DIRO
IFT 2421

EXAMEN INTRA

Max Mignotte

DIRO, Département d'Informatique et de Recherche Opérationnelle, local 2384/2377.
http : //www.iro.umontreal.ca/∼mignotte/ift2421/
e-mail : mignotte@iro.umontreal.ca

date : *25/02/2003*

I	Système numérique et ordre des opérations (23 pts).
II	Cancellation de CSE et amplification numérique (15 pts).
III	Méthode du point fixe et de Newton (27 pts).
IV	Méthodes itératives pour les systèmes (15 pts).
V	Système d'équations (25 pts).

TOUS DOCUMENTS PERSONNELS, CALCULATRICES ET CALCULATEURS AUTORISÉS

I. Système numérique et ordre des opérations (23 pts)

1. **Système numérique**
   (a) Dans un système numérique pour lequel $b=2$, $s=3$, $m=-9$, $M=9$, combien y a t-il de nombres distincts que l'on peut représenter parfaitement, sans approximation ? <**2 pts**>
   (b) Même question pour le système numérique suivant $b=10$, $s=3$, $m=-9$, $M=9$. <**2 pts**>
   (c) Donner la formule générale en fonction de b, s, m, M exprimant ;
      i. le nombre distinct de valeurs que l'on peut représenter sans approximation, <**2 pts**>
      ii. le plus petit nombre représentable (epsilon machine), <**2 pts**>
      iii. le plus grand nombre représentable. <**4 pts**>
2. **Ordre des opérations**
   Soit le calcul suivant,

$$f(x,y) \quad = \quad \frac{\exp(-x)}{\exp(-y)}, \tag{1}$$

   que l'on désire effectuer pour $x = 120$ et $y = 100$ avec le système numérique suivant ($b = 2, s = 24, m = -126, M = 127$).
   (a) A quel type de problème serons nous confronté lors de l'estimation de $f(x = 120, y = 100)$? <**3 pts**>
   (b) Proposer une expression équivalente pour laquelle le problème ne se pose pas. <**2 pts**>
   (c) Soit le calcul précédent (qui se fait d'abord en estimant $\exp(-x)$, puis $\exp(-y)$, puis la division de ces deux nombres) et le calcul suivant équivalent analytiquement,

$$\begin{aligned} u(x,y) &= \quad x - y, \\ f(u) &= \exp(-u). \end{aligned} \tag{2}$$

   En supposant que le problème numérique de la question précédente ne se pose plus maintenant. Laquelle de ces deux stratégies assure le minimum d'amplification d'erreur numérique ? Justifier votre réponse en donnant l'amplification d'erreur dans les deux cas en fonction de l'incertitude sur x et y que l'on notera Δx et Δy. <**6 pts**>

Réponse

1.

(a). Pour le système ($b = 2, s = 3, m = -9, M = 9$), on a $1 \times 2 \times 2 \times 19 = 76$ nombres distincts représentables. (si on compte le signe, il faut multiplier par deux cette expression et à cela on doit ajouter le nombre zéro). <**2 pts**>

(b). Pour le système ($b = 10, s = 3, m = -9, M = 9$), on a $9 \times 10 \times 10 \times 19 = 17100$ nombres distincts ($\times$ 2 si on compte le signe et + le nombre zéro). <**2 pts**>

(c). Pour le systeme (b, s, m, M), la formule générale donnant le nombre distinct de valeurs que l'on peut représenter sans approximation s'écrit,

- **i.** $(b-1)\, b^{s-1}\, (|m| + |M| + 1)$ ($\times$ 2 si on compte le signe et + le zéro). <**2 pts**>
  De même, la formule générale pour trouver l'epsilon machine et le nombre le plus grand représentable sans approximation par ce système numérique s'écrit :
- **ii.** Epsilon machine $= b^{-1}\, b^{m} = b^{m-1}$. <**2 pts**>

– **iii.** Pour calculer le plus grand nombre $\mathcal{B}$, on doit considérer la somme suivante,

$$\begin{aligned}
\mathcal{B} &= \left[(b-1)\,b^{-1} + (b-1)\,b^{-2} + \ldots + (b-1)\,b^{-s}\right] b^{M}, \\
&= \left[(b-1)\,b^{-1}\,(1 + b^{-1} + b^{-2} + \ldots + b^{-(s-1)})\right] b^{M}, \\
&= \left[(b-1)\,b^{-1}\,\frac{1-b^{-s}}{1-b^{-1}}\right] b^{M}, \\
&= (b-1)\,b^{M-1}\left(\frac{1-b^{-s}}{1-b^{-1}}\right). \quad < \mathbf{4\ pts} >
\end{aligned}$$

2.

(a). $\exp(-100)$ et $\exp(-120)$ sont inférieur à l'epsilon machine de ce système de représentation numérique (epsilon machine $= 2^{-127}$). L'estimation de $f(x = 120, y = 100)$ conduit donc à la forme indéterminée du type $\frac{0}{0}$ (l'ordinateur pourrait répondre par une erreure du style "division par zéro" ou "NAN" (qui signifie Not A Number)) $<$**3 pts**$>$.

(b). Une expression equivalente pour laquelle le probleme ne se pose pas est de considérer,

$$f(x, y) \;=\; \exp(-(x-y)),$$

dans ce cas, l'opération soustraction $(120 - 100 = 20)$ puis l'estimation de l'exponentiel $(\exp(-20))$ pourra se faire $(\exp(-20) >$ Epsilon machine). $<$**2 pts**$>$

(c). Pour le premier calcul, on a pour les trois opérations, avec la convention *valeur approchée* $\pm$ *incertitude*,

$$\begin{aligned}
&\text{op}(1); && \exp(-100) \pm |-\exp(-100)|\,\Delta x, \\
&\text{op}(2); && \exp(-80) \pm |-\exp(-80)|\,\Delta y, \\
&\text{op}(3); && \exp(-20) \pm \left(\left|\frac{1}{\exp(-80)}\right|\,|-\exp(-100)|\,\Delta x + \left|-\frac{\exp(-100)}{\exp(-160)}\right|\,|-\exp(-80)|\right)\Delta y, \\
& && = \exp(-20) \pm |\exp(-20)|\,(\Delta x + \Delta y). \quad < \mathbf{2.5pts} >
\end{aligned}$$

Pour le deuxieme calcul, on a pour les deux opérations,

$$\begin{aligned}
&\text{op}(1); && 20 \pm \left(\Delta x + \Delta y\right), \\
&\text{op}(2); && \exp(-20) \pm |\exp(-20)|\,(\Delta x + \Delta y). \quad < \mathbf{2.5pts} >
\end{aligned}$$

Les deux façon de calculer assurent la même amplification numérique. Le problème de la première expression est seulement lié au fait que les deux calculs intermédiaires seront plus petit que l'epsilon machine. $<$**1 pt**$>$

II. Cancellation de CSE et amplification numérique (15 pts)

1. Soit la fonction $f(x, y) = x - y$. Quelle valeur approchée et incertitude (erreur absolue et relative) a t-on sur $f(x, y)$ lorsque $x = 101 \overset{+}{-} 2$ et $y = 100 \overset{+}{-} 1$. Comparer le nombre de cse et l'incertitude relative que l'on a sur x et y avec le nombre de cse et l'incertitude relative de $f(x, y)$. $<$**4 pts**$>$
2. Montrer que les trois fonctions suivantes sont équivalentes, $<$**3 pts**$>$

$$
\begin{aligned}
f_1(x) &= 1 - \cos x, \qquad (3)\\
f_2(x) &= \frac{\sin^2 x}{1 + \cos x}, \qquad (4)\\
f_3(x) &= 2\cos^2 \frac{x}{2}. \qquad (5)
\end{aligned}
$$

3. Donner en le justifiant l'expression qui aura le plus de problème d'erreur numérique lorsque $x \to 0$. On rappelle que $\cos x \approx 1 - x$ quand $x \to 0$ (developpement limité d'ordre 1 de la fonction $\cos x$ au voisinage de zéro). $<$**2 pts**$>$
4. Donner les différentielles exprimant l'amplification des erreurs numériques de ces trois expressions et leur comportement quand $x \to 0$, (on rappelle que $\sin x \approx x$ quand $x \to 0$). $<$**6 pts**$>$

Réponse

1. $f(x, y) = 1 \overset{+}{-} 3$, $<$**1 pt**$>$ (l'erreur absolue d'une différence est la somme des erreurs absolues) avec aucun chiffre significatif pour la valeur approchée de $f(x, y)$ $<$**0.5 pt**$>$ et une erreur relative $\epsilon_r(f(x, y)) = 3/1 = 300\%$ (!) $<$**0.5 pt**$>$.

À comparer avec $\epsilon_r(x) = 2/101 \approx 2\%$ $<$**0.5 pt**$>$ et $\epsilon_r(y) = 1/100 = 1\%$ $<$**0.5 pt**$>$ d'erreur relative de x et y respectivement connu avec deux chiffres significatifs chacun ($x = \underline{10}1 \overset{+}{-} 2$ $<$**0.5 pt**$>$ et $y = \underline{10}0 \overset{+}{-} 1$ $<$**0.5 pt**$>$). C'est le problème de la cancellation de cse dans le cas de la soustraction de valeur presqu'égale.

2. On a $(1 - \cos x)(1 + \cos x) = (1 - \cos^2 x) = \sin^2 x$ et donc $f_1(x) = \frac{\sin^2 x}{1+\cos x} = f_2(x)$ $<$**1.5 pt**$>$. De plus on a la relation, $\cos^2 x = (1/2)(1 - \cos 2x)$ soit $2\cos^2 \frac{x}{2} = f_3(x) = 1 - \cos x = f_2(x)$ $<$**1.5 pt**$>$.

3. L'expression qui aura le plus de problème quand $x \to 0$ est l'expression (3) car $\cos x \approx 1 - x$ quand $x \to 0$ et cela se traduira donc par la soustraction de deux valeurs presqu'égale (cancellation de cse).
$<$**2 pts**$>$

4.

$$
\begin{aligned}
\Delta f_1(x) &= |\sin x|\, \Delta x \underset{x\to 0}{\approx} |x|\, \Delta x, \quad < \mathbf{1pt} >\\
\Delta f_2(x) &= \Big|\frac{(2\sin x \cos x)(1 + \cos x) - \sin^3 x}{(1 + \cos x)}\Big| \Delta x \underset{x\to 0}{\approx} \Big|\frac{2x(1 - x)(2 - x) - x^3}{(2 - x)^2}\Big| \Delta x \underset{x\to 0}{\approx} |x|\, \Delta x, \quad < \mathbf{2pts} >\\
\Delta f_3(x) &= \big|-2\cos\frac{x}{2} \sin\frac{x}{2}\big| \Delta x \underset{x\to 0}{\approx} \big|-2(1 - \frac{x}{2})(\frac{x}{2})\big| \Delta x \underset{x\to 0}{\approx} |-x|\, \Delta x. \quad < \mathbf{2pts} >
\end{aligned}
$$

Les trois expressions se comportent de la même façon (quand $x \to 0$) face au problème de l'amplification des erreurs numériques. Le seul problème est donc donnée par l'expression (1) qui reste très sensible au problème de cancellation de cse. $<$**1 pt**$>$

III. Méthode du point fixe et de Newton (27 pts)
On se propose de trouver une valeur approchée d'une des deux racines de l'équation,

$$f(x) = \exp(x) - 4x = 0. \tag{6}$$

1. **Méthode du point fixe**.
   (a) Montrer qu'il existe une racine unique r pour cette équation (6) dans l'intervalle $J = [0,1]$. En remarquant que l'équation proposée est équivalente à $g_1(x) = x$ avec $g_1(x) = (1/4)\exp(x)$, montrer que cet intervalle est un intervalle sur lequel la convergence vers une solution unique par la méthode du pont fixe est assurée. $<$**3 pts**$>$
   (b) Déterminer analytiquement, grâce au théorème des accroissements finis (ou de la valeur moyenne), une majoration du type $|r_n - r| \leq K$, ou r_n désigne la valeur approchée, à la n-ième itération, de cette racine par la méthode itérative du point fixe. En déduire le nombre d'itérations n_* nécessaire pour obtenir, par cette méthode, une valeur approchée à 10^{-4} près de cette racine. $<$**7 pts**$>$
   (c) Calculer les 8 premières estimées $r_1, \ldots, r_8$, en partant de $r_0 = 0.5$. $<$**3 pts**$>$
   (d) Pour améliorer la vitesse de convergence du point fixe, la méthode de steffenson se propose de faire les étapes suivantes
      i. À partir de r_0, calculer r_1 et r_2 par la méthode du point fixe traditionelle.
      ii. Calculer $\hat{r}_0 = r_0 - (r_1 - r_0)^2/(r_2 - 2r_1 + r_0)$
      iii. Poser $r_0 = \hat{r}_0$ et recommencer à l'étape i. $<$**3 pts**$>$

      Calculer deux itérations par la méthode de Steffensen.
2. **Méthode de Newton**.
   (a) Soit $g_2(x)$, la fonction intervenant dans la méthode itérative de Newton pour la résolution de l'équation (6). Donner $g_2(x)$ ainsi que la relation itérative $r_{n+1} = g_2(r_n)$. On prendra $r_0 = 0.5$ pour amorcer cette relation itérative. Rappeler quelle doit être les conditions de convergence sur cette relation itérative pour que celle-ci converge. $<$**5 pts**$>$
   (b) En déduire une valeur approchée de r après 3 itérations (i.e., donner r_1, r_2, r_3) (avec toujours $r_0 = 0.5$). $<$**3 pts**$>$
   (c) Sans faire d'itérations, déterminer combien vous devriez en faire pour calculer cette racine à l'aide de la méthode de la bissection, avec une précision de 10^{-4} en partant de l'intervalle $J = [0,1]$. $<$**3 pts**$>$

Réponse

1.

(a). L'étude des variations de la fonction $f(x)$ sur $J = [0,1]$ montre que la fonction est continue et décroissante sur cet intervalle (donc monotone) ($f'(x) = \exp(x) - 4$ et $f'(x) < 0 \;\; \forall x \in [0,1]$) $<$**1 pt**$>$. De plus, on a $f(0) = 1$ et $f(1) = e - 4 \approx -1.28$ $<$**1 pt**$>$. Il existe donc une racine r unique dans cet intervalle. $g_1'(x) \in [e/4, (1/4)]$ pour $x \in J$ la fonction est donc contractante sur J et la convergence est assurée.
$<$**1 pt**$>$

(b). En utilisant le théorème de la valeur moyenne, on obtient, puisque $g_1(r) = r$ et $r_n = g_1(r_{n-1})$ avec r_n la valeur approchée de la racine à la n-ième itération,

$$\begin{aligned}
(r_n - r) &= \frac{g_1(r_{n-1}) - g_1(r)}{(r_{n-1} - r)} \times (r_{n-1} - r), \\
&= g_1'(\zeta) \times (r_{n-1} - r), \qquad \text{avec} \quad \zeta \in J.
\end{aligned}$$

En utilisant l'inégalité $|g_1'(x)| < \frac{e}{4}$ <**1 pt**>, (i.e., la borne la plus pessimiste), on obtient les inégalités suivantes,

$$\begin{aligned}
|r_n - r| \leq \frac{e}{4}\,|r_{n-1} - r| &\leq \left(\frac{e}{4}\right)^2 |r_{n-2} - r|\,, \\
&\leq \ldots \\
&\leq \left(\frac{e}{4}\right)^n |r_0 - r|\,, \\
&\leq \left(\frac{e}{4}\right)^n. \quad < \mathbf{4pts} >
\end{aligned}$$

On obtiendra obligatoirement donc $|r_n - r| < 10^{-4}$ des que,

$$\begin{aligned}
\left(\frac{e}{4}\right)^n &< 10^{-4}, \\
n \ln\left(\frac{e}{4}\right) &< \ln\left(10^{-4}\right).
\end{aligned}$$

Soit pour tout $n \geqslant 24$. <**2 pts**>

(c). En partant de $r_0 = 0.5$, on a, $r_n = g_1(r_{n-1})$ et,

$$\begin{aligned}
r_0 &= 0.5, \\
r_1 &= 0.4121803177, \\
r_2 &= 0.3775266777, \\
r_3 &= 0.3646680893, \\
r_4 &= 0.3600089914, \\
r_5 &= 0.3583355456, \\
r_6 &= 0.3577364326, \\
r_7 &= 0.3575221615, \\
r_8 &= 0.3574455631,
\end{aligned}$$

< **3pts** >

(d). Une itération de la méthode de steffensen permet d'estimer la racine en utilisant r_0, r_1, r_2. On obtient $\hat{r}_0 = 0.3549394263$ <**1.5 pts**>. En utilisant cette estimée comme point de départ ($\hat{r}_0 = r_0$) à une méthode du point fixe, on peut estimer r_1 et r_2, i.e., $r_1 = 0.356523567$, $r_2 = 0.3708798$ et on calcule la $2e$ estimée par la méthode de Steffenson $\hat{r}_0 = 0.3574023548$ <**1.5 pts**>.

2.

(a). La fonction $f(x)$ est dérivable sur J et on a,

$$g_2(x) = x - \frac{f(x)}{f'(x)} = x - \frac{\exp(x) - 4x}{\exp(x) - 4}.$$

On a donc la formule itérative suivante,

$$r_{n+1} = r_n - \frac{\exp(r_n) - 4r_n}{\exp(r_n) - 4} \qquad < \mathbf{3pts} >.$$

Condition de convergence : Pour $r \in [0, 1]$, $\exp(r_n) - 4 \neq 0$, on aura donc aucune division par zéro et en partant de $r_0 = 0.5$, c'est à dire pas trop loin de la solution (en tout cas, en étant pas séparé de la solution par un minimum local ce qui est le cas puisque on a vu precedemment que $f(x)$ est monotone sur $[0, 1]$) <**2 pts**>.

la deuxième solution consisterait à raisonner comme dans le cas d'un point fixe et de démontrer que $|g_2'(x)| < 1 \; \forall x \in J$ ce qui serait beaucoup plus ardu.

(b). En partant de $r_0 = 0.5$, on a, $r_n = g_2(r_{n-1})$ et,

$$\begin{aligned} r_0 &= 0.5, \\ r_1 &= 0.3506009837, \\ r_2 &= 0.357390196, \\ r_3 &= 0.3574029561. \quad < \mathbf{3pts} > \end{aligned}$$

(c). Si on avait utiliser la méthode de la bissection, et en partant de l'intervalle $[0, 1]$, il aurait fallu faire au moins n itérations avec,

$$\frac{b-a}{2^{n+1}} \leq 10^{-4},$$

Soit au moins $n \geqslant 13$ itérations $<$**3 pts**$>$.

IV. Méthodes itératives pour les systèmes (15 pts)

1. **Amélioration itérative d'un système**
   Soit le système $Ax = b$ à résoudre. Supposons que $x^{[0]}$ soit une solution estimée numériquement par une méthode quelconque (Gaus ou autre) entachée d'une erreur numérique.
   (a) Utiliser la méthode du point fixe pour trouver la relation $x^{[n+1]} = G(x^{[n]})$ nécessaire pour améliorer itérativement cette solution (i.e., exprimer $G(.)$ en fonction de A, b et $x^{[n]}$ et itérer cette relation $x = g(x)$). $<$**5 pts**$>$
   (b) Rappeler les deux conditions qui sont utiles à la convergence de cette méthode de rafinement itérative. $<$**2 pts**$>$
2. **Méthode itératives de résolution**
   Soit le système suivant,

$$\begin{aligned} x_1 + x_2 - 5x_3 &= -12, \\ 4x_1 + x_2 + 2x_3 &= 12, \\ x_1 + 8x_2 + x_3 &= 20. \end{aligned} \qquad (7)$$

   (a) Donner la relation itérative (sans calculer la solution de ce système) permettant de résoudre le système (7) avec la méthode de Gauss-Seidel. $<$**4 pts**$>$
   (b) Discuter de la convergence de ce système. Est elle assurée ? $<$**2 pts**$>$
   (c) Indiquer quels sont les avantages computationnelles de la méthode de Jacobi sur la méthode de Gauss-Seidel. $<$**2 pts**$>$

Réponse

1.

(a). Nous devons mettre $Ax = b$ sous la forme $x = G(x)$, ce qui peut être fait en écrivant,

$$\begin{aligned} A\,x &= b, \\ x &= A^{-1}b + x - x, \\ x &= A^{-1}b + x - A^{-1}A\,x, \\ x &= \underbrace{x + A^{-1}(b - A\,x)}_{g(x)}. \end{aligned}$$

La relation itérative s'écrit donc,

$$x^{[n+1]} = G(x^{[n]}) = x^{[n]} + \underbrace{A^{-1} \underbrace{(b - Ax^{[n]})}_{\text{Résidu à l'itér. n}}}_{\text{Terme correcteur}} . \quad < \mathbf{5\ pts} >$$

(b). Cette technique converge lorsque $\text{cond}(A)$ n'est pas trop grand et lorsque le calcul du terme correcteur est fait en double précision. <**2 pts**>. Une autre solution consisterait a considérer la condition de convergence du point fixe et de démontrer que $|G'(x)| < 1\, \forall x$ ce qui serait beaucoup plus ardu.

2.

(a). On doit d'abord ré-arranger le système pour que celui-ci soit diagonalement dominant, i.e.,

$$\begin{aligned} 4x_1 + x_2 + 2x_3 &= 12, \\ x_1 + 8x_2 + x_3 &= 20, \\ x_1 + x_2 - 5x_3 &= -12. \end{aligned}$$

Puis ré-écrire le système sous la forme $x^{[n+1]} = G(x^{[n]}, x^{[n+1]})$,

$$\begin{aligned} x_1^{[n+1]} &= 3 - (1/4)x_2^{[n]} - (1/2)x_3^{[n]}, \\ x_2^{[n+1]} &= (5/2) - (1/8)x_1^{[n+1]} - (1/8)x_3^{[n]}, \\ x_3^{[n+1]} &= -(12/5) - (1/5)x_1^{[n+1]} - (1/5)x_2^{[n+1]}. \quad < \mathbf{4pts} > \end{aligned}$$

(b). La convergence est assurée (condition suffisante de convergence) si $|a_{ii}| > \sum_{j=1, j\neq i}^{j=3} |a_{ij}| \quad \forall\, i = 1, 2, 3$, <**1 pt**>, ce qui est le cas içi. <**1 pt**>

(c). la méthode de Jacobi est particulièrement intéressante (en terme de coût calculatoire) si elle est implémenter sur un ordinateur massivement parallèle, ce qui est possible car $x^{[n+1]}$ ne dépend que de $x^{[n]}$. <**2 pts**>

V. Système d'équations (25 pts)

Soit la matrice A suivante,

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 3 & 5 & 2 \end{pmatrix}. \tag{8}$$

1. Faîte deux décompositions LU de cette matrice, une pour laquelle determinant (L)= 1 et l'autre assurant determinant (U)= 1 (n'utilisez ni le pivotage partiel ou total dans les deux cas et indiquez le nom de la technique utilisée). <**7 pts**>
2. Calculer le déterminant de la matrice A. <**2 pts**>
3. Utiliser l'une des deux décompositions LU faîtes precedemment pour calculer la solution du système linéaire, <**4 pts**>

$$Ax = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

4. Expliquez comment vous pouvez utiliser efficacement les étapes précédentes plus quelques calculs que vous détailerez pour calculer la matrice inverse (A^{-1}). Calculer cette matrice inverse. <**6 pts**>
5. Si la matrice A est remplacé par une matrice légèrement perturbée A_ϵ, i.e.,

$$A_\epsilon = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3+\epsilon & 0 \\ 3 & 5 & 2 \end{pmatrix}.$$

Donner l'errreur relative (suivant la norme ∞) que vous ferez sur le vecteur solution en fonction de ϵ. Faîtes une application numérique pour $\epsilon = 10^{-3}$. <**6 pts**>

Réponse

1. La décomposition LU unique assurant des 1 sur la diagonale de la matrice L est la décomposition obtenue par la méthode de Gauss <**1 pt**>. L'opération $\text{ligne}_3 = \text{ligne}_3 - (3)\text{ligne}_1$ donne,

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & -1 & 2 \end{pmatrix},$$

suivie de l'opération $\text{ligne}_3 = \text{ligne}_3 - (-1/3)\text{ligne}_2$ donne,

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Ce qui nous donne la décomposition LU suivante,

$$A \;=\; \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & -(1/3) & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{U}. \qquad < \mathbf{2.5\ pts} >$$

La décomposition LU unique assurant des 1 sur la diagonale de la matrice U est la décomposition obtenue par la méthode de factorisation directe (ou réduction de crout) <**1 pt**> dont le résultat est (presque) immédiat,

$$A \;=\; \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 3 & -1 & 2 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{U}. \qquad < \mathbf{2.5\ pts} >$$

2. Dans les deux cas,

$$\det(\mathrm{A}) \;=\; \det(\mathrm{L}) \times \det(\mathrm{U}) = 6. \quad < \mathbf{2pts} >$$

3. En prenant la première décomposition LU, on doit résoudre le système suivant,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & -(1/3) & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Par substitution avant puis arrière on trouve, $Ly = b$ avec $y = (1,\, 0,\, -3)^t$, puis $Ux = y$ avec finalement $x = (1,\, 0, -(3/2))^t$. <**4 pts**>

4. Les colonnes de la matrices inverse correspondent aux 3 solutions des systèmes linéaires $Ax = b$ où $b = (1, 0, 0)$ (déjà calculé précedemment) et des deux autres vecteurs $b = (0, 1, 0)$ et $b = (0, 0, 1)$. <**1 pt**>

Pour le deuxième vecteur, par substitution avant puis arrière on trouve, $Ly = b$ avec $y = (0,\ 1,\ (1/3))^t$, puis $Ux = y$ avec finalement $x = ((-2/3),\ (1/3),\ (1/6))^t$. <**2 pts**>

Pour le troisième vecteur, on trouve, $Ly = b$ avec $y = (0,\ 0,\ 1)^t$, puis $Ux = y$ avec finalement $x = (0,\ 0,\ (1/2))^t$. <**2 pts**>

La matrice inverse s'écrit donc,

$$(A^{-1}) = \begin{pmatrix} 1 & -(2/3) & 0 \\ 0 & (1/3) & 0 \\ -(3/2) & (1/6) & (1/2) \end{pmatrix}. \quad < \textbf{1 pt} >$$

5.

On calcule d'abord le nombre de conditionnement de la matrice $\text{cond}(A) = |A|\,|A^{-1}| \approx 10 \times \max((5/3), (1/3), (13/6)) = 10 \times (13/6) = 130/6 \approx 21.66$, <**1 pt**> On utilise ensuite la relation vue en cours,

$$\begin{aligned} \frac{\| x^* - x \|_\infty}{\| x \|_\infty} &\leq \text{cond}(A)\, \frac{\| A - A^* \|_\infty}{\| A \|_\infty}, \quad < \textbf{2 pts} > \\ &\leq 21.66 \times \frac{\epsilon}{10} \approx 2.166\epsilon. \quad < \textbf{2 pts} > \end{aligned}$$

pour $\epsilon = 10^{-3}$, on trouve,

$$\epsilon_r(x) \leq 2.16610^{-3}. \quad < \textbf{1 pt} >$$