DIRO
IFT 2425

EXAMEN INTRA

Max Mignotte

DIRO, Département d'Informatique et de Recherche Opérationnelle, local 2377
Http : //www.iro.umontreal.ca/∼mignotte/ift2425/
E-mail : mignotte@iro.umontreal.ca

Date : *02/03/2012*

I	Mesure d'Incertitude et Amplification d'Erreur (29 pts)
II	Erreur en Arithmétique Flottante (39 pts)
III	Méthode du Point Fixe (36 pts)
IV	Factorisation LU (13 pts)
Total	116 points.

Tous documents personnels, calculatrices et calculateurs autorisés

I. Mesure d'Incertitude et Amplification d'Erreur (29 pts)

La dépendence à la température de la viscosité d'un gaz peut être exprimé par l'équation de Sutherland

$$\mu = \frac{b\,T^{3/2}}{T+s}$$

où μ est la viscosité dynamique (en Pa·s), T est la température (en Kelvin) et b et s sont deux constantes empiriques. Plus précisement, s est la constante de Sutherland pour le gaz en question. On supposera ici que le gaz considéré est l'air et que cette constante est égale à $s = 120$ en considérant que tous les chiffres donnés sont considérés comme significatifs "exact" (cse). De plus, on considérera que l'on connait la valeur de b sans erreur ($b = 1.5$) et que l'on connait une mesure approximée pour T, i.e., $T \approx 300$ mais pas l'incertitude faite sur cette variable.

On vous demande de :

1. Exprimer, pour la constante s, l'incertitude que l'on a sur elle en la mettant sous la forme $s = s^\star \pm \Delta s$ où $s^\star$ et Δs seront déterminés. Préciser aussi l'erreure relative que l'on fait sur cette constante.
   <5 pts>
2. Donner l'approximation de la viscosité μ et la borne supérieure de l'erreur absolue (i.e., incertitude totale de $\mu(T,s)$ ou $\Delta\mu(T,s)$) que l'on obtiendrait par la méthode de propagation d'erreur (utilisant l'approximation de Taylor de la fonction au premier ordre) en fonction de Δs et ΔT.
   <8 pts>
3. Quelle incertitude maximale sur T (i.e., ΔT_{max}) pourrait-on faire pour que sa contribution sur l'incertitude total de $\mu(T,s)$ ne soit pas plus grande que l'incertitude engendrée par l'imprécision sur s ? Donner ensuite l'erreur relative maximale que l'on peut faire sur cette variable T. Dans ce cas (où les deux incertitudes sur s et T sont équivalentes), indiquer le nombre de chiffre significatifs de l'approximation de la viscosité μ.
   <10 pts>
4. Supposons maintenant que $T = 300 \pm \Delta T_{\text{max}}$. Peut on utiliser la méthode de la fourchette pour obtenir $\Delta\mu(T,s)$? Si oui expliquer pourquoi et utiliser cette méthode pour obtenir l'intervalle d'incertitude résultante pour $\mu(T,s)$ que l'on obtiendrait en fonction de ΔT_{max}.
   <6 pts>

Réponse

1.

Le chiffre des unités du nombre 120 étant significatif, i.e., "0" (de rang 0), cela nous permet de trouver une borne supérieure de l'erreur absolue ou de l'incertitude que l'on a sur cette constante s, i.e., $\Delta s = 0.5 \times 10^0$ et donc nous permet d'écrire $s = \underline{120} \pm 0.5$. **<3 pts>**

Puisque $\Delta s = 0.5$, l'erreur relative que l'on fait sur cette constante est de l'ordre de $\Delta s/s = 0.5/120 \approx 0.0042$ ou 0.42% d'erreur relative. **<2 pts>**

2.

On a pour valeur approchée de μ

$$\mu^* = 1.5 \cdot \frac{300^{3/2}}{300+120} \approx 18.56 \quad (\text{Pa·s}) \qquad \textbf{< 2 pts >}$$

Pour calculer l'incertitude de μ, on doit calculer la différentielle $\Delta\mu(T,s)$, i.e.,

$$\begin{aligned}\Delta\mu(T,s) &= \left|\frac{\frac{3b}{2}T^{1/2}(T+s) - bT^{3/2}}{(T+s)^2}\right| \cdot \Delta T + \left|-\frac{b\,T^{3/2}}{(T+s)^2}\right| \cdot \Delta s \\ &= \left|\frac{b\,T^{3/2} + 3\,b\,s\,T^{1/2}}{2\,(T+s)^2}\right| \cdot \Delta T + \left|-\frac{b\,T^{3/2}}{(T+s)^2}\right| \cdot \Delta s \qquad \textbf{< 6 pts >}\end{aligned}$$

$\boxed{\textbf{3.}}$

Numériquement, l'expression précédente permet d'écrire

$$\begin{aligned}\Delta\mu(T,s) &= \left|\frac{1.5 \times 300^{3/2} + 3 \times 1.5 \times 120 \times 300^{1/2}}{2 \times (300+120)^2}\right| \cdot \Delta T + \left|-\frac{1.5 \times 300^{3/2}}{(300+120)^2}\right| \cdot \Delta s \\ &\approx 0.0486 \cdot \Delta T \;+\; 0.0442 \cdot \Delta s\end{aligned}$$

Le deuxieme terme, i.e., $0.0442\,\Delta s \approx 0.0442 \times 0.5 = 0.0221$ est la contribution de l'imprécision sur s sur l'incertitude total de $\mu(T,s)$. Pour que la contribution de l'imprécision sur T ne soit pas plus grande que l'incertitude engendrée par l'imprécision sur s, il faut donc que

$$\begin{aligned}0.0486 \cdot \Delta T_{\max} &\leq 0.0221 \\ \Delta T_{\max} &\lessapprox 0.4547 \qquad \textbf{< 6 pts >}\end{aligned}$$

Soit une erreur relative maximale pour T de $E_r(T) = 0.4547/300 \approx 0.0015157 = 0.15\%$ **<2 pts>**

Lorsque l'incertitude sur s et T sont équivalentes ; $\Delta\mu(T,s) = 0.0221 \times 2 = 0.0442 < 0.5 \times 10^{-1}$ et le rang du dernier chiffre significatif dans l'aproximation de μ est le chiffre des dixième. On obtient donc $\mu^* = \underline{18.5}6$. **<2 pts>**

$\boxed{\textbf{4.}}$

La méthode de la fourchette peut être utilisée ici car la fonction $\mu(T,s) = b\,T^{3/2}/(T+s)$ est strictement monotone sur chacun des intervalles dans lequel se trouve les variables incertaines T et s (i.e., strictement croissante en T sur l'interval dans lequel évolue T et strictement décroissante sur l'interval dans lequel évolue s). Il est aussi possible, en utilisant les valeurs minimales et maximales pour T et s, de déterminer la borne minimale et maximale pour μ. Dans notre cas :

$$\mu \in \left[b\frac{T_{\min}^{3/2}}{T_{\max} + s_{\max}}, \cdots, b\frac{T_{\max}^{3/2}}{T_{\min} + s_{\min}}\right]$$

Numériquement, avec $T_{\min} \in [300 - \Delta T_{\max}, \ldots, 300 + \Delta T_{\max}]$ et $s \in [119.5, \ldots, 120.5]$

$$\mu \in \left[1.5\,\frac{(300 - \Delta T_{\max})^{3/2}}{(420.5 + \Delta T_{\max})}, \ldots, 1.5\,\frac{(300 + \Delta T_{\max})^{3/2}}{(419.5 - \Delta T_{\max})}\right] \qquad \textbf{< 6 pts >}$$

Nota : En considérant $\Delta T_{\max} = 0.9361$, on obtient l'intervalle

$$\mu \in [18.40795371, \cdots, 18.7085842]$$

ou $\underline{18}.55 \pm 0.3 < 0.5 \times 10^0$, ce que nous donne approximativement la même chose que la méthode utilisant la série de Taylor.

II. Erreur en Arithmétique Flottante (39 pts)

1. Soit la représentation flottante binaire 32 bits avec un format du type $\pm 0.xxxx\ldots$ (avec 24 x, i.e , 24 bits pour la mantisse) dont un bit caché permettant de créer le bit de signe et un exposant sur 8 bits permettant d'exprimer les valeurs de l'exposant de -126 à 127.
   Donner la première valeur flottante immédiatement supérieure à 2.0 qui aura une représentation exacte (i.e., sans erreur d'affectation). Plus généralement, donner la première valeur flottante immédiatement supérieure à 2^p ($p > 0$) qui aura une representation exacte.
   <7 pts>
2. Souvenez vous du problème de la défaillance du missile US patriote présenté en cours (à la fin du chapitre sur la propagation des erreurs numériques) et du problème créé par le fait que l'horloge interne de ce missile était incrémentée 10 fois par seconde (par pas et incrément de 0.1 seconde). Quelle(s) solution(s) simple(s) aurait permis à ce missile de ne pas défaillir.
   <6 pts>
3. Supposons maintenant que l'on veuille faire une procédure qui nécessite de diviser un intervalle de longueur 1 (entre $[0.0, \ldots, 1.0]$ en 20 sous-intervalles de même longueur. Un étudiant à programmé en pseudo-code la routine suivante :

```
. float NbInterv=20.0 ;
. float l ;
. for( l=0.0 ; l != 1.0 ; l+=1.0/20.0 )      { ...}
```

   Expliquer, en justifiant bien votre réponse, quel problème numérique sera confrontée cette routine. Indiquer aussi comment résoudre ce problème.
   <6 pts>
4. Soit la routine suivante, dans laquelle on suppose que l'ensemble des valeurs du tableau Tab[.] sont des flottants (compris entre 10^{-4} et 2×10^{-4}). On supposera aussi que pour des raisons techniques on ne peut utiliser des doubles.

```
. long int ite ;
. float Sum=0.0 ;
. for(ite=0.0 ; ite<1000000000 ; ite++)     Sum+=Tab[ite] ;
. printf(" [Sum=%f]", Sum) ;
```

   Indiquer, en justifiant bien votre réponse, à quel type de problème numérique on sera confronté et proposer une solution algorithmique pour y remedier (i.e., qui permettrait de minimiser l'erreur d'estimation mis en évidence précédemment).
   <8 pts>
5. Pour chacune de ces fonctions, il existe un voisinage au point x_0 pour lequel $f(x)$ ne pourra être évalué très précisement (i.e. avec une bonne précision) et pour lequel des problèmes numériques pourraient survenir.
   (a) $\sqrt{1+x^2} - \sqrt{1-x^2}$
   (b) $\ln x - \ln(1/x) \qquad \text{pour} \;\; x \neq 0$
   (c) $\exp(x) - \exp(-2x)$

   Trouver ce point x_0 et expliquer pourquoi (i.e., identifier et citer le problème numérique associé) et proposer une formule équivalente plus fiable qui permettrait d'éviter ces problèmes numériques et qui permettrait d'augmenter la précision des calculs de ces trois expressions dans un voisinage de x_0.
   <12 pts>

Réponse

1.

La valeur flottante, immédiatement plus grande que 2.0 et sans erreur d'affectation sera la valeur

$$0.\underbrace{1000000000000000000001}_{24 \text{ bits}} \times 2^2 = 2^{(-1+2)} + 2^{(-24+2)} = 2^1 + 2^{-22} = 2 + 2^{-22}$$

Et plus généralement, la valeur flottante, immédiatement plus grande que 2^p ($p > 0$) et sans erreur d'affectation sera la valeur

$$0.\underbrace{1000000000000000000001}_{24 \text{ bits}} \times 2^{(p+1)} = 2^{(-1+p+1)} + 2^{(-24+p+1)} = 2^p + 2^{(p-23)}$$

<7 pts>

2.

Je rappelle que le missile patriote comptait le temps par pas de 10 coups par seconde puis multipliait ce compteur par la valeur flottante "0.1" (stocké dans un registre), qui n'a pas de représentation exacte en binaire (et dont l'erreur d'affectation en base 10, de cette représentation flottante est de $\epsilon = 0.000000095$). Après 100 heures, l'erreur entre l'horloge de la station qui guidait le missile et l'horloge interne du missile fut de $100 \times 60 \times 60 \times 10 \times \epsilon \approx 0.34$ seconde, source du problème de notre missile qui s'écrasa sur une caserne US au lieu de stopper le missile SCUD adverse.

Parmis les solutions numériques atténuant ou réglant ce problème (l'idée de ne pas faire de missile du tout, donné par un étudiant, ne fait pas partie des solutions numériques demandées dans cette question), on peut noter que l'utilisation de DOUBLE au lieu de FLOAT ou un recalage régulier de l'horloge du missile sur les données d'une horloge atomique aurait permis de minimiser le problème mais pas de le résoudre.

Une solution plus judicieuse et sans coût aurait été de compter par incrément et pas de 1/8 ou 1/16, ces deux pas (ou tout autre pas) ayant la propriété d'avoir une représentation exacte en base 2.

<6 pts>

3.

La boucle For(.) va incrémenter par pas de $1.0/20.0 = 0.05$ qui n'a pas de représentation exacte en base 2. En fait 0.05 doit être représenté par un nombre infini de bits comme la valeur 0.1 (0.05 étant la valeur divisée par deux de 0.1 ou décalée d'un bit à droite de la representation binaire de 0.1, qui elle exige un nombre infini de bits pour une representation parfaite en binaire). De ce fait, après 20 incréments de 0.05, on aura une somme qui (du fait de l'erreur d'affectation de 0.05) nous donnera une valeur flottante, proche mais néanmoins différente de 1.0. La condition d'arrêt de cette boucle va ainsi comparer deux flottants ; 1.0 qui a une representation exacte en flottant et une valeur flottante, proche mais différente de 1.0 et la boucle ne va jamais s'arrêter.

<3 pts>

Une solution simple pour résoudre ce probleme consiste à implémenter la boucle de cette façon :

```
. for(l=0.0, int n=0 ; n<=20 ; l+=1.0/20.0, n++)  { ... }
```

ou de cette façon

```
. for(l=0.0, int n=0 ; n<=20 ; l=n/20.0, n++)  { ... }
```

<3 pts>

4.

Dans cette routine qui implique l'estimation d'une somme d'un milliard de termes, on sera confronté, pour une valeur de ITE assez grande, à la perte de précision due à l'addition de deux nombres d'ordre de grandeur différente (erreur de décalage).

<**4 pts**>

Une solution simple pour résoudre ce probleme consiste à incrémenter la variable SUM du résultat des différentes sommes partielles (par exemple par bloc de 10000 valeurs consécutives de ITE), du tableau TAB[.].

```
. LONG INT ITE ;
. FLOAT SUM=0.0, SUMPAR=0.0 ;
. FOR( ITE=0.0 ; ITE < 1000000000 ; ITE++ )
    . { SUMPAR+=TAB[ITE] ;
    . IF (!(ITE % 10000))    { SUM+=SUMPAR ;   SUMPAR=0.0 ; }    }
. PRINTF(" [SUM=%F]", SUM) ;
```

<**4 pts**>

Nota : Si le tableau TAB[.] est remplie de valeur flottante égale à $0.1/1024$, la véritable somme, sans erreur, est SUM= 97656.25. En utilisant la version non corrigée de cette routine (n'utilisant pas de somme partielle), on obtient sum=2048 ! ! ! ! et en utilisant la version corrigée, ci-dessus, on obtient sum=97655.031250.

Nota : Pour solutionner le problème, au lieu d'une somme partielle, un trie par ordre croissant du tableau TAB[.] serait aussi possible et éviterait la perte de précision due à l'addition de deux nombres d'ordre de grandeur différente (erreur de décalage). Un BONUS de $\boxed{+2}$ pour celui qui m'a indiquer ces deux possibilités.

5.

Dans les trois cas, on aura un problème d'annulation de cse due à la soustraction de deux nombres approximés quasi égaux. Pour éviter ce problème, on peut écrire ces relations de telle façon que plus aucune soustraction de deux nombres incertains (i.e., imprécis) quasi égaux apparaissent.

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

plus de problème de précision au voisinage de $x_0=0$

$\ln(x)-\ln(1/x)=\ln x^2=2\ln(x)$ plus de problème de précision au voisinage de $x_0=1$

$$\exp(x)-\exp(-2x)=\Big(1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots\Big)-\Big(1-2x+\frac{4x^2}{2!}-\frac{8x^3}{3!}+\dots\Big)=3x-\frac{3}{2}x^2+\frac{3}{2}x^3-\dots$$

plus de problème de précision au voisinage de $x_0=0$, une alternative est d'écrire cette expression

$$\frac{\exp(3x)-1}{\exp(2x)}=\exp(-2x)\Big(3x+\frac{9}{2!}x^2+\frac{27}{3!}\,x^3+\dots\Big)$$

<**12 pts**>

III. Méthode du Point Fixe (36 pts)

On se propose de trouver numériquement dans R^+ une valeur approchée d'une des racines de la fonction

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

1. Montrer qu'il existe une racine unique r pour cette Eq. (1) dans l'intervalle $J=[0.5,1.0]$. En remarquant que l'équation $f(x)=0$ est équivalente à $g_1(x)=x$ avec $g_1(x)=\exp(-x)$, montrer que l' intervalle

J est un intervalle sur lequel la convergence vers une solution unique par la méthode du point fixe est assurée.
<**6 pts**>

2. Utiliser le résultat de la question précédente pour calculer les 6 premières estimées $r_1, \ldots, r_6$, en partant de $r_0 = 0.75$.
<**6 pts**>

3. Trouver une fonction $g_2(x)$ équivalente à l'équation $f(x) = 0$ et pour laquelle vous demontrerez qu'elle n'est pas contractante sur J (i.e., pour laquelle la convergence vers une solution unique par la méthode du point fixe n'est pas assurée).
<**5 pts**>

4. Pour obtenir à l'avance le nombre d'itérations nécessaires de la méthode du point fixe pour arriver à la racine souhaitée avec la précision voulue, on peut essayer de trouver (en utilisant le 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 ni-ème itération, de cette racine. Trouver cette majoration et en déduire le nombre d'itérations nécessaire pour obtenir, par cette méthode, une valeur approchée à 10^{-2} près de cette racine.
<**5 pts**>

5. Une autre possibilité pour, cette fois-ci, arrêter la procédure itérative du point fixe au moment où on est arrivé à la racine souhaitée avec la précision voulue, consiste à trouver une majoration du type

$$|r_n - r| \leq K \, |r_{n+1} - r_n| \, ,$$

où K est un majorant de $\left|\frac{1}{1-g'(x)}\right|$ sur J et $(r_{n+1} - r_n) = (r_{n+1} - r) - (r_n - r)$ est trouvé à l'aide du théorème des accroissement finis ou de la valeur moyenne. En déduire, avec cette méthode le nombre d'itérations nécessaire pour obtenir une valeur approchée à 10^{-2} près de la racine r.
<**8 pts**>

6. Si on avait utilisé la méthode de la bissection, avec combien d'itérations serait-on arrivé à une valeur approchée de la racine à 10^{-2} près ? (Nota : on vous demande d'obtenir une estimation de ce nombre en utilisant les propriété de la méthode de la bissection mais sans calculer les différentes estimation $r_0, r_1, \ldots,$ donné par cette méthode.)
<**6 pts**>

Réponse

$\boxed{\textbf{1.}}$

L'étude des variations de la fonction f sur $J = [0.5, 1.0]$ montre que la fonction est continue et croissante sur J (donc monotone) (car $f'(x) = \exp(x) + 1/x^2 > 0$ sur J). <**1.5 pts**>
De plus, on a $f(0.5) \approx -0.35$ et $f(1.0) \approx 1.6$ donc $f(0)f(1) < 0$ et puisque la fonction f est continue et croissante sur J, il existe donc une racine r unique dans cet intervalle. <**1.5 pts**>

De plus $f(x) = 0$ est équivalent à l'équation $x = \exp(-x) = g_1(x)$ avec

$$|g_1'(x)| = |-\exp(-x)| = |\exp(-x)| < 1 \quad \forall x \in J = [0.5, 1.0]$$

parce que $g_1''(x) = \exp(-x) > 0$ sur J, donc g_1' croissante sur J et $g_1'(0.5) \approx -0.61$ et $g_1'(1.0) \approx 0.36$.
La fonction $g_1(x)$ est donc contractante sur J et la convergence est assurée.
<**3 pts**>

$\boxed{\textbf{2.}}$

En partant de $r_0 = 0.75$, on a, $r_n = g_1(r_{n-1})$ et,

$$\begin{aligned} r_1 &= 0.4723665527 \\ r_2 &= 0.6235249163 \\ r_3 &= 0.5360515666 \\ r_4 &= 0.5850537436 \\ r_5 &= 0.5570759217 \\ r_6 &= 0.5728817682 \end{aligned}$$

qui va converger doucement vers la valeur $r = 0.5671432904$.

<6 pts>

$\boxed{\textbf{3.}}$

$f(x) = 0$ est équivalent à $x = g_2(x) = x^2 \exp(x)$. Sur J, $g_2(x)$ n'est pas contractante car $|g_2'(x)| = |x^2 \exp(x) + 2x \exp(x)|$ et $|g_2'(1)| \approx 8.15 > 1$.

<5 pts>

Nota :

- $f(x) = 0$ est équivalent aussi à $x = g_2(x) = -\ln(x)$. Sur J, cette forme de $g_2(x)$ n'est pas contractante car $|g_2'(x)| = |-1/x|$ et $|g_2'(0.5)| = 2 > 1$.
- $f(x) = 0$ est équivalent aussi à $x = g_2(x) = x\exp(x) - 1$. Sur J, cette forme de $g_2(x)$ n'est pas contractante car $|g_2'(x)| = |x\exp(x) + \exp(x)|$ et $|g_2'(1.0)| = 5.43 > 1$.
- $f(x) = 0$ est équivalent aussi à $x = g_2(x) = \exp(x) - (1/x) + x$. Sur J, cette forme de $g_2(x)$ n'est pas contractante car $|g_2'(x)| = |\exp(x) + 1 + (1/x^2)|$ et $|g_2'(1.0)| \approx 4.72 > 1$.
- $f(x) = 0$ est équivalent aussi à $x = g_2(x) = x^2 \exp(x)$. Sur J, cette forme de $g_2(x)$ n'est pas contractante car $|g_2'(x)| = |2x\exp(x) + x^2\exp(x)|$ et $|g_2'(1.0)| \approx 8.15 > 1$.

$\boxed{\textbf{4.}}$

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

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

En utilisant l'inégalité $|g'(\xi)| < 0.61$, on obtient les inégalités suivantes

$$|r_n - r| \leq (0.61)\,|r_{n-1} - r| \quad \leq \quad (0.61)^2\,|r_{n-2} - r| \leq \ldots \leq (0.61)^n \left(\frac{1}{2}\right)$$

On obtiendra donc $|r_n - r| < 10^{-2}$ dès que $0.5 \times (0.61)^n < 10^{-2}$, i.e., dès que $n = 8$.

<5 pts>

Nota : Dans notre cas $r_8 = 0.5689$ et plus d'itérations ne permet pas de changer le chiffre des centième de cette estimation qui est aussi significatif au sens du larousse.

$\boxed{\textbf{5.}}$

En utilisant le théorème des accroissement finis, on a

$$\begin{aligned} r_{n+1} - r_n &= (r_{n+1} - r) - (r_n - r) \\ &= (g(r_n) - g(r)) - (r_n - r) \\ &= \frac{(g(r_n) - g(r)) - (r_n - r)}{(r_n - r)} \times (r_n - r) \\ &= (g'(\zeta) - 1) \times (r_n - r) \quad \text{avec } \zeta \text{ compris entre } r_n \text{ et } r \end{aligned}$$

De ce fait, on a $\quad |r_n - r| = \dfrac{1}{|1 - g'(\zeta)|} \times |r_{n+1} - r_n| \quad$ avec $\quad \forall x \in J \qquad |g'(x)| < 0.61$

d'où $\quad \dfrac{1}{|1 - g'(\zeta)|} \leq 2.57 \quad$ et on peut écrire $\quad |r_n - r| \leq 2.57 \times |r_{n+1} - r_n|$

Si on veut $|r_n - r|$ inférieur à 10^{-2}, il suffit de choisir n tel que $|r_{n+1} - r_n| < \frac{10^{-2}}{2.57} \approx 0.00389 \approx 4 \times 10^{-3}$.

<8 pts>

Nota : Dans notre, on arrive à $r_7 \approx 0.5639$, $r_8 \approx 0.5689$ et $r_9 \approx 0.5661$ la différence $|r_8 - r_7| \approx 5 \times 10^{-3}$ n'atteint pas le critère souhaité. Par contre la différence $|r_9 - r_8| \approx 2.8 \times 10^{-3} < 4 \times 10^{-3}$ et la méthode de cette question nous indique que l'on est arrivé avec $n = 9$, i.e., avec r_9 à la précision voulue.

$\boxed{\textbf{6.}}$

La méthode de la bissection permet de construire à partir de l'intervalle $[0.5, 1.0]$ contenant r, un nouvel intervalle de longueur moitié contenant r. En appliquant n fois consécutives la méthode, on obtient un intervalle de longueur $0.5/2^n$ contenant r. A l'itération n, on a un majorant de l'erreur absolu donné par $\Delta r = \frac{|0.5|}{2^n}$. Or on veut $\Delta r < 10^{-2}$, donc $2^n > 0.5 \times 10^2 = 50$, ce qui, dans notre cas est vrai des que $n = 6$. Ce nombre d'itération est à comparer avec le $n = 8$ ou $n = 9$ obtenu par les deux critères différents utilisés pour la méthode itérative du point fixe précédente.

<6 pts>

Nota : La méthode de la bissection, bien qu'ultra-simple à comprendre et à implémenter est tout aussi (si ce n'est quelque-fois plus) efficace que la méthode du point fixe ! Dans cette méthode les six itérations de cet algorithme aurait conduit aux intervalles 1 : $[0.50 : 0.75]$ 2 : $[0.50 : 0.625]$ 3 : $[0.5625 : 0.625]$ 4 : $[0.5625 : 0.59375]$ 4 : $[0.5625 : 0.578125]$ 5 : $[0.5625 : 0.5703125]$ 6 : $[0.56640625 : 0.5703125]$
i.e., une estimation $\approx 0.5683 \pm 0.002$.

IV. Système d'Équations (13 pts)

Soit la matrice A suivante, dans lequel α est un paramètre strictement positif

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

1. Faîte deux décompositions LU de cette matrice, une pour laquelle determinant (L)= 1 et l'autre assurant determinant (U)= 1 (sans pivotage dans les deux cas et indiquez le nom de la technique utilisée).
   <10 pts>
2. Calculer le déterminant de la matrice A.
   <3 pts>

Réponse

$\boxed{\textbf{1.}}$

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

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

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

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

Ce qui nous donne la décomposition LU suivante

$$A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & -(1/\alpha) & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 2 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{U} \qquad <\textbf{4 pts}>$$

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 si on respecte l'ordre de calcul des différents coefficients de la matrice tel que vu en cours.

$$A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 3 & -1 & 2 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{U} \qquad <\textbf{4 pts}>$$

$\boxed{\textbf{2.}}$

Dans les deux cas (ce qui nous permet de vérifier que l'on a vraissemblablement pas fait d'ereur

$$\det(\mathrm{A}) = \det(\mathrm{L}) \times \det(\mathrm{U}) = 2\alpha \quad <\textbf{3 pts}>$$

Nota : On vérifie et on s'assure que pour les deux décomposition, il est bien sur, comme il se doit, identique...