2}\begin{bmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix}$$
2. On admet que les vecteurs propres de la matrice $\mathbf{A}$ sont les vecteurs $\mathbf{u}^{(k)}$ dont les composantes sont données par :
$$u_i^{(k)} = \sin\left(\frac{ik\pi}{n+1}\right)$$
et qu'ils sont ordonnés dans l'ordre croissant des valeurs propres.

(a) Déterminer la plus petite en module valeur propre λ_1 de $\mathbf{A}$ en fonction de n.
Indication : on rappelle que $\sin(2a) = 2\sin a \cdot \cos a$.
SOL.- La multiplication $Au^{(1)} = \lambda_1 u^{(1)}$ donne
$$\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1^1 \\ u_2^1 \\ \vdots \\ \vdots \\ u_n^1 \end{pmatrix} = \lambda_1 \begin{pmatrix} u_1^1 \\ u_2^1 \\ \vdots \\ \vdots \\ u_n^1 \end{pmatrix}$$
soit pour la première ligne
$$\begin{aligned} 2u_1^1 - u_2^1 &= \lambda_1 u_1^1 \\ &\Leftrightarrow \lambda_1 u_1^1 = 2\sin\left(\frac{\pi}{n+1}\right) - \sin\left(\frac{2\pi}{n+1}\right) \end{aligned}$$
or $\sin\left(\frac{2\pi}{n+1}\right) = 2\sin\left(\frac{\pi}{n+1}\right)\cos\left(\frac{\pi}{n+1}\right)$ donc
$$\begin{aligned} \lambda_1 \sin\left(\frac{\pi}{n+1}\right) &= 2\sin\left(\frac{\pi}{n+1}\right) - 2\sin\left(\frac{\pi}{n+1}\right)\cos\left(\frac{\pi}{n+1}\right) \\ &\Leftrightarrow \lambda_1 = 2 - 2\cos\left(\frac{\pi}{n+1}\right) \end{aligned}$$

(b) On admet que la plus grande en module valeur propre λ_n de $\mathbf{A}$ s'exprime en fonction de n par :

$$\lambda_n = 2 + 2\cos\left(\frac{\pi}{n+1}\right)$$

En déduire le conditionnement de $\mathbf{A}$ en norme 2.

SOL.- Du fait que la matrice A est symétrique, le conditionnement se calcule par

$$\begin{aligned} cond_2(A) &= \frac{|\lambda_1|}{|\lambda_n|} = \left[2 - 2\cos(\frac{\pi}{n+1})\right] / \left[2 + 2\cos(\frac{\pi}{n+1})\right] \\ &= \left[1 - 1\cos(\frac{\pi}{n+1})\right] / \left[1 + 1\cos(\frac{\pi}{n+1})\right] \end{aligned}$$

3. À l'aide d'une égalité matricielle liant $\mathbf{A}$ et $\mathbf{J}$, déterminer la relation existant entre les valeurs propres extrémales de $\mathbf{J}$ et celles de $\mathbf{A}$.

   SOL.- On a $2J = -A + 2I$ soit $J = -\frac{1}{2}A + I$

4. En déduire que le rayon spectral de $\mathbf{J}$ est donné par

$$\rho(\mathbf{J}) = \cos\frac{\pi}{n+1}$$

   SOL.- La relation entre les valeurs propres de J et celles de A est déduite de la relation précédente : si $Jv = \mu v$ alors $\left(-\frac{1}{2}A + I\right)v = \mu v \Leftrightarrow -\frac{1}{2}Av + v = \mu v \Leftrightarrow -\frac{1}{2}Av = (\mu - 1)v \Leftrightarrow Av = -2(\mu - 1)v$. Si λ est valeur propre de A alors $\lambda = -2(\mu - 1)$ avec μ valeur propre de J. Soit $\mu = 1 - \frac{\lambda}{2}$.

   La plus grande valeur propre de J est donc donnée par l'intermédiaire de la plus petite valeur propre de A et donc

$$\mu_n = 1 - \frac{\lambda_1}{2} = 1 - 1 - \cos\left(\frac{\pi}{n+1}\right) = -\cos\left(\frac{\pi}{n+1}\right)$$

   La plus petite valeur propre de J se calcule à partir de λ_n, et est égale en module. En module on obtient donc

$$\rho_J = \left|\cos\left(\frac{\pi}{n+1}\right)\right|$$

5. La méthode de Jacobi est-elle convergente pour cette matrice ? Justifier brièvement votre réponse.

   SOL.- La méthode converge puisqu'elle est consistante et on a $\rho_J < 1$

6. Déterminer la matrice d'itération de Gauss-Seidel, notée $\mathcal{L}_1$.

   SOL.- $\mathcal{L}_1 = (D - E)^{-1}F$ soit

$$\begin{aligned} \mathcal{L}_1 &= \begin{pmatrix} 2 & 0 & & & 0 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 0 \\ 0 & & & -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & & & 0 \\ 0 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2^{-1} & 0 & & & 0 \\ 2^{-2} & 2^{-1} & \ddots & & \\ 2^{-3} & 2^{-2} & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 0 \\ 2^{-n} & \cdots & 2^{-3} & 2^{-2} & 2^{-1} \end{pmatrix} \begin{pmatrix} 0 & 1 & & & 0 \\ 0 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2^{-1} & 0 & & 0 \\ 0 & 2^{-2} & 2^{-1} & \ddots & \vdots \\ & 2^{-3} & 2^{-2} & \ddots & 0 \\ & \vdots & & \ddots & 2^{-1} \\ 0 & 2^{-n} & 2^{-n+1} & \cdots & 2^{-2} \end{pmatrix} \end{aligned}$$

7. Quelle relation lie $\rho(\mathbf{J})$ et $\rho(\mathcal{L}_1)$?
   SOL.- $\rho(J)^2 = \rho(\mathcal{L}_1)$ car A est tridiagonale.
8. Déterminer $\rho(\mathcal{L}_1)$ et en conclure si la méméthode de Gauss-Seidel converge.
   SOL.- De $\rho(J) = \cos\frac{\pi}{n+1}$ on déduit : $\rho(\mathcal{L}_1) = \cos^2\frac{\pi}{n+1}$
9. Laquelle des méthodes de Jacobi et de Gauss-Seidel est la plus rapide pour résoudre le système $\mathbf{Ax} = \mathbf{b}$? Justifier brièvement votre réponse.
   SOL.- Le cosinus élevé au carré est plus petit que le cosinus donc $\rho(\mathcal{L}_1) < \rho(J)$ donc la méthode de Gauss-Seidel est plus rapide.

Exercice 3
Soit la matrice

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 0 & 6 & 1 \\ 1 & 0 & -5 \end{bmatrix}$$

Estimer ses valeurs propres sans faire leur calcul explicite.

SOL.- On utilise les disques de Geršgorin et on trouve une valeur propre dans le disque de centre -5 et de rayon 1 et deux valeurs propres dans les deux disques qui se recoupent avec centres 5 et 6 et rayon 1.

Exercice 4
Soit la matrice

$$\mathbf{A} = \begin{bmatrix} -9 & -6 & -2 & -4 \\ -8 & -6 & -3 & -1 \\ 20 & 15 & 8 & 5 \\ 32 & 21 & 7 & 12 \end{bmatrix}$$

Sans faire le calcul des vecteurs propres, déterminer parmi les vecteurs suivants

$$\mathbf{u}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \mathbf{u}_3 = \begin{bmatrix} 0 \\ 1 \\ -3 \\ 0 \end{bmatrix}$$

ceux qui sont vecteurs propres de $\mathbf{A}$ et, pour chacun de ces vecteurs propres, identifier la valeur propre associée.

SOL.- $\mathbf{u}$ est vecteur propre de $\mathbf{A}$ si $\mathbf{Au} = \boldsymbol{\lambda}\mathbf{u}$. Donc il faut que $\mathbf{Au}$ soit un multiple de $\boldsymbol{\lambda}\mathbf{u}$. Ici $\mathbf{u}_1$ et $\mathbf{u}_3$ sont des multiples de $\mathbf{Au}$. De plus on a $\mathbf{Au}_1{=}1\mathbf{u}_1$ et $\mathbf{Au}_3{=}3\mathbf{u}_3$.