Dauphine | PSL
UNIVERSITÉ PARIS

Département MIDO

Méthodes numériques
année 2022/2023

Corrigé (succinct) de l'examen du 28 juin 2023

Exercice 1 (résolution numérique de l'équation de Kepler, 4 points). En astronomie, l'équation de Kepler

$$x - e\sin(x) = M,$$

lie l'anomalie excentrique x, servant à calculer la position d'un astre dont le mouvement suit une orbite képlérienne, à l'excentricité e et l'anomalie moyenne M. On s'intéresse dans cet exercice à la résolution numérique de l'équation lorsque l'orbite est elliptique et périodique, c'est-à-dire quand e et M sont donnés et appartiennent respectivement à $]0,1[$ et $[0,2\pi[$.

1. Montrer que l'équation de Kepler admet une unique solution ξ contenue entre $M-e$ et $M+e$.

Posons $f(x) = x - e\sin(x) - M$ de façon à ramener la résolution de l'équation de Kepler à la recherche des zéros de f. La fonction f est dérivable sur $\mathbb{R}$ en tant que combinaison linéaire d'une fonction polynomiale et d'une fonction trigonométrique, de dérivée valant $f'(x) = 1 - e\cos(x)$. L'excentricité e étant strictement comprise entre 0 et 1, la dérivée de f est strictement positive sur $\mathbb{R}$ et la fonction f est donc strictement croissante et continue, ce qui en fait une bijection de $\mathbb{R}$ dans $\mathbb{R}$. On en déduit alors que f admet un unique zéro ξ. En observant que $\xi = M + e\sin(\xi)$ et en utilisant $-1 \leq \sin(\xi) \leq 1$, on trouve enfin que $M - e \leq \xi \leq M + e$.

2. Montrer la méthode de point fixe basée sur la relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = e\,\sin(x^{(k)}) + M,$$

converge globalement vers ξ, c'est-à-dire quel que soit le choix de l'initialisation $x^{(0)}$ dans $\mathbb{R}$.

La fonction de point fixe $g(x) = e\,\sin(x^{(k)}) + M$ est une fonction de classe $\mathscr{C}^1$ sur $\mathbb{R}$, à valeurs dans $[M-e, M+e] \subset \mathbb{R}$. On a par ailleurs

$$\forall x \in \mathbb{R},\ \left|g'(x)\right| = e\,|\cos(x)| < 1,$$

puisque $0 < e < 1$, ce qui fait de g une appllication contractante sur $\mathbb{R}$. On sait alors par un résultat du cours que la méthode de point fixe converge pour toute initialisation $x^{(0)}$ vers l'unique point fixe de la fonction g, qui n'est autre que ξ.

3. Donner la relation de récurrence de la méthode de Newton–Raphson pour la résolution de l'équation de Kepler et proposer (en la justifiant) une valeur pertinente pour l'initialisation de cette méthode lorsque la valeur de l'excentricité e est proche de 0.

En utilisant la reformulation du problème utilisant la fonction f introduite dans la première question, on trouve la relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} = x^{(k)} + \frac{M + e\sin(x^{(k)}) - x^{(k)}}{1 - e\cos(x^{(k)})}.$$

La méthode de Newton–Raphson converge en général seulement localement, ce qui signifie que l'initialisation $x^{(0)}$ doit être choisie proche de ξ. Lorsque l'excentricité e est proche de 0, ξ est proche de l'anomalie moyenne M et on peut alors supposer que le choix $x^{(0)} = M$ est avisé.

Exercice 2 (règle du trapèze pour l'intégrale d'une fonction convexe). Soit $[a,b]$ un intervalle borné et non vide de $\mathbb{R}$ et f une fonction continue et convexe sur $[a,b]$. On souhaite calculer une valeur approchée de l'intégrale

$$I(f) = \int_a^b f(x)\,\mathrm{d}x.$$

1. Montrer [1] que l'on a

$$\forall x \in [a,b],\ f(x) \leq \Pi_1 f(x),$$

où $\Pi_1 f$ est le polynôme d'interpolation de Lagrange de la fonction f associé aux nœuds $x_0 = a$ et $x_1 = b$.

1. On rappelle que la courbe représentative d'une fonction convexe sur un intervalle réel se trouve au dessous de tout segment du plan joignant deux points de cette courbe.

En paramétrant l'intervalle $[a,b]$ par la variable t prenant ses valeurs dans l'intervalle $[0,1]$ et en utilisant le fait que la fonction f est convexe sur $[a,b]$, on a l'inégalité

$$\forall t \in [0,1],\ f((1-t)a+tb) \leq (1-t)f(a)+tf(b).$$

Par ailleurs, on a par définition

$$\forall x \in \mathbb{R},\ \Pi_1 f(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}.$$

En remarquant que, pour tout x dans l'intervalle $[a,b]$, on a

$$0 \leq \frac{x-a}{b-a} \leq 1 \text{ et } 1-\frac{x-a}{b-a} = \frac{x-b}{a-b},$$

il suffit de poser $t = \frac{x-b}{a-b}$ pour conclure.

2. En se servant de cette observation, expliquer pourquoi la règle du trapèze composée fournira toujours, c'est-à-dire quel que soit le nombre de sous-intervalles utilisés pour subdiviser $[a,b]$, une approximation par excès de la valeur de $I(f)$.

La règle du trapèze composée utilise une subdivision de l'intervalle d'intégration en sous-intervalles sur chacun desquels la règle du trapèze est employée, c'est-à-dire que l'intégrale de f sur chaque sous-intervalle est approchée par celle du polynôme d'interpolation de f associé aux deux bornes du sous-intervalle considéré. Pour une subdivision de $[a,b]$ en m sous-intervalles $[x_{j-1},x_j]$, $j=1,\dots,m$, de même longueur $\frac{b-a}{m}$, notons $\Pi_{j,1}f$ le polynôme d'interpolation de Lagrange de f associé aux nœuds x_{j-1} et x_j. On a, d'après la précédente question,

$$\forall x \in [x_{j-1},x_j],\ f(x) \leq \Pi_{j,1}f(x),\ j=1,\dots,m,$$

d'où

$$I(f) = \sum_{j=1}^{m}\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x \leq \sum_{j=1}^{m}\int_{x_{j-1}}^{x_j} \Pi_{j,1}f(x)\,\mathrm{d}x = I_{m,1}(f).$$

3. On choisit $f(x) = \frac{1}{x}$, $a=1$ et $b=2$. Obtenir une approximation de $\ln(2)$ en appliquant la règle du trapèze composée à deux sous-intervalles au calcul approché de $I(f)$.

Dans ce cas, on a

$$I(f) = \int_1^2 \frac{\mathrm{d}x}{x} = [\ln(x)]_1^2 = \ln(2)$$

et la formule de quadrature approche donc directement $\ln(2)$ par la valeur

$$I_{2,1}(f) = \frac{b-a}{2}\left(\frac{1}{2}f(a) + f\left(\frac{a+b}{2}\right) + \frac{1}{2}f(b)\right) = \frac{1}{2}\left(\frac{1}{2}+\frac{2}{3}+\frac{1}{4}\right) = \frac{17}{24}.$$

Soit m un entier naturel non nul. On rappelle que, si la fonction f est de classe $\mathscr{C}^2$, on rappelle que l'erreur de quadrature de la règle du trapèze composée à m sous-intervalles a pour expression

$$E_{m,1}(f) = -\frac{(b-a)^3}{12\,m^2}f''(\xi),\ \text{avec } \xi \in]a,b[.$$

4. Donner une estimation du nombre de sous-intervalles à utiliser pour obtenir une approximation de $\ln(2)$ de précision inférieure ou égale à 10^{-8}.

La fonction f est de classe $\mathscr{C}^2$ sur l'intervalle $[a,b]$ choisi, telle que $f''(x) = \frac{2}{x^3}$, la formule pour l'erreur de quadrature composée fournit donc dans ce cas l'estimation

$$\left|E_{m,1}\right| = \frac{1}{12m^2}\left|\frac{2}{\xi^3}\right| \leq \frac{1}{6m^2}.$$

On cherche donc le plus petit entier naturel m tel que $\frac{1}{6m^2} \leq 10^{-8}$, soit encore $m \geq \frac{10^4}{\sqrt{6}}$. Un calcul numérique (non demandé) montre que $m = 4083$.

Exercice 3 (une méthode itérative de résolution de système linéaire). Soit n un entier naturel non nul. Le but de cet exercice est d'étudier une méthode de résolution numérique d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec A une matrice réelle symétrique définie positive d'ordre n que l'on suppose pouvoir écrire

$$A = M - N = P - Q,$$

les matrices M et P étant inversibles. À partir d'une initialisation $\boldsymbol{x}^{(0)}$ arbitraire, on considère la méthode itérative définie par les relations de récurrence

$$\forall k \in \mathbb{N},\ M\,\boldsymbol{x}^{(k+\frac{1}{2})} = N\,\boldsymbol{x}^{(k)} + \boldsymbol{b},\ P\,\boldsymbol{x}^{(k+1)} = Q\,\boldsymbol{x}^{(k+\frac{1}{2})} + \boldsymbol{b}.$$

Dans la suite, on pose

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x},\ \boldsymbol{\varepsilon}^{(k)} = M^{-1}A\boldsymbol{e}^{(k)},\ \boldsymbol{e}^{(k+\frac{1}{2})} = \boldsymbol{x}^{(k+\frac{1}{2})} - \boldsymbol{x} \text{ et } \boldsymbol{\varepsilon}^{(k+\frac{1}{2})} = M^{-1}A\boldsymbol{e}^{(k+\frac{1}{2})},$$

et l'on désigne par $\|\cdot\|_A$ la norme de $M_{n,1}(\mathbb{R})$ associée au produit scalaire défini par la matrice A, *i.e.*

$$\forall \boldsymbol{u} \in M_{n,1}(\mathbb{R}),\ \|\boldsymbol{u}\|_A{}^2 = \langle A\boldsymbol{u}, \boldsymbol{u}\rangle,$$

avec $\langle\cdot,\cdot\rangle$ le produit scalaire canonique de $M_{n,1}(\mathbb{R})$.

1. Montrer que les matrices $M + N^\top$ et $P + Q^\top$ sont symétriques.

On a

$$(M+N^\top)^\top = M^\top + N = (A+N)^\top + N = A^\top + N^\top + N = A + N^\top + N = M - N + N + N^\top = M + N^\top,$$

un calcul similaire montrant que $P + Q^\top$ est symétrique.

2. Vérifier que, $\forall k \in \mathbb{N}$, $\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k+\frac{1}{2})}$.

La matrice M étant inversible, on a, par définition de la méthode,

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k+\frac{1}{2})} = \boldsymbol{x}^{(k+\frac{1}{2})} - \boldsymbol{x} = M^{-1}(N\boldsymbol{x}^{(k)} + \boldsymbol{b}) - M^{-1}(N\boldsymbol{x} + \boldsymbol{b}) = M^{-1}N(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = M^{-1}N\boldsymbol{e}^{(k)},$$

d'où

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k+\frac{1}{2})} = (I_n - M^{-1}N)\boldsymbol{e}^{(k)} = M^{-1}(M-N)\boldsymbol{e}^{(k)} = M^{-1}A\boldsymbol{e}^{(k)} = \boldsymbol{\varepsilon}^{(k)}.$$

3. Montrer que, $\forall k \in \mathbb{N}$, $\|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 - \|\boldsymbol{e}^{(k)}\|_A{}^2 = -\left\langle (M+N^\top)\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle$.

En utilisant le résultat de la question précédente, la bilinéarité et la symétrie du produit scalaire, il vient

$$\begin{aligned}
\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 - \|\boldsymbol{e}^{(k)}\|_A{}^2 &= \left\langle A\boldsymbol{e}^{(k+\frac{1}{2})}, \boldsymbol{e}^{(k+\frac{1}{2})}\right\rangle - \left\langle A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}\right\rangle \\
&= \left\langle A(\boldsymbol{e}^{(k)} - \boldsymbol{\varepsilon}^{(k)}), (\boldsymbol{e}^{(k)} - \boldsymbol{\varepsilon}^{(k)})\right\rangle - \left\langle A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}\right\rangle \\
&= \left\langle A\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle - 2\left\langle A\boldsymbol{e}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle \\
&= \left\langle A\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle - 2\left\langle M\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle \\
&= \left\langle (A-2M)\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle = -\left\langle (M+N)\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle.
\end{aligned}$$

On conclut en utilisant que

$$\forall \boldsymbol{u} \in M_{n,1}(\mathbb{R}),\ \langle N\,\boldsymbol{u}, \boldsymbol{u}\rangle = \langle \boldsymbol{u}, N\boldsymbol{u}\rangle = \left\langle N^\top\,\boldsymbol{u}, \boldsymbol{u}\right\rangle.$$

4. En déduire que l'on a aussi, $\forall k \in \mathbb{N}$, $\|\boldsymbol{e}^{(k+1)}\|_A{}^2 - \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 = -\left\langle (P+Q^\top)\,\boldsymbol{\varepsilon}^{(k+\frac{1}{2})}, \boldsymbol{\varepsilon}^{(k+\frac{1}{2})}\right\rangle$.

La question se traite comme la précédente, en remplaçant $\boldsymbol{e}^{(k+\frac{1}{2})}$ par $\boldsymbol{e}^{(k+1)}$, $\boldsymbol{e}^{(k)}$ par $\boldsymbol{e}^{(k+\frac{1}{2})}$, $\boldsymbol{\varepsilon}^{(k)}$ par $\boldsymbol{\varepsilon}^{(k+\frac{1}{2})}$, M par P et N par Q.
On suppose à partir de maintenant que les matrices symétriques $M + N^\top$ et $P + Q^\top$ sont définies positives.

5. Montrer que la suite $\left(\|\boldsymbol{e}^{(k)}\|_A{}^2\right)_{k\in\mathbb{N}}$ converge. On note ℓ sa limite.

En utilisant les deux précédentes questions, il vient

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\|_A{}^2 - \|\boldsymbol{e}^{(k)}\|_A{}^2 = \|\boldsymbol{e}^{(k+1)}\|_A{}^2 - \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 + \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 - \|\boldsymbol{e}^{(k)}\|_A{}^2 = -\left\langle (P+Q^\top)\,\boldsymbol{\varepsilon}^{(k+\frac{1}{2})}, \boldsymbol{\varepsilon}^{(k+\frac{1}{2})}\right\rangle - \left\langle (M+N^\top)\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)}\right\rangle,$$

le dernier membres de droite étant négatif puisque les matrices $M + N^\top$ et $P + Q^\top$ sont définies positives. La suite $\left(\|\boldsymbol{e}^{(k)}\|_A{}^2\right)_{k\in\mathbb{N}}$ est ainsi positive et décroissante, donc convergente.

6. En déduire que la suite $\left(\|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2\right)_{k\in\mathbb{N}}$ est convergente, de limite égale à ℓ.

Il découle de la question 3 et du fait que la matrice $M+N^\top$ est définie positive que

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2 \leq \|\boldsymbol{e}^{(k)}\|_A{}^2.$$

De la même façon, il découle de la question 4 et du fait que la matrice $P+Q^\top$ est définie positive que

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\|_A{}^2 \leq \|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2.$$

La convergence de la suite $\left(\|\boldsymbol{e}^{(k)}\|_A{}^2\right)_{k\in\mathbb{N}}$ et le théorème des gendarmes impliquent alors que la suite $\left(\|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2\right)_{k\in\mathbb{N}}$ est convergente, de même limite que la suite $\left(\|\boldsymbol{e}^{(k)}\|_A{}^2\right)_{k\in\mathbb{N}}$.

7. En déduire que la suite $\left(\boldsymbol{\varepsilon}^{(k)}\right)_{k\in\mathbb{N}}$ converge vers le vecteur nul.

La convergence des suites $\left(\|\boldsymbol{e}^{(k)}\|_A{}^2\right)_{k\in\mathbb{N}}$ et $\left(\|\boldsymbol{e}^{(k+\frac{1}{2})}\|_A{}^2\right)_{k\in\mathbb{N}}$ et l'égalité établie à la question 3 impliquent que

$$\lim_{k\to+\infty} \left\langle (M+N^\top)\,\boldsymbol{\varepsilon}^{(k)}, \boldsymbol{\varepsilon}^{(k)} \right\rangle = 0.$$

La matrice $M+N^\top$ étant symétrique définie positive, elle définit un produit scalaire et on en déduit alors, par continuité, que

$$\lim_{k\to+\infty} \boldsymbol{\varepsilon}^{(k)} = \mathbf{0}.$$

8. En conclure que la méthode est convergente.

On déduit de la question précédente que

$$\lim_{k\to+\infty} M^{-1}A\boldsymbol{e}^{(k)} = \mathbf{0},$$

ce qui permet de conclure que

$$\lim_{k\to+\infty} \boldsymbol{e}^{(k)} = \mathbf{0}$$

puisque la matrice $M^{-1}A$ est inversible.

Exercice 4 (mise en œuvre de la méthode de sur-relaxation successive).

1. Rappeler la relation de récurrence définissant la méthode de sur-relaxation successive pour la résolution d'un système linéaire de matrice A et de second membre $\boldsymbol{b}$.

On a

$$\forall k \in \mathbb{N},\ \forall i \in \{1,\dots,n\},\ x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\,x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\,x_j^{(k)}\right) + (1-\omega)\,x_i^{(k)},$$

avec ω un réel non nul.

2. Modifier le code Python de la fonction ci-dessous, qui met en œuvre la méthode de Gauss–Seidel, de manière à ce qu'elle mette en œuvre la méthode de sur-relaxation successive. On indiquera, en donnant des explications, quelles sont la (ou les) ligne(s) modifiée(s) ou supprimé(s) et la position relative de toute ligne éventuellement ajoutée en utilisant la numérotation des lignes de code.

```
1  def gaussseidel(A,b,x0,tol,itermax):
2      m,n=A.shape
3      if n!=m:
4          raise ValueError('La␣matrice␣doit␣être␣carrée.')
5      iter=0
6      x=x0.copy()
7      r=b-dot(A,x)
8      nr0=norm(r)
9      relnr=norm(r)/nr0
10     while (relnr>tol) & (iter<itermax):
11         iter=iter+1
12         for i in range(n):
13             r[i]=r[i]/A[i,i]
14             for j in range(i+1,n):
15                 r[j]=r[j]-A[j,i]*r[i]
16             x[i]=x[i]+r[i]
```

```
17          r=b-dot(A,x)
18          relnr=norm(r)/nr0
19      if (relnr>tol):
20          print('Nombre␣maximum␣d\'itérations␣atteint.')
21      return x,iter
```

Les modifications à faire sont les suivantes :

— On modifie la ligne 1 en `def sor(A,b,omega,x0,tol,itermax):` pour changer le nom de la fonction et ajouter le paramètre ω en entrée.

— Après la ligne 4, on ajoute un test sur le paramètre ω : `if omega==0.: raise ValueError('Le paramètre de relaxation doit être non nul.')`.

— On modifie la ligne 13 en `r[i]=omega*r[i]/A[i,i]` pour tenir compte de la modification de la relation de récurrence utilisant le résidu. On a en effet

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + (D-E)^{-1}\boldsymbol{r}^{(k)},$$

avec D et E les matrices issues de A, respectivement diagonale et triangulaire inférieure stricte, introduites en cours, pour la méthode de Gauss-Seidel contre

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \left(\frac{1}{\omega}D-E\right)^{-1}\boldsymbol{r}^{(k)},$$

pour la méthode SOR.