Corrigés de travaux dirigés

Méthodes de résolution des équations non linéaires

Version du 22 février 2023.

Exercice 1 (analyse asymptotique de la méthode Illinois).

1. a. Après une étape non modifiée, on a, par définition de la méthode, $y^{(k)} = f(x^{(k)})$ et $y^{(k-1)} = f(x^{(k-1)})$, d'où

$$x^{(k+1)} = \frac{x^{(k-1)}f(x^{(k)}) - x^{(k)}f(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})} = x^{(k-1)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k-1)}).$$

b. Pour tout réel x strictement compris entre $x^{(k-1)}$ et $x^{(k)}$, la fonction ϕ est continûment dérivable en tout réel t compris entre $x^{(k-1)}$ et $x^{(k)}$ en tant que somme de fonctions continûment dérivables et sa dérivée seconde ϕ'' est définie en tout réel t compris entre $x^{(k-1)}$ et $x^{(k)}$. De plus, on peut facilement vérifier que $\phi(x^{(k-1)}) = \phi(x^{(k)}) = \phi(x) = 0$. Par le théorème de Rolle, il existe donc deux points entre $x^{(k-1)}$ et $x^{(k)}$, plus précisément l'un entre $x^{(k-1)}$ et x et l'autre entre x et $x^{(k-1)}$, en lesquels la dérivée de ϕ s'annule. On peut alors appliquer de nouveau le théorème de Rolle, cette fois à la fonction ϕ', pour obtenir l'existence d'un point $\theta^{(k+1)}$ compris entre $x^{(k-1)}$ et $x^{(k)}$ en lequel

$$\phi''(\theta^{(k+1)}) = 0.$$

On a par ailleurs

$$\phi''(t) = f''(t) + 2\frac{f(x^{(k-1)}) - f(x)}{(x^{(k)} - \xi)(x^{(k-1)} - \xi)} - 2\frac{f(x^{(k)}) - f(x^{(k-1)})}{(x^{(k)} - x^{(k-1)})(x^{(k)} - x)},$$

d'où

$$0 = \phi''(\theta^{(k+1)}) = f''(\theta^{(k+1)}) + 2\frac{f(x^{(k-1)}) - f(x)}{(x^{(k)} - \xi)(x^{(k-1)} - \xi)} - 2\frac{f(x^{(k)}) - f(x^{(k-1)})}{(x^{(k)} - x^{(k-1)})(x^{(k)} - x)},$$

soit encore

$$f(x) = f(x^{(k-1)}) - \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x^{(k-1)} - x) + \frac{1}{2}f''(\theta^{(k+1)})(x^{(k)} - \xi)(x^{(k-1)} - \xi).$$

Puisqu'un tel réel $\theta^{(k+1)}$ existe pour chaque choix de x strictement compris entre $x^{(k-1)}$ et $x^{(k)}$, on peut poser $x = \xi$, ce qui conduit à avoir

$$0 = f(\xi) = f(x^{(k-1)}) - \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x^{(k-1)} - \xi) + \frac{1}{2}f''(\theta^{(k+1)})(x^{(k)} - \xi)(x^{(k-1)} - \xi).$$

En utilisant l'égalité satisfaite par $x^{(k+1)}$ trouvée dans la première question, il vient alors

$$-\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x^{(k-1)} - \xi + x^{(k+1)} - x^{(k-1)}) + \frac{1}{2}f''(\theta^{(k+1)})(x^{(k)} - \xi)(x^{(k-1)} - \xi) = 0,$$

d'où le résultat.

c. Il découle du théorème des accroissements finis appliqué à la fonction f entre $x^{(k-1)}$ et $x^{(k)}$ qu'il existe un réel $\eta^{(k+1)}$, strictement compris entre $x^{(k-1)}$ et $x^{(k)}$, tel que

$$f'(\eta^{(k+1)}) = \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}.$$

En utilisant cette égalité dans celle obtenue dans la question précédente, on trouve que

$$f'(\eta^{(k+1)})(x^{(k+1)} - \xi) = \frac{1}{2}f''(\theta^{(k+1)})(x^{(k)} - \xi)(x^{(k-1)} - \xi).$$

On conclut en utilisant que la fonction f' ne s'annule pas, ce qui permet de diviser par $f'(\eta^{(k+1)})$.

d. La fonction f étant supposée trois fois dérivable, on a, en utilisant la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ vers ξ et le fait que les réels $\theta^{(k+1)}$ et $\eta^{(k+1)}$ sont strictement compris entre $x^{(k-1)}$ et $x^{(k)}$,

$$\begin{aligned}\frac{f''(\theta^{(k+1)})}{f'(\eta^{(k+1)})} &= \frac{f''(\xi)+O(\theta^{(k+1)}-\xi)}{f'(\xi)+O(\eta^{(k+1)}-\xi)}\\ &= \frac{f''(\xi)}{f'(\xi)+O(\eta^{(k+1)}-\xi)} + O(x^{(k)}-\xi)+O(x^{(k-1)}-\xi)\\ &= \frac{f''(\xi)}{f'(\xi)} + \frac{O(\eta^{(k+1)}-\xi)}{f'(\xi)(f'(\xi)+O(\eta^{(k+1)}-\xi))} + O(x^{(k)}-\xi)+O(x^{(k-1)}-\xi)\\ &= \frac{f''(\xi)}{f'(\xi)} + O(x^{(k)}-\xi)+O(x^{(k-1)}-\xi).\end{aligned}$$

On en déduit alors que

$$\frac{1}{2}\frac{f''(\theta^{(k+1)})}{f'(\eta^{(k+1)})}(x^{(k)}-\xi)(x^{(k-1)}-\xi) = \frac{1}{2}\frac{f''(\xi)}{f'(\xi)} + o((x^{(k)}-\xi)(x^{(k-1)}-\xi)).$$

2. En supposant que le comportement de l'erreur de la méthode est celui du régime asymptotique, on peut dresser un tableau de signe de l'erreur et du type de l'étape correspondante. En utilisant que les deux premiers termes de la suite des approximations encadrent ξ, on trouve respectivement

0	1	2	3	4	5	6	7	8	9	…
−	+	−	−	+	−	−	+	−	−	…
	U	U	M	U	U	M	U	U	M	…

si $\dfrac{f''(\xi)}{f'(\xi)} > 0$, et

0	1	2	3	4	5	6	7	8	…
−	+	+	−	+	+	−	+	+	…
	U	M	U	U	M	U	U	M	…

si $\dfrac{f''(\xi)}{f'(\xi)} < 0$.

3. Sur un motif, on a asymptotiquement

$$x^{(k+2)}-\xi \approx \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}(x^{(k+1)}-\xi)(x^{(k)}-\xi),\ x^{(k+1)}-\xi \approx \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}(x^{(k)}-\xi)(x^{(k-1)}-\xi) \text{ et } x^{(k)}-\xi \approx -(x^{(k-1)}-\xi),$$

d'où

$$x^{(k+2)}-\xi \approx \left(\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right)^2 (x^{(k)}-\xi)^2(x^{(k-1)}-\xi) \approx \left(\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right)^2 (-(x^{(k-1)}-\xi))^2(x^{(k-1)}-\xi) = \left(\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right)^2 (x^{(k-1)}-\xi)^3.$$

4. D'après la question précédente, la méthode est d'ordre égal à trois sur un motif, ce dernier comportant trois étapes, ce qui nécessite trois évaluations de la fonction f. L'indice d'efficacité computationnelle de la méthode vaut donc $3^{\frac{1}{3}} \approx 1,44$.

Exercice 2 (étude de convergence de la méthode de point fixe).

1. Pour tout x de $[a,b]$, posons $f(x) = g(x) - x$. La fonction f est continue de $[a,b]$ dans $\mathbb{R}$. De plus, puisque $g(a) \geq a$ et $g(b) \leq b$, il vient que $f(a) \geq 0$ et $f(b) \leq 0$. Ainsi, par le théorème des valeurs intermédiaires, il existe un certain ξ dans $[a,b]$ tel que $f(\xi) = 0$, *i.e.* $g(\xi) = \xi$. Le réel ξ est donc un point fixe de g.
Remarque : le point fixe ξ n'est pas forcément unique. Par exemple, si $g(x) = x$, tout point de $[a,b]$ est un point fixe (dans ce cas, $f \equiv 0$).

2. a. La fonction g étant de classe $\mathscr{C}^1$ sur l'intervalle $[\xi - h, \xi + h]$, sa dérivée g' est continue sur $[\xi - h, \xi + h]$. Notamment, $g'(x)$ tend vers $g'(\xi)$ quand x tend vers ξ. D'après l'hypothèse $|g'(\xi)| < 1$, on a de plus que $1 - |g'(\xi)| > 0$. En revenant à la définition de la limite, on a alors que, pour tout $\varepsilon > 0$, il existe un réel δ_ε, strictement positif et nécessairement plus petit que h, tel que

$$x \in [\xi - \delta_\varepsilon, \xi + \delta_\varepsilon] \Rightarrow \left|g'(x) - g'(\xi)\right| \leq \varepsilon.$$

En prenant $\varepsilon = \frac{1}{2}(1 - |g'(\xi)|)$, on obtenu l'existence d'un réel δ stictement positif tel que

$$x \in [\xi - \delta, \xi + \delta] \Rightarrow \left|g'(x) - g'(\xi)\right| \leq \frac{1}{2}(1 - |g'(\xi)|).$$

b. En utilisant l'inégalité triangulaire et la question précédente, on a, pour tout x dans I_δ,

$$|g'(x)| = |g'(x) - g'(\xi) + g'(\xi)| \leq |g'(x) - g'(\xi)| + |g'(\xi)| \leq \frac{1}{2}(1 - |g'(\xi)|) + |g'(\xi)| \leq \frac{1}{2}(1 + |g'(\xi)|).$$

Le résultat annoncé est donc vérifié avec $L = \frac{1}{2}(1 + |g'(\xi)|)$, qui est bien une constante strictement plus petite que 1 puisque $|g'(\xi)| < 1$.

c. On suppose que $x^{(k)}$ appartient à l'intervalle I_δ, *i.e.* $|x^{(k)} - \xi| \leq \delta$. On a donc notamment que, pour tout réel t compris entre ξ et $x^{(k)}$, $|t - \xi| \leq \delta$, *i.e.* t appartient à I_δ, et donc $|g'(t)| \leq L$. Comme la fonction g' est continue sur I_δ, on peut appliquer le théorème fondamental de l'analyse entre ξ et $x^{(k)}$ pour obtenir que

$$|x^{(k+1)} - \xi| = |g(x^{(k)}) - g(\xi)| = \left| \int_\xi^{x^{(k)}} g'(t)\,\mathrm{d}t \right| \leq \left| \int_\xi^{x^{(k)}} |g'(t)|\,\mathrm{d}t \right| \leq L\,|x^{(k)} - \xi|$$

(on rappelle que $g(\xi) = \xi$ puisque ξ est un point fixe). Comme $L < 1$, on a $|x^{(k+1)} - \xi| \leq |\xi - x^{(k)}| \leq \delta$ et $x^{(k+1)}$ appartient à I_δ. Si on l'on suppose que l'initialisation $x^{(0)}$ appartient à I_δ, on remarque que l'argument précédent utilisé avec $k = 0$ donne $|x^{(1)} - \xi| \leq L|x^{(0)} - \xi|$ et $x^{(1)}$ appartient à I_δ.

Ce que l'on a fait précédemment est exactement la propriété d'hérédité pour démontrer par récurrence la propriété :

$$\forall k \in \mathbb{N},\ x^{(k)} \in I_\delta,\ |x^{(k+1)} - \xi| \leq L|x^{(k)} - \xi|.$$

Ainsi, ayant déjà initialisé la récurrence, cette dernière propriété est alors vraie pour tout entier naturel k. Par une récurrence descendante (c'est-à-dire en remplaçant successivement à droite $|x^{(k)} - \xi|$ par $L|x^{(k-1)} - \xi|$, et ainsi de suite jusqu'à arriver à $|x^{(0)} - \xi|$, ce qui nécessite de le faire k fois), il est alors très facile d'en déduire que

$$\forall k \in \mathbb{N},\ |x^{(k)} - \xi| \leq L^k |x^{(0)} - \xi|.$$

d. Comme $0 < L < 1$, on a que L^k tend vers 0 quand k tend $+\infty$, et on en déduit, par le théorème des gendarmes, que $|x^{(k)} - \xi|$ tend vers 0 quand k tend vers $+\infty$, *i.e.* $x^{(k)}$ tend vers ξ quand k tend vers $+\infty$.

Remarques :

— Si l'on regarde bien la preuve, il n'est pas nécessaire de supposer g' continue sur un voisinage de ξ, mais seulement que g' est continue au point ξ (car dans ce cas, g' est intégrable sur un voisinage de ξ et le théorème fondamental de l'analyse s'applique).

— Puisque $|g'(\xi)| < 1$, il y a unicité du point fixe de g dans $[\xi - \delta, \xi + \delta]$.

3. Reprenons les étapes de l'argument précédent. La fonction g étant de classe $\mathscr{C}^1$ sur l'intervalle $[\xi - h, \xi + h]$, sa dérivée g' est continue sur $[\xi - h, \xi + h]$. Notamment, $g'(x)$ tend vers $g'(\xi)$ quand x tend vers ξ. L'hypothèse $|g'(\xi)| > 1$ implique que $|g'(\xi)| - 1 > 0$ et, par définition de la limite, pour tout réel ε strictement positif, il existe un réel δ strictement positif, tel que

$$x \in [\xi - \delta, \xi + \delta] \Rightarrow \left| g'(x) - g'(\xi) \right| \leq \varepsilon.$$

En choisissant $\varepsilon = \frac{|g'(\xi)| - 1}{2}$, on obtient l'existence d'un réel $\delta > 0$ tel que

$$x \in [\xi - \delta, \xi + \delta] \Rightarrow \left| g'(x) - g'(\xi) \right| \leq \frac{|g'(\xi)| - 1}{2}.$$

En utilisant la seconde inégalité triangulaire, on a alors, pour tout x appartenant à I_δ,

$$\left| g'(x) \right| = \left| g'(x) - g'(\xi) + g'(\xi) \right| \geq \left| g'(\xi) \right| - \left| g'(x) - g'(\xi) \right| \geq \left| g'(\xi) \right| - \frac{|g'(\xi)| - 1}{2} \geq \frac{|g'(\xi)| + 1}{2},$$

dont on déduit que $|g'(x)| \geq L$, avec $L = \frac{|g'(\xi)| + 1}{2} > 1$, pour tout élément x de I_δ.

Considérons une initialisation $x^{(0)}$ appartenant à $I_\delta \backslash \{\xi\}$ et raisonnons par l'absurde. Supposons que pour tout entier naturel k, l'itéré $x^{(k)}$ appartient à I_δ, *i.e.* $|x^{(k)} - \xi| \leq \delta$. On a donc notamment que, pour tout t compris entre ξ et $x^{(k)}$], $|t - \xi| \leq \delta$, c'est-à-dire que t appartient à I_δ et donc $|g'(t)| \geq L$. De plus, quitte à réduire un peu δ, on peut supposer que $g'(t)$ est du même signe que $g'(\xi)$ sur I_δ. La fonction g' étant continue sur I_δ, on peut appliquer le théorème fondamental de l'analyse pour obtenir que

$$|x^{(k+1)} - \xi| = |g(x^{(k)}) - g(\xi)| = \left| \int_\xi^{x^{(k)}} g'(t)\,\mathrm{d}t \right| = \left| \int_\xi^{x^{(k)}} |g'(t)|\,\mathrm{d}t \right| \geq L|x^{(k)} - \xi|$$

(on rappelle que ξ est un point fixe de g et que $g'(t)$ est du même signe que $g'(\xi)$, de telle sorte que la troisième égalité est vraie).

Compte tenu de l'hypothèse faite sur les itérés, une récurrence descendante permet de déduire que

$$\forall k \in \mathbb{N},\ |x^{(k)} - \xi| \geq L^k |x^{(0)} - \xi|.$$

Puisque $x^{(0)} \neq \xi$, on a $|x^{(0)} - \xi| \neq 0$ et, comme $L > 1$, on a L^k tend vers $+\infty$ lorsque k tend vers $+\infty$. Notamment, il existe un entier naturel non nul k_0 pour lequel $|x^{(k_0)} - \xi| \geq 2\,\delta$, ce qui signifie encore que $x^{(k_0)}$ n'appartient pas à I_δ, d'où une contradiction.

Ainsi, il existe un rang k_0 pour lequel la suite $(x^{(k)})_{k\in\mathbb{N}}$ sort de I_δ. On a alors les possibilités suivantes.

— Pour tout entier k plus grand que k_0, $x^{(k)}$ n'appartient pas à I_δ. Dans ce cas, la suite ne peut converger vers ξ.

— Soit il existe un entier k_0', strictement plus grand que k_0, tel que $x^{(k_0')}$ appartient à I_δ.

- Si $x^{(k_0')} \neq \xi$, on peut alors faire exactement le même raisonnement que précédemment à partir du rang k_0' et en déduire l'existence d'un entier $k_1 > k_0'$ tel que $x^{(k_1)}$ n'appartient pas à I_δ. On est ramené à la situation initiale.
- Si en revanche $x^{(k_0')} = \xi$, la suite devient stationnaire et on a donc convergence.

4. La fonction g_1 est continue et dérivable sur $]-1,+\infty[$, de dérivée valant $g_1'(x) = \frac{1}{1+x}$. La fonction $x \mapsto g_1(x) - x$ est donc continue, strictement croissante sur $]-1,0[$ et strictement décroissante sur $]0,+\infty[$. De plus, on a $\lim_{x\to -1,\ x>-1} g_1(x) - x = -\infty$, $g_1(0) = 0,2$ et $g_1(1) - 1 = \ln(2) + 0,2 - 1 \approx -0,10685 < 0$, la fonction admet donc exactement deux points fixes : ξ_- appartenant à l'intervalle $]-1,0[$, qui est répulsif, et ξ_+ appartenant à l'intervalle $]0,1[$, qui est attractif.

   On déduit alors de la deuxième question que, pour toute initialisation $x^{(0)}$ choisie dans $]0,+\infty[$, la méthode est convergente vers le point fixe attractif ξ_+.

   Pour tout initialisation $x^{(0)}$ dans l'intervalle $]\xi_-,0]$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ est strictement croissante et converge vers le point fixe attractif ξ_+.

   Pour $x^{(0)} = \xi_-$, la suite est stationnaire.

   Enfin, pour toute initialisation choisie dans l'intervalle $]-1,\xi_-[$, la stricte décroissance de la suite conduit à $x^{(k)} \leq -1$ pour un certain entier k supérieur à 1 et la méthode n'est alors plus définie.

   La méthode est donc convergente pour toute initialisation choisie dans l'intervalle $[\xi_-,+\infty[$.

   Pour la fonction g_2, on remarque tout d'abord que $g_2(x) = x \Leftrightarrow x^2 + c - 2x = 0$. Les points fixes de g_2 sont donc les racines de ce polynôme du second degré. Celles-ci sont réelles et distinctes (car c est contenu dans $[0,1[$) et valent

$$\xi_- = 1 - \sqrt{1-c} \text{ et } \xi_+ = 1 + \sqrt{1-c}.$$

   Puisque $g_2'(x) = x$, on en déduit que ξ_- est un point fixe attractif et ξ_+ un point fixe répulsif. On a par ailleurs

$$\left|g_2'(x)\right| < 1 \Leftrightarrow x \in]-1,1[.$$

   On déduit de la deuxième question que la méthode de point fixe est convergente (vers le point fixe attractif ξ_-) si l'initialisation $x^{(0)}$ appartient à l'intervalle $]-1,1[$.

   Pour $x^{(0)}$ appartenant à $]-\xi_+,-1] \cup [1,\xi_+[$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ est positive et strictement décroissante au plus tard à partir du rang un, donc convergente vers le point fixe ξ_-.

   Enfin, pour $\left|x^{(0)}\right| = \xi_+$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ est stationnaire et égale à ξ_+ à partir du rang un au plus tard, donc convergente vers ξ_+.

   Ainsi, la méthode est convergente pour toute initialisation appartenant à l'intervalle $[-\xi_+,\xi_+]$. La suite construite est divergente (vers $+\infty$) pour tout autre choix d'initialisation.

   La fonction g_3 est continue et dérivable sur $\mathbb{R}_+^*$, sa dérivée vaut $g_3'(x) = -\frac{1}{x}$. La fonction $x \mapsto g_3(x) - x$ est donc strictement décroissante sur $\mathbb{R}_+^*$, tend vers $+\infty$ quand x tend vers 0 par valeurs positives, et telle que que $g_3(1) - 1 = -1$. On en déduit que la fonction g_3 possède un unique point fixe ξ appartenant à l'intervalle $]0,1[$, qui est répulsif. Dans ce cas, quelle que soit l'initialisation de la méthode dans $]0,1[\backslash\{\xi\}$, on peut montrer que la suite des itérés finit par sortir de l'intervalle de définition du logarithme, mettant fin aux itérations sans convergence vers ξ.

Exercice 3.

1. Soit ξ un zero de la fonction f, c'est-à-dire tel que $\xi^3 - 2 = 0$, soit encore $\frac{1}{\xi^2} = \frac{\xi}{2}$. On a alors :

$$\begin{aligned} g(\xi) &= \left(1 - \frac{\omega}{3}\right)\xi + (1-\omega)\,\xi^3 + \frac{2\omega}{3\,\xi^2} + 2(\omega - 1) \\ &= \left(1 - \frac{\omega}{3}\right)\xi + 2(1-\omega) + \frac{\omega}{3}\,\xi + 2(\omega - 1) = \xi, \end{aligned}$$

   et ξ est un point fixe de g pour toute valeur du paramètre ω.

2. Soit $(x^{(k)})_{k\in\mathbb{N}}$ une suite convergeant vers ξ, avec $x^{(0)}$ un réel donné et, pour tout entier naturel k, $x^{(k+1)} = g(x^{(k)})$. On sait que la méthode est au moins ordre deux si

$$\lim_{k\to\infty} \left| \frac{x^{(k+1)} - \xi}{x^{(k)} - \xi} \right| = 0.$$

On a

$$x^{(k+1)}-\xi=\left(1-\frac{\omega}{3}\right)(x^{(k)}-\xi)+(1-\omega)\left((x^{(k)})^3-\xi^3\right)+\frac{2\omega}{3}\left(\frac{1}{(x^{(k)})^2}-\frac{1}{\xi^2}\right)$$
$$=(x^{(k)}-\xi)\left[1-\frac{\omega}{3}+(1-\omega)\left(x^{(k)^2}+\xi^2+x^{(k)}\xi\right)-\frac{2\omega}{3}\frac{x^{(k)}+\xi}{x^{(k)^2}\xi^2},\right]$$

d'où, en utilisant que $\lim_{k\to+\infty} x^{(k)}=\xi$,

$$\lim_{k\to\infty}\left|\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi}\right|=\left|1-\frac{\omega}{3}+(1-\omega)\left(\xi^2+\xi^2+\xi\,\xi\right)-\frac{2\omega}{3}\frac{\xi+\xi}{\xi^2\xi^2}\right|=\left|(1-\omega)(1-3\xi^2)\right|,$$

qui est nulle si et seulement si $\omega=1$.
On notera qu'on peut obtenir directement ce résultat en utilisant qu'une méthode de point fixe, de fonction g et de point fixe ξ, est d'ordre $p+1$, avec p un entier naturel supérieur ou égal à 1, si $g^{(i)}(\xi)=0$, $i=1,\dots,p$, et $g^{(p+1)}(\xi)\neq 0$. Pour que la méthode soit au moins d'ordre deux, il suffit donc que la dérivée première de g soit nulle au point ξ. On a ici

$$g'(x)=\left(1-\frac{\omega}{3}\right)+3(1-\omega)\,x^2-\frac{4\omega}{3\,x^3},$$

d'où

$$g'(\xi)=\left(1-\frac{\omega}{3}\right)+3(1-\omega)\,\xi^2-\frac{2\omega}{3}=(1-\omega)(1-3\xi^2).$$

3. La méthode est d'ordre supérieur à deux si la dérivée seconde de g s'annule au point ξ, en plus de la dérivée première. On a $g''(x)=6(1-\omega)x+\frac{12\omega}{3x^4}$ et donc, si $\omega=1$, il vient que $g''(\xi)=\frac{12\omega}{3\xi^4}\neq 0$. On en déduit qu'il n'existe pas de valeur de ω pour laquelle la méthode est d'ordre supérieur à deux.

Exercice 4 (exemple de divergence de la méthode de Newton–Raphson). On considère la fonction $f(x)=\arctan(x)$, qui a pour zéro $\xi=0$.

1. La dérivée de la fonction arctan est la fonction $x\mapsto\frac{1}{1+x^2}$ et la relation de récurrence de la méthode de Newton–Raphson s'écrit alors
$$\forall k\in\mathbb{N},\ x^{(k+1)}=x^{(k)}-(1+(x^{(k)})^2)\arctan(x^{(k)}).$$
La fonction de point fixe g ainsi définie est
$$g(x)=x-(1+x^2)\arctan(x).$$

2. En utilisant la seconde inégalité triangulaire et le fait que la fonction arctan est impaire, on a
$$|g(x)|=\left|x-(1+x^2)\arctan(x)\right|\geq\left||x|-\left|(1+x^2)\arctan(x)\right|\right|=\left||x|-(1+x^2)\arctan(|x|)\right|.$$
Il découle alors de l'hypothèse que
$$|g(x)|\leq(1+x^2)\arctan(|x|)-|x|>2\,|x|-|x|=|x|\,.$$

3. Posons $\phi(x)=(1+x^2)\arctan(x)-2\,x$ et étudions la fonction ϕ sur $[0,+\infty[$. Celle-ci est de classe $\mathscr{C}^\infty$ et $\phi'(x)=2x\arctan(x)-1$. Notons α le réel strictement positif tel que $\alpha\arctan(\alpha)=\frac{1}{2}$. On a le tableau de variations suivant

x	0		α		$+\infty$
$\phi'(x)$		$-$	0	$+$	
$\phi(x)$	0	↘	$\phi(\alpha)$	↗	$+\infty$

On voit que $\phi(\alpha)<0$ et, la fonction ϕ étant strictement croissante sur $]\alpha,+\infty[$, il existe un unique réel β ($\beta\approx 1,3917452$) strictement plus grand que α pour lequel
$$x>\beta\Rightarrow\phi(x)>\phi(\beta)=0.$$
On en déduit que
$$|x|>\beta\Rightarrow\arctan(|x|)>\frac{2\,|x|}{1+x^2}$$
et, en utilisant le résultat de la question précédente, on obtient que
$$|x|>\beta\Rightarrow|g(x)|>|x|\,,$$
d'où $|g(x)|>\beta$ et donc
$$\arctan(|g(x)|)>\frac{2\,|g(x)|}{1+g(x)^2}.$$

4. Si
$$\arctan(|x^{(0)}|) > \frac{2|x^{(0)}|}{1+(x^{(0)})^2},$$
alors on déduit de la précédente question que
$$|x^{(1)}| = |g(x^{(0)})| > |x^{(0)}|.$$
On montre alors par récurrence que la suite $(|x^{(k)}|)_{k\in\mathbb{N}}$ est strictement croissante. Supposons qu'elle reste bornée. Elle est alors convergente et sa limite ℓ vérifie $\ell > \beta$ et $|g(\ell)| = \ell$, ce qui est absurde. La suite est donc divergente et l'on a
$$\lim_{k\to+\infty} |x^{(k)}| = +\infty.$$

Exercice 5 (convergence globale de la méthode de Newton–Raphson).

1. Les hypothèses de changement de signe de la fonction continue f et de signe constant de sa dérivée f', également continue, sur $[a,b]$ impliquent qu'il existe un unique zéro ξ appartenant à $[a,b]$.
2. Si $f(x^{(0)})f''(x^{(0)}) = 0$, cela signifie que $f(x^{(0)}) = 0$ et donc, d'après la précédente question, on a $x^{(0)} = \xi$. La méthode est alors (trivialement, la suite construite étant constante) convergente vers ξ.
3. a. Si la fonction f'' est strictement positive sur $[a,b]$, alors la condition sur l'initialisation impose que $f(x^{(0)}) > 0$. Si de plus, $\forall x \in [a,b]$, $f'(x) > 0$, on a alors, $\forall x \in [a,\xi[$, $f(x) < 0$ et, $\forall x \in]\xi,b]$, $f(x) > 0$, d'où $x^{(0)} \in]\xi,b]$. On vérifie alors que
$$\forall x \in]\xi,b],\ g'(x) = \frac{f(x)f''(x)}{(f'(x))^2} > 0,$$
la fonction g étant la fonction de point fixe définissant la méthode. Il en découle que g est strictement croissante sur $]\xi,b]$. On en déduit d'une part que
$$\xi = g(\xi) < g(x^{(0)}) = x^{(1)},$$
et d'autre part que
$$x^{(1)} = g(x^{(0)}) = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} < x^{(0)},$$
d'où $\xi \le x^{(1)} < x^{(0)}$. En raisonnant par récurrence, on obtient que la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode est strictement décroissante et minorée par ξ. Elle est donc convergente et a pour limite l'unique point fixe de la fonction g, ξ. Si, $\forall x \in [a,b]$, $f'(x) < 0$, un raisonnement identique conduit au fait que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est strictement croissante et majorée par ξ. De nouveau, cette suite est convergente et a pour limite ξ.
   b. Si la fonction f'' est strictement négative sur $[a,b]$ (et donc si $f(x^{(0)}) < 0$), il suffit de reprendre la preuve ci-dessus en remplaçant f par $-f$ pour établir la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$.
4. On déduit des questions précédentes que la méthode de Newton–Raphson est globalement convergente dans ce cas.

Exercice 6 (étude de convergence de la méthode de Newton–Raphson vers un zéro multiple).

1. Puisque $f'(\xi) = 0$, la fonction g n'est pas définie au point ξ. Le zéro ξ étant d'ordre m, on peut vérifier[1] que la fonction f peut s'écrire sous la forme
$$f(x) = (x-\xi)^m h(x),$$
où h est une fonction régulière telle que $h(\xi) \neq 0$. En dérivant cette égalité, on obtient
$$f'(x) = m(x-\xi)^{m-1}h(x) + (x-\xi)^m h'(x).$$
On a alors
$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{(x-\xi)^m h(x)}{m(x-\xi)^{m-1}h(x) + (x-\xi)^m h'(x)} = x - \frac{(x-\xi)h(x)}{m\,h(x) + (x-\xi)h'(x)}.$$
Utilisons la dernière égalité pour calculer la limite de g en ξ. Par continuité de h et de h', on a
$$\lim_{x\to\xi} h(x) = h(\xi) \neq 0,\ \lim_{x\to\xi}(x-\xi)h(x) = 0 \text{ et } \lim_{x\to\xi}(x-\xi)h'(x) = 0,$$
et par conséquent
$$\lim_{x\to\xi} g(x) = \xi.$$
On peut donc prolonger la fonction g au point ξ par continuité en posant $g(\xi) = \xi$. Ceci permet de calculer la limite du taux d'accroissement $\frac{g(x)-g(\xi)}{x-\xi}$. On a en effet
$$\lim_{x\to\xi} \frac{g(x)-g(\xi)}{x-\xi} = \lim_{x\to\xi}\left(1 - \frac{h(x)}{m\,h(x) + (x-\xi)h'(x)}\right) = 1 - \frac{h(\xi)}{m\,h(\xi)} = 1 - \frac{1}{m}.$$
On en déduit que le prolongement par continuité de la fonction g est dérivable au point ξ.

1. On a notamment que la fonction h est de classe $\mathscr{C}^1$, telle que $h(\xi) = \frac{f^{(m)}(\xi)}{m!}$ et $h'(\xi) = \frac{f^{(m+1)}(\xi)}{(m+1)!}$.

2. La fonction g est de classe $\mathscr{C}^1$ et l'on a, d'après la question précédente, $|g'(\xi)| < 1$. Par continuité de g', il existe alors un réel $K > 0$ tel que, pour tout x appartenant à $[\xi - K, \xi + K]$, $|g'(x)| < 1$. On peut ainsi, toujours par continuité, poser $M := \max_{x\in[\xi-K,\xi+K]} |g'(x)| < 1$.

   Considérons une initialisation $x^{(0)}$ appartenant à $[\xi - K, \xi + K]$ et montrons, en raisonnant par récurrence sur l'entier naturel k, que tout terme de la suite $(x^{(k)})_{k\in\mathbb{N}}$ appartient à $[\xi - K, \xi + K]$ et que $\left|x^{(k)} - \xi\right| \leq M^k \left|x^{(0)} - \xi\right|$. L'affirmation est évidemment vraie pour $k = 0$. Supposons qu'elle soit vraie pour un entier naturel k quelconque. Par le théorème des accroissements finis, il existe un réel $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ tel que

$$x^{(k+1)} = g(x^{(k)}) = g(\xi) + g'(\eta^{(k)})(x^{(k)} - \xi),$$

   ce qui implique que, comme $g(\xi) = \xi$,

$$x^{(k+1)} - \xi = g'(\eta^{(k)})(x^{(k)} - \xi).$$

   Le réel $\eta^{(k)}$ étant compris entre $x^{(k)}$ et ξ et puisque $x^{(k)}$ appartient $[\xi - K, \xi + K]$ par hypothèse, on a que $\eta^{(k)}$ appartient à $[\xi - K, \xi + K]$ et donc $\left|g'(\eta^{(k)})\right| \leq M$. Il en résulte que

$$|x^{(k+1)} - \xi| \leq M\,|x^{(k)} - \xi| \leq M^{k+1}\,|x^{(0)} - \xi|.$$

   Ceci implique en particulier que $x^{(k+1)}$ appartient à $[\xi - K, \xi + K]$, puisque $M < 1$.
   On peut ainsi établir la convergence de la méthode par le théorème des gendarmes :

$$0 \leq \lim_{k\to+\infty} |x^{(k)} - \xi| \leq \lim_{k\to+\infty} M^k |x^{(0)} - \xi| = |x^{(0)} - \xi| \lim_{k\to+\infty} M^k = 0.$$

   Pour montrer que la convergence est linéaire, on calcule

$$\lim_{k\to+\infty} \frac{x^{(k+1)} - \xi}{x^{(k)} - \xi} = \lim_{k\to+\infty} \frac{g(x^{(k)}) - g(\xi)}{x^{(k)} - \xi} = \lim_{k\to+\infty} g'(\eta^{(k)}) = g'(\xi) = 1 - \frac{1}{m} \neq 0,$$

   où l'on s'est servi du fait que $\eta^{(k)}$ tend vers ξ lorsque k tend vers $+\infty$, puisque $\eta^{(k)}$ est compris entre $x^{(k)}$ et ξ et que $x^{(k)}$ converge vers ξ lorsque k tend vers $+\infty$.

3. Montrons que l'ordre de convergence de la méthode modifiée est au moins quadratique, comme c'est le cas pour la méthode de Newton-Raphson lorsque le zéro est simple.
   Définissons la fonction $\tilde{g}$ par

$$\tilde{g}(x) = x - m\,\frac{f(x)}{f'(x)},$$

   de sorte que l'itération de la méthode modifiée s'écrive

$$\forall k \in \mathbb{N},\ x^{(k+1)} = \tilde{g}(x^{(k)}).$$

   En reprenant les étapes de la première question, on montre que

$$\lim_{x\to\xi} \tilde{g}(x) = \lim_{x\to\xi} x - \frac{m\,(x-\xi)h(x)}{m\,h(x) + (x-\xi)h'(x)} = \xi,$$

   et la fonction $\tilde{g}$ peut être prolongée par continuité en ξ en posant $\tilde{g}(\xi) = \xi$. On montre ensuite que

$$\lim_{x\to\xi} \frac{\tilde{g}(x) - \tilde{g}(\xi)}{x - \xi} = \lim_{x\to\xi} \left(1 - \frac{m\,h(x)}{m\,h(x) + (x-\xi)h'(x)}\right) = 1 - \frac{m\,h(\xi)}{m\,h(\xi)} = 1 - 1 = 0.$$

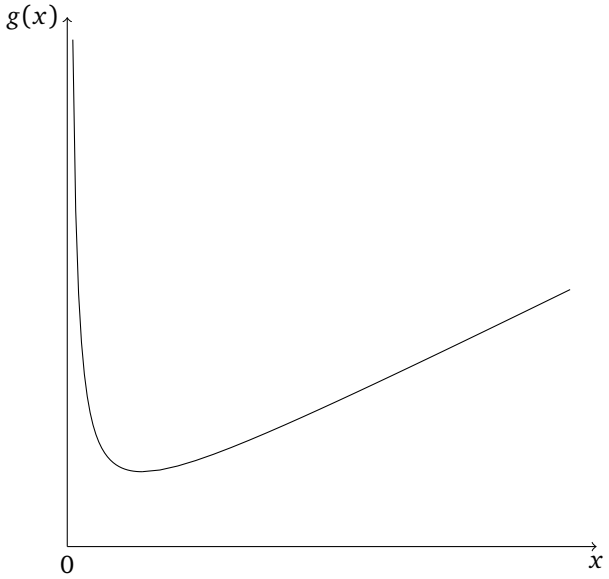
   La fonction $\tilde{g}$ est donc dérivable en ξ et la convergence de la méthode est au moins quadratique, puisque la valeur de la dérivée en ce point est nulle.

Exercice 7 (étude de la méthode de Héron pour le calcul de $\sqrt{2}$).

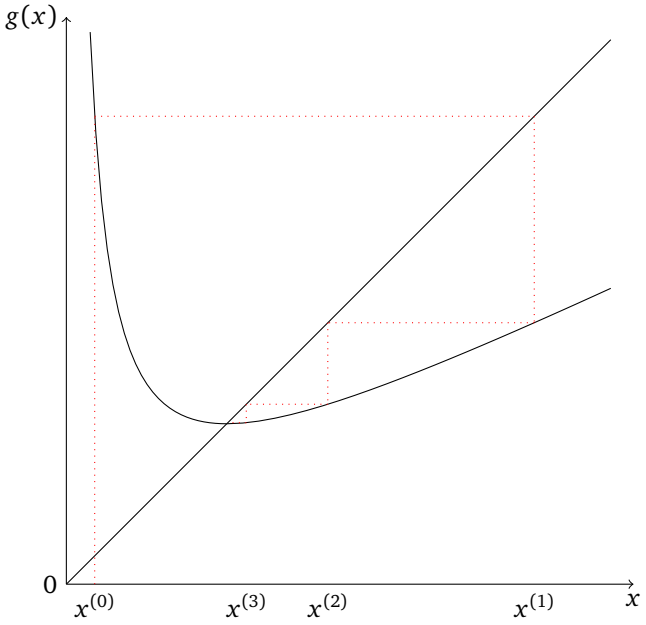
1. La fonction g est dérivable sur $\mathbb{R}_+^*$ et sa dérivée vaut $g'(x) = \frac{1}{2} - \frac{1}{x^2}$. On a alors le tableau de variations suivant :

x	0		$\sqrt{2}$		$+\infty$
$g'(x)$	‖	$-$	0	$+$	
$g(x)$	$+\infty$	↘	$\sqrt{2}$	↗	$+\infty$

et le graphe



2. Pour une représentation graphique plus claire, on choisit $x^{(0)} = \frac{1}{4}$.



3. On vérifie tout d'abord que $g([1,2]) \subset [1,2]$. On a $g(1) = g(2) = \frac{3}{2}$ et, d'après l'étude effectuée dans la première question, g est décroissante sur $[1,\sqrt{2}]$, puis croissante sur $[\sqrt{2},2]$, avec un minimum valant $\sqrt{2}$.
Montrons ensuite que la fonction g est lipschitzienne. On a

$$\forall (x,y) \in [1,2]^2,\ |g(x)-g(y)| = \frac{1}{2}\left|x+\frac{2}{x}-y-\frac{2}{y}\right| = \frac{1}{2}\left|x-y+2\,\frac{y-x}{xy}\right|.$$

Pour $1 \leq x, y \leq 2$, on a $1 \leq xy \leq 4$, d'où $-1 \leq 1-\frac{2}{xy} \leq \frac{1}{2}$. Il vient donc

$$\forall (x,y) \in [1,2]^2,\ |g(x)-g(y)| = \frac{1}{2}\left|1-\frac{2}{xy}\right||x-y| \leq \frac{1}{2}|x-y|,$$

et la fonction g est lipschitzienne, de constante de Lipschitz valant $\frac{1}{2}$, sur $[1,2]$, elle est donc contractante sur cet intervalle.
L'initialisation $x^{(0)}$ étant égale à 1, elle appartient à l'intervalle $[1,2]$ et, par un théorème du cours, la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers l'unique point fixe ξ de g dans $[1,2]$. On a alors

$$g(\xi) = \xi \Leftrightarrow \frac{1}{2}\left(\xi+\frac{2}{\xi}\right) = \xi \Leftrightarrow \xi^2 = 2,$$

d'où $\xi = \sqrt{2}$.

4. D'après la question précédente, on a convergence pour toute initialisation $x^{(0)}$ choisie dans l'intervalle $[1,2]$, mais on peut étendre ce résultat à un plus grand intervalle.

   En effet, en reprenant le raisonnement précédent, on voit tout d'abord que $g([1,+\infty[) \subset [1,+\infty[$ (en utilisant le fait que $xy \geq 1$). On montre ensuite que la fonction g est contractante sur cet intervalle. D'après le théorème des accroissements finis, on a

$$\forall (x,y) \in [1,+\infty[^2,\ \exists z \in [1,+\infty[,\ \frac{g(y)-g(x)}{y-x} = g'(z),$$

   et donc

$$\forall (x,y) \in [1,+\infty[^2,\ |g(y)-g(x)| \leq \max_{t\in[1,+\infty[} \left|g'(t)\right| \, |y-x|.$$

   On a par ailleurs $g''(x) = \frac{2}{x^3}$, d'où le tableau de variations

x	1		$+\infty$
$g''(x)$	$+$		
$g'(x)$	$-\frac{1}{2}$	↗	$\frac{1}{2}$

   et la fonction g est donc lipschitzienne, de constante de Lipshitz valant $\frac{1}{2}$, sur l'intervalle $[1,+\infty[$.

   Il reste enfin à considérer l'intervalle $]0,1[$. Il est clair que la dérivée de g n'y est pas bornée, mais on peut néanmoins voir ce qu'il se passe pour une initialisation $x^{(0)}$ choisie dans cet intervalle. On a

$$x^{(1)} = \frac{1}{2}\left(x^{(0)} + \frac{2}{x^{(0)}}\right) \geq \frac{3}{2},$$

   et l'itéré $x^{(1)}$ appartient donc à l'intervalle $[1,+\infty[$. On se trouve donc ramené, après une itération, sur un intervalle pour lequel on a prouvé que la méthode était convergente.

   La méthode est donc convergente pour toute initialisation choisie dand l'intervalle $]0,+\infty[$: la convergence est *globale*.

5. On a $g'(\xi) = \frac{1}{2}\left(1 - \frac{2}{(\sqrt{2})^2}\right) = 0$ et $g''(\xi) = \frac{2}{(\sqrt{2})^3} = \frac{1}{\sqrt{2}} \neq 0$. La convergence de la méthode est donc quadratique.

   On remarque que cette méthode est un cas particulier de la méthode de Newton–Raphson appliquée à la résolution de $f(x) = 0$ avec $f(x) = x^2 - 2$, fonction pour laquelle $\sqrt{2}$ est un zéro simple.

   On notera que la méthode peut plus généralement être utilisée pour le calcul de la racine carrée de tout réel a strictement positif en utilisant la fonction $f(x) = x^2 - a$, c'est-à-dire $g(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$.

Dauphine | PSL
UNIVERSITÉ PARIS
Département MIDO

Méthodes numériques
année 2022/2023

Corrigés de travaux dirigés

Interpolation polynomiale

Version du 22 février 2023.

Dans toute cette feuille, on identifie polynôme et fonction polynomiale associée.

Exercice 1. Soit f une fonction réelle et continue sur l'intervalle $[-1,1]$. Par définition, on a

$$\Pi_1 f(x) = f(-1)\, l_0(x) + f(1)\, l_1(x),$$

avec $l_0(x) = -\frac{x-1}{2}$ et $l_1(x) = \frac{x+1}{2}$.

Posons à présent $E(x) = f(x) - \Pi_1 f(x)$ pour tout x appartenant à $]-1,1[$ (les inégalités sont trivialement vérifiées aux points $x = \pm 1$) et introduisons

$$G(t) = E(t) - \omega(t)\frac{E(x)}{\omega(x)},$$

avec $\omega(t) = (t+1)(t-1) = t^2 - 1$. Puisque f que f est de classe $\mathscr{C}^2$ sur $[-1,1]$, on a que G est de classe $\mathscr{C}^2$ sur $[-1,1]$ et possède au moins trois zéros distincts dans $[-1,1]$. En effet, si on pose $x_0 = -1$ et $x_1 = 1$, on a que

$$\forall i \in \{0,1\},\ G(x_i) = \underbrace{E(x_i)}_{=0} + \underbrace{\omega(x_i)}_{=0}\frac{E(x)}{\omega(x)} = 0,$$

et de plus $G(x) = E(x) - \omega(x)\frac{E(x)}{\omega(x)} = 0$. Par le théorème de Rolle, la fonction G' admet au moins deux zéros distincts dans l'intervalle $[-1,1]$ et, par suite, la fonction G'' admet au moins un zéro dans ce même intervalle, que l'on va à noter c. On a par ailleurs

$$G''(t) = E''(t) - \omega''(t)\frac{E(x)}{\omega(x)} = f''(t) - 2\frac{E(x)}{\omega(x)},$$

d'où $E(x) = \frac{f''(c)}{2}\omega(x)$. On a ainsi obtenu que

$$|E(x)| = \frac{|f''(c)|}{2}\,|\omega(x)| \leq \frac{M_2}{2}(x^2-1) \leq \frac{M_2}{2},$$

et l'on a égalité pour toute fonction telle que f'' est constante sur $[-1,1]$.

Exercice 2. Par définition, on a

$$\Pi_1 f(x) = f(0)\, l_0(x) + f(a)\, l_1(x) = a^3\,\frac{x}{a} = a^2\, x.$$

En procédant comme dans l'exercice précédent, on peut montrer que, pour tout x appartenant à $]0,a[$, il existe un réel c appartenant à $[0,a]$ tel que l'on a

$$f(x) - \Pi_1 f(x) = \frac{f''(c)}{2}x(x-a),$$

où c annule la fonction

$$G''(t) = 6t - 2\frac{x^3 - a^2 x}{x(x-a)},$$

c'est-à-dire que $c = \frac{2}{6}\,\frac{x^3-a^2x}{x(x-a)} = \frac{x+a}{3}$.

Si on considère la fonction $f(x) = (2x-a)^4$, on aura cette fois que $f'(x) = 8(2x-a)^3$ et $f''(x) = 48(2x-a)^2$ et alors

$$48(2c-a)^2 = 2\frac{(2x-a)^4 - a^4}{x(x-a)} \Leftrightarrow c = \frac{1}{2}\left(a \pm \frac{1}{2}\sqrt{\frac{(2x-a)^4 - a^4}{6x(x-a)}}\right).$$

Exercice 3. Le polynôme de Lagrange Π_{n+1} de degré $n+1$ associé aux couples $\{(x_i, y_i)\}_{0\leq i\leq n+1}$ est l'unique polynôme de degré au plus $n+1$ tel que $\Pi_{n+1}(x_i) = y_i$ pour tout entier i dans $\{0,\ldots,n+1\}$.

Montrons que $p = \Pi_{n+1}$. Comme $\Pi_{0,\ldots,n}$ et $\Pi_{1,\ldots,n+1}$ sont des polynômes de degré au plus n, on a bien que p est un polynôme de degré au plus $n+1$ et il suffit donc de vérifier que $p(x_i) = y_i$ pour tout entier i de $\{0,\ldots,n+1\}$.

On a tout d'abord

$$p(x_0) = -\frac{(x_0 - x_{n+1})\Pi_{0,\dots,n}(x_0)}{(x_{n+1} - x_0)} = \Pi_{0,\dots,n}(x_0) = y_0$$

puisque $\Pi_{0,\dots,n}$ est le polynôme de Lagrange associé aux couples $\{(x_i, y_i)\}_{0\le i\le n}$. De la même manière, on trouve que

$$p(x_{n+1}) = \Pi_{1,\dots,n+1}(x_{n+1}) = y_{n+1}.$$

Enfin, pour $1 \le i \le n$, on a

$$p(x_i) = \frac{(x_i - x_0)\Pi_{1,\dots,n+1}(x_i) - (x_i - x_{n+1})\Pi_{0,\dots,n}(x_i)}{(x_{n+1} - x_0)} = \frac{(x_i - x_0)y_i - (x_i - x_{n+1})y_i}{(x_{n+1} - x_0)} = y_i.$$

Exercice 4 (forme de Newton du polynôme d'interpolation).

1. Le polynôme q_j est de degré inférieur ou égal à j par construction et tel que $q_j(x_i) = \pi_j(x_i) - \pi_{j-1}(x_i) = y_i - y_i = 0$ pour tout i appartenant à $\{0, \dots, j-1\}$. Il est par conséquent égal à $a_j \prod_{i=0}^{j-1}(x - x_i)$, la constante a_j étant obtenue en écrivant que

   $$y_j = \Pi_j(x_j) = \Pi_{j-1}(x_j) + a_j\, \omega_j(x_j),$$

   d'où

   $$a_j = \frac{y_j - \Pi_{j-1}(x_j)}{\omega_j(x_j)}.$$

2. Montrons la formule par récurrence sur l'entier j. Pour $j = 0$, on a $[x_0]\boldsymbol{y} = y_0$ et $\omega_0 \equiv 1$, d'où $[x_0]\boldsymbol{y}\, \omega_0 \equiv y_0 \equiv \Pi_0$. Fixons à présent j dans $\{1, \dots, n\}$ et supposons que la formule est vraie au rang $j-1$:

   $$\Pi_{j-1}(x) = \sum_{k=0}^{j-1} [x_0, \dots, x_k]\boldsymbol{y}\, \omega_k(x).$$

   En utilisant le résultat de la question précédente en conjonction avec cette hypothèse de récurrence, il vient alors

   $$\Pi_j(x) = \Pi_{j-1}(x) + [x_0, \dots, x_j]\boldsymbol{y}\, \omega_j(x) = \sum_{k=0}^{j-1} [x_0, \dots, x_k]\boldsymbol{y}\, \omega_k(x) + [x_0, \dots, x_j]\boldsymbol{y}\, \omega_j(x) = \sum_{k=0}^{j} [x_0, \dots, x_k]\boldsymbol{y}\, \omega_k(x).$$

3. Fixons l'entier j dans $\{0, \dots, n\}$. Par définition du polynôme de Newton ω_{j+1}, on a

   $$\begin{aligned}\sum_{i=0}^{j} \frac{\omega_{j+1}(x)}{(x - x_i)\,\omega'_{j+1}(x_i)}\, y_i &= \sum_{i=0}^{j} \frac{\prod_{k=0}^{j}(x - x_k)}{(x - x_i)\,\omega'_{j+1}(x_i)}\, y_i \\ &= \sum_{i=0}^{j} \frac{\prod_{\substack{k=0\\k\neq i}}^{j}(x - x_k)}{\omega'_{j+1}(x_i)}\, y_i.\end{aligned}$$

   Ainsi, pour $x = x_\ell$, $\ell \in \{0, \dots, j\}$, on trouve

   $$\begin{aligned}\sum_{i=0}^{j} \frac{\omega_{j+1}(x_\ell)}{(x_\ell - x_i)\,\omega'_{j+1}(x_i)}\, y_i &= \sum_{i=0}^{j} \frac{\prod_{\substack{k=0\\k\neq i}}^{j}(x_\ell - x_k)}{\omega'_{j+1}(x_i)}\, y_i \\ &= \frac{\prod_{\substack{k=0\\k\neq \ell}}^{j}(x_\ell - x_k)}{\omega'_{j+1}(x_\ell)}\, y_\ell.\end{aligned}$$

   Or, on a

   $$\forall x \in \mathbb{R},\ \omega'_{j+1}(x) = \sum_{i=0}^{j} \prod_{\substack{k=0\\k\neq i}}^{j} (x - x_k),$$

   d'où $\omega'_{j+1}(x_\ell) = \prod_{\substack{k=0\\k\neq \ell}}^{j}(x_\ell - x_k)$. On a alors vérifié que le polynôme $\sum_{i=0}^{j} \frac{\omega_{j+1}(x)}{(x - x_i)\,\omega'_{j+1}(x_i)}\, y_i$, de degré égal à j, prend la valeur y_ℓ en chaque nœud x_ℓ, pour tout entier ℓ appartenant à $\{0, \dots, j\}$. Par unicité du polynôme d'interpolation de Lagrange, on en déduit qu'il coïncide avec Π_j.
   Par suite, en remarquant (en se servant de la forme de Newton) que le coefficient du monôme de plus haut degré de Π_j est égal à $[x_0, \dots, x_j]\boldsymbol{y}$, on obtient par identification que

   $$[x_0, \dots, x_j]\boldsymbol{y} = \sum_{i=0}^{j} \frac{y_i}{\omega'_{j+1}(x_i)}.$$

4. Pour $j=1$, on a

$$[x_0,x_1]\boldsymbol{y}=\sum_{i=0}^{1}\frac{y_i}{\omega_2'(x_i)}=\frac{y_0}{\omega_2'(x_0)}+\frac{y_1}{\omega_2'(x_1)}=\frac{y_0}{x_0-x_1}+\frac{y_1}{x_1-x_0}=\frac{y_1-y_0}{x_1-x_0}=\frac{[x_1]\boldsymbol{y}-[x_0]\boldsymbol{y}}{x_1-x_0}.$$

Fixons l'entier j dans $\{1,\dots,n\}$ et supposons la formule vérifiée au rang $j-1$. On a

$$\begin{aligned}
[x_0,\dots,x_j]\boldsymbol{y}&=\sum_{i=0}^{j}\frac{y_i}{\omega_{j+1}'(x_i)}\\
&=\sum_{i=0}^{j}\frac{y_i}{\prod_{\substack{k=0\\k\neq i}}^{j}(x_i-x_k)}\\
&=\frac{y_0}{\prod_{k=1}^{j}(x_0-x_k)}+\sum_{i=1}^{j-1}\frac{y_i}{\prod_{\substack{k=0\\k\neq i}}^{j}(x_i-x_k)}+\frac{y_j}{\prod_{k=0}^{j-1}(x_j-x_k)}\\
&=\frac{y_0}{(x_0-x_j)\prod_{k=1}^{j-1}(x_0-x_k)}+\sum_{i=1}^{j-1}\frac{y_i}{(x_i-x_0)(x_i-x_j)\prod_{\substack{k=1\\k\neq i}}^{j-1}(x_i-x_k)}+\frac{y_j}{(x_j-x_0)\prod_{k=1}^{j-1}(x_j-x_k)}\\
&=\frac{1}{x_j-x_0}\left(-\frac{y_0}{\prod_{k=1}^{j-1}(x_0-x_k)}+\sum_{i=1}^{j-1}\frac{x_j-x_0}{(x_i-x_0)(x_i-x_j)}\frac{y_i}{\prod_{\substack{k=1\\k\neq i}}^{j-1}(x_i-x_k)}+\frac{y_j}{\prod_{k=1}^{j-1}(x_j-x_k)}\right)\\
&=\frac{1}{x_j-x_0}\left(-\frac{y_0}{\prod_{k=1}^{j-1}(x_0-x_k)}+\sum_{i=1}^{j-1}\left(\frac{1}{x_i-x_j}-\frac{1}{x_i-x_0}\right)\frac{y_i}{\prod_{\substack{k=1\\k\neq i}}^{j-1}(x_i-x_k)}+\frac{y_j}{\prod_{k=1}^{j-1}(x_j-x_k)}\right)\\
&=\frac{1}{x_j-x_0}\left(\sum_{i=1}^{j}\frac{y_i}{\prod_{\substack{k=1\\k\neq i}}^{j}(x_i-x_k)}-\sum_{i=0}^{j-1}\frac{y_i}{\prod_{\substack{k=0\\k\neq i}}^{j-1}(x_i-x_k)}\right)\\
&=\frac{1}{x_j-x_0}\big([x_1,\dots,x_j]\boldsymbol{y}-[x_0,\dots,x_{j-1}]\boldsymbol{y}\big).
\end{aligned}$$

Note : L'égalité

$$[x_1,\dots,x_j]\boldsymbol{y}=\sum_{i=1}^{j}\frac{y_i}{\prod_{\substack{k=1\\k\neq i}}^{j}(x_i-x_j)}$$

est obtenue comme celle pour $[x_0,\dots,x_{j-1}]\boldsymbol{y}$ en considérant les nœuds $x_1,\dots,x_j$ en place de $x_0,\dots,x_{j-1}$.

On peut plus généralement établir la formule

$$\forall i\in\{0,\dots,j\},\forall k\in\{0,\dots,i\},\ [x_{i-k},\dots,x_i]\boldsymbol{y}=\frac{[x_{i-k+1},\dots,x_i]\boldsymbol{y}-[x_{i-k},\dots,x_{i-1}]\boldsymbol{y}}{x_i-x_{i-k}},$$

qui fournit un procédé de calcul effectif des différences divisées par construction du tableau suivant

$$\begin{array}{c|ccccc}
x_0 & [x_0]\boldsymbol{y} & & & & \\
x_1 & [x_1]\boldsymbol{y} & [x_0,x_1]\boldsymbol{y} & & & \\
x_2 & [x_2]\boldsymbol{y} & [x_1,x_2]\boldsymbol{y} & [x_0,x_1,x_2]\boldsymbol{y} & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & \\
x_j & [x_j]\boldsymbol{y} & [x_{j-1},x_j]\boldsymbol{y} & [x_{j-2},x_{j-1},x_j]\boldsymbol{y} & \cdots & [x_0,\dots,x_j]\boldsymbol{y}
\end{array}$$

au sein duquel les différences divisées sont disposées de manière à ce que leur évaluation se fasse de proche en proche par la règle suivante : la valeur d'une différence est obtenue en soustrayant à la différence placée immédiatement à sa gauche celle située au dessus de cette dernière, puis en divisant le résultat par la différence entre les deux points de l'ensemble $\{x_i\}_{i=0,\dots,j}$ situés respectivement sur la ligne de la différence à calculer et sur la dernière ligne atteinte en remontant diagonalement dans le tableau à partir de cette même différence.

Les différences divisées apparaissant dans la forme de Newton du polynôme d'interpolation de Lagrange sont les coefficients diagonaux de ce tableau. On voit donc que pour construire Π_j à partir de Π_{j-1} en se servant de la forme de Newton, il suffit de calculer le coefficient supplémentaire $[x_0,\dots,x_j]\boldsymbol{y}$, c'est-à-dire d'ajouter une dernière ligne au tableau, ce qui requiert $2\,j$ soustractions et j divisions.

Ceci constitue un avantage de la forme de Newton sur la forme de Lagrange, cette dernière impliquant de recalculer l'ensemble de la base de polynômes de Lagrange, et donc l'ensemble des coefficients du polynôme d'interpolation dans cette base, pour construire Π_j connaissant Π_{j-1}.

Exercice 5 (polynômes de Chebyshev et meilleurs points d'interpolation).

1. Remarquons que le réel $\cos(x)$ appartient à l'intervalle $[-1,1]$ pour tout réel x. Supposons tout d'abord que x appartient à $[0,\pi]$. La fonction arccos étant une bijection de $[-1,1]$ dans $[0,\pi]$, on a dans ce cas $\arccos(\cos(x)) = x$ et par conséquent
$$T_n(\cos(x)) = \cos(n\arccos(\cos(x))) = \cos(nx).$$
Supposons maintenant que x appartient à $[\pi, 2\pi]$. Alors, $2\pi - x$ appartient à $[0,\pi]$ et, en utilisant la parité et la 2π-périodicité du cosinus, on a $\cos(2\pi - x) = \cos(x)$, dont on déduit que
$$\arccos(\cos(x)) = \arccos(\cos(2\pi - x)) = 2\pi - x.$$
Ainsi, en utilisant encore une fois la parité et la 2π-périodicité du cosinus, on en déduit que
$$T_n(\cos(x)) = \cos(n(2\pi - x)) = \cos(-nx) = \cos(nx).$$
La formule est donc prouvée sur $[0, 2\pi]$.
Si x n'appartient pas à l'intervalle $[0,2\pi]$, on se ramène à l'intervalle $[0,2\pi]$ en posant k le plus grand entier inférieur ou égal à $\frac{x}{2\pi}$. On a alors
$$k \leq \frac{x}{2\pi} \leq k+1,$$
ce qui peut se réécrire, via des manipulations algébriques élémentaires,
$$0 \leq x - 2k\pi \leq 2\pi,$$
d'où $x - 2k\pi$ appartient à $[0,2\pi]$. On s'est ainsi ramené au cas précédent et on en déduit, en utilisant encore une fois la parité et la 2π-périodicité du cosinus, que
$$T_n(\cos(x)) = \cos(n\arccos(\cos(x - 2k\pi))) = \cos(n(x - 2k\pi)) = \cos(nx).$$
Montrons enfin la relation de récurrence donnée dans l'énoncé. Soit x appartenant à $[-1,1]$. Il existe un unique y dans $[0,\pi]$ tel que $x = \cos(y)$. Grâce à l'identité précédente et en utilisant une formule de trigonométrie, on a alors
$$\begin{aligned} T_{n+2}(x) + T_n(x) &= \cos((n+2)y) + \cos(ny) \\ &= 2\cos\left(\frac{(n+2)y + ny}{2}\right)\cos\left(\frac{(n+2)y - ny}{2}\right) \\ &= 2\cos((n+1)y)\cos(y) = 2T_{n+1}(x)\,x, \end{aligned}$$
d'où le résultat voulu.

2. Tout d'abord, $T_0(x) = 1$ et $T_1(x) = \cos(\arccos(x)) = x$ pour tout x appartenant à $[-1,1]$. Ainsi, T_0 est un polynôme de degré 0 et T_1 est un polynôme de degré 1, de monôme de plus haut degré valant $x = 2^{1-1}x^1$. On remarque aussi que, par la formule de récurrence obtenue dans la question précédente, $T_2(x) = 2x^2 - 1$. Ainsi, T_2 est un polynôme de degré 2, de monôme de plus haut degré $2x^2 = 2^{2-1}x^2$.
On peut alors raisonner par récurrence. L'hypothèse de récurrence au rang n à faire est : pour tout entier k de $\{1,\dots,n\}$, T_k est un polynôme de degré k et son monôme de plus haut degré est $2^{k-1}x^k$. Au rang $n+1$ (avec $n \geq 2$, puisque les cas $n = 1$ et 2 ont été traités), la relation de récurrence
$$T_{n+1}(x) = 2xT_n(x) + T_{n-1}(x)$$
assure que T_{n+1} est un polynôme (ce qui n'a rien d'évident au départ si l'on se base sur la définition de T_n donnée dans l'énoncé). De plus, par hypothèse de récurrence, T_{n-1} est de degré $n-1$, alors que $2x\,T_n$ est de degré $n+1$. C'est donc ce dernier terme qui fournit le degré de T_{n+1}, qui est égal à $n+1$. Le monôme de plus haut degré de T_{n+1} est alors $(2x)(2^{n-1}X^n) = 2^n x^{n+1}$. Ceci achève la récurrence.
On remarque enfin que, pour tout entier n supérieur ou égal à 1, la formule $T_n(x) = \cos(n\arccos(x))$ permet immédiatement de trouver les racines de T_n. En effet, on a
$$T_n(x) = 0 \Leftrightarrow n\arccos(x) = \frac{\pi}{2} + 2k\pi,\ k \in \mathbb{Z} \Leftrightarrow \arccos(x) = \frac{(2k+1)\pi}{2n},\ k \in \mathbb{Z}.$$
Cette équation n'a de solutions que si
$$\frac{(2k+1)\pi}{2n} \in [0,\pi].$$
Compte tenu du fait que k doit être entier, ceci équivaut à ce que k appartienne à $\{0,\dots,n-1\}$. Les racines de T_n sont donc bien de la forme voulue et on peut vérifier qu'elles sont distinctes.

3. On applique la formule de dérivation des fonctions composées à l'identité définissant T_n sur $[-1,1]$ pour trouver que

$$T_n'(x) = \frac{n\sin(n\arccos(x))}{\sqrt{1-x^2}}.$$

On a ainsi

$$T_n'(x) = 0 \Leftrightarrow n\arccos(x) = k\pi,\ k\in\mathbb{Z} \Leftrightarrow \arccos(x) = \frac{k\pi}{n},\ k\in\mathbb{Z}.$$

Cette équation n'a de solutions que si

$$\frac{k\pi}{n} \in [0,\pi],$$

i.e. l'entier k appartient à $\{0,\dots,n\}$, et les solutions sont alors données par

$$x_k' = \cos\left(\frac{k\pi}{n}\right),\ k=0,\dots,n,$$

et sont distinctes. On remarque alors que

$$T_n(x_k') = \cos(n\arccos\left(\cos\left(\frac{k}{n\pi}\right)\right) = \cos\left(n\frac{k\pi}{n}\right) = \cos(k\pi) = (-1)^k.$$

Il est alors aisé d'en déduire que les points x_k' sont des extremas globaux, puisque, par définition, on a que $T_n(x)$ appartient à $[-1,1]$ pour tout réel x appartenant à $[-1,1]$.

4. Introduisons le polynôme

$$\tilde{T}_{n+1}(x) = \prod_{i=0}^{n}(x-x_k),$$

qui n'est rien d'autre que le polynôme T_n normalisé de telle sorte que le coefficient du terme de plus haut degré soit égal à 1, autrement dit, d'après la deuxième question, $\tilde{T}_{n+1} = \frac{1}{2^n}T_{n+1}$.
D'après la question précédente, puisque $|T_{n+1}(x)| \leq 1$ et que la valeur 1 est atteinte, on en déduit que

$$\sup_{x\in[-1,1]}\left|\prod_{k=0}^{n}(x-x_k)\right| = \frac{1}{2^n}\|T_{n+1}\|_{\infty,[-1,1]} = \frac{1}{2^n}.$$

Ainsi, on se trouve ramené à montrer que, pour tout polynôme Q de $\mathbb{R}_{n+1}[X]$ unitaire, on a

$$\frac{1}{2^n} \leq \|Q\|_{\infty,[-1,1]}.$$

Raisonnons par l'absurde et supposons qu'il existe un polynôme Q de $\mathbb{R}_{n+1}[X]$ unitaire tel que

$$\frac{1}{2^n} > \|Q\|_{\infty,[-1,1]}.$$

On remarque que $Q-\tilde{T}_{n+1}$ appartient à $\mathbb{R}_n[X]$ (Q et $\tilde{T}_{n+1}$ étant tous deux unitaires, les termes de plus haut degré s'annulent quand on fait leur différence). De plus, si les points x_k', $k=0,\dots,n+1$, sont ceux déterminés dans la précédente question pour le polynôme T_{n+1}, on a

$$(Q-\tilde{T}_{n+1})(x_k') = Q(x_k') - \frac{(-1)^k}{2^n},$$

Puisque $\left|Q(x_k')\right| < \frac{1}{2^n}$. Ainsi, si k est pair,

$$(Q-\tilde{T}_{n+1})(x_k') < \frac{1}{2^n} - \frac{1}{2^n} < 0,$$

et, si k est impair,

$$(Q-\tilde{T}_{n+1})(x_k') > -\frac{1}{2^n} + \frac{1}{2^n} > 0.$$

Par le théorème des valeurs intermédiaires, sur chaque intervalle de la forme $]x_k', x_{k+1}'[$ avec $k\in\{0,\dots,n\}$, $Q-\tilde{T}_{n+1}$ va s'annuler au moins une fois. Les intervalles étant tous disjoints et au nombre de $n+1$, on en déduit que $Q-\tilde{T}_{n+1}$ s'annule au moins $n+1$ fois, autrement dit il possède au moins $n+1$ racines. Comme c'est un polynôme de degré inférieur ou égal à n, il est donc identiquement nul, *i.e.* $Q=\tilde{T}_{n+1}$. On a alors $\|Q\|_{\infty,[-1,1]} = \|\tilde{T}_{n+1}\|_{\infty,[-1,1]}$, ce qui contredit l'hypothèse de départ

$$\frac{1}{2^n} = \|\tilde{T}_{n+1}\|_{\infty,[-1,1]} > \|Q\|_{\infty,[-1,1]},$$

et permet de conclure.

Dauphine | PSL
UNIVERSITÉ PARIS
Département MIDO

Méthodes numériques
année 2022/2023

Corrigés de travaux dirigés

Formules de quadrature

Version du 22 février 2023.

Dans toute cette feuille, on identifie polynôme et fonction polynomiale associée.

Exercice 1.

1. Posons $I(f)=\int_{-1}^{1} f(x)\,\mathrm{d}x$. On cherche les poids α_0, α_1 et α_2 tels que $I_{ap}(f)=I(f)$ pour tout polynôme f de degré inférieur ou égal à deux.
   Comme tout polynôme de degré inférieur ou égal à deux est de la forme $a+b\,x+c\,x^2$ et comme I et I_{ap} sont linéaires au sens où
   $$I(x\mapsto a+b\,x+c\,x^2)=a\,I(x\mapsto 1)+b\,I(x\mapsto x)+c\,I(x\mapsto x^2)$$
   et
   $$I_{ap}(x\mapsto a+b\,x+c\,x^2)=a\,I_{ap}(x\mapsto 1)+b\,I_{ap}(x\mapsto x)+c\,I_{ap}(x\mapsto x^2),$$
   il suffit de garantir que
   $$I_{ap}(x\mapsto 1)=I(x\mapsto 1),\ I_{ap}(x\mapsto x)=I(x\mapsto x) \text{ et } I_{ap}(x\mapsto x^2)=I(x\mapsto x^2).$$
   On explicite alors ces trois égalités :
   $$\alpha_0+\alpha_1+\alpha_2=2,\ -\alpha_0+\alpha_2=0 \text{ et } \frac{\alpha_0}{4}+\frac{\alpha_2}{4}=\frac{2}{3},$$
   et on résout le système linéaire pour trouver
   $$\alpha_0=\frac{4}{3},\ \alpha_1=-\frac{2}{3},\ \alpha_2=\frac{4}{3}.$$

2. Le degré d'exactitude est le plus grand entier p tel que, pour tout polynôme f de degré inférieur ou égal à p, on a $I(f)=I_{ap}(f)$. D'après la question précédente, on sait le degré d'exactitude de la formule est au moins égal à deux. Il sera au moins égal à trois si de plus
   $$I_{ap}(x\mapsto x^3)=I(x\mapsto x^3).$$
   Voyons si cette égalité est satisfaite. On a $I(x\mapsto x^3)=0$ d'une part et
   $$I_{ap}(x\mapsto x^3)=\alpha_0\left(-\frac{1}{2}\right)^3+\alpha_2\left(\frac{1}{2}\right)^3=0$$
   d'autre part, puisque $\alpha_2=\alpha_0$. Le degré d'exactitude de la formule est donc au moins égal à trois. Pour qu'il soit égal à quatre, il faudrait de plus que $I_{ap}(x\mapsto x^4)=I(x\mapsto x^4)$. Or, on a
   $$I(x\mapsto x^4)=\frac{2}{5} \text{ et } I_{ap}(x\mapsto x^4)=\alpha_0\left(-\frac{1}{2}\right)^4+\alpha_2\left(-\frac{1}{2}\right)^4=\frac{1}{6}.$$
   Le degré d'exactitude de la formule est donc égal à trois.

Exercice 2.

1. Les polynômes de Lagrange associés aux nœuds x_0 et x_1 sont
   $$l_0(x)=\frac{x-x_1}{x_0-x_1} \text{ et } l_1(x)=\frac{x-x_0}{x_1-x_0}.$$

2. Pour que la formule de quadrature soit exacte pour chacun de ces polynômes, on veut que

$$I_{ap}(l_0) = \int_{-1}^{1} l_0(x)\,\mathrm{d}x \text{ et } I_{ap}(l_1) = \int_{-1}^{1} l_1(x)\,\mathrm{d}x,$$

ce qu'on peut encore écrire

$$\begin{aligned} & I_{ap}(l_0) = \int_{-1}^{1} \frac{x-x_1}{x_0-x_1}\,\mathrm{d}x \\ \Leftrightarrow\ & \alpha_0\, l_0(x_0) + \alpha_1\, l_0(x_1) = \left[\frac{(x-x_1)^2}{2(x_0-x_1)}\right]_{-1}^{1} \\ \Leftrightarrow\ & \alpha_0 = -\frac{2x_1}{x_0-x_1} \end{aligned}$$

et

$$\begin{aligned} & I_{ap}(l_1) = \int_{-1}^{1} \frac{x-x_0}{x_1-x_0}\,\mathrm{d}x \\ \Leftrightarrow\ & \alpha_0\, l_1(x_0) + \alpha_1\, l_1(x_1) = \left[\frac{(x-x_0)^2}{2(x_1-x_0)}\right]_{-1}^{1} \\ \Leftrightarrow\ & \alpha_1 = -\frac{2x_0}{x_1-x_0}. \end{aligned}$$

Les polynômes l_0 et l_1 constituant une base de de $\mathbb{P}_1$ (c'est résultat établi dans le cours), la formule de quadrature dont on vient de déterminer les poids se trouve être exacte pour tout polynôme de degré inférieur ou égal à un.

3. Ayant obtenu des conditions sur les poids pour que la formule de quadrature soit de degré d'exactitude au moins égale à un, déterminons une condition sur les nœuds de la formule pour que le degré soit au moins égal à deux. Pour cela, il suffit qu'en plus

$$\begin{aligned} & I_{ap}(x \mapsto x^2) = \int_{-1}^{1} x^2\,\mathrm{d}x \\ \Leftrightarrow\ & \alpha_0\,(x_0)^2 + \alpha_1\,(x_1)^2 = \frac{2}{3} \\ \Leftrightarrow\ & -\frac{2(x_0)^2 x_1}{x_0-x_1} - \frac{2x_0(x_1)^2}{x_1-x_0} = \frac{2}{3} \\ \Leftrightarrow\ & \frac{(x_0)^2 x_1 - x_0(x_1)^2}{x_1-x_0} = \frac{1}{3} \\ \Leftrightarrow\ & x_0 x_1 \frac{x_0-x_1}{x_1-x_0} = \frac{1}{3} \\ \Leftrightarrow\ & x_0 x_1 = -\frac{1}{3}. \end{aligned}$$

4. Une formule de quadrature satisfaisant aux trois conditions obtenues est exacte pour les polynômes de degré inférieur ou égal à deux. Pour que le degré d'exactitude soit au moins égal à trois, il faut de plus que

$$\begin{aligned} & I_{ap}(x \mapsto x^3) = \int_{-1}^{1} x^3\,\mathrm{d}x \\ \Leftrightarrow\ & \alpha_0\,(x_0)^3 + \alpha_1\,(x_1)^3 = 0 \\ \Leftrightarrow\ & -\frac{2(x_0)^3 x_1}{x_0-x_1} - \frac{2x_0(x_1)^3}{x_1-x_0} = 0 \\ \Leftrightarrow\ & x_0 x_1 (x_0+x_1) = 0 \\ \Leftrightarrow\ & -\frac{1}{3}(x_0+x_1) = 0 \\ \Leftrightarrow\ & x_0 + x_1 = 0. \end{aligned}$$

5. Pour montrer que le degré d'exactitude n'est pas égal à quatre, on peut procéder comme précédemment avec la fonction $x \mapsto x^4$ et montrer que l'égalité voulue n'est pas compatible avec les conditions obtenues, mais on peut utiliser un contre-exemple comme suggéré. Pour cela, supposons le degré d'exactitude de la formule égal à quatre. La formule devrait alors être exacte pour le polynôme $\omega(x) = ((x-x_0)(x-x_1))^2$. On a

$$I_{ap}(\omega) = 0.$$

Or, l'intégrale $\int_{-1}^{1} \omega(x)\mathrm{d}x$ est strictement positive, car ω est un polynôme positif et non identiquement nul sur $[-1,1]$. On arrive ainsi à une absurdité.

6. Utilisons les conditions trouvées pour déterminer les noeuds de quadrature. On a

$$x_0 x_1 = -\frac{1}{3} \text{ et } x_0 + x_1 = 0, \text{ avec } x_0 < x_1.$$

On trouve alors

$$x_0 = -\frac{1}{\sqrt{3}} \text{ et } x_1 = \frac{1}{\sqrt{3}}.$$

En utilisant les conditions restantes sur les poids de quadrature, on trouve enfin que

$$\alpha_0 = \frac{2}{\sqrt{3}\left(\frac{2}{\sqrt{3}}\right)} = 1 \text{ et } \alpha_1 = \frac{-2}{\sqrt{3}\left(\frac{-2}{\sqrt{3}}\right)} = 1.$$

La formule de quadrature à deux nœuds sur l'intervalle $[-1,1]$ et de degré d'exactitude égal à trois obtenue est donc

$$I_{ap}(f) = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Exercice 3 (erreur pour la formule de quadrature de Simpson).

1. Par définition, on a $\Pi_2 f(x) = \sum_{i=0}^{2} f(x_i) l_i(x)$, les polynômes de Lagrange associés aux nœuds x_0, x_1 et x_2 sont respectivement

$$l_0(x) = \frac{(x-\frac{a+b}{2})(x-b)}{(a-\frac{a+b}{2})(a-b)},\ l_1(x) = \frac{(x-a)(x-b)}{(\frac{a+b}{2}-a)(\frac{a+b}{2}-b)} \text{ et } l_2(x) = \frac{(x-\frac{a+b}{2})(x-a)}{(b-\frac{a+b}{2})(b-a)}.$$

On a donc

$$\begin{aligned}
\Pi_2 f(x) &= f(a)\frac{(x-\frac{a+b}{2})(x-b)}{(a-\frac{a+b}{2})(a-b)} + f(\tfrac{a+b}{2})\frac{(x-a)(x-b)}{(\frac{a+b}{2}-a)(\frac{a+b}{2}-b)} + f(b)\frac{(x-\frac{a+b}{2})(x-a)}{(b-\frac{a+b}{2})(b-a)} \\
&= f(a)\frac{(x-\frac{a+b}{2})(x-b)}{\frac{1}{2}(a-b)(a-b)} + f(\tfrac{a+b}{2})\frac{(x-a)(x-b)}{\frac{1}{4}(b-a)(a-b)} + f(b)\frac{(x-\frac{a+b}{2})(x-a)}{\frac{1}{2}(b-a)(b-a)} \\
&= \frac{2}{(a-b)^2}\left(f(a)(x-\tfrac{a+b}{2})(x-b) - 2f(\tfrac{a+b}{2}))(x-a)(x-b) + f(b)(x-\tfrac{a+b}{2})(x-a)\right).
\end{aligned}$$

En intégrant entre a et b, on obtient

$$\begin{aligned}
I_2(f) &= \int_a^b \frac{2}{(a-b)^2}\left(f(a)(x-\tfrac{a+b}{2})(x-b) - 2f(\tfrac{a+b}{2}))(x-a)(x-b) + f(b)(x-\tfrac{a+b}{2})(x-a)\right)\mathrm{d}x \\
&= \frac{2f(a)}{(a-b)^2}\int_a^b (x-\tfrac{a+b}{2})(x-b)\,\mathrm{d}x - \frac{4f(\frac{a+b}{2})}{(a-b)^2}\int_a^b (x-a)(x-b)\,\mathrm{d}x \\
&\quad + \frac{2f(b)}{(a-b)^2}\int_a^b (x-\tfrac{a+b}{2})(x-a)\,\mathrm{d}x,
\end{aligned}$$

d'où

$$\begin{aligned}
\alpha_0 &= \frac{2}{(a-b)^2}\int_a^b (x-\tfrac{a+b}{2})(x-b)\,\mathrm{d}x \\
&= \frac{1}{(a-b)^2}\left(\int_a^b (x-b)^2 + (x-a)(x-b)\,\mathrm{d}x\right) \\
&= \frac{1}{(a-b)^2}\left[\frac{(x-b)^3}{3} + \frac{x^3}{3} - \frac{(b+a)x^2}{2} + ab\,x\right]_a^b \\
&= \frac{1}{6}(b-a),
\end{aligned}$$

$$\begin{aligned}
\alpha_1 &= -\frac{4}{(a-b)^2}\int_a^b (x-a)(x-b)\,\mathrm{d}x \\
&= -\frac{4}{(a-b)^2}\left[\frac{x^3}{3} - \frac{(b+a)x^2}{2} + ab\,x\right]_a^b \\
&= \frac{4}{6}(b-a),
\end{aligned}$$

$$\begin{aligned}\alpha_2 &= \frac{2}{(a-b)^2}\int_a^b (x-\tfrac{a+b}{2})(x-a)\,\mathrm{d}x\\ &= \frac{1}{(a-b)^2}\left(\int_a^b (x-a)^2+(x-a)(x-b)\,\mathrm{d}x\right)\\ &= \frac{1}{(a-b)^2}\left[\frac{(x-a)^3}{3}+\frac{x^3}{3}-\frac{(b+a)\,x^2}{2}+ab\,x\right]_a^b\\ &= \frac{1}{6}(b-a).\end{aligned}$$

2. a. Par définition de la fonction G, on a

$$G(1)=\int_{-1}^{1} f(\tfrac{a+b}{2}+\tfrac{b-a}{2}\,t)\,\mathrm{d}t-\frac{1}{3}\left[F(-1)+4F(0)+F(1)\right].$$

On fait dans l'intégrale le changement de variable $x=\frac{a+b}{2}+\frac{b-a}{2}\,t$, d'où $\mathrm{d}t=\frac{2}{b-a}\,\mathrm{d}x$ et

$$\int_{-1}^{1} f(\tfrac{a+b}{2}+\tfrac{b-a}{2}\,t)\,\mathrm{d}t=\frac{2}{b-a}\int_a^b f(x)\,\mathrm{d}x$$

On a aussi directement, à partir de la definition de F,

$$F(-1)=f(a),\ F(0)=f(\tfrac{a+b}{2})\text{ et }F(1)=f(b).$$

En utilisant la première question dans la définition de $E_2(f)$ et en comparant avec ces expressions, on arrive à

$$\begin{aligned}E_2(f) &= \int_a^b (f(x)-\Pi_2 f(x))\,\mathrm{d}x\\ &= \int_a^b f(x)\,\mathrm{d}x-\frac{b-a}{6}\left(f(a)+4f(\tfrac{a+b}{2})+f(b)\right)\\ &= \frac{b-a}{2}\left(\frac{2}{b-a}\int_a^b f(x)\,\mathrm{d}x-\frac{1}{3}\left(f(a)+4f(\tfrac{a+b}{2})+f(b)\right)\right)\\ &= \frac{b-a}{2}\,G(1).\end{aligned}$$

b. On étudie maintenant la fonction $H(t)=G(t)-t^5\,G(1)$, ce qui va permettre d'estimer $G(1)$ et donc l'erreur de quadrature. On va pour cela successivement appliquer le théorème de Rolle à H et à ses dérivées première et seconde.

On a supposé que la fonction f est de classe $\mathscr{C}^4$, la fonction F l'est donc aussi, ainsi que les fonctions G et H. On a

$$H(0)=G(0)=0,\ H(1)=G(1)-G(1)=0.$$

En appliquant le théorème de Rolle à la fonction H sur l'intervalle $[0,1]$, on obtient l'existence d'un réel x_1 dans $]0,1[$ tel que $H'(x_1)=0$.

Calculons maintenant la dérivée [1] de H. On a

$$\begin{aligned}H'(t) &= G'(t)-5t^4\,G(1)\\ &= F(t)+F(-t)-\frac{1}{3}\left[F(-t)+4F(0)+F(t)\right]+\frac{t}{3}\left[F'(-t)-F'(t)\right]-5t^4\,G(1),\end{aligned}$$

et donc

$$H'(0)=F(0)+F(0)-\frac{1}{3}\left[F(0)+4F(0)+F(0)\right]=0.$$

On applique alors le théorème de Rolle à H' sur l'intervalle $[0,x_1]$, on obtient l'existence d'un réel x_2 dans $]0,x_1[$ tel que $H''(x_2)=0$.

Calcule enfin la dérivée seconde de H. On a

$$\begin{aligned}H''(t) &= F'(t)-F'(-t)+\frac{1}{3}\left[F'(-t)-F'(t)\right]+\frac{1}{3}\left[F'(-t)-F'(t)\right]-\frac{t}{3}\left[F''(-t)+F''(t)\right]-20t^3\,G(1)\\ &= \frac{1}{3}\left[F'(t)-F'(-t)\right]-\frac{t}{3}\left[F''(-t)+F''(t)\right]-20t^3\,G(1).\end{aligned}$$

ce qui donne immediatement $H''(0)=0$. En appliquant alors le théorème de Rolle à H'' sur l'intervalle $[0,x_2]$, on obtient l'existence d'un réel ζ dans $]0,x_2[\subset]0,1[$ (par construction) tel que $H'''(\zeta)=0$.

1. On rappelle que $(t\to\int_{-t}^{t}\lambda(u)\,\mathrm{d}u)'=(t\to\Lambda(t)-\Lambda(-t))'=\lambda(t)+\lambda(-t)$, où Λ est une primitive de λ.

c. Calculons la dérivée troisième de H. On a

$$\begin{aligned} H'''(t) &= \frac{1}{3}\big[F''(t)+F''(-t)\big]-\frac{1}{3}\big[F''(-t)+F''(t)\big]-\frac{t}{3}\big[F'''(t)-F'''(-t)\big]-60t^2\,G(1) \\ &= -\frac{t}{3}\big[F'''(t)-F'''(-t)\big]-60t^2\,G(1). \end{aligned}$$

en appliquant le théorème des accroissements finis à F''', qui est continue et dérivable sur $]-\zeta,\zeta[$, on obtient l'existence d'un réel ξ dans $]-\zeta,\zeta[$ tel que $F'''(\zeta)-F'''(-\zeta)=2\zeta\,F^{(4)}(\xi)$. En substituant dans $H'''(\zeta)$, on trouve

$$H'''(\zeta) = -\frac{\zeta}{3}\big[F'''(\zeta)-F'''(-\zeta)\big]-60\zeta^2\,G(1) = -\frac{\zeta}{3}\,2\zeta\,F^{(4)}(\xi)-60\zeta^2\,G(1) = -\frac{2\zeta^2}{3}\,(F^{(4)}(\xi)-90\,G(1)).$$

On a vu précédemment que $H'''(\zeta)=0$. On trouve alors que

$$G(1) = -\frac{1}{90}F^{(4)}(\xi) = -\frac{(b-a)^4}{1440}f^{(4)}(c),$$

où $c=\frac{a+b}{2}+\frac{b-a}{2}\,\xi$. Le réel c appartient bien à l'intervalle $]a,b[$ car ξ appartient à l'intervalle $]-\zeta,\zeta[$ qui est inclus dans $]-1,1[$.
On en déduit finalement que

$$E_2(f) = \frac{b-a}{2}\,G(1) = -\frac{(b-a)^5}{2880}f^{(4)}(c).$$

3. Il découle de l'égalité ci-dessus que la formule de quadrature de Simpson est exacte pour les fonctions dont la dérivée quatrième est identiquement nulle. C'est le cas pour les fonctions polynomiales de degré inferieur ou égal à trois. Le degré d'exactitude de cette formule de quadrature vaut donc trois.

Exercice 4 (degré d'exactitude maximal d'une formule de quadrature interpolatoire).

1. On veut montrer par double implication que les coefficients w_i, $i=0,\dots,n$, forment la solution du système linéaire

$$\sum_{i=1}^{n} P_k(x_i)w_i = \begin{cases} \langle P_0,P_0\rangle_w & \text{si } k=0, \\ 0 & \text{si } k=1,\dots,n, \end{cases} \tag{1}$$

si et seulement s'ils sont strictement positifs et que

$$\forall P\in\mathscr{P}_{2n+1}(a,b),\ \int_a^b P(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{n} w_i P(x_i).$$

a. i. Le système linéaire dont les coefficients $w_0,\dots,w_n$ sont solution s'écrit matriciellement

$$\begin{pmatrix} P_0(x_0) & \dots & P_0(x_n) \\ P_1(x_0) & \dots & P_1(x_n) \\ \vdots & & \vdots \\ P_n(x_0) & \dots & P_n(x_n) \end{pmatrix}\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \langle P_0,P_0\rangle_w \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Les fonctions $P_0,\dots,P_n$ forment une famille orthogonale, donc libre, et les points $x_0,\dots,x_n$ étant deux à deux distincts, on en déduit que la matrice associée au système est inversible.

ii. Soit P une fonction de $\mathscr{P}_{2n+1}(a,b)$. En effectuant une division euclidienne de P par P_{n+1} on obtient la décomposition

$$P = QP_{n+1}+R,$$

où le quotient Q est de degré au plus égal à $2n+1-(n+1)=n$ et le reste R est de degré strictement inférieur à $n+1$, soit encore au plus égal à n. On conclut écrivant Q et R dans la base $\{P_0,\dots,P_n\}$.

iii. En utilisant le résultat de la précédente question, le fait que $P_0\equiv 1$ et l'orthogonalité de la famille des polynômes, on trouve que, pour toute fonction polynomiale P de $\mathscr{P}_{2n+1}(a,b)$,

$$\begin{aligned} \int_a^b P(x)w(x)\,\mathrm{d}x &= \int_a^b Q(x)P_{n+1}(x)w(x)\,\mathrm{d}x+\int_a^b R(x)w(x)\,\mathrm{d}x \\ &= \langle Q,P_{n+1}\rangle_w+\langle R,P_0\rangle_w \\ &= \sum_{i=0}^{n}\alpha_i\,\langle P_i,P_{n+1}\rangle_w+\sum_{i=0}^{n}\beta_i\,\langle P_i,P_0\rangle_w \\ &= \beta_0\,\langle P_0,P_0\rangle_w\,. \end{aligned}$$

Par ailleurs, on a

$$\forall i \in \{0,\dots,n\},\ P_{n+1}(x_i) = 0,$$

d'où

$$\sum_{i=0}^{n} P(x_i)w_i = \sum_{i=0}^{n} R(x_i)w_i = \sum_{i=0}^{n}\left(\sum_{j=0}^{n}\beta_j P_j(x_i)\right)wi = \sum_{j=0}^{n}\beta_j\left(\sum_{i=0}^{n} P_j(x_i)wi\right).$$

iv. Les coefficients étant solution du système linéaire, on a

$$\forall j \in \{0,\dots,n\},\ \sum_{i=0}^{n} P_j(x_i)wi = \langle P_0, P_0\rangle_w \delta_{0j},$$

d'où

$$\forall P \in \mathscr{P}_{2n+1}(a,b),\ \sum_{i=0}^{n} P(x_i)w_i = \langle P_0, P_0\rangle_w \sum_{j=0}^{n}\beta_j\delta_{0j} = \beta_0 \langle P_0, P_0\rangle_w .$$

On a ainsi montré que

$$\forall P \in \mathscr{P}_{2n+1}(a,b),\ \int_a^b P(x)w(x)\,\mathrm{d}x = \beta_0 \langle P_0, P_0\rangle_w = \sum_{i=0}^{n} P(x_i)w_i.$$

Soit enfin j un entier entre 0 et n. Pour montrer que le coefficient w_j est strictement positif, on considère la fonction de $\mathscr{P}_{2n}(a,b)$ définie par

$$P(x) = \prod_{\substack{k=0\\k\neq j}}^{n}(x-x_k)^2,$$

pour laquelle

$$0 < \int_a^b P(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{n} P(x_i)w_i = w_j \prod_{\substack{k=0\\k\neq j}}^{n}(x_j-x_k)^2.$$

b. i. En utilisant le fait que $P_0 \equiv 1$ et l'orthogonalité de la famille des polynômes, on trouve que

$$\forall k \in \{0,\dots,n\},\ \sum_{i=0}^{n} P_k(x_i)w_i = \int_a^b P_k(x)w(x)\,\mathrm{d}x = \langle P_k, P_0\rangle_w = \langle P_0, P_0\rangle_w \delta_{k0}.$$

ii. Pour tout entier k de $\{0,\dots,n\}$, la fonction polynomiale $P = P_{n+1}P_k$ est de degré $n+k+1 \leq 2n+1$ et on a donc

$$\int_a^b P_{n+1}(x)P_k(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{n} P_k(x_i)P_{n+1}(x_i)w_i.$$

On a également, par orthogonalité,

$$\forall k \in \{0,\dots,n\},\ \int_a^b P_{n+1}(x)P_k(x)w(x)\,\mathrm{d}x = \langle P_{n+1}, P_k\rangle_w = 0,$$

d'où le résultat.

iii. On a montré que les poids de quadrature étaient solution du système linéaire de l'énoncé, il reste donc à montrer que les nœuds de quadrature constituent l'ensemble des racines de P_{n+1}. Il découle de la précédente question que le système linéaire

$$\begin{pmatrix} P_0(x_0) & \dots & P_0(x_n)\\ \vdots & & \vdots\\ P_n(x_0) & \dots & P_n(x_n)\end{pmatrix}\begin{pmatrix} P_{n+1}(x_0)w_0\\ \vdots\\ P_{n+1}(x_n)w_n\end{pmatrix} = \begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}$$

est vérifié. Sa matrice étant inversible, les coefficients de sa solution sont tous nuls et l'on peut conclure en utilisant que les poids $sw_0,\dots,w_n$ sont strictement positifs.

2. Il suffit de considérer la fonction de $\mathscr{P}_{2n+1}(a,b)$ définie par

$$P(x) = \prod_{k=0}^{n}(x-x_k)^2,$$

pour laquelle

$$\int_a^b P(x)w(x)\,\mathrm{d}x > 0 \text{ et } \sum_{i=0}^{n} P(x_i)w_i = 0.$$

Exercice 5 (erreur de quadrature et noyau de Peano).

1. Les trois polynômes de Lagrange associés aux points d'interpolation -1, 0 et 1 sont

$$l_0(x)=\frac{1}{2}\,x(x-1),\ l_1(x)=-(x+1)(x-1) \text{ et } l_2(x)=\frac{1}{2}\,(x+1)x.$$

Le polynôme d'interpolation $\Pi_2 f$ étant de degré deux, on a $\Pi_2 f = f$ pour toute fonction f de $\mathbb{P}_2$, et les poids de la formule de quadrature recherchés sont simplement les intégrales des polynômes de Lagrange entre -1 et 1. On trouve alors :

$$\alpha_0=\int_{-1}^{1} l_0(t)\,\mathrm{d}t=\frac{1}{3},\ \alpha_1=\int_{-1}^{1} l_1(t)\,\mathrm{d}t=\frac{4}{3} \text{ et } \alpha_2=\int_{-1}^{1} l_2(t)\,\mathrm{d}t=\frac{1}{3}.$$

2. Par propriété du polynôme d'interpolation de f aux points -1, 0 et 1, on a $s(-1)=f(-1)-\Pi_2 f(-1)=0$, $s(0)=f(0)-\Pi_2 f(0)=0$ et $s(1)=f(1)-\Pi_2 f(1)=0$. On peut donc factoriser s par le polynôme $(x+1)x(x-1)=(x^2-1)x$, qui est un polynôme de degré trois. Puisque s est également un polynôme de degré inférieur ou égal à trois, le quotient de la division (exacte) de $s(x)$ par $(x^2-1)x$ est un réel, que l'on note a. En remarquant à présent que

$$\int_{-1}^{1} s(t)\,\mathrm{d}t = a\int_{-1}^{1}(t^2-1)t\,\mathrm{d}t=0,$$

on en déduit que, pour f appartenant à $\mathbb{P}_3$,

$$E_2(f)=E_2(\Pi_2 f)+a\left(\int_{-1}^{1} s(t)\,\mathrm{d}t-\frac{1}{3}s(-1)-\frac{4}{3}s(0)-\frac{1}{3}s(1)\right)=E_2(\Pi_2 f)=0,$$

la formule de quadrature étant exacte pour tout polynôme de degré inférieur ou égal à deux.

3. Par définition de l'erreur de quadrature $E_2(f)$, on a

$$E_2(f)=\int_{-1}^{1} f(t)\,\mathrm{d}t-\frac{1}{3}f(-1)-\frac{4}{3}f(0)-\frac{1}{3}f(1),$$

d'où, par utilisation de la formule de Taylor avec reste intégral,

$$\begin{aligned}E_2(f)=E_2(p)+\frac{1}{6}\int_{-1}^{1}\left(\int_{-1}^{1} f^{(4)}(t)(x-t)_+^3\,\mathrm{d}t\right)\mathrm{d}x-\frac{1}{18}\int_{-1}^{1} f^{(4)}(t)(-1-t)_+^3\,\mathrm{d}t\\ -\frac{2}{9}\int_{-1}^{1} f^{(4)}(t)(-t)_+^3\,\mathrm{d}t-\frac{1}{18}\int_{-1}^{1} f^{(4)}(t)(1-t)_+^3\,\mathrm{d}t.\end{aligned}$$

Ayant montré à la question précédente que la formule de quadrature était en fait exacte pour tout polynôme de degré inférieur ou égal à trois, on sait que $E_2(p)=0$ et par conséquent, après avoir inversé l'ordre des intégrales (ce qu'on ne demande pas de justifier), on trouve

$$E_2(f)=\frac{1}{6}\int_{-1}^{1}\left(\int_{-1}^{1}(x-t)_+^3\,\mathrm{d}x-\frac{1}{3}(-1-t)_+^3-\frac{4}{3}(-t)_+^3-\frac{1}{3}(1-t)_+^3\right)f^{(4)}(t)\,\mathrm{d}t,$$

dont on déduit l'expression générale du noyau de Peano,

$$\forall t\in[-1,1],\ K(t)=\int_{-1}^{1}(x-t)_+^3\,\mathrm{d}x-\frac{1}{3}(-1-t)_+^3-\frac{4}{3}(-t)_+^3-\frac{1}{3}(1-t)_+^3.$$

4. On admet que la fonction K est paire. Pour tout t dans l'intervalle $[0,1]$, on a $(-1-t)_+=0$, $(-t)_+=0$, $(1-t)_+=(1-t)$ et

$$\int_{-1}^{1}(x-t)_+^3\,\mathrm{d}x=\int_{t}^{1}(x-t)^3\,\mathrm{d}x=\frac{(1-t)^4}{4},$$

d'où

$$\forall t\in[0,1],\ K(t)=\frac{(1-t)^4}{4}-\frac{1}{3}(1-t)^3=\frac{1}{12}(1-t)^3(3(1-t)-4)=-\frac{1}{12}(1-t)^3(1+3t).$$

5. L'expression ci-dessus montre que $K(t)$ reste négative pour t appartenant à $[0,1]$ et il en est de même pour t appartenant à $[-1,0]$ par parité. La fonction $f^{(4)}$ étant continue sur $[-1,1]$, on a alors l'encadrement

$$\max_{\theta\in[-1,1]} f^{(4)}(\theta)\int_{-1}^{1} K(t)\,\mathrm{d}t\le\int_{-1}^{1} K(t)f^{(4)}(t)\,\mathrm{d}t\le\min_{\theta\in[-1,1]} f^{(4)}(\theta)\int_{-1}^{1} K(t)\,\mathrm{d}t,$$

et, par suite,

$$\max_{\theta\in[-1,1]} f^{(4)}(\theta) \leq \frac{E_2(f)}{\frac{1}{6}\int_{-1}^{1} K(t)\,\mathrm{d}t} \leq \min_{\theta\in[-1,1]} f^{(4)}(\theta).$$

Le théorème des valeurs intermédiaires assure alors qu'il existe un réel ξ appartenant à $[-1,1]$ tel que

$$f^{(4)}(\xi) = \frac{E_2(f)}{\frac{1}{6}\int_{-1}^{1} K(t)\,\mathrm{d}t},$$

d'où la première égalité demandée. Enfin, on a

$$\int_{-1}^{1} K(t)\,\mathrm{d}t = 2\int_{0}^{1} K(t)\,\mathrm{d}t = -\frac{1}{6}\int_{0}^{1} (1-t)^3(1+3t)\,\mathrm{d}t = -\frac{1}{15},$$

dont on déduit la seconde égalité demandée.

Dauphine | PSL
UNIVERSITÉ PARIS
Département MIDO

Méthodes numériques
année 2022/2023

Corrigés de travaux dirigés

Méthodes de factorisation pour la résolution de systèmes linéaires

Version du 22 février 2023.

Exercice 1.

1. On a
$$A^{(0)} = \begin{pmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{pmatrix} \text{ et } L^{(0)} = I_4,$$
$$A^{(1)} = \begin{pmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{pmatrix} \text{ et } L^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$
$$A^{(2)} = \begin{pmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix} = U \text{ et } L^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{pmatrix} = L.$$
On observe que deux étapes suffisent pour factoriser la matrice.
2. On doit tout d'abord résoudre le système linéaire triangulaire inférieur $LY = B$:
$$\begin{cases} y_1 = 5 \\ y_2 = 9 - 2 \times 5 = -1 \\ y_3 = 1 + 5 - (-1) = 7 \\ y_4 = -2 - 3 \times (-1) = 1 \end{cases},$$
puis le système linéaire triangulaire supérieur $UX = Y$:
$$\begin{cases} x_4 = 1 \\ x_3 = \dfrac{1}{5}(7-2) = 1 \\ x_2 = -1 + 3 - 1 = 1 \\ x_1 = \dfrac{1}{2}(5 + 1 - 4) = 1 \end{cases}.$$
3. On a $\det(A) = \det(LU) = \det(L)\det(U) = \det(U) = \prod_{i=1}^{4} u_{ii} = 2 \times 1 \times 5 \times 1 = 10$.

Exercice 2. On rapelle la définition des matrices de transvection servant à l'élimination de Gaussdans cet exercice : pour $(i, j) \in \{1, \dots, 4\}^2$, $i \neq j$, et λ un réel non nul,
$$T_{ij}(\lambda) = I_4 + \lambda E_{ij}.$$

1. On a
$$A^{(0)} = \begin{pmatrix} 2 & 2 & 1 \\ -2 & 6 & 4 \\ 1 & 0 & 2 \end{pmatrix} \text{ et } L^{(0)} = I_4,$$
$$A^{(1)} = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 8 & 5 \\ 0 & -1 & \frac{3}{2} \end{pmatrix},\ L^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix} \text{ et } E^{(1)} = T_{31}(-\frac{1}{2})T_{21}(1),$$
$$A^{(2)} = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 8 & 5 \\ 0 & 0 & \frac{17}{8} \end{pmatrix} = U,\ L^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \frac{1}{2} & -\frac{1}{8} & 1 \end{pmatrix} = L \text{ et } E^{(2)} = T_{32}(\frac{1}{8}).$$

2. On a

$$A^{(0)} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 2 & 4 & 5 & 1 \\ 0 & 4 & 0 & 4 \\ 1 & 0 & 0 & 2 \end{pmatrix} \text{ et } L^{(0)} = I_4,$$

$$A^{(1)} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 4 & 3 & 1 \\ 0 & 4 & 0 & 4 \\ 0 & 0 & -1 & 2 \end{pmatrix},\ L^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ et } E^{(1)} = T_{41}(-1)T_{21}(-2),$$

$$A^{(2)} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 4 & 3 & 1 \\ 0 & 0 & -3 & 3 \\ 0 & 0 & -1 & 2 \end{pmatrix},\ L^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ et } E^{(2)} = T_{32}(-1),$$

$$A^{(3)} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 4 & 3 & 1 \\ 0 & 0 & -3 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix} = U,\ L^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 1 \end{pmatrix} = L \text{ et } E^{(2)} = T_{42}(-\frac{1}{3}).$$

3. On a

$$A^{(0)} = \begin{pmatrix} a & a & a & a \\ a & b & b & b \\ a & b & c & c \\ a & b & c & d \end{pmatrix} \text{ et } L^{(0)} = I_4,$$

si $a \neq 0$,

$$A^{(1)} = \begin{pmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & b-a & c-a & c-a \\ 0 & b-a & c-a & d-a \end{pmatrix},\ L^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ et } E^{(1)} = T_{41}(-1)T_{31}(-1)T_{21}(-1),$$

si $b-a \neq 0$,

$$A^{(2)} = \begin{pmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & c-b & d-b \end{pmatrix},\ L^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \text{ et } E^{(2)} = T_{42}(-1)T_{32}(-1),$$

si $c-b \neq 0$,

$$A^{(3)} = \begin{pmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & 0 & d-c \end{pmatrix} = U,\ L^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} = L \text{ et } E^{(2)} = T_{43}(-1).$$

Exercice 3 (factorisation LU d'une matrice bande). L'intérêt de cet exercice est de montrer qu'une structure de stockage adaptée au fait que A est une matrice bande (seuls les éléments contenus dans la bande sont effectivement stockés) peut recevoir la factorisation LU de A sans aucune modification, en stockant L, sans sa diagonale, dans la partie inférieure stricte de A et U dans la partie supérieure de A.

Raisonnons par récurrence pour montrer que la matrice $A^{(n-1)} = U$ est une matrice bande de même largeur de bande que A. Pour cela, on va montrer chacune des matrices de la suite $(A^{(k)})_{k=0,\dots,n-1}$, obtenue par application du procédé d'élimination de Gauss, est une matrice bande, de même largeur de bande que A.

On suppose que la matrice A est bande de largeur $2p+1$, c'est-à-dire que

$$a_{ij} = 0 \text{ si } |i-j| > p.$$

Pour $k = 0$, $A^{(0)} = A$ est une matrice bande de largeur $2p+1$. Pour un entier naturel k inférieur ou égal à $n-2$, on suppose que la matrice $A^{(k)}$ est une matrice bande de largeur $2p+1$. On rappelle que les coefficients de la matrice $A^{(k+1)}$ sont donnés par

$$\forall (i,j) \in \{1,\dots,n\}^2,\ a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} - \frac{a_{ik+1}^{(k)}}{a_{k+1k+1}^{(k)}} a_{k+1j}^{(k)} & \text{si } k+2 \leq i \leq n \text{ et } k+1 \leq j \leq n, \\ a_{ij}^{(k)} & \text{sinon.} \end{cases}$$

On s'intéresse aux coefficients matriciels se trouvant hors de la bande, c'est-à-dire dont les indices sont tels que

$$(i,j) \in \{1,\dots,n\}^2,\ |i-j| > p \iff (i,j) \in \{1,\dots,n\}^2,\ |i-(k+1)-(j-(k+1))| > p.$$

On considère alors deux cas :

— soit $1 \leq i \leq k+1$ ou $1 \leq j \leq k$ (c'est-à-dire qu'on se trouve hors de la bande et dans une partie de la matrice qui n'est plus modifiée lors de l'élimination), et alors $a_{ij}^{(k+1)} = a_{ij}^{(k)} = 0$.

— soit $k+2 \leq i \leq n$ et $k+1 \leq j \leq n$ (c'est-à-dire qu'on se trouve hors de la bande dans le bloc modifié par l'élimination), et alors $i-(k+1) > p$ ou $j-(+1k) > p$, d'où $a_{ik+1}^{(k)} = 0$ ou $a_{k+1j}^{(k)} = 0$ et donc $a_{ik}^{(k)} a_{k+1j}^{(k)} = 0$. On a ainsi $a_{ij}^{(k+1)} = a_{ij}^{(k)} = 0$.

La matrice $A^{(k+1)}$ est donc une matrice bande, de même largeur de bande que $A^{(k)}$, ce qui achève le raisonnement par récurrence. On a montré que la matrice U était une matrice bande.

On rappelle que les coefficients de la matrice triangulaire inférieure L de la factorisation $A = LU$ sont donnés par

$$\forall (i,j) \in \{1,\dots,n\}^2,\ l_{ij} = \begin{cases} \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}} & \text{si } j < i, \\ 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

Chacune des matrices $A^{(k)}$, $k = 0,\dots,n-1$, étant bande de largeur $2p-1$, on a $a_{ij}^{(j-1)} = 0$ pour $i-j > p$, d'où $l_{ij} = 0$ pour $i-j > p$.

Exercice 4 (factorisation LU d'une matrice tridiagonale – algorithme de Thomas).

1. Il suffit de vérifier que le produit LU est tridiagonal, puis de conclure en utilisant l'unicité de la factorisation.
2. On identifie les coefficients *a priori* non nuls de la matrice A avec ceux du produit LU. On trouve alors

$$a_1 = u_1,\ d_i = l_i u_{i-1},\ a_i = u_i + l_i c_{i-1},\ 2 \leq i \leq n,$$

dont on déduit

$$u_1 = a_1,\ l_i = \frac{d_i}{u_{i-1}},\ u_i = a_i - l_i c_{i-1},\ 2 \leq i \leq n.$$

3. On doit résoudre successivement les systèmes linéaires triangulaires $LY = B$ et $UX = Y$ par descente puis remontée, c'est-à-dire

$$y_1 = b_1,\ y_i = b_i - l_i y_{i-1},\ i = 2,\dots,n,$$

puis

$$x_n = \frac{y_n}{u_n},\ x_j = \frac{y_j - c_j x_{j+1}}{u_j},\ j = n-1,\dots,1.$$

C'est l'*algorithme de Thomas*.

4. Le compte d'opérations donne : $n-1$ soustractions, $n-1$ multiplications et $n-1$ divisions pour la factorisation, $n-1$ soustractions et $n-1$ multiplications pour la descente, $n-1$ soustractions, $n-1$ multiplications et n divisions pour la remontée, soit au total $8n-7$ opérations.

Exercice 5 (factorisation LU d'une matrice à diagonale strictement dominante).

1. Si la matrice A est non inversible, alors il existe une matrice colonne X de $M_{n,1}(\mathbb{R})$ non nulle telle que $AX = 0$, ce qui signifie encore que

$$\forall i \in \{1,\dots,n\},\ \sum_{j=1}^{n} a_{ij} x_j = 0.$$

La matrice X étant non nulle, il existe un indice i_0 tel que $\left|x_{i_0}\right| = \max_{i \in \{1,\dots,n\}} |x_i| \neq 0$. On a alors

$$-a_{i_0 i_0} x_{i_0} = \sum_{\substack{j=1 \\ j \neq i_0}}^{n} a_{i_0 j} x_j,$$

d'où

$$\left|a_{i_0 i_0}\right| x_{i_0} \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{n} \left|a_{i_0 j}\right| \left|\frac{x_j}{x_{i_0}}\right| \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{n} \left|a_{i_0 j}\right|,$$

ce qui contredit le fait que A est à diagonale strictement dominante par lignes.

2. On suppose que A admet une factorisation LU. On a

$$A = LU \iff A^\top = U^\top L^\top,$$

avec U inversible, puisque $\det(U) = \det(A) \neq 0$. Soit Δ la matrice diagonale ayant pour éléments diagonaux ceux de U. La matrice $\widetilde{L} = U^\top \Delta^{-1}$ est triangulaire inférieure à éléments diagonaux égaux à 1. La matrice $\widetilde{U} = \Delta L^\top$ est une matrice triangulaire supérieure inversible car $\det(\widetilde{U}) = \det(\Delta)\det(L) = \det(U) = \det(A)$. On a enfin $A^\top = U^\top \Delta^{-1} \Delta L^\top = \widetilde{L}\widetilde{U}$.

3. On a
$$\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \frac{1}{a}V & I_{n-1} \end{array}\right)\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \mathbf{0} & B \end{array}\right)\left(\begin{array}{c|c} a & W^\top \\ \hline \mathbf{0} & I_{n-1} \end{array}\right)=\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \frac{1}{a}V & B \end{array}\right)\left(\begin{array}{c|c} a & W^\top \\ \hline \mathbf{0} & I_{n-1} \end{array}\right)=\left(\begin{array}{c|c} a & W^\top \\ \hline V & \frac{1}{a}VW^\top+B \end{array}\right).$$
En identifiant avec la partition par blocs de A, il vient $A_1=\frac{1}{a}VW^\top+B$, d'où $B=A_1-\frac{1}{a}VW^\top$.

4. Si B admet une factorisation LU, il existe des matrices triangulaires L_B et U_B, vérifiant les propriétés requises, telles que $B=L_BU_B$ et on a alors
$$A=\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \frac{1}{a}V & L_B \end{array}\right)\left(\begin{array}{c|c} a & W^\top \\ \hline \mathbf{0} & U_B \end{array}\right)$$
la première matrice dans le membre de droite de l'égalité étant triangulaire inférieure à coefficients diagonaux égaux à 1 et la seconde étant triangulaire supérieure.

5. a. Si A est à diagonale strictement dominante par colonnes, alors $A^\top$ est à diagonale strictement dominante par lignes et on a
$$\forall i\in\{1,\dots,n\},\ |a_{ii}|>\sum_{\substack{j=1\\ j\neq i}}^{n}\left|a_{ji}\right|.$$
En utilisant les notations de la décomposition par blocs de $A^\top$,
$$A^\top=\left(\begin{array}{c|c} a & V^\top \\ \hline W & {A_1}^\top \end{array}\right),$$
ces inégalités s'écrivent
$$|a|>\sum_{i=1}^{n-1}|v_i|,$$
ce qui implique
$$\forall j\in\{1,\dots,n-1\},\ |a|-\left|v_j\right|>\sum_{\substack{i=1\\ i\neq j}}^{n-1}|v_i|, \tag{2}$$
et
$$\forall j\in\{1,\dots,n-1\},\ \left|(A_1)_{jj}\right|>\sum_{\substack{i=1\\ i\neq j}}^{n-1}\left|(A_1)_{ij}\right|+\left|w_j\right|. \tag{3}$$
On a enfin $B=A_1-\frac{1}{a}VW^\top$, d'où
$$\begin{aligned}\forall j\in\{1,\dots,n-1\},\ \sum_{\substack{i=1\\ i\neq j}}^{n-1}\left|b_{ij}\right|&=\sum_{\substack{i=1\\ i\neq j}}^{n-1}\left|(A_1)_{ij}-\frac{1}{a}v_iw_j\right|\\ &\leq\sum_{\substack{i=1\\ i\neq j}}^{n-1}\left|(A_1)_{ij}\right|+\left|\frac{w_j}{a}\right|\sum_{\substack{i=1\\ i\neq j}}^{n-1}|v_i|\\ &<\left|(A_1)_{jj}\right|-\left|w_j\right|+\left|\frac{w_j}{a}\right|(|a|-\left|v_j\right|)\\ &=\left|(A_1)_{jj}\right|-\left|\frac{w_j}{a}\right|\left|v_j\right|\\ &\leq\left|b_{jj}\right|.\end{aligned}$$
La matrice B est donc à diagonale strictement dominante par colonnes.

b. Si la matrice A est telle que $A^\top$ soit à diagonale strictement dominante par lignes alors il existe une matrice B telle que $B^\top$ est à diagonale strictement dominante par lignes et
$$A=L_1\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \mathbf{0} & B \end{array}\right)U_1,$$
où $L_1=\left(\begin{array}{c|c} 1 & \mathbf{0}^\top \\ \hline \frac{1}{a}V & I_{n-1} \end{array}\right)$ est une matrice triangulaire inférieure et $U_1=\left(\begin{array}{c|c} a & W^\top \\ \hline \mathbf{0} & I_{n-1} \end{array}\right)$ est une matrice triangulaire supérieure. En appliquant le procédé à B et raisonnant par récurrence sur l'ordre du bloc, on arrive, après $n-1$ étapes, à
$$A=L_1L_2\dots L_{n-1}\left(\begin{array}{c|c} I_{n-1} & \mathbf{0}^\top \\ \hline \mathbf{0} & b \end{array}\right)U_{n-1}\dots U_2U_1,$$
avec b un réel non nul. Il suffit alors de poser $L_n=I_n$ et $U_n=\left(\begin{array}{c|c} I_{n-1} & \mathbf{0}^\top \\ \hline \mathbf{0} & b \end{array}\right)$.

6. Si A est à diagonale strictement dominante par lignes, alors A est inversible et $A^\top$, qui est à diagonale strictement dominante par colonnes, admet une factorisation LU. On déduit alors de la deuxième question que A admet une factorisation LU.
7. (a) On vérifie que la matrice A est à diagonale strictement dominante par lignes. Elle admet donc une factorisation LU :
$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{5} & 1 & 0 \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 0 & \frac{15}{4} & \frac{3}{4} & \frac{3}{4} \\ 0 & 0 & \frac{18}{5} & \frac{3}{5} \\ 0 & 0 & 0 & \frac{21}{6} \end{pmatrix}.$$
(b) La matrice A est triangulaire supérieure. Elle admet donc trivialement une factorisation LU avec $L = I_4$ et $U = A$.
(c) La matrice A n'est pas inversible (deux de ses colonnes sont identiques), mais elle peut néanmoins être factorisée LU :
$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{3} & 1 & 0 \\ \frac{3}{4} & -1 & 0 & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 0 & \frac{15}{4} & \frac{3}{4} & \frac{3}{4} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$
0n notera qu'il n'y a pas unicité de cette décomposition, l'élément l_{43} de la matrice pouvant être choisi arbitrairement.

Exercice 6 (existence et unicité de la factorisation de Cholesky).

1. Une matrice réelle symétrique définie positive n'a que des valeurs propres réelles et strictement positives, 0 n'est donc pas valeur propre de A et la matrice est inversible.
2. Si la matrice A admet une factorisation de Cholesky, alors $A^\top = (BB^\top)^\top = (B^\top)^\top B^\top = BB^\top = A$, A est donc réelle symétrique. On a par ailleurs
$$\forall X \in M_{n,1}(\mathbb{R}),\ X^\top AX = X^\top BB^\top X = (B^\top X)^\top (B^\top X) = \|B^\top X\|_2^{\,2} \geq 0.$$
Enfin, la matrice B étant inversible, c'est aussi le cas pour $B^\top$ et donc $\|B^\top X\|_2^{\,2} = 0_{M_{n,1}(\mathbb{R})}$ implique que $X = 0$.
3. Si la matrice A admet une factorisation de Cholesky, alors A est réelle symétrique définie positive d'après la question précédente. En vertu du critère de Sylvester, tous les mineurs principaux dominants de A sont strictement positifs, ce qui implique que A admet une factorisation LU, que l'on peut encore écrire LDV, avec D une matrice diagonale ayant pour éléments diagonaux ceux de U et $V = D^{-1}U$ à coefficients diagonaux égaux à 1. Puisque A est symétrique, il vient $(LDV)^\top = V^\top DL^\top = LDV$, d'où $V = L^\top$ et donc $A = LDL^\top$.
Soit $A = BB^\top$ et Δ la matrice diagonale ayant pour éléments diagonaux ceux de B. Si ces coefficients sont strictement positifs, alors Δ est inversible et on a
$$A = BB^\top = (B\Delta^{-1})\Delta^2(\Delta^{-1}B^\top) = LDL^\top,$$
en posant $L = B\Delta^{-1}$ et $D = \Delta^2$. La factorisation est unique.
4. a. En vertu du critère de Sylvester, tous les mineurs principaux dominants de A sont strictement positifs. Les mineurs principaux dominants de A_{n-1} étant des mineurs principaux dominants de A et cette condition étant aussi suffisante, on en déduit que A_{n-1} est définie positive.
b. On a
$$BB^\top = \begin{pmatrix} B_{n-1}B_{n-1}^\top & B_{n-1}M \\ M^\top B_{n-1}^\top & M^\top M + b^2 \end{pmatrix},$$
d'où, par identification avec la décomposition par blocs de A, $B_{n-1}M = V$ et $b^2 = a_{nn} - M^\top M$. La matrice B_{n-1} étant inversible, M est définie de manière unique. De plus, on a
$$0 < \det(A) = \det(BB^\top) = \det(B)\det(B^\top) = (\det(B_{n-1}))^2 b^2,$$
d'où $b^2 > 0$, ce qui permet de définir b de façon unique en imposant que b soit de plus positif.
c. En raisonnant par récurrence sur l'ordre du bloc A_k, on finit par arriver à $A_1 = a_{11} > 0$ et il suffit alors de poser $B_1 = \sqrt{a_{11}}$.
5.
6. L'algorithme de factorisation décrit ci-dessus nécessite $\frac{1}{6}n(n^2-1)$ additions/soustractions, $\frac{1}{6}n(n^2-1)$ multiplications, $\frac{1}{2}n(n-1)$ divisions et n extractions de racines carrées, et la résolution des deux systèmes triangulaires résultant nécessite $n(n-1)$ additions/soustractions, $n(n-1)$ multiplications et $2n$ divisions. Au total, on a donc besoin de l'ordre de $\frac{n^3}{6}$ additions/soustractions, $\frac{n^3}{6}$ multiplications, $\frac{n^2}{2}$ divisions et n extractions de racines carrées à comparer à de l'ordre de $\frac{n^3}{3}$ additions/soustractions, $\frac{n^3}{3}$ multiplications et $\frac{n^2}{2}$ divisions pour une résolution par l'élimination de Gauss.

7. On a :

$$A = BB^\top \iff \forall (i,j) \in \{1,\dots,n\}^2,\ a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{\min(i,j)} b_{ik}b_{jk},$$

car B est triangulaire supérieure. La matrice A étant symétrique, il suffit, par exemple, que les relations ci-dessus soient vérifiées pour $j \leq i$ et l'on construit alors les colonnes de la matrice B à partir de celles de A. En fixant l'indice j à 1 et en faisant varier l'indice i de 1 à n, on trouve

$$\begin{array}{rcl@{\quad}rcl} a_{11} &=& (b_{11})^2, & \text{d'où} & b_{11} &=& \sqrt{a_{11}}, \\ a_{21} &=& b_{11}b_{21}, & \text{d'où} & b_{21} &=& \dfrac{a_{21}}{b_{11}}, \\ \vdots & & & & \vdots & & \\ a_{n1} &=& b_{11}b_{n1}, & \text{d'où} & b_{n1} &=& \dfrac{a_{n1}}{b_{11}}, \end{array}$$

ce qui permet la détermination de la première colonne de B. Les coefficients de la j^{e} colonne de B, $2 \leq j \leq n$, s'obtiennent en utilisant les relations

$$\begin{array}{rcl@{\quad}rcl} a_{jj} &=& (b_{j1})^2 + (b_{j2})^2 + \cdots + (b_{jj})^2, & \text{d'où} & b_{jj} &=& \sqrt{a_{jj} - \sum_{k=1}^{j-1}(b_{jk})^2}, \\ a_{j+1j} &=& b_{j1}b_{j+11} + b_{j2}b_{j+12} + \cdots + b_{jj}b_{j+1j}, & \text{d'où} & b_{j+1j} &=& \dfrac{a_{j+1j} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}}, \\ \vdots & & & & \vdots & & \\ a_{nj} &=& b_{j1}b_{n1} + b_{j2}b_{n2} + \cdots + b_{jj}b_{nj}, & \text{d'où} & b_{nj} &=& \dfrac{a_{nj} - \sum_{k=1}^{j-1} b_{jk}b_{nk}}{b_{jj}}, \end{array}$$

après avoir préalablement déterminé les $j-1$ premières colonnes.

(a) On trouve $B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 2 & 0 & 1 & 4 \end{pmatrix}$.

(b) On trouve $B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -5 & 1 & 0 \\ 4 & 2 & 3 & 4 \end{pmatrix}$.

(c) On trouve $B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 3 & -2 & 1 & 0 \\ 1 & -\frac{3}{2} & -2 & \frac{\sqrt{3}}{2} \end{pmatrix}$.

Exercice 7.

1. La matrice A étant réelle symétrique, on peut se servir du critère de Sylvester. La matrice A est donc définie positive si et seulement si

$$\varepsilon > 0,\ \begin{vmatrix} \varepsilon & 1 \\ 1 & 3 \end{vmatrix} = 3\varepsilon - 1 > 0 \text{ et } \det(A) = 8\varepsilon - 11 > 0.$$

c'est-à-dire si et seulement si

$$\varepsilon > \frac{11}{8}.$$

2. Pour $\varepsilon = 0$, la matrice A n'est pas définie positive, ce qui élimine la factorisation de Cholesky. Le premier pivot de la matrice étant par ailleurs nul dans ce cas, on doit appliquer l'élimination de Gauss avec échange, c'est-à-dire une factorisation de type $PA = LU$.

3. a. On a $\varepsilon = 2 > \frac{11}{8}$. On trouve :

$$b_{11} = \sqrt{a_{11}} = \sqrt{2},\ b_{21} = \frac{a_{12}}{b_{11}} = \frac{1}{\sqrt{2}},\ b_{22} = \sqrt{a_{22} - {b_{21}}^2} = \sqrt{3 - \frac{1}{2}} = \sqrt{\frac{5}{2}},$$

$$b_{31} = \frac{a_{13}}{b_{11}} = \sqrt{2},\ b_{32} = \frac{a_{23} - b_{21}b_{31}}{b_{22}} = \frac{1 - \frac{\sqrt{2}}{\sqrt{2}}}{\sqrt{\frac{5}{2}}} = 0 \text{ et } b_{33} = \sqrt{a_{33} - {b_{31}}^2 - {b_{32}}^2} = \sqrt{3-2} = 1.$$

b. On a $BB^\top x = \begin{pmatrix}1\\1\\1\end{pmatrix}$ et on pose $y = B^\top x$, d'où $By = \begin{pmatrix}1\\1\\1\end{pmatrix}$. On trouve ainsi $y = \begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{2}\sqrt{\frac{2}{5}}\\ 0\end{pmatrix}$, puis $x = \begin{pmatrix}\frac{2}{5}\\ \frac{1}{5}\\ 0\end{pmatrix}$

Exercice 8 (factorisation QR).

1. On montre facilement que la matrice $A^\top A$ est réelle symétrique définie positive, elle admet donc une factorisation de Cholesky : il existe une matrice réelle B triangulaire inférieure à diagonale strictement positive telle que $A = BB^\top$. Il suffit alors de poser $R = B^\top$.
2. On a $A = (A^\top)^{-1}R^\top R$, ce qui conduit à poser $Q = (A^\top)^{-1}R^\top$. On a alors
$$QQ^\top = (A^\top)^{-1}R^\top R((A^\top)^{-1})^\top = (A^\top)^{-1}(A^\top A)A^{-1} = I_n.$$
3. On suppose qu'il existe deux factorisation QR de A : $A = Q_1R_1 = Q_2R_2$ et alors $R_1{R_2}^{-1} = {Q_1}^\top Q_2$. En posant $B = (R_1{R_2}^{-1})^\top$, on a que B est une matrice réelle triangulaire inférieure à diagonale strictement positive, telle que $BB^\top = {Q_2}^\top Q_1{Q_1}^\top Q_2 = I_n$. La dernière égalité montre que la matrice B est issue de la factorisation de Cholesky de la matrice identité, d'où $B = I_n$.
4. a. On a
$$A = QR \iff \forall (i,j) \in \{1,\dots,n\}^2,\ a_{ij} = \sum_{k=1}^{n} q_{ik}r_{kj} = \sum_{k=1}^{j} q_{ik}r_{kj},$$
puisque la matrice R est triangulaire supérieure. On en déduit que
$$\forall j \in \{1,\dots,n\},\ A_j = \sum_{k=1}^{j} r_{kj}Q_k.$$
b. On va utiliser le fait que la matrice Q est orthogonale, ses colonnes formant une famille orthonormée. Pour $j = 1$, on a
$$A_1 = r_{11}Q_1 \iff Q_1 = \frac{A_1}{r_{11}},$$
d'où $r_{11} = \|A_1\|_2$. Pour $j = 2$, on a
$$A_2 = r_{12}Q_1 + r_{22}Q_2 \iff Q_2 = \frac{A_2 - r_{12}Q_1}{r_{22}},$$
d'où $r_{12} = \langle A_2, Q_1\rangle_2$ et $r_{22} = \|A_2 - \langle A_2, Q_1\rangle_2 Q_1\|_2$.
En raisonnant par récurrence sur l'entier j, on trouve que
$$\forall j \in \{1,\dots,n\}, \forall k \in \{1,\dots,j-1\},\ r_{kj} = \left\langle A_j, Q_k\right\rangle_2,\ r_{jj} = \|A_j - \sum_{i=1}^{j-1}\left\langle A_j, Q_k\right\rangle_2\|_2.$$
La factorisation peut par conséquent être obtenue en appliquant le procédé d'orthonormalisation de Gram–Schmidt aux colonnes de la matrice A.

Dauphine | PSL
UNIVERSITÉ PARIS

Département MIDO

Méthodes numériques
année 2022/2023

Corrigés de travaux dirigés

Méthodes itératives pour la résolution de systèmes linéaires

Version du 22 février 2023.

Dans toute cette feuille, on identifie tout vecteur de $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, avec la matrice colonne de $M_{n,1}(\mathbb{K})$ ayant les mêmes éléments.

Exercice 1 (normes matricielles subordonnées).

1. La matrice A étant à coefficients dans le corps $\mathbb{C}$, il existe nécessairement un vecteur propre x associé à une valeur propre λ de plus grand module, *i.e.*, $\exists \boldsymbol{x} \neq \mathbf{0}$, $A\boldsymbol{x} = \lambda x$, avec $|\lambda| = \rho(A)$. On a alors, par propriétés d'une norme matricielle,
$$\rho(A)\|\boldsymbol{x}\|_p = \|\lambda \boldsymbol{x}\|_p = \|A\boldsymbol{x}\|_p \leq \|A\|_p \|\boldsymbol{x}\|_p,$$
d'où
$$\rho(A) \leq \|A\|_p.$$
Note : si la matrice A est à coefficients réels, ses valeurs propres peuvent être complexes et ses vecteurs propres ne sont alors pas nécessairement des vecteurs réels. Si c'est le cas pour la valeur propre associée au rayon spectral, la technique de preuve employée ci-dessus doit être modifiée.
Pour cela, on fait le choix d'une norme $\|\cdot\|$ sur $\mathbb{C}^n$ et l'on note de la même manière la norme subordonnée qui lui est associée sur $M_n(\mathbb{C})$. Bien entendu, la restriction de cette dernière norme à $M_n(\mathbb{R})$, encore notée $\|\cdot\|$, est une norme sur $M_n(\mathbb{R})$, équivalente à toute autre norme sur $M_n(\mathbb{R})$. En particulier, il existe une constante strictement positive C telle que
$$\forall M \in M_n(\mathbb{R}),\ \|M\| \leq C\,\|M\|_p.$$
Par un raisonnement par récurrence sur l'entier naturel k, on a d'une part $\rho(A)^k = \rho(A^k)$ et d'autre part $\|A^k\|_p \leq \|A\|_p{}^k$. En se servant de la majoration obtenue dans le cas complexe, on trouve alors que
$$\rho(A)^k = \rho(A^k) \leq \|A^k\| \leq C\,\|A^k\|_p \leq C\,\|A\|_p{}^k,$$
ce qui implique que
$$\rho(A) \leq C^{\frac{1}{k}}\,\|A\|_p.$$
Il suffit alors de faire tendre l'entier k vers $+\infty$ dans cette dernière inégalité, et d'utiliser que $\lim_{k\to+\infty} C^{\frac{1}{k}} = 1$, pour obtenir que
$$\rho(A) \leq \|A\|_p.$$

2. Pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, on a
$$\|A\boldsymbol{v}\|_1 = \sum_{i=1}^n \left|\sum_{j=1}^n a_{ij}v_j\right| \leq \sum_{j=1}^n |v_j| \sum_{i=1}^n |a_{ij}| \leq \left(\max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|\right)\|\boldsymbol{v}\|_1.$$
On construit ensuite un vecteur non nul (qui dépend de la matrice A) de sorte à avoir une égalité dans l'inégalité ci-dessus. Il suffit pour cela de considérer le vecteur de composantes
$$v_i = 0 \text{ pour } i \neq j_0,\ v_{j_0} = 1,$$
où j_0 est un indice vérifiant
$$\max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}| = \sum_{i=1}^n |a_{ij_0}|.$$

3. Pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, on a
$$\|A\boldsymbol{v}\|_\infty = \max_{1\leq i\leq n} \left|\sum_{j=1}^n a_{ij}v_j\right| \leq \left(\max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|\right)\|\boldsymbol{v}\|_\infty.$$

On obtient une égalité dans cette dernière inégalité en choisissant le vecteur $\boldsymbol{v}$ tel que

$$v_j = \frac{\overline{a_{i_0 j}}}{\left|a_{i_0 j}\right|} \text{ si } a_{i_0 j} \neq 0,\ v_j = 1 \text{ sinon,}$$

avec i_0 un indice vérifiant

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} \left|a_{ij}\right| = \sum_{j=1}^{n} \left|a_{i_0 j}\right|.$$

4. Puisque $U^*U = UU^* = I_n$, on a

$$\|UA\|_2{}^2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|UA\boldsymbol{v}\|_2{}^2}{\|\boldsymbol{v}\|_2{}^2} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\langle U^*UA\boldsymbol{v}, A\boldsymbol{v}\rangle}{\|\boldsymbol{v}\|_2{}^2} = \|A\|_2.$$

Le changement de variable $\boldsymbol{u} = U\boldsymbol{v}$ vérifiant $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2$, on a par ailleurs

$$\|AU\|_2{}^2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|AU\boldsymbol{v}\|_2{}^2}{\|\boldsymbol{v}\|_2{}^2} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2{}^2}{\|U^*\boldsymbol{u}\|_2{}^2} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2{}^2}{\|\boldsymbol{u}\|_2{}^2} = \|A\|_2.$$

On a enfin

$$\forall \boldsymbol{v}\in\mathbb{C}^n,\ \|U^*AU\boldsymbol{v}\|_2{}^2 = \langle U^*AU\boldsymbol{v}, U^*AU\boldsymbol{v}\rangle = \langle UU^*AU\boldsymbol{v}, AU\boldsymbol{v}\rangle = \langle AU\boldsymbol{v}, AU\boldsymbol{v}\rangle = \|AU\boldsymbol{v}\|_2,$$

dont on déduit la dernière égalité.

5. La matrice A^*A étant hermitienne, il existe une matrice unitaire U telle que la matrice U^*A^*AU est une matrice diagonale dont les éléments sont les valeurs propres, par ailleurs positives, μ_i, $i = 1,\dots,n$, de A^*A. En posant $\boldsymbol{w} = U^*\boldsymbol{v}$, on a alors

$$\|A\|_2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \sqrt{\frac{\langle A^*A\boldsymbol{v}, \boldsymbol{v}\rangle}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\frac{\langle U^*A^*AU\boldsymbol{w}, \boldsymbol{w}\rangle}{\langle \boldsymbol{w}, \boldsymbol{w}\rangle}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\sum_{i=1}^{n} \mu_i \frac{|w_i|^2}{\sum_{j=1}^{n}\left|w_j\right|^2}} = \sqrt{\max_{1\leq i\leq n} \mu_i}.$$

D'autre part, en utilisant l'inégalité de Cauchy–Schwarz, on trouve, pour tout vecteur $\boldsymbol{v}$ non nul,

$$\frac{\|A\boldsymbol{v}\|_2{}^2}{\|\boldsymbol{v}\|_2{}^2} = \frac{\langle A^*A\boldsymbol{v}, \boldsymbol{v}\rangle}{\|\boldsymbol{v}\|_2{}^2} \leq \frac{\|A^*A\boldsymbol{v}\|_2\|\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2{}^2} \leq \|A^*A\|_2 \leq \|A^*\|_2\|A\|_2,$$

d'où $\|A\|_2 \leq \|A^*\|_2$. En appliquant cette inégalité à A^*, on obtient l'égalité $\|A\|_2 = \|A^*\|_2 = \sqrt{\rho(AA^*)}$.

Exercice 2 (rayon spectral et série de Neumann). Note : l'énoncé ne le précisant pas, on suppose que les coefficients de la matrice A sont à valeurs dans un corps commutatif sur lequel tout polynôme est scindé, comme $\mathbb{C}$. Ceci simplifie la preuve de la première question, puisque l'existence d'un vecteur propre associé à la valeur propre de plus grand module est assurée dans ce cas.

1. Montrons l'implication $\lim_{k\to+\infty} A^k = 0 \Rightarrow \rho(A) < 1$. Pour cela, supposons que $\lim_{k\to+\infty} A^k = 0$ et raisonnons par l'absurde. Pour tout couple de valeur et vecteur propres (λ, v) de A, on a, pour tout entier naturel k non nul, $\|A^k v\| = \|\lambda^k v\| = |\lambda|^k \|v\|$, d'où $|\lambda|^k \leq \|A^k\|$. Cette inégalité est en particulier vraie pour la valeur propre de plus grand module et on a donc $(\rho(A))^k \leq \|A^k\|$ pour tout entier naturel k non nul. Ainsi, si $\rho(A) \geq 1$, alors on a $\|A^k\| \geq 1$ et la suite réelle $(\|A^k\|)_{k\in\mathbb{N}}$ ne peut converger vers 0. Il en découle que la suite de matrices $(A^k)_{k\in\mathbb{N}}$ ne peut converger vers la matrice nulle, d'où une contradiction.

   Prouvons ensuite l'implication $\rho(A) < 1 \to \lim_{k\to+\infty} A^k = 0$. Supposons que $\rho(A) < 1$. il existe alors un réel ε strictement positif tel que $\rho(A) + \varepsilon < 1$ (il suffit par exemple de prendre $\varepsilon = \frac{1}{2}(1-\rho(A))$). Par un résultat du cours, il existe une norme matricielle $|||\cdot|||$, dépendant de A et de ε, telle que

$$|||A||| \leq \rho(A) + \varepsilon < 1.$$

   Puisque $|||A||| < 1$, la suite réelle $(|||A|||^k)_{k\in\mathbb{N}}$ converge vers 0 et, comme $\left|\left|\left|A^k\right|\right|\right| \leq |||A|||^k$ pour tout entier naturel k, la suite réelle $(\left|\left|\left|A^k\right|\right|\right|)_{k\in\mathbb{N}}$ également. Ceci signifie encore que $\lim_{k\to+\infty} A^k = 0$.

2. On a $\sigma(I_n - A) = \{1-\lambda \,|\, \lambda\in\sigma(A)\}$ et $\sigma(I_n + A) = \{1+\lambda \,|\, \lambda\in\sigma(A)\}$. Si $\rho(A) < 1$, on a alors, pour toute valeur propre λ de A, $0 < 1 - |\lambda| \leq |1 \pm \lambda| \leq 1 + |\lambda|$. Leurs valeurs propres étant non nulles, les matrices $I_n - A$ et $I_n + A$ sont inversibles.

3. Montrons l'implication $\sum_{k=0}^{+\infty} A^k$ converge $\Rightarrow \rho(A) < 1$. Si la série est convergente, alors, pour tout tout couple de valeur et vecteur propres (λ, v) de A, on a $\left(\sum_{k=0}^{+\infty} A^k\right) v = \left(\sum_{k=0}^{+\infty} \lambda^k\right) v$. La série $\sum_{k=0}^{+\infty} \lambda^k$ est donc convergente pour toute valeur propre λ de A, ce qui implique que $|\lambda| < 1$. On en déduit en particulier que $\rho(A) < 1$.

Montrons enfin l'implication $\rho(A) < 1 \Rightarrow \sum_{k=0}^{+\infty} A^k$ converge. Supposons que $\rho(A) < 1$. On sait d'après la question précédente que la matrice $I_n - A$ est inversible. Posons alors

$$S_k = I_n + A + \cdots + A^k.$$

On a alors

$$AS_k = A + A^2 + \cdots + A^{k+1}$$

d'où

$$(I_n - A)S_k = I_n - A^{k+1}.$$

En faisant tendre l'entier k vers $+\infty$ dans cette dernière égalité et en utilisant la première question, on obtient que

$$(I_n - A) \lim_{k\to+\infty} S_k = I_n, \text{ soit encore } \lim_{k\to+\infty} S_k = (I_n - A)^{-1}, \text{ c'est-à-dire } \sum_{k=0}^{+\infty} A^k = (I_n - A)^{-1}.$$

Exercice 3 (convergence de méthodes itératives pour les matrices à diagonale strictement dominante). Prouvons par l'absurde que la matrice A est inversible. Si elle est non inversible, alors son noyau n'est pas réduit à zéro et il existe un vecteur $\boldsymbol{x}$ de $\mathbb{R}^n$ non nul tel que $A\boldsymbol{x} = \mathbf{0}$. Ceci implique que

$$\forall i \in \{1, \ldots, n\}, \ \sum_{j=1}^{n} a_{ij}\, x_j = 0.$$

Le vecteur $\boldsymbol{x}$ étant non nul, il existe un indice i_0 dans $\{1, \ldots, n\}$ tel que $0 \neq \left|x_{i_0}\right| = \max_{1\leq i\leq n} |x_i|$ et l'on a alors

$$-a_{i_0 i_0} x_{i_0} = \sum_{\substack{j=1 \\ j\neq i_0}}^{n} a_{i_0 j} x_j,$$

d'où

$$\left|a_{i_0 i_0}\right| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} \left|a_{i_0 j}\right| \frac{\left|x_j\right|}{\left|x_{i_0}\right|} \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} \left|a_{i_0 j}\right|,$$

ce qui contredit le fait que A est à diagonale strictement dominante par lignes.

Montrons maintenant que les méthodes sont convergentes. Pour la méthode de Jacobi, on pose

$$r = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|,$$

et on observe alors que $\|B_J\|_\infty = r < 1$. Il s'ensuit que la méthode est convergente.

Pour la méthode de Gauss–Seidel, on considère l'erreur à l'itération $k+1$, $k \in \mathbb{N}$, qui vérifie

$$e_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}\, e_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}\, e_j^{(k)}, \ 1 \leq i \leq n.$$

On va maintenant établir que

$$\forall k \in \mathbb{N}, \ \|\boldsymbol{e}^{(k+1)}\|_\infty \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty,$$

en raisonnant par récurrence sur l'indice i, $1 \leq i \leq n$, des composantes du vecteur. Pour $i = 1$, on a

$$e_1^{(k+1)} = -\sum_{j=2}^{n} \frac{a_{1j}}{a_{11}}\, e_j^{(k)}, \text{ d'où } |e_1^{(k+1)}| \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty.$$

Supposons que $|e_j^{(k+1)}| \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty$ pour $j = 1, \ldots, i-1$. On a alors

$$|e_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k+1)}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k)}| \leq \|\boldsymbol{e}^{(k)}\|_\infty \left(r \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \right) < \|\boldsymbol{e}^{(k)}\|_\infty \sum_{\substack{i=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|,$$

d'où $|e_i^{(k+1)}| \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty$, ce qui achève la preuve par récurrence. On a par conséquent

$$\|\boldsymbol{e}^{(k)}\|_\infty \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty \leq \cdots \leq r^k\, \|\boldsymbol{e}^{(0)}\|_\infty,$$

et, par suite,

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

ce qui prouve la convergence de la méthode.

Exercice 4.

1. La matrice A est à diagonale strictement dominante par lignes, les méthodes de Gauss–Seidel et de Jacobi sont donc convergentes pour cette matrice.
2. D'une part, on a
$$B_J = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{pmatrix}.$$
Les valeurs propres de cette matrice sont 0 et $\pm\frac{\sqrt{2}}{4}$ et $\rho(B_J)$ vaut donc $\frac{\sqrt{2}}{4}$. D'autre part, on a
$$B_{GS} = \begin{pmatrix} 4 & 0 & 0 \\ -1 & 4 & 0 \\ 0 & -1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ \frac{1}{16} & \frac{1}{4} & 0 \\ \frac{1}{64} & \frac{1}{16} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{16} & \frac{1}{4} \\ 0 & \frac{1}{64} & \frac{1}{16} \end{pmatrix}.$$
Les valeurs propres de cette matrice sont 0 et $\frac{1}{8}$ (d'ordre de multiplicité algébrique égal à deux) et $\rho(B_{GS})$ vaut donc $\frac{1}{8}$. On trouve bien l'égalité attendue et la méthode convergeant le plus rapidement est la méthode de Gauss-Seidel, puisque c'est celle dont la matrice d'itération a le plus petit rayon spectral.

Exercice 5. Les valeurs propres de la matrice A sont α et $\alpha \pm 1$. Cette matrice est donc inversible si et seulement si $\alpha \notin \{-1, 0, 1\}$. Pour $|\alpha| > 1$, la matrice est à diagonale strictement dominante par lignes et les méthodes de Jacobi et Gauss–Seidel sont alors convergentes. Si $0 < |\alpha| < 1$, il faut étudier les matrices d'itération respectives des méthodes, et plus particulièrement déterminer leur rayons spectraux.

Pour la méthode de Jacobi, on a
$$B_J = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -\frac{1}{\alpha} \\ 0 & 0 & 0 \\ -\frac{1}{\alpha} & 0 & 0 \end{pmatrix}.$$
Les valeurs propres de cette matrice sont 0 et $\pm\frac{1}{\alpha}$ et $\rho(B_J)$ vaut donc $\frac{1}{|\alpha|} > 1$. Cette méthode est donc divergente dans ce cas.

Pour la méthode de Gauss–Seidel, on a
$$B_{GS} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 1 & 0 & \alpha \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha} & 0 & 0 \\ 0 & \frac{1}{\alpha} & 0 \\ -\frac{1}{\alpha^2} & 0 & \frac{1}{\alpha} \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -\frac{1}{\alpha} \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha^2} \end{pmatrix}.$$
Les valeurs propres de cette matrice sont 0 (d'ordre de multiplicité algébrique égal à deux) et $\frac{1}{\alpha^2}$, et $\rho(B_{GS})$ vaut donc $\frac{1}{\alpha^2} > 1$. Cette méthode est également divergente dans ce cas.

Exercice 6.

1. Pour $\alpha = 0$, on a $A_0 = 2\,I_3$, qui est définie positive. Si $\alpha \neq 0$, on pose $\alpha\mu = 2 - \lambda$ et l'on a
$$\det(A_\alpha - \lambda\, I_3) = \alpha^3 \begin{vmatrix} \mu & 1 & 0 \\ 1 & \mu & 1 \\ 0 & 1 & \mu \end{vmatrix} = \alpha^3 \mu(\mu^2 - 2),$$
d'où $\det(A_\alpha - \lambda\, I_3) = \alpha^3\mu(\mu+\sqrt{2})(\mu-\sqrt{2})$. Les valeurs propres de A_α sont donc obtenues pour $\mu = 0, \pm\sqrt{2}$, c'est-à-dire pour $\lambda = 2, 2 \mp \alpha\sqrt{2}$. La matrice est donc définie positive si et seulement si $-\sqrt{2} < \alpha < \sqrt{2}$.
Pour $\beta = 0$, on a $C_0 = I_3$, qui est définie positive. Si $\beta \neq 0$, on pose $\beta\eta = 1 - \lambda$ et l'on a
$$\det(C_\beta - \lambda\, I_3) = \beta^3 \begin{vmatrix} \eta & 1 & 1 \\ 1 & \eta & 1 \\ 1 & 1 & \eta \end{vmatrix} = \beta^3(\eta^3 - 3\,\eta + 2),$$
d'où $\det(C_\beta - \lambda\, I_3) = \beta^3(\eta + 2)(\eta - 1)^2$. Les valeurs propres de C_β sont donc obtenues pour $\eta = -2, 1$, c'est-à-dire pour $\lambda = 1 + 2\,\beta, 1 - \beta$. La matrice est donc définie positive si et seulement si $-\frac{1}{2} < \beta < 1$.
2. La matrice d'itération de la méthode de Jacobi associée à A_α est
$$B_J = -\frac{\alpha}{2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$
Ses valeurs propres sont 0 et $\pm\frac{\sqrt{2}}{2}\alpha$ et son rayon spectral vaut $\rho(B_J) = \frac{\sqrt{2}}{2}|\alpha|$. La méthode converge donc si $-\sqrt{2} < \alpha < \sqrt{2}$.

La matrice d'itération de la méthode de Jacobi associée à C_β est

$$B_J = -\beta \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Ses valeurs propres sont $-\beta$ et 2β et son rayon spectral vaut $\rho(B_J) = 2\,|\beta|$. La méthode converge donc si $-\frac{1}{2} < \beta < \frac{1}{2}$.

3. La matrice d'itération de la méthode de Gauss–Seidel associée à A_α est

$$B_{GS} = \begin{pmatrix} 2 & 0 & 0 \\ \alpha & 2 & 0 \\ 0 & \alpha & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{\alpha}{4} & \frac{1}{2} & 0 \\ \frac{\alpha^2}{8} & -\frac{\alpha}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix} = -\frac{\alpha}{2} \begin{pmatrix} 0 & 1 & 0 \\ 0 & -\frac{\alpha}{2} & 1 \\ 0 & -\frac{\alpha}{2} & -\frac{\alpha^2}{4} \end{pmatrix}.$$

La matrice A_α étant tridiagonale, on sait que $\rho(B_{GS}) = \rho(B_J)^2$. On déduit alors de la précédente question que la méthode converge si et seulement si $-\sqrt{2} < \alpha < \sqrt{2}$.

Exercice 7.

1. La matrice d'itération de la méthode de Jacobi associée à la matrice A est

$$B_J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}.$$

On a

$$\chi_{B_J}(X) = \begin{vmatrix} X & 2 & -2 \\ 1 & X & 1 \\ 2 & 2 & X \end{vmatrix} = X^3,$$

d'où $\rho(B_J) = 0$.
La matrice d'itération de la méthode de Jacobi associée à la matrice A est

$$B_{GS} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}.$$

On a

$$\chi_{B_{GS}}(X) = \begin{vmatrix} X & 2 & -2 \\ 0 & X-2 & 3 \\ 0 & 0 & X-2 \end{vmatrix} = X(X-2)^2,$$

d'où $\rho(B_{GS}) = 2$.
On a donc $\rho(B_J) < 1 < \rho(B_{GS})$.

2. De la même manière, la matrice d'itération de la méthode de Jacobi associée à la matrice A est

$$B_J = \frac{1}{2} \begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{pmatrix}.$$

On a

$$\chi_{B_J}(X) = \frac{1}{8} \begin{vmatrix} 2X & -1 & 1 \\ 2 & 2X & 2 \\ -1 & -1 & 2X \end{vmatrix} = \frac{1}{8}(8X^3 + 10X) = X\left(X + \mathrm{i}\frac{\sqrt{5}}{2}\right)\left(X - \mathrm{i}\frac{\sqrt{5}}{2}\right),$$

d'où $\rho(B_J) = \frac{\sqrt{5}}{2} \approx 1,118$.
La matrice d'itération de la méthode de Jacobi associée à la matrice A est

$$B_{GS} = \frac{1}{2} \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix}.$$

On a

$$\chi_{B_{GS}}(X) = \frac{1}{8} \begin{vmatrix} 2X & -1 & 1 \\ 0 & 2X+1 & 1 \\ 0 & 0 & 1+2X \end{vmatrix} = \frac{1}{4} X(2X+1)^2,$$

d'où $\rho(B_{GS}) = \frac{1}{2}$.
On a donc $\rho(B_{GS}) < 1 < \rho(B_J)$.

Exercice 8 (une méthode de relaxation).

1. En isolant $x^{(k+\frac{1}{2})}$ dans la première équation et en subtituant dans la seconde équation, on trouve

$$x^{(k+1)} = (\omega(D-E)^{-1}F + (1-\omega)I_n)\,x^{(k)} + \omega(D-E)^{-1}b,$$

d'où $B(\omega) = \omega(D-E)^{-1}F+(1-\omega)I_n$ et $c(\omega) = \omega(D-E)^{-1}b$. Pour $\omega = 1$, on a $B(1) = (D-E)^{-1}F$ et $c(1) = (D-E)^{-1}b$, ce qui correspond bien à la méthode de Gauss-Seidel.

2. Pour ces choix de A et b, on a

$$B(\omega) = \omega\begin{pmatrix}2 & 0\\ 0 & 5\end{pmatrix}^{-1}\begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix} + (1-\omega)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} = \begin{pmatrix}1-\omega & -\frac{\omega}{2}\\ 0 & 1-\omega\end{pmatrix}.$$

On a $\chi_{B(\omega)}(X) = (X-(1-\omega))^2$, d'où $\rho(B(\omega)) = |1-\omega|$. La méthode est donc convergente si et seulement si $0 < \omega < 2$.

Exercice 9.

1. On a

$$\det(A) = \begin{vmatrix}1 & 2 & 0\\ 1 & 1 & 0\\ 1 & 1 & 1\end{vmatrix} = -1,$$

donc la matrice A est inversible.

2. On a

$$\det\left(\frac{1}{\omega}I_3 - E\right)) = \begin{vmatrix}\frac{1}{\omega} & 2 & 0\\ 1 & \frac{1}{\omega} & 0\\ 0 & 0 & \frac{1}{\omega}\end{vmatrix} = \frac{1}{\omega}\left(\frac{1}{\omega^2} - 2\right).$$

Le paramètre ω appartenant à l'intervalle $]0,2[$, ce déterminant est non nul si et seulement si $\frac{1}{\omega^2} - 2$ est non nul, c'est-à-dire si $\omega \neq \frac{\sqrt{2}}{2}$.

3. On a

$$\chi_{B(\omega)}(X) = \det\left(X\,I_3 - \left(\frac{1}{\omega}I_3 - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}I_3\right)\right) = \frac{\det((X-1+\omega)I_3 - X\omega E - \omega F)}{\omega^3\det\left(\frac{1}{\omega}I_3 - E\right)}.$$

Les valeurs propres de $B(\omega)$ sont donc les racines de $\det((X-1+\omega)I_3 - X\omega E - \omega F)$. Un calcul donne

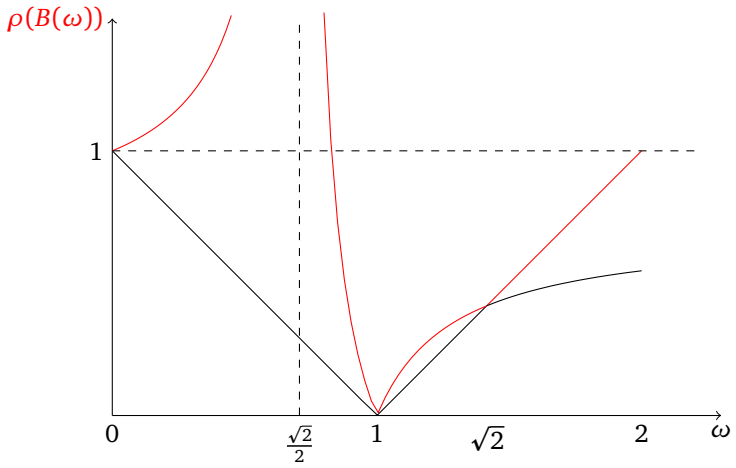
$$\begin{vmatrix}X-1+\omega & 2\omega X & 0\\ \omega X & X-1+\omega & 0\\ \omega & \omega & X-1+\omega\end{vmatrix} = (X-1+\omega)((1+\sqrt{2}\omega)X-1+\omega)((1-\sqrt{2}\omega)X-1+\omega).$$

Le spectre de $B(\omega)$ est donc l'ensemble $\left\{1-\omega, \frac{1-\omega}{1+\sqrt{2}\omega}, \frac{1-\omega}{1-\sqrt{2}\omega}\right\}$. Pour tout ω appartenant à $]0,2[$, on remarque que $\left|1+\sqrt{2}\omega\right| > 1$ et donc $|1-\omega| > \left|\frac{1-\omega}{1+\sqrt{2}\omega}\right|$. On doit donc comparer $|1-\omega|$ et $\left|\frac{1-\omega}{1-\sqrt{2}\omega}\right|$. Pour cela, on réalise l'étude de ces deux fonctions :

ω	0		1		2
$\lvert 1-\omega\rvert$	1	↘	0	↗	1

ω	0		$\frac{\sqrt{2}}{2}$		1		2
$\left\lvert\frac{1-\omega}{1-\sqrt{2}\omega}\right\rvert$	1	↗	$+\infty$	↘	0	↗	$\frac{1}{2\sqrt{2}-1}$

Comme on le voit sur le dessin ci-dessous, les graphes de ces deux fonctions se coupent en un point dont l'abscisse α est strictement comprise entre 1 et 2 et vérifie $|1-\omega| = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right|$, c'est-à-dire pour $\alpha = \sqrt{2}$.



On en déduit donc que

$$\rho(B(\omega)) = \begin{cases} \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| & \text{si } \omega \in]0, \sqrt{2}], \\ \omega - 1 & \text{si } \omega \in [\sqrt{2}, 2[. \end{cases}$$

4. La méthode étant convergente si et seulement si $\rho(B(\omega)) < 1$. Pour ω appartenant à $\left]0, \frac{\sqrt{2}}{2}\right[$, la méthode n'est pas convergente. Pour ω appartenant à $\left]\frac{\sqrt{2}}{2}, 1\right[$, la convergence de la méthode équivaut à avoir

$$\frac{1-\omega}{\sqrt{2}\omega - 1} < 1, \text{ soit encore } \frac{2}{1+\sqrt{2}} < \omega.$$

Pour ω appartenant à $[1, \sqrt{2}[$, la convergence équivaut à avoir $\omega - 1 < \sqrt{2}\omega - 1$, soit encore $1 < \sqrt{2}$, ce qui est le cas. Enfin, pour ω appartenant à $[\sqrt{2}, 2[$, la convergence équivaut à avoir $\omega - 1 < 1$, soit encore $\omega < 2$, ce qui n'implique aucune contrainte supplémentaire sur le paramètre.

En conclusion, la méthode est convergente pour ω appartenant $\left]\frac{2}{1+\sqrt{2}}, 2\right[$

5. Le minimum de $\rho(B(\omega))$ est atteint en $\omega_0 = 1$ et vaut 0.

Exercice 10 (méthode de Richardson).

1. Supposons que la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ converge et notons x^* sa limite. En passant alors à la limite dans la relation de récurrence définissant la méthode, il vient

$$x^* = x^* - \alpha(b - Ax^*),$$

soit encore $Ax^* = b$, puisque le réel α est non nul.

2. La matrice d'itération de la méthode est $B_R(\alpha) = I_n - \alpha A$ et on sait que la méthode est convergente si et seulement si son rayon spectral est strictement inférieur à 1. Les valeurs propres de cette matrice sont par ailleurs de la forme $1 - \alpha\lambda$, avec λ une valeur propre de A. Les lignes de la matrice A étant linéairement dépendantes, 0 est une valeur propre de A et 1 est donc une valeur propre de $B_R(\alpha)$ pour toute valeur de α. Le rayon spectral de $B_R(\alpha)$ est donc supérieur ou égal à 1, quelle que soit la valeur de α.

3. Pour que la méthode soit convergente, il faut et il suffit que

$$\forall \lambda \in \sigma(A),\ -1 < 1 - \alpha\lambda < 1 \Leftrightarrow 0 < \alpha\lambda < 2.$$

Si la matrice A est symétrique définie positive, son spectre est contenu dans $\mathbb{R}_+^*$, et cette condition devient

$$0 < \alpha < \frac{2}{\rho(A)}.$$

Exercice 11 (convergence de la méthode de Richardson pour une matrice symétrique définie positive).

1. Soit x un vecteur de $\mathrm{Ker}(A)$. On a $Ax = 0$, ce qui implique que $0 = \langle Ax, x\rangle \geq c\,\|x\|_2^{\,2}$. On en déduit que $\|x\|_2 = 0$ et donc que $x = 0$. Le noyau de la matrice A est donc réduit au vecteur nul et celle-ci est inversible.

2. a. On a, par les définitions de la méthode et du résidu,

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} + \alpha\, r^{(k)}$$

d'où

$$\forall k \in \mathbb{N},\ Ax^{(k+1)} - b = Ax^{(k)} - b + \alpha A r^{(k)} \iff r^{(k+1)} = r^{(k)} - \alpha A r^{(k)} = (I_n - \alpha A) r^{(k)}.$$

Par un raisonnement par récurrence évident, on montre alors que, pour tout entier naturel k, on a $r^{(k)} = (I_n - \alpha A)^k r^{(0)}$.

b. Le paramètre α étant strictement positif, on obtient en développant

$$\begin{aligned}\|(I_n - \alpha A)x\|_2^{\,2} &= \langle x - \alpha A\,x, x - \alpha A\,x\rangle \\ &= \|x\|_2^{\,2} - 2\alpha\,\langle Ax, x\rangle + \alpha^2\|Ax\|_2^{\,2} \\ &\leq \|x\|_2^{\,2} - 2\alpha\,\langle Ax, x\rangle + \alpha^2\|A\|_2^{\,2}\|x\|_2^{\,2} \\ &= 1 - 2\alpha\,\langle Ax, x\rangle + \alpha^2\|A\|_2^{\,2} \\ &\leq 1 - 2\alpha c + \alpha^2\|A\|_2^{\,2}.\end{aligned}$$

c. Par définition d'une norme subordonnée, avoir $\|I_n - \alpha A\|_2 < 1$ équivaut à avoir $\|x - \alpha A\,x\|_2 < 1$ pour tout vecteur x de norme euclidienne unitaire. D'après la précédente question, il est pour cela suffisant que le paramètre α soit tel que

$$1 - 2\alpha c + \alpha^2\|A\|_2^{\,2} < 1,$$

c'est-à-dire que $0 < \alpha < \frac{2c}{\|A\|_2^{\,2}}$.

3. La matrice A étant symétrique, la matrice d'itération de la méthode de Richardson, $I_n - \alpha A$ est symétrique et l'on a $\rho(I_n - \alpha A) = \|I_n - \alpha A\|_2$. La condition obtenue dans la question précédente est donc une condition suffisante de convergence.

Dauphine | PSL
UNIVERSITÉ PARIS

Département MIDO

Méthodes numériques
année 2022/2023

Corrigés de travaux dirigés

Calcul de valeurs et de vecteurs propres

Version du 22 février 2023.

Exercice 1 (localisation des valeurs propres).

1. Supposons que le nombre complexe λ soit une valeur propre de A. Il existe alors un vecteur non nul $\boldsymbol{v}$ tel que $A\boldsymbol{v} = \lambda\boldsymbol{v}$, c'est-à-dire
$$\forall i \in \{1,\dots,n\},\ \sum_{j=1}^{n} a_{ij} v_j = \lambda\, v_i.$$
Soit v_k, $k \in \{1,\dots,n\}$, la composante de $\boldsymbol{v}$ ayant le plus grand module (ou l'une des composantes de plus grand module s'il y en a plusieurs). On a d'une part $v_k \neq 0$, puisque $\boldsymbol{v}$ est non nul par hypothèse, et d'autre part
$$|\lambda - a_{kk}|\,|v_k| = |\lambda v_k - a_{kk} v_k| = \left|\sum_{j=1}^{n} a_{kj} v_j - a_{kk} v_k\right| = \left|\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj} v_j\right| \leq |v_k| \left|\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}\right| \leq |v_k| \sum_{\substack{j=1\\ j\neq k}}^{n} \left|a_{kj}\right|,$$
ce qui prouve, après division par $|v_k|$, que la valeur propre λ est contenue dans le disque de Gershgorin D_k, d'où le résultat.
2. La transposée $A^\top$ de A possède le même spectre que A. On peut donc facilement améliorer le résultat précédent : *Si A est une matrice d'ordre n, alors*
$$\sigma(A) \subseteq \left(\bigcup_{i=1}^{n} D_i\right) \cap \left(\bigcup_{j=1}^{n} D'_j\right),$$
où les ensembles D_i, $i = 1,\dots,n$, sont définis par
$$D_i = \left\{ z \in \mathbb{C} \,|\, |z - a_{ii}| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} \left|a_{ij}\right| \right\},$$
et les ensembles D'_j, $j = 1,\dots,n$, sont tels que
$$D'_j = \left\{ z \in \mathbb{C} \,|\, \left|z - a_{jj}\right| \leq \sum_{\substack{i=1\\ i\neq j}}^{n} \left|a_{ij}\right| \right\}.$$

Exercice 2 (déflation de Wielandt).

1. On a
$$\widetilde{A}\boldsymbol{v}_j = A\boldsymbol{v}_j - \boldsymbol{v}_j \boldsymbol{u}_j{}^\top \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j - \lambda_j \boldsymbol{v}_j = \boldsymbol{0}.$$
2. On a
$$\widetilde{A}(\boldsymbol{v}_i + \alpha_i \boldsymbol{v}_j) = A\boldsymbol{v}_i + \alpha_i A\boldsymbol{v}_j - \boldsymbol{v}_j \boldsymbol{u}_j{}^\top \boldsymbol{v}_i - \alpha_i \boldsymbol{v}_j \boldsymbol{u}_j{}^\top \boldsymbol{v}_j = \lambda_i \boldsymbol{v}_i + \alpha_i \lambda_j \boldsymbol{v}_j - \boldsymbol{v}_j \boldsymbol{u}_j{}^\top \boldsymbol{v}_i - \alpha_i \lambda_j \boldsymbol{v}_j = \lambda_i \left(\boldsymbol{v}_i - \frac{\boldsymbol{u}_j{}^\top \boldsymbol{v}_i}{\lambda_i} \boldsymbol{v}_j \right),$$
puisque $\lambda_i \neq 0$ (A étant inversible). Le vecteur $\boldsymbol{v}_i + \alpha_i \boldsymbol{v}_j$ est donc un vecteur propre de $\widetilde{A}$, associé à λ_i, si $\alpha_i = -\frac{\boldsymbol{u}_j{}^\top \boldsymbol{v}_i}{\lambda_i}$.
3. Les valeurs propres de $\widetilde{A}$ sont donc λ_i, pour tout i appartenant à $\{1,\dots,n\}\backslash\{j\}$, de vecteur propre associé $\boldsymbol{v}_i - \frac{\boldsymbol{u}_j{}^\top \boldsymbol{v}_i}{\lambda_i} \boldsymbol{v}_j$, et 0, de vecteur propre associé $\boldsymbol{v}_j$.

4. Les vecteurs propres $\boldsymbol{v}_i$, $i = 1, \ldots, n$, de la matrice A étant supposés former une base orthogonale, on a pour ce choix ${\boldsymbol{u}_j}^\top \boldsymbol{v}_i = \lambda_j \delta_{ij}$. Dans ce cas particulier, les vecteurs propres de $\widetilde{A}$ sont exactement ceux de A : c'est la déflation de Hotelling, vue en cours.

Exercice 3 (convergence de la méthode de la puissance pour une matrice réelle symétrique). La matrice A étant réelle symétrique, ses valeurs propres sont réelles et il existe une base orthonormée $\{\boldsymbol{v}_j\}_{j=1,\ldots,n}$ formée de vecteurs propres de A. Par définition de la méthode, on a, en utilisant la décomposition du vecteur unitaire $\boldsymbol{q}^{(0)}$ dans la base $\{\boldsymbol{v}_j\}_{j=1,\ldots,n}$ ($\boldsymbol{q}^{(0)} = \sum_{j=1}^n \alpha_j \boldsymbol{v}_j$ avec $\sum_{j=1}^n {\alpha_j}^2 = 1$) et le fait que A est symétrique,

$$\forall k \in \mathbb{N},\ \nu^{(k)} = (\boldsymbol{q}^{(k)})^\top A \boldsymbol{q}^{(k)} = \frac{(\boldsymbol{q}^{(0)})^\top A^{2k+1} \boldsymbol{q}^{(0)}}{(\boldsymbol{q}^{(0)})^\top A^{2k} \boldsymbol{q}^{(0)}} = \frac{\sum_{j=1}^n {\alpha_j}^2 {\lambda_j}^{2k+1}}{\sum_{j=1}^n {\alpha_j}^2 {\lambda_j}^{2k}},$$

et donc

$$\forall k \in \mathbb{N},\ \left|\nu^{(k)} - \lambda_n\right| = \left|\frac{\sum_{j=1}^{n-1} {\alpha_j}^2 {\lambda_j}^{2k}(\lambda_j - \lambda_n)}{\sum_{j=1}^n {\alpha_j}^2 {\lambda_j}^{2k}}\right| \leq \max_{i \in \{1,\ldots,n-1\}} |\lambda_i - \lambda_n| \frac{\sum_{j=1}^{n-1} {\alpha_j}^2 {\lambda_j}^{2k}}{\sum_{j=1}^n {\alpha_j}^2 {\lambda_j}^{2k}}.$$

Par hypothèse sur $\boldsymbol{q}^{(0)}$, il vient alors

$$\frac{\sum_{j=1}^{n-1} {\alpha_j}^2 {\lambda_j}^{2k}}{\sum_{j=1}^n {\alpha_j}^2 {\lambda_j}^{2k}} \leq \frac{\sum_{j=1}^{n-1} {\alpha_j}^2 {\lambda_j}^{2k}}{{\alpha_n}^2 {\lambda_n}^{2k}} \leq \frac{1}{{\alpha_n}^2}\left(\sum_{j=1}^{n-1} {\alpha_j}^2\right)\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k} = \frac{1 - {\alpha_n}^2}{{\alpha_n}^2}\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k},$$

d'où l'inégalité voulue.

Exercice 4 (réduction d'une matrice symétrique à la forme tridiagonale par la méthode de Householder).

1. — La matrice $H(\boldsymbol{u})$ est symétrique car

$$H(\boldsymbol{u})^\top = {I_n}^\top - \frac{2}{{\|\boldsymbol{u}\|_2}^2}(\boldsymbol{u}\boldsymbol{u}^\top)^\top = I_n - \frac{2}{{\|\boldsymbol{u}\|_2}^2}(\boldsymbol{u}^\top)^\top \boldsymbol{u}^\top = H(\boldsymbol{u}).$$

— La matrice $H(\boldsymbol{u})$ est orthogonale car

$$H(\boldsymbol{u})^\top H(\boldsymbol{u}) = I_n - \frac{4}{{\|\boldsymbol{u}\|_2}^2}\boldsymbol{u}\boldsymbol{u}^\top - \frac{4}{{\|\boldsymbol{u}\|_2}^4}(\boldsymbol{u}\boldsymbol{u}^\top)^2 = I_n$$

puisque $\boldsymbol{u}^\top \boldsymbol{u} = {\|\boldsymbol{u}\|_2}^2$.

— La matrice $H(\boldsymbol{u})$ est inversible, car elle est orthogonale, et on a

$$H(\boldsymbol{u})^{-1} = H(\boldsymbol{u})^\top = H(\boldsymbol{u}).$$

— Enfin, on a

$$H(\boldsymbol{u})\boldsymbol{u} = \boldsymbol{u} - \frac{2}{{\|\boldsymbol{u}\|_2}^2}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{u} = \boldsymbol{u} - 2\,\boldsymbol{u} = -\boldsymbol{u}.$$

2. a. Un calcul donne :

$$\boldsymbol{c} = \begin{pmatrix} 0 \\ a_{21} + \mathrm{sgn}(a_{21})\|\boldsymbol{a}\|_2 \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix}.$$

Si $\boldsymbol{c} = \boldsymbol{0}$, ceci implique que $a_{i1} = 0$ pour tout entier i appartenant à $\{3, \ldots, n\}$.

b. i. On a $\langle \boldsymbol{u}, \boldsymbol{e}_1 \rangle = \boldsymbol{u}^\top \boldsymbol{e}_1 = \frac{1}{\|\boldsymbol{c}\|_2}\boldsymbol{c}^\top \boldsymbol{e}_1 = 0$ d'après l'expression obtenue dans la précédente question. On a ainsi que $H(\boldsymbol{u})\boldsymbol{e}_1 = \boldsymbol{e}_1 - \frac{2}{{\|\boldsymbol{u}\|_2}^2}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{e}_1 = \boldsymbol{e}_1$.

ii. En utilisant le résultat de la précédente question, on trouve $B\boldsymbol{e}_1 = H(\boldsymbol{u})AH(\boldsymbol{u})\boldsymbol{e}_1 = H(\boldsymbol{u})A\boldsymbol{e}_1 = H(\boldsymbol{u})\boldsymbol{a}_1$.

iii. On a d'une part

$$\boldsymbol{u}^\top(\boldsymbol{a} - \boldsymbol{b}) = \frac{1}{\|\boldsymbol{a} + \boldsymbol{b}\|_2}(\boldsymbol{a} + \boldsymbol{b})^\top(\boldsymbol{a} - \boldsymbol{b}) = \frac{1}{\|\boldsymbol{a} + \boldsymbol{b}\|_2}\left({\|\boldsymbol{a}\|_2}^2 - {\|\boldsymbol{b}\|_2}^2\right) = 0,$$

car $\|\boldsymbol{b}\|_2 = |\mathrm{sgn}(a_{21})|\,\|\boldsymbol{a}\|_2 \|\boldsymbol{e}_2\|_2 = |\mathrm{sgn}(a_{21})|\,\|\boldsymbol{a}\|_2$. On en déduit que $H(\boldsymbol{u})(\boldsymbol{a} - \boldsymbol{b}) = \boldsymbol{a} - \boldsymbol{b}$. D'autre part, on a $H(\boldsymbol{u})(\boldsymbol{a} + \boldsymbol{b}) = H\left(\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}\right)\boldsymbol{c} = -\boldsymbol{c} = -\boldsymbol{a} - \boldsymbol{b}$.

iv. On a $H(\boldsymbol{u})(\boldsymbol{a}-\boldsymbol{b}) = H(\boldsymbol{u})\boldsymbol{a} - H(\boldsymbol{u})\boldsymbol{b} = \boldsymbol{a}-\boldsymbol{b}$ et $H(\boldsymbol{u})(\boldsymbol{a}+\boldsymbol{b}) = H(\boldsymbol{u})\boldsymbol{a} + H(\boldsymbol{u})\boldsymbol{b} = -\boldsymbol{a}-\boldsymbol{b}$, d'où $H(\boldsymbol{u})\boldsymbol{a} = -\boldsymbol{b}$. Par ailleurs, on a $H(\boldsymbol{u})\boldsymbol{a} = H(\boldsymbol{u})(\boldsymbol{a}_1 - a_{11}\boldsymbol{e}_1) = H(\boldsymbol{u})\boldsymbol{a}_1 - a_{11}H(\boldsymbol{u})\boldsymbol{e}_1$, et donc $H(\boldsymbol{u})\boldsymbol{a}_1 = H(\boldsymbol{u})\boldsymbol{a} + a_{11}H(\boldsymbol{u})\boldsymbol{e}_1 = -\boldsymbol{b} + a_{11}\,\boldsymbol{e_1}$. On en déduit que

$$B\boldsymbol{e}_1 = H(\boldsymbol{u})\boldsymbol{a}_1 = a_{11}\,\boldsymbol{e_1} - \boldsymbol{b} = \begin{pmatrix} a_{11} \\ -\mathrm{sgn}(a_{21})\|\boldsymbol{a}\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

c. Si le vecteur $\boldsymbol{c}$ est nul, il suffit de poser $P_1 = I_n$ et $B = A$. Sinon, on pose $P_1 = H(\boldsymbol{u})$ et $B = H(\boldsymbol{u})AH(\boldsymbol{u})$. Dans les deux cas, la matrice P_1 est symétrique et orthogonale, et la matrice B est semblable à A, telle que $b_{i1} = 0$ pour tout entier i appartenant à $\{3,\dots,n\}$.

d. Soit n un entier naturel supérieur ou égal à 4. On suppose le résultat vrai pour les matrices symétriques d'ordre $n-1$. D'après la question précédente, on sait qu'il existe une matrice symétrique et orthogonale P_1 telle que $B = P_1AP_1$ soit symétrique et de la forme :

$$\begin{pmatrix} a_{11} & b_{12} & 0 & \dots & 0 \\ b_{21} & b_{2,2} & b_{23} & \dots & b_{2n} \\ 0 & b_{32} & b_{33} & \dots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & b_{2n} & b_{3n} & \dots & b_{nn} \end{pmatrix}.$$

Soit B_1 la sous-matrice d'ordre $n-1$ extraite de B suivante :

$$B_1 = \begin{pmatrix} b_{22} & b_{23} & \dots & b_{2n} \\ b_{32} & b_{33} & \dots & b_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{2n} & b_{3n} & \dots & b_{nn} \end{pmatrix}.$$

D'après l'hypothèse de récurrence, il existe $n-3$ matrices symétriques et orthogonales $Q_2, Q_3, \dots, Q_{n-2}$ telles que la matrice $T_1 = Q_{n-2}\dots Q_3Q_2B_1Q_2Q_3\dots Q_{n-2}$ soit tridiagonale. En posant alors, pour tout entier i appartenant à $\{2,\dots,n-2\}$,

$$P_i = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & Q_i & \\ 0 & & & \end{pmatrix}$$

et

$$T = \begin{pmatrix} a_{11} & b_{12} & 0 & \dots & 0 \\ b_{21} & & & & \\ 0 & & & & \\ \vdots & & T_1 & & \\ 0 & & & & \end{pmatrix},$$

on a

$$T = P_{n-2}\dots P_2BP_2\dots P_{n-2} = P_{n-2}\dots P_2P_1AP_1P_2\dots P_{n-2}.$$