1.4 Normes et conditionnement d'une matrice

Dans ce paragraphe, nous allons définir la notion de conditionnement d'une matrice, qui peut servir à établir une majoration des erreurs d'arrondi dues aux erreurs sur les données. Malheureusement, nous verrons également que cette majoration n'est pas forcément très utile dans des cas pratiques, et nous nous efforcerons d'y remédier. La notion de conditionnement est également utilisée dans l'étude des méthodes itératives que nous verrons plus loin. Pour l'étude du conditionnement comme pour l'étude des erreurs, nous avons tout d'abord besoin de la notion de norme et de rayon spectral, que nous rappelons maintenant.

1.4.1 Normes, rayon spectral

Définition 1.27 (Norme matricielle, norme induite)**.** *On note $\mathcal{M}_n(\mathbb{R})$ l'espace vectoriel (sur $\mathbb{R}$) des matrices carrées d'ordre n.*

1. *On appelle norme matricielle sur $\mathcal{M}_n(\mathbb{R})$ une norme $\|\cdot\|$ sur $\mathcal{M}_n(\mathbb{R})$ t.q.*
$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_n(\mathbb{R}) \tag{1.56}$$
2. *On considère $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$. On appelle norme matricielle induite (ou norme induite) sur $\mathcal{M}_n(\mathbb{R})$ par la norme $\|\cdot\|$, encore notée $\|\cdot\|$, la norme sur $\mathcal{M}_n(\mathbb{R})$ définie par :*
$$\|A\| = \sup\{\|A\boldsymbol{x}\|;\ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}, \ \forall A \in \mathcal{M}_n(\mathbb{R}) \tag{1.57}$$

Proposition 1.28 (Propriétés des normes induites)**.** *Soit $\mathcal{M}_n(\mathbb{R})$ muni d'une norme induite $\|\cdot\|$. Alors pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$, on a :*

1. $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$,
2. $\|A\| = \max\{\|A\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \ \boldsymbol{x} \in \mathbb{R}^n\}$,
3. $\|A\| = \max\left\{\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \ ; \ \boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}\right\}$.
4. $\|\cdot\|$ *est une norme matricielle.*

DÉMONSTRATION –

1. Soit $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$, posons $\boldsymbol{y} = \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, alors $\|\boldsymbol{y}\| = 1$ donc $\|A\boldsymbol{y}\| \leq \|A\|$. On en déduit que $\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|$ et donc que $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$. Si maintenant $\boldsymbol{x} = 0$, alors $A\boldsymbol{x} = 0$, et donc $\|\boldsymbol{x}\| = 0$ et $\|A\boldsymbol{x}\| = 0$; l'inégalité $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$ est encore vérifiée.
2. L'application φ définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par : $\varphi(\boldsymbol{x}) = \|A\boldsymbol{x}\|$ est continue sur la sphère unité $S_1 = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$ qui est un compact de $\mathbb{R}^n$. Donc φ est bornée et atteint ses bornes : il existe $\boldsymbol{x}_0 \in \mathbb{R}^n$ tel que $\|A\| = \|A\boldsymbol{x}_0\|$.
3. Cette égalité résulte du fait que
$$\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|A\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\| \text{ et } \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in S_1 \text{ et } \boldsymbol{x} \neq 0.$$
4. Soient A et $B \in \mathcal{M}_n(\mathbb{R})$, on a $\|AB\| = \max\{\|AB\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \ \boldsymbol{x} \in \mathbb{R}^n\}$. Or
$$\|AB\boldsymbol{x}\| \leq \|A\|\|B\boldsymbol{x}\| \leq \|A\|\|B\|\|\boldsymbol{x}\| \leq \|A\|\|B\|.$$
On en déduit que $\|\cdot\|$ est une norme matricielle.

■

Définition 1.29 (Rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On appelle rayon spectral de A la quantité $\rho(A) = \max\{|\lambda|;\ \lambda \in \mathbb{C},\ \lambda$ valeur propre de $A\}$.*

La proposition suivante caractérise les principales normes matricielles induites.

Proposition 1.30 (Caractérisation de normes induites)**.** *Soit $A = (a_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$.*

1. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Alors*

$$\|A\|_\infty = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |a_{i,j}|. \tag{1.58}$$

2. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Alors*

$$\|A\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^{n} |a_{i,j}| \tag{1.59}$$

3. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$.*

$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.60}$$

En particulier, si A est symétrique, $\|A\|_2 = \rho(A)$.

DÉMONSTRATION – La démonstration des points 1 et 2 fait l'objet de l'exercice 29 page 71. On démontre ici uniquement le point 3.

Par définition de la norme 2, on a :

$$\|A\|_2^2 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A\boldsymbol{x}\cdot A\boldsymbol{x} = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A^tA\boldsymbol{x}\cdot\boldsymbol{x}.$$

Comme A^tA est une matrice symétrique positive (car $A^tA\boldsymbol{x}\cdot\boldsymbol{x} = A\boldsymbol{x}\cdot A\boldsymbol{x} \geq 0$), il existe une base orthonormée $(\boldsymbol{f}_i)_{i=1,\dots,n}$ et des valeurs propres $(\mu_i)_{i=1,\dots,n}$, avec $0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ tels que $A\boldsymbol{f}_i = \mu_i\boldsymbol{f}_i$ pour tout $i \in \{1,\dots,n\}$. Soit $\boldsymbol{x} = \sum_{i=1,\dots,n} \alpha_i\boldsymbol{f}_i \in \mathbb{R}^n$. On a donc :

$$A^tA\boldsymbol{x}\cdot\boldsymbol{x} = \Big(\sum_{i=1,\dots,n} \mu_i\alpha_i\boldsymbol{f}_i\Big)\cdot\Big(\sum_{i=1,\dots,n} \alpha_i\boldsymbol{f}_i\Big) = \sum_{i=1,\dots,n} \alpha_i^2\mu_i \leq \mu_n\|\boldsymbol{x}\|_2^2.$$

On en déduit que $\|A\|_2^2 \leq \rho(A^tA)$.

Pour montrer qu'on a égalité, il suffit de considérer le vecteur $\boldsymbol{x} = \boldsymbol{f}_n$; on a en effet $\|\boldsymbol{f}_n\|_2 = 1$, et $\|A\boldsymbol{f}_n\|_2^2 = A^tA\boldsymbol{f}_n\cdot\boldsymbol{f}_n = \mu_n = \rho(A^tA)$.

■

Nous allons maintenant comparer le rayon spectral d'une matrice avec des normes. Rappelons d'abord le théorème de triangularisation (ou trigonalisation) des matrices complexes. On rappelle d'abord qu'une matrice unitaire $Q \in \mathcal{M}_n(\mathbb{C})$ est une matrice inversible telle que $Q^* = Q^{-1}$; ceci est équivalent à dire que les colonnes de Q forment une base orthonormale de $\mathbb{C}^n$. Une matrice carrée orthogonale est une matrice unitaire à coefficients réels ; on a dans ce cas $Q^* = Q^t$, et les colonnes de Q forment une base orthonormale de $\mathbb{R}^n$.

Théorème 1.31 (Décomposition de Schur, triangularisation d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ ou $\mathcal{M}_n(\mathbb{C})$ une matrice carrée quelconque, réelle ou complexe ; alors il existe une matrice complexe Q unitaire (c.à.d. une matrice telle que $Q^t = Q^{-1}$ et une matrice complexe triangulaire supérieure T telles que $A = QTQ^{-1}$.*

Ce résultat s'énonce de manière équivalente de la manière suivante : Soit ψ une application linéaire de E dans E, où E est un espace vectoriel normé de dimension finie n sur $\mathbb{C}$. Alors il existe une base $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de $\mathbb{C}$ et une famille de complexes $(t_{i,j})_{i=1,\ldots,n, j=1,\ldots,n, j\geq i}$ telles que $\psi(\boldsymbol{f}_i) = t_{i,i}\boldsymbol{f}_i + \sum_{k<i} t_{k,i}\boldsymbol{f}_k$. De plus $t_{i,i}$ est valeur propre de ψ et de A pour tout $i \in \{1, \ldots, n\}$.
Les deux énoncés sont équivalents au sens où la matrice A de l'application linéaire ψ s'écrit $A = QTQ^{-}1$, où T est la matrice triangulaire supérieure de coefficients $(t_{i,j})_{i,j=1,\ldots,n,j\geq i}$ et Q la matrice inversible dont la colonne j est le vecteur $\boldsymbol{f}_j$).

DÉMONSTRATION – On démontre cette propriété par récurrence sur n. Elle est évidemment vraie pour $n = 1$. Soit $n \geq 1$, on suppose la propriété vraie pour n et on la démontre pour $n+1$. Soint donc E un espace vectoriel sur $\mathbb{C}$ de dimension $n+1$ et ψ une application linéaire de E dans $\mathbb{C}$. On sait qu'il existe $\lambda \in \mathbb{C}$ (qui résulte du caractère algébriquement clos de $\mathbb{C}$) et $\boldsymbol{f}_1 \in E$ tels que $\psi(\boldsymbol{f}_1) = \lambda \boldsymbol{f}_1$ et $\|\boldsymbol{f}_1\| = 1$; on pose $t_{1,1} = \lambda$ et on note F le sous espace vectoriel de E supplémentaire orthogonal de $\mathbb{C}\boldsymbol{f}_1$. Soit $\boldsymbol{u} \in F$, il existe un unique couple $(\mu, \boldsymbol{v}) \in \mathbb{C} \times F$ tel que $\psi(\boldsymbol{u}) = \mu \boldsymbol{f}_1 + \boldsymbol{v}$. On note $\tilde{\psi}$ l'application qui à $\boldsymbol{u}$ associe $\boldsymbol{v}$. On peut appliquer l'hypothèse de récurrence à $\tilde{\psi}$ (car $\tilde{\psi}$ est une application linéaire de F dans F, et F est de dimension n). Il existe donc une base orthonormée $\boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n+1}$ de F et $(t_{i,j})_{j\geq i\geq 2}$ tels que

$$\tilde{\psi}(\boldsymbol{f}_i) = \sum_{2\leq j\leq i} t_{j,i}\boldsymbol{f}_j, \ i = 2, \ldots, n+1.$$

On en déduit que

$$\psi(\boldsymbol{f}_i) = \sum_{1\leq j\leq i\leq n} t_{j,i}\boldsymbol{f}_j, \ i = 1, \ldots, n+1.$$

■

Dans la proposition suivante, nous montrons qu'on peut toujours trouver une norme (qui dépend de la matrice) pour approcher son rayon spectral d'aussi près que l'on veut par valeurs supérieures.

Théorème 1.32 (Approximation du rayon spectral par une norme induite)**.**
1. Soit $\|\cdot\|$ une norme induite. Alors

$$\rho(A) \leq \|A\|, \text{ pour tout } A \in \mathcal{M}_n(\mathbb{R}).$$

2. Soient maintenant $A \in \mathcal{M}_n(\mathbb{R})$ et $\varepsilon > 0$, alors il existe une norme sur $\mathbb{R}^n$ (qui dépend de A et ε) telle que la norme induite sur $\mathcal{M}_n(\mathbb{R})$, notée $\|\cdot\|_{A,\varepsilon}$, vérifie $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$.

DÉMONSTRATION – 1. Soit $\lambda \in \mathbb{C}$ valeur propre de A et $\boldsymbol{x}$ un vecteur propre associé, alors $A\boldsymbol{x} = \lambda\boldsymbol{x}$, et comme $\|\cdot\|$ est une norme induite, on a :

$$\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\| = \|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{x}\|.$$

On en déduit que toute valeur propre λ vérifie $\lambda \leq \|A\|$ et donc $\rho(A) \leq \|A\|$.

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$, alors par le théorème de triangularisation de Schur (théorème 1.31 ppécédent), il existe une base $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i,j=1,\ldots,n,j\geq i}$ telles que $A\boldsymbol{f}_i = \sum_{j\leq i} t_{j,i}\boldsymbol{f}_j$. Soit $\eta \in]0,1[$, qu'on choisira plus précisément plus tard. Pour $i = 1, \ldots, n$, on définit $\boldsymbol{e}_i = \eta^{i-1}\boldsymbol{f}_i$. La famille $(\boldsymbol{e}_i)_{i=1,\ldots,n}$ forme une base de $\mathbb{C}^n$. On définit alors une norme sur $\mathbb{R}^n$ par $\|\boldsymbol{x}\| = (\sum_{i=1}^n \alpha_i\overline{\alpha_i})^{1/2}$, où les α_i sont les composantes de $\boldsymbol{x}$ dans la base $(\boldsymbol{e}_i)_{i=1,\ldots,n}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon > 0$; montrons que pour η bien choisi, on a $\|A\| \leq \rho(A) + \varepsilon$. Remarquons d'abord que

$$A\boldsymbol{e}_i = A(\eta^{i-1}\boldsymbol{f}_i) = \eta^{i-1}A\boldsymbol{f}_i\eta^{i-1}\sum_{j\leq i} t_{k,i}\boldsymbol{f}_j = \eta^{i-1}\sum_{j\leq i} t_{j,i}\eta^{1-j}\boldsymbol{e}_j = \sum_{1\leq j\leq i} \eta^{i-j}t_{j,i}\boldsymbol{e}_j,$$

Soit maintenant $\boldsymbol{x} = \sum_{i=1,\dots,n} \alpha_i \boldsymbol{e}_i$. On a

$$A\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i A\boldsymbol{e}_i = \sum_{i=1}^{n} \sum_{1\le j\le i} \eta^{i-j} t_{j,i}\alpha_i \boldsymbol{e}_j = \sum_{j=1}^{n} \big(\sum_{i=j}^{n} \eta^{i-j}\lambda_{i,j}\alpha_i\big)\boldsymbol{e}_j.$$

On en déduit que

$$\begin{aligned}
\|A\boldsymbol{x}\|^2 &= \sum_{j=1}^{n} \big(\sum_{i=j}^{n} \eta^{i-j} t_{j,i}\alpha_i\big)\big(\sum_{i=j}^{n} \eta^{i-j}\overline{t}_{j,i}\overline{\alpha}_i\big),\\
&= \sum_{j=1}^{n} t_{j,j}\overline{t_{j,j}}\alpha_j\overline{\alpha_j} + \sum_{j=1}^{n} \sum_{\substack{k,\ell\ge j\\(k,\ell)\neq(j,j)}} \eta^{k+\ell-2j} t_{j,k}\overline{t_{j,\ell}}\alpha_k\overline{\alpha_\ell}\\
&\le \rho(A)^2\|\boldsymbol{x}\|^2 + \max_{k=1,\dots,n} |\alpha_k|^2 \sum_{j=1}^{n} \sum_{\substack{k,\ell\ge j\\(k,\ell)\neq(j,j)}} \eta^{k+\ell-2j} t_{j,k}\overline{t_{j,\ell}}.
\end{aligned}$$

Comme $\eta \in [0,1]$ et $k+\ell-2j \ge 1$ dans la dernière sommation, on a

$$\sum_{j=1}^{n} \sum_{\substack{k,\ell\ge j\\(k,\ell)\neq(j,j)}} \eta^{k+\ell-2j} t_{j,k}\overline{t_{j,\ell}} \le \eta C_T n^3,$$

où $C_T = \max_{j,k,\ell=1,\dots,n} |t_{j,k}||t_{j,k}|$ ne dépend que de la matrice T, qui elle même ne dépend que de A. Comme

$$\max_{k=1,\dots,n} |\alpha_k|^2 \le \sum_{k=1,\dots,n} |\alpha_k|^2 = \|\boldsymbol{x}\|^2,$$

on a donc

$$\frac{\|A\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} \le \rho(A)^2 + \eta C_T n^3.$$

On en conclut que :

$$\|A\| \le \rho(A)\left(1 + \frac{\eta C_T n^3}{\rho(A)^2}\right) \le \rho(A) + \frac{\eta C_T n^3}{2\rho(A)},$$

D'où le résultat, en prenant $\|\cdot\|_{A,\varepsilon} = \|\cdot\|$ et η tel que $\eta = \min\left(1, \dfrac{2\rho(A)\varepsilon}{C_T n^3}\right)$..

■

Corollaire 1.33 (Convergence et rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

$$\rho(A) < 1 \textit{ si et seulement si } A^k \to 0 \textit{ quand } k \to \infty.$$

DÉMONSTRATION – Si $\rho(A) < 1$, grâce au résultat d'approximation du rayon spectral de la proposition précédente, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$ et une norme induite $\|.\|_{A,\varepsilon}$ tels que $\|A\|_{A,\varepsilon} = \mu \le \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon} \le \mu^k \to 0$ lorsque $k \to \infty$. Comme l'espace $\mathcal{M}_n(\mathbb{R})$ est de dimension finie, toutes les normes sont équivalentes, et on a donc $\|A^k\| \to 0$ lorsque $k \to \infty$.

Montrons maintenant la réciproque : supposons que $A^k \to 0$ lorsque $k \to \infty$, et montrons que $\rho(A) < 1$. Soient λ une valeur propre de A et $\boldsymbol{x}$ un vecteur propre associé. Alors $A^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$, et si $A^k \to 0$, alors $A^k\boldsymbol{x} \to 0$, et donc $\lambda^k\boldsymbol{x} \to 0$, ce qui n'est possible que si $|\lambda| < 1$.

■

Remarque 1.34 (Convergence des suites)**.** *Une conséquence immédiate du corollaire précédent est que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ définie par $\boldsymbol{x}^{(k+1)} = A\boldsymbol{x}^{(k)}$ converge vers $\mathbf{0}$ (le vecteur nul) pour tout $\boldsymbol{x}^{(0)}$ donné si et seulement si $\rho(A) < 1$.*

Proposition 1.35 (Convergence et rayon spectral)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.61}$$

DÉMONSTRATION – La démonstration se fait par des arguments d'homogénéité, en trois étapes. Rappelons tout d'abord que

$$\limsup_{k\to+\infty} u_k = \lim_{k\to+\infty} \sup_{n\geq k} u_n,$$
$$\liminf_{k\to+\infty} u_k = \lim_{k\to+\infty} \inf_{n\geq k} u_n,$$

et que si $\limsup_{k\to+\infty} u_k \leq \liminf_{k\to+\infty} u_k$, alors la suite $(u_k)_{k\in\mathbb{N}}$ converge vers $\lim_{k\to+\infty} u_k = \liminf_{k\to+\infty} u_k = \limsup_{k\to+\infty} u_k$.

Etape 1. On montre que

$$\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1. \tag{1.62}$$

En effet, si i $\rho(A) < 1$, d'après le corollaire 1.33 on a : $\|A^k\| \to 0$ donc il existe $K \in \mathbb{N}$ tel que pour $k \geq K$, $\|A^k\| < 1$. On en déduit que pour $k \geq K$, $\|A^k\|^{1/k} < 1$, et donc en passant à la limite sup sur k, on obtient bien que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq 1.$$

Etape 2. On montre maintenant que

$$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1.. \tag{1.63}$$

Pour démontrer cette assertion, rappelons que pour toute suite $(u_k)_{k\in\mathbb{N}}$ d'éléments de $\mathbb{R}$ ou $\mathbb{R}^n$, la limite inférieure $\liminf_{k\to+\infty} u_k$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, donc qu'il existe une suite extraite $(u_{k_n})_{n\in\mathbb{N}}$ telle que $u_{k_n} \to \liminf_{k\to+\infty} u_k$ lorsque $k \to +\infty$. Or $\liminf_{k\to+\infty} \|A^k\|^{1/k} < 1$; donc il existe une sous-suite $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ telle que $\|A^{k_n}\|^{1/k_n} \to \ell < 1$ lorsque $n \to +\infty$, et donc il existe n tel que pour $n \geq n$, $\|A^{k_n}\|^{1/k_n} \leq \eta$, avec $\eta \in]0,1[$. On en déduit que pour $n \geq n$, $\|A^{k_n}\| \leq \eta^{k_n}$, et donc que $A^{k_n} \to 0$ lorsque $n \to +\infty$. Soient λ une valeur propre de A et x un vecteur propre associé, on a : $A^{k_n}x = \lambda^{k_n}x$; on en déduit que $|\lambda| < 1$, et donc que $\rho(A) < 1$.

Etape 3. On montre que $\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.
Soit $\alpha \in \mathbb{R}_+$ tel que $\rho(A) < \alpha$. Alors $\rho(\frac{1}{\alpha}A) < 1$, et donc grâce à (1.62),

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \alpha,\ \forall \alpha > \rho(A).$$

En faisant tendre α vers $\rho(A)$, on obtient donc :

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A). \tag{1.64}$$

Soit maintenant $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. On a alors $\liminf_{k\to+\infty} \|(\frac{1}{\beta}A)^k\|^{\frac{1}{k}} < 1$ et donc en vertu de (1.63), $\rho(\frac{1}{\beta}A) < 1$, donc $\rho(A) < \beta$ pour tout $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. En faisant tendre β vers $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}$, on obtient donc

$$\rho(A) \leq \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.65}$$

De (1.64) et (1.65), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.66}$$

■

Un corollaire important de la proposition 1.35 est le suivant.

Corollaire 1.36 (Comparaison rayon spectral et norme)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme **matricielle**, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

$$\rho(A) \leq \|A\|.$$

Par conséquent, si $M \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, pour montrer que la suite $\boldsymbol{x}^{(k)}$ définie par $\boldsymbol{x}^{(k)} = M^k\boldsymbol{x}^{(0)}$ converge vers $\mathbf{0}$ dans $\mathbb{R}^n$, il suffit de trouver une norme matricielle $\|\cdot\|$ telle que $\|M\| < 1$.

DÉMONSTRATION – Si $\|.\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc par la caractérisation (1.61) du rayon spectral donnée dans la proposition précédente, on obtient que $\rho(A) \leq \|A\|$. ∎

Ce dernier résultat est évidemment bien utile pour montrer la convergence de la suite A^k, ou de suites de la forme $A^k \boldsymbol{x}^{(0)}$ avec $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. Une fois qu'on a trouvé une norme matricielle pour laquelle A est de norme strictement inférieure à 1, on a gagné. Attention cependant au piège suivant : pour toute matrice A, on peut toujours trouver une norme pour laquelle $\|A\| < 1$, alors que la série de terme général A^k peut ne pas être convergente.
Prenons un exemple dans $\mathbb{R}$, $\|x\| = \frac{1}{4}|x|$. Pour $x = 2$ on a $\|x\| = \frac{1}{2} < 1$. Et pourtant la série de terme général x^k n'est pas convergente ; le problème ici est que la norme choisie n'est pas une norme matricielle (on n'a pas $\|xy\| \leq \|x\|\|y\|$).
De même, on peut trouver une matrice et une norme telles que $\|A\| \geq 1$, alors que la série de terme général A^k converge...

Nous donnons maintenant un théorème qui nous sera utile dans l'étude du conditionnement, ainsi que plus tard dans l'étude des méthodes itératives.

Théorème 1.37 (Matrices de la forme $Id + A$)**.**

1. *Soit une norme matricielle induite,* Id *la matrice identité de* $\mathcal{M}_n(\mathbb{R})$ *et* $A \in \mathcal{M}_n(\mathbb{R})$ *telle que* $\|A\| < 1$. *Alors la matrice* $Id + A$ *est inversible et*

$$\|(Id + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

2. *Si une matrice de la forme* $Id + A \in \mathcal{M}_n(\mathbb{R})$ *est singulière, alors* $\|A\| \geq 1$ *pour toute norme matricielle* $\|\cdot\|$.

DÉMONSTRATION –

1. La démonstration du point 1 fait l'objet de l'exercice 34 page 72.
2. Si la matrice $Id + A \in \mathcal{M}_n(\mathbb{R})$ est singulière, alors $\lambda = -1$ est valeur propre, et donc $\rho(A) \geq 1$. En utilisant le corollaire 1.36, on obtient que $\|A\| \geq \rho(A) \geq 1$.

∎

1.4.2 Le problème des erreurs d'arrondis

Soient $A \in \mathcal{M}_n(\mathbb{R})$ inversible et $\boldsymbol{b} \in \mathbb{R}^n$; supposons que les données A et $\boldsymbol{b}$ ne soient connues qu'à une erreur près. Ceci est souvent le cas dans les applications pratiques. Considérons par exemple le problème de la conduction thermique dans une tige métallique de longueur 1, modélisée par l'intervalle $[0, 1]$. Supposons que la température u de la tige soit imposée aux extrémités, $u(0) = u_0$ et $u(1) = u_1$. On suppose que la température dans la tige satisfait à l'équation de conduction de la chaleur, qui s'écrit $(k(x)u'(x))' = 0$, où k est la conductivité thermique. Cette équation différentielle du second ordre peut se discrétiser par exemple par différences finies (on verra une description de la méthode page 11), et donne lieu à un système linéaire de matrice A. Si la conductivité k n'est connue qu'avec une certaine précision, alors la matrice A sera également connue à une erreur près, notée δ_A. On aimerait que l'erreur commise sur les données du modèle (ici la conductivité thermique k) n'ait pas une conséquence trop grave sur le calcul de la solution du modèle (ici la température u). Si par exemple 1% d'erreur sur k entraîne 100% d'erreur sur u, le modèle ne sera pas d'une utilité redoutable...
L'objectif est donc d'estimer les erreurs commises sur $\boldsymbol{x}$ solution de (1.1) à partir des erreurs commises sur $\boldsymbol{b}$ et A. Notons $\boldsymbol{\delta_b} \in \mathbb{R}^n$ l'erreur commise sur $\boldsymbol{b}$ et $\delta_A \in \mathcal{M}_n(\mathbb{R})$ l'erreur commise sur A. On cherche alors à évaluer $\boldsymbol{\delta_x}$

où $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution (si elle existe) du système :

$$\begin{cases} \boldsymbol{x} + \boldsymbol{\delta_x} \in \mathbb{R}^n \\ (A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}. \end{cases} \tag{1.67}$$

On va montrer que si δ_A "n'est pas trop grand", alors la matrice $A + \delta_A$ est inversible, et qu'on peut estimer $\boldsymbol{\delta_x}$ en fonction de δ_A et $\boldsymbol{\delta_b}$.

1.4.3 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.38 (Conditionnement). *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On appelle conditionnement de A par rapport à la norme $\|\cdot\|$ le nombre réel positif $\mathrm{cond}(A)$ défini par :*

$$\mathrm{cond}(A) = \|A\| \, \|A^{-1}\|.$$

Proposition 1.39 (Propriétés générales du conditionnement). *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, alors $\mathrm{cond}(A) \geq 1$.*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\alpha \in \mathbb{R}^*$, alors $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.*
3. *Soient A et $B \in \mathcal{M}_n(\mathbb{R})$ des matrices inversibles, alors $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.*

DÉMONSTRATION – 1. Comme $\|\cdot\|$ est une norme induite, c'est donc une norme matricielle. On a donc pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$,

$$\|\mathrm{Id}\| \leq \|A\| \, \|A^{-1}\|$$

ce qui prouve que $\mathrm{cond}(A) \geq 1$.

2. Par définition,

$$\begin{aligned} \mathrm{cond}(\alpha A) &= \|\alpha A\| \, \|(\alpha A)^{-1}\| \\ &= |\alpha| \, \|A\| \frac{1}{|\alpha|} \|A^{-1}\| = \mathrm{cond}(A) \end{aligned}$$

3. Soient A et B des matrices inversibles, alors AB est une matrice inversible et comme $\|\cdot\|$ est une norme matricielle,

$$\begin{aligned} \mathrm{cond}(AB) &= \|AB\| \, \|(AB)^{-1}\| \\ &= \|AB\| \, \|B^{-1}A^{-1}\| \\ &\leq \|A\| \, \|B\| \, \|B^{-1}\| \, \|A^{-1}\|. \end{aligned}$$

Donc $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$. ∎

Proposition 1.40 (Caractérisation du conditionnement pour la norme 2). *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note σ_n [resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Alors*

$$\mathrm{cond}_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. *Si de plus A une matrice symétrique définie positive, alors*

$$\mathrm{cond}_2(A) = \frac{\lambda_n}{\lambda_1},$$

où λ_n [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.

DÉMONSTRATION – On rappelle que si A a comme valeurs propres $\lambda_1, \ldots, \lambda_n$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$ et A^t a comme valeurs propres $\lambda_1, \ldots, \lambda_n$.
1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$. Or par le point 3. de la proposition 1.30 que $\|A\|_2 = (\rho(A^tA))^{1/2} = \sqrt{\sigma_n}$. On a donc

$$\|A^{-1}\|_2 = (\rho((A^{-1})^tA^{-1}))^{1/2} = \left(\rho(AA^t)^{-1}\right)^{1/2} \text{ ; or } \rho((AA^t)^{-1}) = \frac{1}{\tilde{\sigma}_1},$$

où $\tilde{\sigma}_1$ est la plus petite valeur propre de la matrice AA^t. Mais les valeurs propres de AA^t sont les valeurs propres de A^tA : en effet, si λ est valeur propre de AA^t associée au vecteur propre x alors λ est valeur propre de AA^t associée au vecteur propre A^tx. On a donc

$$cond_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. Si A est s.d.p., alors $A^tA = A^2$ et $\sigma_i = \lambda_i^2$ où λ_i est valeur propre de la matrice A. On a dans ce cas $cond_2(A) = \dfrac{\lambda_n}{\lambda_1}$. ∎

Les propriétés suivantes sont moins fondamentales, mais cependant intéressantes !

Proposition 1.41 (Propriétés du conditionnement pour la norme 2)**.** *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Alors $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.*
3. *Si A et B sont deux matrices symétriques définies positives, alors $\mathrm{cond}_2(A+B) \leq \max(\mathrm{cond}_2(A), \mathrm{cond}_2(B))$.*

La démonstration de la proposition 1.41 fait l'objet de l'exercice 38 page 73.
On va maintenant majorer l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur $\delta_{\boldsymbol{b}}$ sur le second membre $\boldsymbol{b}$.

Proposition 1.42 (Majoration de l'erreur relative pour une erreur sur le second membre)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\boldsymbol{\delta_b} \in \mathbb{R}^n$. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de*

$$A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}, \tag{1.68}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \tag{1.69}$$

DÉMONSTRATION – En retranchant (1.1) à (1.68), on obtient :

$$A\boldsymbol{\delta_x} = \boldsymbol{\delta_b}$$

et donc

$$\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\boldsymbol{\delta_b}\|. \tag{1.70}$$

Cette première estimation n'est pas satisfaisante car elle porte sur l'erreur globale ; or la notion intéressante est celle d'erreur relative. On obtient l'estimation sur l'erreur relative en remarquant que $\boldsymbol{b} = A\boldsymbol{x}$, ce qui entraîne que $\|\boldsymbol{b}\| \leq \|A\|\|\boldsymbol{x}\|$. On en déduit que

$$\frac{1}{\|\boldsymbol{x}\|} \leq \frac{\|A\|}{\|\boldsymbol{b}\|}.$$

En multipliant membre à membre cette dernière inégalité et (1.70), on obtient le résultat souhaité. ∎

Remarquons que l'estimation (1.69) est optimale. En effet, on va démontrer qu'on peut avoir égalité dans (1.69). Pour cela, il faut choisir convenablement $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$. On sait déjà que si $\boldsymbol{x}$ est solution de (1.1) et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de (1.67), alors

$$\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}, \text{ et donc } \|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\|.$$

Soit $\boldsymbol{x} \in \mathbb{R}^n$ tel que $\|\boldsymbol{x}\| = 1$ et $\|A\boldsymbol{x}\| = \|A\|$. Notons qu'un tel $\boldsymbol{x}$ existe parce que

$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\} = \max\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\}$$

(voir proposition 1.28 page 61). On a donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|A\boldsymbol{x}\|}.$$

Posons $\boldsymbol{b} = A\boldsymbol{x}$; on a donc $\|\boldsymbol{b}\| = \|A\|$, et donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|b\|}.$$

De même, grâce à la proposition 1.28, il existe $\boldsymbol{y} \in \mathbb{R}^n$ tel que $\|\boldsymbol{y}\| = 1$, et $\|A^{-1}\boldsymbol{y}\| = \|A^{-1}\|$. On choisit alors $\boldsymbol{\delta_b}$ tel que $\delta_b = \boldsymbol{y}$. Comme $A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}$, on a $\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}$ et donc :

$$\|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\| = \varepsilon\|A^{-1}\boldsymbol{y}\| = \varepsilon\|A^{-1}\| = \|\boldsymbol{\delta_b}\|\ \|A^{-1}\|.$$

On en déduit que

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{\delta_x}\| = \|\boldsymbol{\delta_b}\|\ \|A^{-1}\|\ \frac{\|A\|}{\|\boldsymbol{b}\|} \text{ car } \|\boldsymbol{b}\| = \|A\| \text{ et } \|\boldsymbol{x}\| = 1.$$

Par ce choix de $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$ on a bien égalité dans (1.69) qui est donc optimale.
Majorons maintenant l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur δ_A sur la matrice A.

Proposition 1.43 (Majoration de l'erreur relative pour une erreur sur la matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\delta_A \in \mathbb{R}^n$; on suppose que $A + \delta_A$ est une matrice inversible. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \delta\boldsymbol{x}$ est solution de*

$$(A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} \tag{1.71}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x} + \boldsymbol{\delta_x}\|} \leq \text{cond}(A)\frac{\|\delta_A\|}{\|A\|} \tag{1.72}$$

DÉMONSTRATION – En retranchant (1.1) à (1.71), on obtient :

$$A\delta(\boldsymbol{\delta_x} = -\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x})$$

et donc

$$\boldsymbol{\delta_x} = -A^{-1}\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x}).$$

On en déduit que $\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\delta_A\|\|\boldsymbol{x} + \boldsymbol{\delta_x}\|$, d'où on déduit le résultat souhaité. ∎

On peut en fait majorer l'erreur relative dans le cas où l'on commet à la fois une erreur sur A et une erreur sur $\boldsymbol{b}$. On donne le théorème à cet effet ; la démonstration est toutefois nettement plus compliquée.

Théorème 1.44 (Majoration de l'erreur relative pour une erreur sur matrice et second membre)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq \boldsymbol{0}$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soient $\delta_A \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{\delta_b} \in \mathbb{R}^n$. On suppose que $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$. Alors la matrice $(A+\delta_A)$ est inversible et si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x}+\boldsymbol{\delta_x}$ est solution de* (1.67)*, alors*

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\mathrm{cond}(A)}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|}+\frac{\|\delta_A\|}{\|A\|}\right). \tag{1.73}$$

DÉMONSTRATION – On peut écrire $A+\delta_A = A(\mathrm{Id}+B)$ avec $B = A^{-1}\delta_A$. Or le rayon spectral de B, $\rho(B)$, vérifie $\rho(B) \leq \|B\| \leq \|\delta_A\|\ \|A^{-1}\| < 1$, et donc (voir le théorème 1.37 page 66 et l'exercice 34 page 72) $(\mathrm{Id}+B)$ est inversible et $(\mathrm{Id}+B)^{-1} = \sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(\mathrm{Id}+B)^{-1}\| \leq \sum_{n=0}^{\infty}\|B\|^n = \dfrac{1}{1-\|B\|} \leq \dfrac{1}{1-\|A^{-1}\|\ \|\delta_A\|}$. On en déduit que $A+\delta_A$ est inversible, car $A+\delta_A = A(Id+B)$ et comme A est inversible, $(A+\delta_A)^{-1} = (\mathrm{Id}+B)^{-1}A^{-1}$.

Comme A et $A+\delta_A$ sont inversibles, il existe un unique $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = b$ et il existe un unique $\boldsymbol{\delta_x} \in \mathbb{R}^n$ tel que $(A+\delta_A)(\boldsymbol{x}+\boldsymbol{\delta_x}) = \boldsymbol{b}+\boldsymbol{\delta_b}$. Comme $A\boldsymbol{x} = \boldsymbol{b}$, on a $(A+\delta_A)\boldsymbol{\delta_x} + \delta_A\boldsymbol{x} = \boldsymbol{\delta_b}$ et donc $\boldsymbol{\delta_x} = (A+\delta_A)^{-1}\boldsymbol{\delta_b} - \delta_A\boldsymbol{x})$. Or $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$, on en déduit :

$$\begin{aligned}\|(A+\delta_A)^{-1}\| &\leq \|(\mathrm{Id}+B)^{-1}\|\ \ \|A^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\ \|\delta_A\|}.\end{aligned}$$

On peut donc écrire la majoration suivante :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|x\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\boldsymbol{\delta_b}\|}{\|A\|\ \|\boldsymbol{x}\|}+\frac{\|\delta_A\|}{\|A\|}\right).$$

En utilisant le fait que $\boldsymbol{b} = A\boldsymbol{x}$ et que par suite $\|\boldsymbol{b}\| \leq \|A\|\ \|\boldsymbol{x}\|$, on obtient :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta_A\|}{\|A\|}\right),$$

ce qui termine la démonstration. ∎

1.4.4 Discrétisation d'équations différentielles, conditionnement "efficace"

On suppose encore ici que $\delta_A = 0$. On suppose que la matrice A du système linéaire à résoudre provient de la discrétisation par différences finies du problème de la chaleur unidimensionnel (1.5a). On peut alors montrer (voir exercice 45 page 75) que le conditionnement de A est d'ordre n^2, où n est le nombre de points de discrétisation. Pour $n = 10$, on a donc $\mathrm{cond}(A) \simeq 100$ et l'estimation (1.69) donne :

$$\frac{\|\delta_x\|}{\|x\|} \leq 100\ \frac{\|\delta_b\|}{\|b\|}.$$

Une erreur de 1% sur $\boldsymbol{b}$ peut donc entraîner une erreur de 100% sur x. Autant dire que dans ce cas, il est inutile de rechercher la solution de l'équation discrétisée... Heureusement, on peut montrer que l'estimation (1.69) n'est pas significative pour l'étude de la propagation des erreurs lors de la résolution des systèmes linéraires provenant de la discrétisation d'une équation différentielle ou d'une équation aux dérivées partielles [7]. Pour illustrer notre propos, reprenons l'étude du système linéaire obtenu à partir de la discrétisation de l'équation de la chaleur (1.5a) qu'on écrit : $A\boldsymbol{u} = \boldsymbol{b}$ avec $\boldsymbol{b} = (b_1, \ldots, b_n)$ et A la matrice carrée d'ordre n de coefficients $(a_{i,j})_{i,j=1,n}$ définis par (1.10). On rappelle que A est symétrique définie positive (voir exercice 10 page 19), et que

$$\max_{i=1\ldots n}\{|u_i - u(x_i)|\} \leq \frac{h^2}{96}\|u^{(4)}\|_\infty.$$

7. On appelle équation aux dérivées partielles une équation qui fait intervenir les dérivées partielles de la fonction inconnue, par exemple $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$, où u est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$

En effet, si on note $\overline{u}$ le vecteur de $\mathbb{R}^n$ de composantes $u(x_i)$, $i = 1, \ldots, n$, et R le vecteur de $\mathbb{R}^n$ de composantes R_i, $i = 1, \ldots, n$, on a par définition de R (formule (1.7)) $A(u - \overline{u}) = R$, et donc $\|u - \overline{u}\|_\infty \leq \|A^{-1}\|_\infty \|R\|_\infty$. Or on peut montrer (voir exercice 45 page 75) que $\mathrm{cond}(A) \simeq n^2$. Donc si on augmente le nombre de points, le conditionnement de A augmente aussi. Par exemple si $n = 10^4$, alors $\|\delta_x\|/\|x\| = 10^8 \|\delta_b\|/\|b\|$. Or sur un ordinateur en simple précision, on a $\|\delta_b\|/\|b\| \geq 10^{-7}$, donc l'estimation (1.69) donne une estimation de l'erreur relative $\|\delta_x\|/\|x\|$ de 1000%, ce qui laisse à désirer pour un calcul qu'on espère précis.
En fait, l'estimation (1.69) ne sert à rien pour ce genre de problème, il faut faire une analyse un peu plus poussée, comme c'est fait dans l'exercice 47 page 76. On se rend compte alors que pour f donnée il existe $C \in \mathbb{R}_+$ ne dépendant que de f (mais pas de n) tel que

$$\frac{\|\boldsymbol{\delta_u}\|}{\|\boldsymbol{u}\|} \leq C \frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \text{ avec } \boldsymbol{b} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{1.74}$$

L'estimation (1.74) est évidemment bien meilleure que l'estimation (1.69) puisqu'elle montre que l'erreur relative commise sur u est du même ordre que celle commise sur $\boldsymbol{b}$. En particulier, elle n'augmente pas avec le nombre de points de discrétisation. En conclusion, l'estimation (1.69) est peut-être optimale dans le cas d'une matrice quelconque, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.69)) mais elle n'est pas toujours significative pour l'étude des systèmes linéaires issus de la discrétisation des équations aux dérivées partielles.

1.4.5 Exercices

Exercice 28 (Normes de l'Identité)**.** *Corrigé en page 79*

Soit Id la matrice "Identité" de $\mathcal{M}_n(\mathbb{R})$. Montrer que pour toute norme induite on a $\|Id\| = 1$ et que pour toute norme matricielle on a $\|Id\| \geq 1$.

Exercice 29 (Normes induites particulières)**.** *Suggestions en page 78, corrigé détaillé en page 79.*

Soit $A = (a_{i,j})_{i,j \in \{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$.

1. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Montrer que
$$\|A\|_\infty = \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} |a_{i,j}|.$$
2. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Montrer que
$$\|A\|_1 = \max_{j \in \{1,\ldots,n\}} \sum_{i=1}^{n} |a_{i,j}|.$$

Exercice 30 (Norme non induite)**.** *Corrigé en page 80*

Pour $A = (a_{i,j})_{i,j \in \{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$, on pose $\|A\|_s = (\sum_{i,j=1}^n a_{i,j}^2)^{\frac{1}{2}}$.

1. Montrer que $\|\cdot\|_s$ est une norme matricielle mais n'est pas une norme induite (pour $n > 1$).
2. Montrer que $\|A\|_s^2 = tr(A^t A)$. En déduire que $\|A\|_2 \leq \|A\|_s \leq \sqrt{n}\|A\|_2$ et que $\|Ax\|_2 \leq \|A\|_s \|x\|_2$, pour tout $A \in \mathcal{M}_n(\mathbb{R})$ et tout $x \in \mathbb{R}^n$.
3. Chercher un exemple de norme non matricielle.

Exercice 31 (Valeurs propres d'un produit de matrices)**.** *Corrigé en page 80*

Soient p et n des entiers naturels non nuls tels que $n \leq p$, et soient $A \in \mathcal{M}_{n,p}(\mathbb{R})$ et $B \in \mathcal{M}_{p,n}(\mathbb{R})$. (On rappelle que $\mathcal{M}_{n,p}(\mathbb{R})$ désigne l'ensemble des matrices à n lignes et p colonnes.)

1. Montrer que λ est valeur propre non nulle de AB si et seulement si λ est valeur propre non nulle de BA.

2. Montrer que si $\lambda = 0$ est valeur propre de AB alors λ est valeur propre nulle de BA. *(Il est conseillé de distinguer les cas $Bx \neq 0$ et $Bx = 0$, où x est un vecteur propre associé à la $\lambda = 0$ valeur propre de AB. Pour le deuxième cas, on pourra distinguer selon que $ImA = \mathbb{R}^n$ ou non.)*
3. Montrer en donnant un exemple que λ peut être une valeur propre nulle de BA sans être valeur propre de AB. *(Prendre par exemple $n = 1$, $p = 2$.)*
4. On suppose maintenant que $n = p$, déduire des questions 1 et 2 que l'ensemble des valeurs propres de AB est égal à l'ensemble des valeurs propres de la matrice BA.

Exercice 32 (Rayon spectral)**.** *Corrigé en page 80.*

Soit $A \in \mathcal{M}_n(\mathbb{R})$. Montrer que si A est diagonalisable, il existe une norme induite sur $\mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) = \|A\|$. Montrer par un contre exemple que ceci peut être faux si A n'est pas diagonalisable.

Exercice 33 (Sur le rayon spectral)**.** *Corrigé en page 80*

On définit les matrices carrées d'ordre 2 suivantes :

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ B = \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix},\ C = A + B.$$

Calculer le rayon spectral de chacune des matrices A, B et C et en déduire que le rayon spectral ne peut être ni une norme, ni même une semi-norme sur l'espace vectoriel des matrices.

Exercice 34 (Série de Neumann)**.** *Suggestions en page 78, corrigé détaillé en page 81.*

Soient $A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $\rho(A) < 1$, les matrices $Id - A$ et $Id + A$ sont inversibles.

2. Montrer que la série de terme général A^k converge (vers $(Id - A)^{-1}$) si et seulement si $\rho(A) < 1$.

3. Montrer que si $\rho(A) < 1$, et si $\|\cdot\|$ une norme matricielle telle que $\|A\| < 1$, alors $\|(Id - A)^{-1}\| \leq \frac{1}{1-\|A\|}$ et $\|(Id + A)^{-1}\| \leq \frac{1}{1-\|A\|}$.

Exercice 35 (Normes induites)**.** *Corrigé en page 81*

Soit $\|.\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ par une norme quelconque sur $\mathbb{R}^n$, et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) < 1$ (on rappelle qu'on note $\rho(A)$ le rayon spectral de la matrice A). Pour $x \in \mathbb{R}^n$, on définit $\|x\|_*$ par :

$$\|x\|_* = \sum_{j=0}^{\infty} \|A^j x\|.$$

1. Montrer que l'application définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par $x \mapsto \|x\|_*$ est une norme.
2. Soit $x \in \mathbb{R}^n$ tel que $\|x\|_* = 1$. Calculer $\|Ax\|_*$ en fonction de $\|x\|$, et en déduire que $\|A\|_* < 1$.
3. On ne suppose plus que $\rho(A) < 1$. Soit $\varepsilon > 0$ donné. Construire à partir de la norme $\|.\|$ une norme induite $\|.\|_{**}$ telle que $\|A\|_{**} \leq \rho(A) + \varepsilon$.

Exercice 36 (Un système par blocs)**.** *Corrigé en page 82*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre N inversible, $b, c, f \in \mathbb{R}^n$. Soient α et $\gamma \in \mathbb{R}$. On cherche à résoudre le système suivant (avec $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$) :

$$\begin{aligned} &Ax + b\lambda = f, \\ &c \cdot x + \alpha\lambda = \gamma. \end{aligned} \tag{1.75}$$

1. Ecrire le système (1.75) sous la forme : $My = g$, où M est une matrice carrée d'ordre $n+1$, $y \in \mathbb{R}^{n+1}$, $g \in \mathbb{R}^{n+1}$. Donner l'expression de M, y et g.
2. Donner une relation entre A, b, c et α, qui soit une condition nécessaire et suffisante pour que le système (1.75) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.75) :
   (a) Soient z solution de $Az = b$, et h solution de $Ah = f$.
   (b) $x = h - \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z} z, \qquad \lambda = \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}.$

   Montrer que $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.75).
4. On suppose dans cette question que A est une matrice bande, dont la largeur de bande est p.
   (a) Calculer le coût de la méthode de résolution proposée ci-dessus en utilisant la méthode LU pour la résolution des systèmes linéaires.
   (b) Calculer le coût de la résolution du système $My = g$ par la méthode LU (en profitant ici encore de la structure creuse de la matrice A).
   (c) Comparer et conclure.
   Dans les deux cas, le terme d'ordre supérieur est $2nq^2$, et les coûts sont donc comparables.

Exercice 37 (Calcul de conditionnement)**.** *Corrigé détaillé en page 83.* *Calculer le conditionnement pour la norme 2 de la matrice* $\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$.

Exercice 38 (Propriétés générales du conditionnement)**.** *Corrigé détaillé en page 84.*
On suppose que $\mathbb{R}^n$ *est muni de la norme euclidienne usuelle* $\|\cdot\| = \|\cdot\|_2$ *et* $\mathcal{M}_n(\mathbb{R})$ *de la norme induite (notée aussi* $\|\cdot\|_2$*. On note alors* $\mathrm{cond}_2(A)$ *le conditionnement d'une matrice* A *inversible.*

1. *Soit* $A \in \mathcal{M}_n(\mathbb{R})$ *une matrice inversible. Montrer que* $\mathrm{cond}_2(A) = 1$ *si et seulement si* $A = \alpha Q$ *où* $\alpha \in \mathbb{R}^\star$ *et* Q *est une matrice orthogonale (c'est-à-dire* $Q^t = Q^{-1}$*).*
2. *Soit* $A \in \mathcal{M}_n(\mathbb{R})$ *une matrice inversible. On suppose que* $A = QR$ *où* Q *est une matrice orthogonale. Montrer que* $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
3. *Soit* $A, B \in \mathcal{M}_n(\mathbb{R})$ *deux matrices symétriques définies positives. Montrer que*

$$\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)\}.$$

Exercice 39 (Conditionnement de la matrice transposée)**.** *Corrigé en page 86*
On suppose que $A \in \mathcal{M}_n(\mathbb{R})$ *est inversible.*

1. *Montrer que si* $B \in \mathcal{M}_n(\mathbb{R})$*, on a pour tout* $\lambda \in \mathbb{C}$*,*

$$\det(AB - \lambda Id) = \det(BA - \lambda Id).$$

2. *En déduire que les rayons spectraux des deux matrices* AB *et* BA *sont identiques.*
3. *Montrer que* $\|A^t\|_2 = \|A\|_2$.
4. *En déduire que cond*$_2(A) =$ *cond*$_2(A^t)$.
5. *A-t-on* $\|A^t\|_1 = \|A\|_1$ *?*
6. *Montrer que dans le cas* $n = 2$*, on a toujours cond*$_1(A) =$ *cond*$_1(A^t), \forall A \in M_2(\mathbb{R})$.
7. *Calculer cond*$_1(A)$ *pour* $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ *et conclure.*

Exercice 40 (Conditionnement et normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$)**.** *Corrigé en page 87*

1. On considère la matrice $B = (B_{ij})$ de $\mathcal{M}_n(\mathbb{R})$ définie par $B_{ii} = 1$, $B_{ij} = -1\, i < j$, $B_{ij} = 0$ sinon.
   (a) Calculer B^{-1}.
   (b) En déduire $\mathrm{cond}_1(B)$ et $\mathrm{cond}_\infty(B)$.
2. Soit A une matrice carrée de taille $n \times n$. L'objectif de cette question est de montrer que
$$\frac{1}{n^2}\mathrm{cond}_\infty(A) \leq \mathrm{cond}_1(A) \leq n^2\mathrm{cond}_\infty(A).$$
   (a) Montrer que pour tout $x \in \mathbb{R}^n$,
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty.$$
   (b) En déduire que pour toute matrice carrée de taille $n \times n$
$$\frac{1}{n}\|A\|_\infty \leq \|A\|_1 \leq n\|A\|_\infty.$$
   (c) Conclure.

Exercice 41 (Majoration du conditionnement)**.** *Corrigé en page 88*

Soit $\|\,.\,\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\det(A) \neq 0$.

1. Montrer que si $\|A - B\| < \frac{1}{\|A^{-1}\|}$, alors B est inversible.
2. Montrer que cond $(A) \geq \sup_{\substack{B\in\mathcal{M}_n(\mathbb{R})\\ \det B=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 42 (Minoration du conditionnement)**.** *Corrigé détaillé en page 88.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible, $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$ le conditionnement de A, et soit $\delta_A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $A + \delta_A$ est singulière, alors
$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.76}$$
2. On suppose dans cette question que la norme $\|\cdot\|$ est la norme induite par la norme euclidienne sur $\mathbb{R}^n$. Montrer que la minoration (1.76) est optimale, c'est-à-dire qu'il existe $\delta_A \in \mathcal{M}_n(\mathbb{R})$ telle que $A + \delta_A$ soit singulière et telle que l'égalité soit vérifiée dans (1.76).
   [*On pourra chercher δ_A de la forme*
$$\delta_A = -\frac{y\, x^t}{x^t\, x},$$
   avec $y \in \mathbb{R}^n$ convenablement choisi et $x = A^{-1}y$.]
3. On suppose ici que la norme $\|\cdot\|$ est la norme induite par la norme infinie sur $\mathbb{R}^n$. Soit $\alpha \in]0, 1[$. Utiliser l'inégalité (1.76) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\mathrm{cond}(A)$ pour la matrice
$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \alpha & -\alpha \\ 1 & \alpha & \alpha \end{pmatrix}.$$

Exercice 43 (Conditionnement du carré)**.** *Corrigé en page 89*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice telle que $\det A \neq 0$.

1. Quelle relation existe-t-il en général entre $\mathrm{cond}\,(A^2)$ et $(\mathrm{cond}\,A)^2$?
2. On suppose que A symétrique. Montrer que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$.

3. On suppose que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$. Peut-on conclure que A est symétrique ? (justifier la réponse.)

Exercice 44 (Calcul de l'inverse d'une matrice et conditionnement)**.** *Corrigé détaillé en page 89.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible. On cherche ici des moyens d'évaluer la précision de calcul de l'inverse de A.

1. On suppose qu'on a calculé B, approximation (en raison par exemple d'erreurs d'arrondi) de la matrice A^{-1}. On pose :
$$\left\{\begin{array}{lllll} e_1 &=& \dfrac{\|B-A^{-1}\|}{\|A^{-1}\|}, & e_2 &= \dfrac{\|B^{-1}-A\|}{\|A\|} \\ \\ e_3 &=& \|AB-\mathrm{Id}\|, & e_4 &= \|BA-\mathrm{Id}\| \end{array}\right. \tag{1.77}$$
   (a) Expliquer en quoi les quantités e_1, e_2, e_3 et e_4 mesurent la qualité de l'approximation de A^{-1}.
   (b) On suppose ici que $B = A^{-1} + E$, où $\|E\| \leq \varepsilon\|A^{-1}\|$, et que
$$\varepsilon \mathrm{cond}(A) < 1.$$
   Montrer que dans ce cas,
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon\mathrm{cond}(A)}{1-\varepsilon\mathrm{cond}(A)},\ e_3 \leq \varepsilon\mathrm{cond}(A) \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$
   (c) On suppose maintenant que $AB - Id = E'$ avec $\|E'\| \leq \varepsilon < 1$. Montrer que dans ce cas :
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon}{1-\varepsilon},\ e_3 \leq \varepsilon \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$

2. On suppose maintenant que la matrice A n'est connue qu'à une certaine matrice d'erreurs près, qu'on note δ_A.
   (a) Montrer que la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.
   (b) Montrer que si la matrice $A + \delta_A$ est inversible, alors
$$\frac{\|(A+\delta_A)^{-1} - A^{-1}\|}{\|(A+\delta_A)^{-1}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|}.$$

Exercice 45 (Conditionnement du Laplacien discret 1D)**.** *Suggestions en page 78, corrigé détaillé en page 90.* *Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 12, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.*
Calculer les valeurs propres et les vecteurs propres de A. [On pourra commencer par chercher $\lambda \in \mathbb{R}$ et $\varphi \in C^2(\mathbb{R},\mathbb{R})$ (φ non identiquement nulle) t.q. $-\varphi''(x) = \lambda\varphi(x)$ pour tout $x \in]0,1[$ et $\varphi(0) = \varphi(1) = 0$].
Calculer $\mathrm{cond}_2(A)$ et montrer que $h^2\mathrm{cond}_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 46 (Conditionnement, réaction diffusion 1d.)**.** *Corrigé en page 91*

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème (1.28) étudié à l'exercice 12, qu'on rappelle :
$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.78}$$

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.28) aux points $\left(\frac{j}{n+1}\right)_{j=1,\ldots,n}$. On rappelle que la discrétisation par différences finies de ce problème consiste à chercher U

comme solution du système linéaire $AU = \left(f(\frac{j}{N+1})\right)_{j=1,\dots,n}$ où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est définie par $A = (N+1)^2 B + Id$, Id désigne la matrice identité et

$$B = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

1. (Valeurs propres de la matrice B.)
   On rappelle que le problème aux valeurs propres

$$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.79}$$

   admet la famille $(\lambda_k, u_k)_{k\in\mathbb{N}^*}$, $\lambda_k = (k\pi)^2$ et $u_k(x) = \sin(k\pi x)$ comme solution. Montrer que les vecteurs $U_k = \left(u_k(\frac{j}{n+1})\right)_{j=1,\dots,n}$ sont des vecteurs propres de la matrice B. En déduire toutes les valeurs propres de la matrice B.
2. En déduire les valeurs propres de la matrice A.
3. En déduire le conditionnement pour la norme euclidienne de la matrice A.

Exercice 47 (Conditionnement "efficace"). *Suggestions en page 78. corrigé en page 91.*

Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 12, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.

Pour $u \in \mathbb{R}^n$, on note $u_1, \dots, u_n$ les composantes de u. Pour $u \in \mathbb{R}^n$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1, \dots, n\}$. Pour u, $v \in \mathbb{R}^n$, on note $u \cdot v = \sum_{i=1}^n u_i v_i$.

On munit $\mathbb{R}^n$ de la norme suivante : pour $u \in \mathbb{R}^n$, $\|u\| = \max\{|u_i|,\ i \in \{1, \dots, n\}\}$. On munit alors $\mathcal{M}_n(\mathbb{R})$ de la norme induite, également notée $\|\cdot\|$, c'est-à-dire $\|B\| = \max\{\|Bu\|,\ u \in \mathbb{R}^n \text{ t.q. } \|u\| = 1\}$, pour tout $B \in \mathcal{M}_n(\mathbb{R})$.

Partie I Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de A^{-1}) Soient $b \in \mathbb{R}^n$ et $u \in \mathbb{R}^n$ t.q. $Au = b$. Remarquer que $Au = b$ peut s'écrire :

$$\left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i,\ \forall i \in \{1, \dots, n\}, \\ u_0 = u_{n+1} = 0. \end{array} \right. \tag{1.80}$$

   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considérer $p \in \{0, \dots, n+1\}$ t.q. $u_p = \min\{u_j,\ j \in \{0, \dots, n+1\}$.]

   En déduire que A est inversible.
2. (Préliminaire) On considère la fonction $\varphi \in C([0,1], \mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\boldsymbol{\phi} = (\phi_1, \dots \phi_n) \in \mathbb{R}^n$ par $\phi_i = \varphi(ih)$ pour tout $i \in \{1, \dots, n\}$. Montrer que $(A\boldsymbol{\phi})_i = 1$ pour tout $i \in \{1, \dots, n\}$.
3. (Calcul de $\|A^{-1}\|$) Soient $\boldsymbol{b} \in \mathbb{R}^n$ et $\boldsymbol{u} \in \mathbb{R}^n$ t.q. $A\boldsymbol{u} = \boldsymbol{b}$. Montrer que $\|\boldsymbol{u}\| \leq (1/8)\|\boldsymbol{b}\|$ [Calculer $A(\boldsymbol{u} \pm \|\boldsymbol{b}\|\boldsymbol{\phi})$ avec
   $bfphi$ défini à la question 2 et utiliser la question 1]. En déduire que $\|A^{-1}\| \leq 1/8$ puis montrer que $\|A^{-1}\| = 1/8$.
4. (Calcul de $\|A\|$) Montrer que $\|A\| = \frac{4}{h^2}$.

5. (Conditionnement pour la norme $\|\cdot\|$). Calculer $\|A^{-1}\|\|A\|$. Soient $\boldsymbol{b}$, $\delta_b \in \mathbb{R}^n$ et soient u, $\delta_u \in \mathbb{R}^n$ t.q. $Au = b$ et $A(u+\delta_u) = b+\delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \dfrac{\|\delta_b\|}{\|b\|}$.

   Montrer qu'un choix convenable de $\boldsymbol{b}$ et δ_b donne l'égalité dans l'inégalité précédente.

Partie II Borne réaliste sur l'erreur relative : Conditionnement "efficace"

On se donne maintenant $f \in C([0,1], \mathbb{R})$ et on suppose (pour simplifier...) que $f(x) > 0$ pour tout $x \in]0,1[$. On prend alors, dans cette partie, $b_i = f(ih)$ pour tout $i \in \{1, \ldots, n\}$. On considère aussi le vecteur ϕ défini à la question 2 de la partie I.

1. Montrer que
$$h \sum_{i=1}^{n} b_i \phi_i \to \int_0^1 f(x)\varphi(x)dx \text{ quand } n \to \infty$$
   et que
$$\sum_{i=1}^{n} b_i \phi_i > 0 \text{ pour tout } n \in \mathbb{N}^*.$$
   En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h\sum_{i=1}^{n} b_i\phi_i \geq \alpha$ pour tout $n \in \mathbb{N}^*$.
2. Soit $u \in \mathbb{R}^n$ t.q. $Au = b$. Montrer que $n\|u\| \geq \sum_{i=1}^{n} u_i = u \cdot A\phi \geq \frac{\alpha}{h}$ (avec α donné à la question 1).

   Soit $\delta_b \in \mathbb{R}^n$ et $\delta_u \in \mathbb{R}^n$ t.q. $A(u+\delta_u) = b+\delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha} \dfrac{\|\delta_b\|}{\|b\|}$.
3. Comparer $\|A^{-1}\|\|A\|$ (question I.5) et $\dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$ (question II.2) quand n est "grand" (ou quand $n \to \infty$).

1.5 Méthodes itératives

Les méthodes directes sont très efficaces : elles donnent la solution exacte (aux erreurs d'arrondi près) du système linéaire considéré. Elles ont l'inconvénient de nécessiter une assez grande place mémoire car elles nécessitent le stockage de toute la matrice en mémoire vive. Si la matrice est pleine, c.à.d. si la plupart des coefficients de la matrice sont non nuls et qu'elle est trop grosse pour la mémoire vive de l'ordinateur dont on dispose, il ne reste plus qu'à gérer habilement le "swapping" c'est–à–dire l'échange de données entre mémoire disque et mémoire vive pour pouvoir résoudre le système.
Cependant, si le système a été obtenu à partir de la discrétisation d'équations aux dérivés partielles, il est en général "creux", c.à. d. qu'un grand nombre des coefficients de la matrice du système sont nuls ; de plus la matrice a souvent une structure "bande", *i.e.* les éléments non nuls de la matrice sont localisés sur certaines diagonales. On a vu au chapitre précédent que dans ce cas, la méthode de Choleski "conserve le profil" (voir à ce propos page 43). Si on utilise une méthode directe genre Choleski, on aura donc besoin de la place mémoire pour stocker la strucutre bande.
Lorsqu'on a affaire à de très gros systèmes issus par exemple de l'ingénierie (calcul des structures, mécanique des fluides, ...), où n peut être de l'ordre de plusieurs milliers, on cherche à utiliser des méthodes nécessitant le moins de mémoire possible. On a intérêt dans ce cas à utiliser des méthodes itératives. Ces méthodes ne font appel qu'à des produits matrice vecteur, et ne nécessitent donc pas le stockage du profil de la matrice mais uniquement des termes non nuls. Dans l'exemple précédent, on a 5 diagonales non nulles, donc la place mémoire nécessaire pour un produit matrice vecteur est $5n = 5M^2$. Ainsi pour les gros systèmes, il est souvent avantageux d'utiliser des méthodes itératives qui ne donnent pas toujours la solution exacte du système en un nombre fini d'itérations, mais qui donnent une solution approchée à coût moindre qu'une méthode directe, car elles ne font appel qu'à des produits matrice vecteur.

Remarque 1.45 (Sur la méthode du gradient conjugué)**.**
Il existe une méthode itérative "miraculeuse" de résolution des systèmes linéaires lorsque la matrice A est symétrique définie positive : c'est la méthode du gradient conjugué. Elle est miraculeuse en ce sens qu'elle donne la solution exacte du système $Ax = b$ en un nombre fini d'opérations (en ce sens c'est une méthode directe) : moins de n itérations où n est l'ordre de la matrice A, bien qu'elle ne nécessite que des produits matrice vecteur ou des produits scalaires. La méthode du gradient conjugué est en fait une méthode d'optimisation pour la recherche du minimum dans $\mathbb{R}^n$ de la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par : $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Or on peut montrer que lorsque A est symétrique définie positive, la recherche de x minimisant f dans $\mathbb{R}^n$ est équivalent à la résolution du système $Ax = b$. (Voir paragraphe 3.2.2 page 214.) En fait, la méthode du gradient conjugué n'est pas si miraculeuse que cela en pratique : en effet, le nombre n est en général très grand et on ne peut en géneral pas envisager d'effectuer un tel nombre d'itérations pour résoudre le système. De plus, si on utilise la méthode du gradient conjugué brutalement, non seulement elle ne donne pas la solution en n itérations en raison de l'accumulation des erreurs d'arrondi, mais plus la taille du système croît et plus le nombre d'itérations nécessaires devient élevé. On a alors recours aux techniques de "préconditionnement". Nous reviendrons sur ce point au chapitre 3.
La méthode itérative du gradient à pas fixe, qui est elle aussi obtenue comme méthode de minimisation de la fonction f ci-dessus, fait l'objet de l'exercice 49 page 106 et du théorème **??** *page* **??**.

1.5.1 Définition et propriétés

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\boldsymbol{b} \in \mathbb{R}^n$, on cherche toujours ici à résoudre le système linéaire (1.1) c'est–à–dire à trouver $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = \boldsymbol{b}$, mais de façon itérative, c.à.d. par la construction d'une suite.

Définition 1.46 (Méthode itérative)**.** *On appelle méthode itérative de résolution du système linéaire* (1.1) *une méthode qui construit une suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ (où l'itéré $\boldsymbol{x}^{(k)}$ est calculé à partir des itérés $\boldsymbol{x}^{(0)} \dots \boldsymbol{x}^{(k-1)}$ censée converger vers x solution de* (1.1)*).*

Bien sûr, on souhaite que cette suite converge vers la solution $\boldsymbol{x}$ du système.

Définition 1.47 (Méthode itérative convergente)**.** *On dit qu'une méthode itérative est convergente si pour tout choix initial $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, on a :*

$$\boldsymbol{x}^{(k)} \longrightarrow \boldsymbol{x} \text{ quand } k \to +\infty$$

Enfin, on veut que cette suite soit simple à calculer. Une idée naturelle est de travailler avec une matrice P inversible qui soit "proche" de A, mais plus facile que A à inverser. On appelle matrice de préconditionnement cette matrice P. On écrit alors $A = P - (P - A) = P - N$ (avec $N = P - A$), et on réécrit le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ sous la forme

$$P\boldsymbol{x} = (P - A)\boldsymbol{x} + \boldsymbol{b} = N\boldsymbol{x} + \boldsymbol{b}. \tag{1.90}$$

Cette forme suggère la construction de la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ à partir d'un choix initial $\boldsymbol{x}^{(0)}$ donné, par la formule suivante :

$$\begin{aligned} P\boldsymbol{x}^{(k+1)} &= (P - A)\boldsymbol{x}^{(k)} + \boldsymbol{b} \\ &= N\boldsymbol{x}^{(k)} + \boldsymbol{b}, \end{aligned} \tag{1.91}$$

ce qui peut également s'écrire :.

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B = P^{-1}(P - A) = \mathrm{Id} - P^{-1}A = P^{-1}N \text{ et } \boldsymbol{c} = P^{-1}\boldsymbol{b}. \tag{1.92}$$

Remarque 1.48 (Convergence vers $A^{-1}\boldsymbol{b}$)**.** *Si $P\boldsymbol{x}^{(k+1)} = (P - A)\boldsymbol{x}^{(k)} + b$ pour tout $k \in \mathbb{N}$ et $\boldsymbol{x}^{(k)} \longrightarrow \bar{\boldsymbol{x}}$ quand $k \longrightarrow +\infty$ alors $P\bar{\boldsymbol{x}} = (P - A)\bar{\boldsymbol{x}} + b$, et donc $A\bar{\boldsymbol{x}} = \boldsymbol{b}$, c.à.d. $\bar{\boldsymbol{x}} = \boldsymbol{x}$. En conclusion, si la suite converge, alors elle converge bien vers la solution du système linéaire.*

On introduit l'erreur d'approximation $\boldsymbol{e}^{(k)}$ à l'itération k, définie par

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}, \ k \in \mathbb{N} \tag{1.93}$$

où $\boldsymbol{x}^{(k)}$ est construit par (1.92) et $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Il est facile de vérifier que $\boldsymbol{x}^{(k)} \to \boldsymbol{x} = A^{-1}\boldsymbol{b}$ lorsque $k \to +\infty$ si et seulement si $\boldsymbol{e}^{(k)} \to \boldsymbol{0}$ lorsque $k \to +\infty$

Lemme 1.49. *La suite $(\boldsymbol{e}^{(k)})_{k\in\mathbb{N}}$ définie par* (1.93) *est également définie par*

$$\begin{aligned} \boldsymbol{e}^{(0)} &= \boldsymbol{x}^{(0)} - \boldsymbol{x} \\ \boldsymbol{e}^{(k)} &= B^k\boldsymbol{e}^{(0)} \end{aligned} \tag{1.94}$$

DÉMONSTRATION – Comme $\boldsymbol{c} = P^{-1}\boldsymbol{b} = P^{-1}A\boldsymbol{x}$, on a

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x} = B\boldsymbol{x}^{(k)} - \boldsymbol{x} + P^{-1}A\boldsymbol{x} \tag{1.95}$$

$$= B(\boldsymbol{x}^{(k)} - \boldsymbol{x}). \tag{1.96}$$

Par récurrence sur k,

$$\boldsymbol{e}^{(k)} = B^k(\boldsymbol{x}^{(0)} - \boldsymbol{x}), \quad \forall k \in \mathbb{N}. \tag{1.97}$$

■

Théorème 1.50 (Convergence de la suite)**.** *Soit A et $P \in \mathcal{M}_n(\mathbb{R})$ des matrices inversibles. Soit $\boldsymbol{x}^{(0)}$ donné et soit $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ la suite définie par* (1.92)*.*

1. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ si et seulement si $\rho(B) < 1$.*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge si et seulement si il existe une norme induite notée $\|\cdot\|$ telle que $\|B\| < 1$.*

DÉMONSTRATION –

1. On a vu que la suite $(\boldsymbol{x})^{(k)})_{k\in\mathbb{N}}$ définie par (1.92) converge vers $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ si et seulement si la suite $\boldsymbol{e}^{(k)}$ définie par (1.94) tend vers $\mathbf{0}$. On en déduit par le lemme 1.33 que la suite $(\boldsymbol{x})^{(k)})_{k\in\mathbb{N}}$ converge (vers $\boldsymbol{x}$) si et seulement si $\rho(B) < 1$.
2. Si il existe une norme induite notée $\|\cdot\|$ telle que $\|B\| < 1$, alors en vertu du corollaire 1.33, $\rho(B) < 1$ et donc la méthode converge ce qui précède.
   Réciproquement, si la méthode converge alors $\rho(B) < 1$, et donc il existe $\eta > 0$ tel que $\rho(B) = 1 - \eta$. Prenons maintenant $\varepsilon = \frac{\eta}{2}$ et appliquons la proposition 1.32 : il existe une norme induite $\|\cdot\|$ telle que $\|B\| \leq \rho(B) + \varepsilon < 1$, ce qui démontre le résultat.

■

Pour trouver des méthodes itératives de résolution du système (1.1), on cherche donc une décomposition de la matrice A de la forme : $A = P - (P - A) = P - N$, où P est inversible et telle que le système $Py = d$ soit un système facile à résoudre (par exemple P diagonale ou triangulaire).

Estimation de la vitesse de convergence Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ donné et soit $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ la suite définie par (1.92). On a vu que si $\rho(B) < 1$), $\boldsymbol{x}^{(k)} \to x$ quand $k \to \infty$, où $\boldsymbol{x}$ est la solution du système $A\boldsymbol{x} = b$. On montre à l'exercice 64 page 136 que (sauf cas particuliers)

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|} \longrightarrow \rho(B) \quad \text{lorsque } k \to +\infty,$$

indépendamment de la norme sur $\mathbb{R}^n$. Le rayon spectral $\rho(B)$ de la matrice B est donc une bonne estimation de la vitesse de convergence. Pour estimer cette vitesse de convergence lorsqu'on ne connaît pas $\boldsymbol{x}$, on peut utiliser le fait (voir encore l'exercice 64 page 136) qu'on a aussi

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|} \longrightarrow \rho(B) \ : \ \text{lorsque } k \to +\infty,$$

ce qui permet d'évaluer la vitesse de convergence de la méthode par le calcul des itérés courants.

1.5.2 Quelques exemples de méthodes itératives

Une méthode simpliste

Le choix le plus simple pour le système $Px = (P - A)x + b$ soit facile à résoudre (on rappelle que c'est un objectif dans la construction d'une méthode itérative) est de prendre pour P la matrice identité (qui est très facile à inverser !). Voyons ce que cela donne sur la matrice

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}. \tag{1.98}$$

On a alors $B = P - A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$. Les valeurs propres de B sont 0 et -2 et on a donc $\rho(B) = 2 > 1$. La suite $(\boldsymbol{e}^{(k)})_{n\in\mathbb{N}}$ définie par $\boldsymbol{e}^{(k)} = B^k\boldsymbol{e}^{(0)}$ n'est donc en général pas convergente. En effet, si $\boldsymbol{e}^{(0)} = a\boldsymbol{u}_1 + b\boldsymbol{u}_2$, où $\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ est vecteur propre de A associé à la valeur propre $\lambda = -2$, on a $\boldsymbol{e}^{(k)} = (-2)^k a$ et donc $|\boldsymbol{e}^{(k)}| \to +\infty$ lorsque $k \to \infty$ dès que $a \neq 0$. Cette première idée n'est donc pas si bonne…

La méthode de Richardson

Affinons un peu et prenons maintenant $P = \beta\mathrm{Id}$, avec $\beta \in \mathbb{R}$. On a dans ce cas $P - A = \beta\mathrm{Id} - A$ et $B = \mathrm{Id} - \frac{1}{\beta}A = \mathrm{Id} - \alpha A$ avec $\alpha = \frac{1}{\beta}$. Les valeurs propres de B sont de la forme $1 - \alpha\lambda$, où λ est valeur propre de A. Pour la matrice A définie par (1.98), les valeurs propres de A sont 1 et 3, et les valeurs propres de

$$B = \begin{bmatrix} 1-2\alpha & \alpha \\ \alpha & 1-2\alpha, \end{bmatrix}$$

sont $1-\alpha$ et $1-3\alpha$. Le rayon spectral de la matrice B, qui dépend de α est donc $\rho(B) = \max(|1-\alpha|, |1-3\alpha|)$, qu'on représente sur la figure ci-dessous. La méthode itérative s'écrit

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné }, \\ &\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } \boldsymbol{c} = \alpha\boldsymbol{b}. \end{aligned} \tag{1.99}$$

Pour que la méthode converge, il faut et il suffit que $\rho(B) < 1$, c.à.d. $3\alpha - 1 < 1$, donc $\alpha < \frac{2}{3}$. On voit que le choix $\alpha = 1$ qu'on avait fait au départ n'était pas bon. Mais on peut aussi calculer le meilleur coefficient α pour avoir la meilleure convergence possible : c'est la valeur de α qui minimise le rayon spectral ρ ; il est atteint pour $1 - \alpha = 3\alpha - 1$, ce qui donne $\alpha = \frac{1}{2}$. Cette méthode est connue sous le nom de *méthode de Richardson* [8]. Elle est souvent écrite sous la forme :

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné }, \\ &\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\boldsymbol{r}^{(k)}, \end{aligned}$$

où $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ est le résidu. On vérifie facilement que cette forme est équivalente à la forme (1.99) qu'on vient d'étudier.

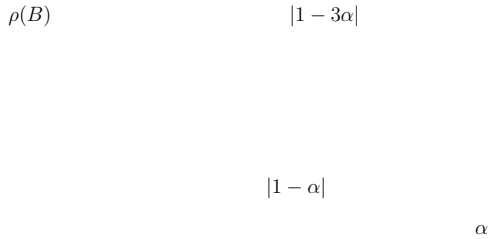

FIGURE 1.4: Rayon spectral de la matrice B de Richardson en fonction du coefficient α.

La méthode de Jacobi

Dans le cas de l'exemple de la matrice A donné par (1.98), la méthode de Richardson avec le coefficient optimal $\alpha = \frac{1}{2}$ revient à prendre comme décomposition de $A = P + A - P$ avec comme matrice $P = D$, où D est la

8. Lewis Fry Richardson, (1881-1953) est un mathématician, physicien, météorologue et psychologue qui a introduit les méthodes mathématiques pour les prévisions météorologiques. Il est également connu pour ses travaux sur les fractals. C'était un pacifiste qui a abandonné ses travaux de météorologie en raison de leur utillisation par l'armée de l'air, pour se tourner vers l'étude des raisons des guerres et de leur prévention.

matrice diagonale dont les coefficients sont les coefficients situés sur la diagonale de A. La *méthode de Jacobi*[9] consiste justement à prendre $P = D$, et ce même si la diagonale de A n'est pas constante.
Elle n'est équivalente à la méthode de Richardson avec coefficient optimal que dans le cas où la diagonale est constante ; c'est le cas de l'exemple (1.98), et donc dans ce cas la méthode de Jacobi s'écrit

$$\begin{aligned} &\boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné ,} \\ &\boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_J \boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_J = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0, \end{bmatrix} \text{ et } \boldsymbol{c} = \frac{1}{2}\boldsymbol{b}. \end{aligned} \tag{1.100}$$

Dans le cas d'une matrice A générale, on décompose A sous la forme $A = D - E - F$, où D représente la diagonale de la matrice A, $(-E)$ la partie triangulaire inférieure et $(-F)$ la partie triangulaire supérieure :

$$D = \begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & & 0 & a_{n,n} \end{bmatrix}, \; -E = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n,1} & \dots & a_{n-1,n} & 0 \end{bmatrix} \text{ et } -F = \begin{bmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ \vdots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & a_{n,n-1} \\ 0 & \dots & 0 & -0 \end{bmatrix}. \tag{1.101}$$

La méthode de Jacobi s'écrit donc :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ D\boldsymbol{x}^{(k+1)} = (E+F)\boldsymbol{x}^{(k)} + \boldsymbol{b}. \end{array} \right. \tag{1.102}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.92) on a $B = D^{-1}(E+F)$; on notera B_J cette matrice :

$$B_J = \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \dots & -\frac{-a_{1,n}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & \ddots & & -\frac{-a_{2,n}}{a_{2,2}} \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{a_{n,1}}{a_{n,n}} & \dots & -\frac{-a_{n-1,n}}{a_{n,n}} & 0 \end{bmatrix}.$$

La méthode de Jacobi s'écrit aussi :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\displaystyle\sum_{j<i} a_{i,j} x_j^{(k)} - \sum_{j>i} a_{i,j} x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{array} \right. \tag{1.103}$$

La méthode de Gauss-Seidel

Dans l'écriture (1.103) de la méthode de Jacobi, on pourrait remplacer les composantes $x_j^{(k)}$ dans la somme pour $j < i$ par les composantes $x_j^{(k+1)}$, puisqu'elles sont déjà calculées au moment où l'on calcule $x_i^{(k+1)}$. C'est l"idée de la méthode de Gauss-Seidel[10] qui consiste à utiliser le calcul des composantes de l'itéré $(k+1)$ dès qu'il est effectué. Par exemple, pour calculer la deuxième composante $x_2^{(k+1)}$ du vecteur $x^{(k+1)}$, on pourrait employer la

9. Carl G. J. Jacobi, (1804 - 1851), mathématicien allemand. Issu d'une famille juive, il étudie à l'Université de Berlin, où il obtient son doctorat à 21 ans. Sa thèse est une discussion analytique de la théorie des fractions. En 1829, il devient professeur de mathématique à l'Université de Königsberg, et ce jusqu'en 1842. Il fait une dépression, et voyage en Italie en 1843. À son retour, il déménage à Berlin où il sera pensionnaire royal jusqu'à sa mort. Sa lettre du 2 juillet 1830 adressée à Legendre est restée célèbre pour la phrase suivante, qui a fait couler beaucoup d'encre : "M. Fourier avait l'opinion que le but principal des mathématiques était l'utilité publique et l'explication des phénomènes naturels ; mais un philosophe comme lui aurait dû savoir que le but unique de la science, c'est l'honneur de l'esprit humain, et que sous ce titre, une question de nombres vaut autant qu'une question du système du monde." C'est une question toujours en discussion. . . .

10. Philipp Ludwig von Seidel (Zweibrücken, Allemagne 1821 – Munich, 13 August 1896) mathématicien allemand dont il est dit qu'il a découvert en 1847 le concept crucial de la convergence uniforme en étudiant une démonstration incorrecte de Cauchy.

"nouvelle" valeur $x_1^{(k+1)}$ qu'on vient de calculer plutôt que la valeur $x_1^{(k)}$ comme dans (1.103) ; de même, dans le calcul de $x_3^{(k+1)}$, on pourrait employer les "nouvelles" valeurs $x_1^{(k+1)}$ et $x_2^{(k+1)}$ plutôt que les valeurs $x_1^{(k)}$ et $x_2^{(k)}$. Cette idée nous suggère de remplacer dans (1.103) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{i<j} a_{i,j} x_j^{(k)} + b_i, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.104}$$

La méthode de Gauss–Seidel s'écrit donc sous la forme $P\boldsymbol{x}^{(k+1)} = (P - A)\boldsymbol{x}^{(k)} + \boldsymbol{b}$, avec $P = D - E$ et $P - A = F$:

$$\left\{ \begin{array}{l} \boldsymbol{x}_0 \in \mathbb{R}^n \\ (D - E)\boldsymbol{x}^{(k+1)} = F\boldsymbol{x}^{(k)} + b. \end{array} \right. \tag{1.105}$$

Si l'on écrit la méthode de Gauss–Seidel sous la forme $\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$, on voit assez vite que $B = (D - E)^{-1}F$; on notera $B_{\rm GS}$ cette matrice, dite matrice de Gauss-Seidel.
Ecrivons la méthode de Gauss-Seidel dans le cas de la matrice A donnée par (1.98) : on a dans ce cas $P = D - E = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix}$, $F = \begin{bmatrix} 0 & 1 \\ 0 & 0, \end{bmatrix}$. L'algorithme de Gauss-Seidel s'écrit donc :

$$\begin{array}{l} \boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné ,} \\ \boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_{GS}\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_{GS} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \text{ et } \boldsymbol{c} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \boldsymbol{b}. \end{array} \tag{1.106}$$

On a donc $\rho(B_{GS}) = \frac{1}{4}$. Sur cet exemple la méthode de Gauss-Seidel converge donc beaucoup plus vite que la méthode de Jacobi : Asymptotiquement, l'erreur est divisée par 4 au lieu de 2 pour la méthode de Jacobi. On peut montrer que c'est le cas pour toutes les matrices tridiagonales, comme c'est énoncé dans le théorème suivant :

Théorème 1.51 (Comparaison de Jacobi et Gauss-Seidel pour les matrices tridiagonales)**.** *On considère une matrice $A \in \mathcal{M}_n(\mathbb{R})$ tridiagonale, c.à.d. telle que $a_{i,j} = 0$ si $|i - j| > 1$; soient $B_{\rm GS}$ et $B_{\rm J}$ les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors :*

$$\rho(B_{\rm GS}) = (\rho(B_{\rm J}))^2.$$

Pour les matrices tridiagonales, la méthode de Gauss–Seidel converge (ou diverge) donc plus vite que celle de Jacobi.

La démonstration de ce résultat se fait en montrant que dans le cas tridiagonal, λ est valeur propre de la matrice d'itération de Jacobi si et seulement si λ^2 est valeur propre de la matrice d'itération de Gauss-Seidel. Elle est laissée à titre d'exercice.

Méthodes SOR et SSOR

L'idée de la méthode de sur-relaxation (SOR = Successive Over Relaxation) est d'utiliser la méthode de Gauss-Seidel pour calculer un itéré intermédiaire $\tilde{x}^{(k+1)}$ qu'on "relaxe" ensuite pour améliorer la vitesse de convergence de la méthode. On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^n \\ a_{i,i}\tilde{x}_i^{(k+1)} = -\displaystyle\sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{i<j} a_{i,j} x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1 - \omega)x_i^{(k)}, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.107}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)
L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $a_{i,i}$ la ligne 3 de l'algorithme (1.107)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = \omega \left[-\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \right] \\ \qquad\qquad +(1-\omega)a_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.108}$$

On obtient donc

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

La matrice d'itération de l'algorithme SOR est donc

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D) = P^{-1}N, \text{ avec } P = \frac{D}{\omega} - E \text{ et } N = F + \left(\frac{1-\omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = P - N$.

Proposition 1.52 (Condition nécessaire de convergence de la méthode SOR)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$ et soiet D, E et F les matrices définies par (1.101)*; on a donc $A = D - E - F$. Soit B_ω la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :*

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1} \left(F + \frac{1-\omega}{\omega} D\right), \quad \omega \neq 0.$$

Si $\rho(B_\omega) < 1$ alors $0 < \omega < 2$.

DÉMONSTRATION – Calculons $\det(B_\omega)$. Par définition,

$$B_\omega = P^{-1}N, \text{ avec } P = \frac{1}{\omega}D - E \text{ et } N = F + \frac{1-\omega}{\omega}D.$$

Donc $\det(B_\omega) = (\det(P))^{-1}\det(N)$. Comme P et N sont des matrices triangulaires, leurs déterminants sont les produits ceofficients diagonaux (voir la remarque 1.59 page 104). On a donc :

$$\det(B_\omega) = \frac{(\frac{1-\omega}{\omega})^n \det(D)}{(\frac{1}{\omega})^n \det(D)} = (1-\omega)^n.$$

Or le déterminant d'une matrice est aussi le produit des valeurs propres de cette matrice (comptées avec leur multiplicités algébriques), dont les valeurs absolues sont toutes, par définition, inférieures au rayon spectral. On a donc : $|\det(B_\omega)| = |(1-\omega)^n| \leq (\rho(B_\omega))^n$, d'où le résultat. ∎

On a un résultat de convergence de la méthode SOR (et donc également de Gauss-Seidel) dans le cas où A est symétrique définie positive, grâce au lemme suivant :

Lemme 1.53 (Condition suffisante de convergence pour la suite définie par (1.92))**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et soient P et $N \in \mathcal{M}_n(\mathbb{R})$ telles que $A = P - N$ et P est inversible. Si la matrice $P^t + N$ est symétrique définie positive alors $\rho(P^{-1}N) = \rho(B) < 1$, et donc la suite définie par* (1.92) *converge.*

DÉMONSTRATION – On rappelle (voir le corollaire (1.36) page 65) que si $B \in \mathcal{M}_n(\mathbb{R})$, et si $\|\cdot\|$ est une norme induite sur $\mathcal{M}_n(\mathbb{R})$ par une norme sur $\mathbb{R}^n$, on a toujours $\rho(B) \leq \|B\|$. On va donc chercher une norme sur $\mathbb{R}^n$, notée $\|\cdot\|_*$ telle que

$$\|P^{-1}N\|_* = \max\{\|P^{-1}N\boldsymbol{x}\|_*,\ \boldsymbol{x} \in \mathbb{R}^n,\ \|\boldsymbol{x}\|_* = 1\} < 1,$$

(où on désigne encore par $\|.\|_*$ la norme induite sur $\mathcal{M}_n(\mathbb{R})$) ou encore :

$$\|P^{-1}N\boldsymbol{x}\|_* < \|x\|_*,\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq 0. \tag{1.109}$$

On définit la norme $\|\cdot\|_*$ par $\|\boldsymbol{x}\|_* = \sqrt{A\boldsymbol{x}\cdot\boldsymbol{x}}$, pour tout $\boldsymbol{x} \in \mathbb{R}^n$. Comme A est symétrique définie positive, $\|\cdot\|_*$ est bien une norme sur $\mathbb{R}^n$, induite par le produit scalaire $(\boldsymbol{x}|\boldsymbol{y})_A = A\boldsymbol{x}\cdot\boldsymbol{y}$. On va montrer que la propriété (1.109) est vérifiée par cette norme. Soit $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$, on a : $\|P^{-1}N\boldsymbol{x}\|_*^2 = AP^{-1}N\boldsymbol{x}\cdot Pt^{-1}N\boldsymbol{x}$. Or $N = P - A$, et donc : $\|P^{-1}N\boldsymbol{x}\|_*^2 = A(\mathrm{Id} - Pt^{-1}A)\boldsymbol{x}\cdot(Id - P^{-1}A)\boldsymbol{x}$. Soit $\boldsymbol{y} = P^{-1}A\boldsymbol{x}$; remarquons que $\boldsymbol{y} \neq 0$ car $\boldsymbol{x} \neq 0$ et $P^{-1}A$ est inversible. Exprimons $\|P^{-1}N\boldsymbol{x}\|_*^2$ à l'aide de $\boldsymbol{y}$.

$$\|P^{-1}N\boldsymbol{x}\|_*^2 = A(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y}) = A\boldsymbol{x}\cdot\boldsymbol{x} - 2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = \|\boldsymbol{x}\|_*^2 - 2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y}.$$

Pour que $\|P^{-1}N\boldsymbol{x}\|_*^2 < \|\boldsymbol{x}\|_*^2$ (et par suite $\rho(Pt^{-1}N) < 1$), il suffit donc de montrer que $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} < 0$. Or, comme $P\boldsymbol{y} = A\boldsymbol{x}$, on a : $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = -2P\boldsymbol{y}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y}$. En écrivant : $P\boldsymbol{y}\cdot\boldsymbol{y} = y\cdot P^t y = P^t y\cdot y$, on obtient donc que : $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = (-P - P^t + A)\boldsymbol{y}\cdot\boldsymbol{y}$, et comme $A = P - N$ on obtient $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = -(P^t + N)\boldsymbol{y}\cdot\boldsymbol{y}$. Comme $P^t + N$ est symétrique définie positive par hypothèse et que $\boldsymbol{y} \neq \mathbf{0}$, on en déduit que $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} < 0$, ce qui termine la démonstration. ∎

Théorème 1.54 (CNS de convergence de la méthode SOR pour les matrices s.d.p.)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et soient D, E et F les matrices définies par (1.101)*; on a donc $A = D - E - F$. Soit B_ω la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :*

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right),\quad \omega \neq 0.$$

Alors :

$$\rho(B_\omega) < 1 \text{ si et seulement si } 0 < \omega < 2.$$

En particulier, si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge.

DÉMONSTRATION – On sait par la proposition 1.52 que si $\rho(B_\omega) < 1$ alors $0 < \omega < 2$. Supposons maintenant que A est une matrice symétrique définie positive, que $0 < \omega < 2$et montrons que $\rho(B_\omega) < 1$. Par le lemme 1.53 page 101, il suffit pour cela de montrer que $P^t + N$ est une matrice symétrique définie positive. Or,

$$P^t = \left(\frac{D}{\omega} - E\right)^t = \frac{D}{\omega} - F,$$

$$P^t + N = \frac{D}{\omega} - F + F + \frac{1-\omega}{\omega}D = \frac{2-\omega}{\omega}D.$$

La matrice $P^t + N$ est donc bien symétrique définie positive. ∎

Remarque 1.55 (Comparaison Gauss–Seidel/Jacobi)**.** *On a vu (théorème 1.54) que si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge. Par contre, même dans le cas où A est symétrique définie positive, il existe des cas où la méthode de Jacobi ne converge pas, voir à ce sujet l'exercice 50 page 107.*

Remarquons que le résultat de convergence des méthodes itératives donné par le théorème précédent n'est que partiel, puisqu'il ne concerne que les matrices symétriques définies positives et que les méthodes Gauss-Seidel et SOR. On a aussi un résultat de convergence de la méthode de Jacobi pour les matrices à diagonale dominante stricte, voir exercice 54 page 108, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.56 donné ci-après. Dans la pratique, il faudra souvent compter sur sa bonne étoile…

Estimation du coefficient de relaxation optimal de SOR La question est ici d'estimer le coefficient de relaxation ω optimal dans la méthode SOR, c.à.d. le coefficient $\omega_0 \in]0,2[$ (condition nécessaire pour que la méthode SOR converge, voir théorème 1.54) tel que

$$\rho(\mathcal{L}_{\omega_0}) < \rho(B_\omega), \forall \omega \in]0,2[.$$

Ce coefficient ω_0 donnera la meilleure convergence possible pour SOR. On sait le faire dans le cas assez restrictif des matrices tridiagonales (ou tridiagonales par blocs, voir paragraphe suivant). On ne fait ici qu'énoncer le résultat dont la démonstration est donnée dans le livre de Ph.Ciarlet conseillé en début de cours.

Théorème 1.56 (Coefficient optimal, matrice tridiagonale)**.** *On considère une matrice $A \in \mathcal{M}_n(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.110 page 104 ; on suppose que la matrice A est tridiagonale par blocs, c.à.d. $A_{i,j} = 0$ si $|i-j| > 1$; soient B_{GS} et B_{J} les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors : On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles ; alors le paramètre de relaxation optimal, c.à.d. le paramètre ω_0 tel que $\rho(B_{\omega_0}) = \min\{\rho(B_\omega), \omega \in]0,2[\}$, s'exprime en fonction du rayon spectral $\rho(B_{\mathrm{J}})$ de la matrice J par la formule :*

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(B_{\mathrm{J}})^2}} > 1,$$

et on a : $\rho(B_{\omega_0}) = \omega_0 - 1$.

La démonstration de ce résultat repose sur la comparaison des valeurs propres des matrices d'itération. On montre que λ est valeur propre de B_ω si et seulement si

$$(\lambda + \omega - 1)^2 = \lambda \omega \mu^2,$$

où μ est valeur propre de B_{J} (voir [Ciarlet] pour plus de détails).

Remarque 1.57 (Méthode de Jacobi relaxée)**.** *On peut aussi appliquer une procédure de relaxation avec comme méthode iérative "de base" la méthode de Jacobi, voir à ce sujet l'exercice 57 page 109). Cette méthode est toutefois beaucoup moins employée en pratique (car moins efficace) que la méthode SOR.*

Méthode SSOR En "symétrisant" le procédé de la méthode SOR, c.à.d. en effectuant les calculs SOR sur les blocs dans l'ordre 1 à n puis dans l'ordre n à 1, on obtient la méthode de sur-relaxation symétrisée (SSOR = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode I avec

$$B_{SSOR} = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1}\left(E + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \quad n \ldots 1}\underbrace{\left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \quad 1 \ldots n}.$$

1.5.3 Les méthodes par blocs

Décomposition par blocs d'une matrice

Dans de nombreux cas pratiques, les matrices des systèmes linéaires à résoudre ont une structure "par blocs", et on se sert alors de cette structure lors de la résolution par une méthode itérative.

Définition 1.58. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible ; une décomposition par blocs de A est définie par un entier $S \leq n$, des entiers $(n_i)_{i=1,\dots,S}$ tels que $\sum_{i=1}^{S} n_i = n$, et S^2 matrices $A_{i,j} \in \mathcal{M}_{n_i,n_j}(\mathbb{R})$ (ensemble des matrices rectangulaires à n_i lignes et n_j colonnes, telles que les matrices $A_{i,i}$ soient inversibles pour $i = 1, \dots, S$ et*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & & \dots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \dots & & \dots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.110}$$

Remarque 1.59.

1. *Si $S = n$ et $n_i = 1$ $\forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on retrouve la structure habituelle d'une matrice. Les méthodes que nous allons décrire maintenant sont alors celles que nous avons vu dans le cas de matrices sans structure particulière.*
2. *Si A est symétrique définie positive, la condition $A_{i,i}$ inversible dans la définition 1.58 est inutile car $A_{i,i}$ est nécessairement symétrique définie positive donc inversible. Pour s'en convaincre, prenons par exemple $i = 1$; soit $y \in \mathbb{R}^{n_1}$, $y \neq 0$ et $x = (y, 0 \dots, 0)^t \in \mathbb{R}^n$. Alors $A_{1,1}y \cdot y = Ax \cdot x > 0$ donc $A_{1,1}$ est symétrique définie positive.*
3. *Si A est une matrice triangulaire par blocs, c.à.d. de la forme* (1.110) *avec $A_{i,j} = 0$ si $j > i$, alors*

$$\det(A) = \prod_{i=1}^{S} \det(A_{i,i}).$$

Par contre si A est décomposée en 2×2 blocs carrés (i.e. tels que $n_i = m_j$, $\forall (i,j) \in \{1,2\}$), on a en général :

$$\det(A) \neq \det(A_{1,1})\det(A_{2,2}) - \det(A_{1,2})\det(A_{2,1}).$$

Méthode de Jacobi

On cherche une matrice P tel que le système $Px = (P - A)x + b$ soit facile à résoudre (on rappelle que c'est un objectif dans la construction d'une méthode itérative). On avait pris pour P une matrice diagonale dans la méthode de Jacobi. La méthode de Jacobi par blocs consiste à prendre pour P la matrice diagonale D formée par les blocs diagonaux de A :

$$D = \begin{bmatrix} A_{1,1} & 0 & \dots & & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & & \dots & 0 & A_{S,S} \end{bmatrix}.$$

Dans la matrice ci-dessus, 0 désigne un bloc nul.

On a alors $N = P - A = E + F$, où E et F sont constitués des blocs triangulaires inférieurs et supérieurs de la matrice A :

$$E = \begin{bmatrix} 0 & 0 & \dots & & \dots & 0 \\ -A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ -A_{S,1} & \dots & & \dots & -A_{S,S-1} & 0 \end{bmatrix}, \; F = \begin{bmatrix} 0 & -A_{1,2} & \dots & & \dots & -A_{1,S} \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & -A_{S-1,S} \\ 0 & \dots & & \dots & 0 & 0 \end{bmatrix}.$$

On a bien $A = P - N$ et avec D, E et F définies comme ci-dessus, la méthode de Jacobi s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ Dx^{(k+1)} = (E + F)x^{(k)} + b. \end{cases} \tag{1.111}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.92) on a $B = D^{-1}(E + F)$; on notera J cette matrice. En introduisant la décomposition par blocs de x, solution recherchée de (1.1), c.à.d. : $x = [x_1, \dots, x_S]^t$, où $x_i \in \mathbb{R}^{n_i}$, on peut aussi écrire la méthode de Jacobi sous la forme :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, S. \end{cases} \tag{1.112}$$

Si $S = n$ et $n_i = 1 \; \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on obtient la méthode de Jacobi par points (aussi appelée méthode de Jacobi), qui s'écrit donc :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{cases} \tag{1.113}$$

Méthode de Gauss-Seidel

La même procédure que dans le cas $S = n$ et $n_i = 1$ donne :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i, \quad i = 1, \dots, S. \end{cases} \tag{1.114}$$

La méthode de Gauss–Seidel s'écrit donc sous forme la forme $Px^{(k+1)} = (P - A)x^{(k)} + b$, $P = D - E$ et $P - A = F$:

$$\begin{cases} x_0 \in \mathbb{R}^n \\ (D - E)x^{(k+1)} = Fx^{(k)} + b. \end{cases} \tag{1.115}$$

Si l'on écrit la méthode de Gauss–Seidel sous la forme $x^{(k+1)} = Bx^{(k)} + c$, on voit assez vite que $B = (D - E)^{-1}F$; on notera B_{GS} cette matrice, dite matrice de Gauss-Seidel.

Méthodes SOR et SSOR

La méthode SOR s'écrit aussi par blocs : on se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1 - \omega)x_i^{(k)}, \quad i = 1, \dots, S. \end{cases} \tag{1.116}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)
L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $A_{i,i}$ la ligne 3 de l'algorithme (1.107)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = \omega \left[-\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \right] \\ \qquad\qquad +(1-\omega)A_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.117}$$

On obtient donc

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

L'algorithme SOR s'écrit donc comme une méthode II avec

$$P = \frac{D}{\omega} - E \text{ et } N = F + \left(\frac{1-\omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = P - N$.

L'algorithme SOR s'écrit aussi comme une méthode I avec

$$B = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D).$$

On notera $\mathcal{L}_\omega$ cette matrice.

Remarque 1.60 (Méthode de Jacobi relaxée)**.** *On peut aussi appliquer une procédure de relaxation avec comme méthode iérative "de base" la méthode de Jacobi, voir à ce sujet l'exercice 57 page 109). Cette méthode est toutefois beaucoup moins employée en pratique (car moins efficace) que la méthode SOR.*

En "symétrisant" le procédé de la méthode SOR, c.à.d. en effectuant les calculs SOR sur les blocs dans l'ordre 1 à n puis dans l'ordre n à 1, on obtient la méthode de sur-relaxation symétrisée (SSOR = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode I avec

$$B = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1} \left(E + \frac{1-\omega}{\omega} D\right)}_{\text{calcul dans l'ordre } S \ldots 1} \underbrace{\left(\frac{D}{\omega} - E\right)^{-1} \left(F + \frac{1-\omega}{\omega} D\right)}_{\text{calcul dans l'ordre } 1 \ldots S}.$$

1.5.4 Exercices, énoncés

Exercice 48 (Convergence de suites)**.** *Corrigé en page 115*

Etudier la convergence de la suite $(\boldsymbol{x}^{(k)})_{n\in\mathbb{N}} \subset \mathbb{R}^n$ définie par $\boldsymbol{x}^{(0)}$ donné, $\boldsymbol{x}^{(k)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$ dans les cas suivants :

$$\text{(a)} \quad B = \begin{bmatrix} \frac{2}{3} & 1 \\ 0 & \frac{2}{3} \end{bmatrix}, \ \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \text{(b)} \quad B = \begin{bmatrix} \frac{2}{3} & 1 \\ 0 & 2 \end{bmatrix}, \ \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Exercice 49 (Méthode de Richardson)**.** *Suggestions en page 114, corrigé en page 115*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, $b \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$. Pour trouver la solution de $Ax = b$, on considère la méthode itérative suivante :

- Initialisation : $x^{(0)} \in \mathbb{R}^n$,
- Iterations : $x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)})$.

1. Pour quelles valeurs de α (en fonction des valeurs propres de A) la méthode est-elle convergente ?

2. Calculer α_0 (en fonction des valeurs propres de A) t.q. $\rho(Id - \alpha_0 A) = \min\{\rho(Id - \alpha A), \alpha \in \mathbb{R}\}$.

Commentaire sur la méthode de Richardson : On peut la voir comme une méthode de gradient à pas fixe pour la minimisation de la fonction f définie de $\mathbb{R}^N$ dans $\mathbb{R}$ par : $\boldsymbol{x} \mapsto f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}$, qui sera étudiée au chapitre Optimisation. On verra en effet que grâce qu caractère symétrique définie positif de A, la fonction $\boldsymbol{f}$ admet un unique minimum, caractérisé par l'annulation du gradient de f en ce point. Or $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$, et annuler le gradient consiste à résoudre le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.

Exercice 50 (Non convergence de la méthode de Jacobi)**.** *Suggestions en page 114. Corrigé en page 116.*

Soit $a \in \mathbb{R}$ et

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

Montrer que A est symétrique définie positive si et seulement si $-1/2 < a < 1$ et que la méthode de Jacobi converge si et seulement si $-1/2 < a < 1/2$.

Exercice 51 (Jacobi et Gauss–Seidel : cas des matrices tridiagonales)**.** *Corrigé en page 116.**Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n inversible et tridiagonale ; on note $a_{i,j}$ le coefficient de la ligne i et la ligne j de la matrice A. On décompose en $A = D - E - F$, où D représente la diagonale de la matrice A, $(-E)$ la partie triangulaire inférieure stricte et $(-F)$ la partie triangulaire supérieure stricte.*
On note B_J et B_{GS} les matrices d'itération des méthodes de Jacobi et Gauss-Seidel pour la résolution d'un système linéaire de matrice A.

1. *Calculer les matrices B_J et B_{GS} pour la matrice particulère $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ et calculer leurs rayons spectraux.Montrer que les méthodes convergent, et citez les résultat du cours qui s'appliquent pour cette matrice.*
2. *Montrer que λ est valeur propre de B_J si et seulement s'il existe un vecteur complexe $\boldsymbol{x} = (x_1, \dots x_n) \in \mathbb{C}^n$, $\boldsymbol{x} \neq \boldsymbol{0}$, tel que*
$$-a_{p,p-1}x_{p-1} - a_{p,p+1}x_{p+1} = \lambda a_{p,p}x_p, \quad p = 1, \dots, n.$$
avec $x_0 = x_{n+1} = 0$.
3. *Soit $\boldsymbol{y} = (y_1, \dots y_n) \in \mathbb{C}^n$ défini par $y_p = \lambda^{-p}x_p$, où λ est une valeur propre non nulle de B_J et $\boldsymbol{x} = (x_1, \dots, x_n)$ un vecteur propre associé. On pose $y_0 = y_{n+1} = 0$. Montrer que*
$$-a_{p,p-1}y_{p-1} - \lambda^2 a_{p,p+1}y_{p+1} = \lambda^2 a_{p,p}y_p, \quad p = 1, \dots, n.$$
4. *Montrer que μ est valeur propre de B_{GS} associée à un vecteur propre $\boldsymbol{z} \neq \boldsymbol{0}$ si et seulement si*
$$(F - \mu(D - E))\,\boldsymbol{z} = \boldsymbol{0}.$$
5. *Montrer que λ est valeur propre non nulle de B_J si et seulement si λ^2 est valeur propre de B_{GS}, et en déduire que $\rho(B_{GS}) = \rho(B_J)^2$.*
6. *On considère la matrice :*
$$A = \begin{bmatrix} 1 & \frac{3}{4} & \frac{3}{4} \\ \frac{3}{4} & 1 & \frac{3}{4} \\ \frac{3}{4} & \frac{3}{4} & 1 \end{bmatrix}$$
Montrer que cette matrice est symétrique définie positive. Montrer que $\rho(B_{GS}) \neq \rho(B_J)$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 52 (Jacobi pour une matrice 3×3 particulière)**.** *Corrigé en page 117.**Soit $A = \begin{bmatrix} a & 0 & \alpha \\ 0 & b & 0 \\ \alpha & 0 & c \end{bmatrix}$. On suppose que A est symétrique définie positive. Montrer que la méthode de Jacobi converge pour n'importe quel second membre et n'importe quel choix initial.*

Exercice 53 (Une matrice cyclique)**.** *Suggestions en page 114 Corrigé en page 117*

Soit $\alpha \in \mathbb{R}$ et soit $A \in \mathcal{M}_4(\mathbb{R})$ la matrice définie par

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

Cette matrice est dite cyclique : chaque ligne de la matrice peut être déduite de la précédente en décalant chaque coefficient d'une position.

1. Déterminer les valeurs propres de A.
2. Pour quelles valeurs de α la matrice A est-elle symétrique définie positive ? singulière ?
3. On suppose ici que $\alpha \neq 0$. Soit $b = (b_1, b_2, b_3, b_4)^t \in \mathbb{R}^4$ donné. On considère la méthode de Jacobi pour la résolution du système $Ax = b$. Soit $(x^{(k)})_{n\in\mathbb{N}}$ la suite de vecteurs donnés par l'algorithme. On note $x_i^{(k)}$ pour $i = 1, \ldots, 4$ les composantes de $x^{(k)}$. Donner l'expression de $x_i^{(k+1)}$, $i = 1, \ldots, 4$, en fonction de $x_i^{(k)}$ et $b_i^{(k)}$, $i = 1, \ldots, 4$. Pour quelles valeurs de α la méthode de Jacobi converge-t-elle ?
4. On suppose maintenant que A est symétrique définie positive. Reprendre la question précédente pour la méthode de Gauss-Seidel.

Exercice 54 (Jacobi pour les matrices à diagonale dominante stricte)**.** *Suggestions en page 114, corrigé en page 119*

Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice à diagonale dominante stricte (c'est-à-dire $|a_{i,i}| > \sum_{j\neq i} |a_{i,j}|$ pour tout $i = 1, \ldots, n$). Montrer que A est inversible et que la méthode de Jacobi (pour calculer la solution de $Ax = b$) converge.

Exercice 55 (Jacobi pour pour un problème de diffusion)**.** *Corrigé en page 120.*

Soit $f \in C([0, 1])$; on considère le système linéaire $Ax = b$ issu de la discrétisation par différences finies de pas uniforme égal à $h = \frac{1}{n+1}$ du problème suivant :

$$\begin{cases} -u''(x) + \alpha u(x) = f(x),\ x \in [0, 1], \\ u(0) = 0, u(1) = 1, \end{cases} \tag{1.118}$$

où $\alpha \geq 0$.

1. Donner l'expression de A et $\boldsymbol{b}$.
2. Montrer que la méthode de Jacobi appliquée à la résolution de ce système converge (distinguer les cas $\alpha > 0$ et $\alpha = 0$).

Exercice 56 (Jacobi et diagonale dominance forte)**.** *Corrigé en page 121*

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive.
   (a) Montrer que tous les coefficients diagonaux de A sont strictement positifs.
   (b) En déduire que la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, est bien définie.

Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, avec $n > 1$. On dit que la matrice M est irréductible si :

$$\text{pour tous ensembles d'indices } I \subset \{1, \ldots, n\}, I \neq \emptyset, \text{ et } J = \{1 \ldots, n\} \setminus I, J \neq \emptyset,\ \exists i \in I, \exists j \in J; a_{i,j} \neq 0. \tag{1.119}$$

2 (a) Montrer qu'une matrice diagonale n'est pas irréductible. En déduire qu'une matrice inversible n'est pas forcément irréductible.

2 (b) Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, qui s'écrit sous la forme :

$$M = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}$$

où A et C sont des matrices carrées d'ordre p et q, avec $p+q=n$, et $B \in \mathcal{M}_{q,p}(\mathbb{R})$. La matrice M peut-elle être irréductible ?

3. Soit $A \in \mathcal{M}_n(\mathbb{R})$, $n > 1$ une matrice irréductible qui vérifie de plus la propriété suivante :

$$\forall i = 1, \ldots, n, a_{i,i} \geq \sum_{j \neq i} |a_{i,j}| \tag{1.120}$$

(On dit que la matrice est à diagonale dominante). Montrer que la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, est bien définie.

4. Soit $A \in \mathcal{M}_n(\mathbb{R})$, $n > 1$ une matrice irréductible qui vérifie la propriété (1.120). On note B_{J} la matrice d'itération de la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, et $\rho(B_{\mathrm{J}})$ son rayon spectral. On suppose que A vérifie la propriété supplémentaire suivante :

$$\exists i_0; a_{i_0,i_0} > \sum_{j \neq i_0} |a_{i,j}|. \tag{1.121}$$

(a) Montrer que $\rho(B_{\mathrm{J}}) \leq 1$.

(b) Montrer que si $Jx = \lambda x$ avec $|\lambda| = 1$, alors $|x_i| = \|x\|_\infty$, $\forall i = 1, \ldots, n$, où $\|x\|_\infty = \max_{k=1,\ldots,N} |x_k|$. En déduire que $x = 0$ et que la méthode de Jacobi converge.

(c) Retrouver ainsi le résultat de la question 2 de l'exercice 55.

5. En déduire que si A est une matrice qui vérifie les propriétés (1.119), (1.120) et (1.121), alors A est inversible.

6. Montrer que la matrice A suivante est symétrique définie positive et vérifie les propriétés (1.120) et (1.121).

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

La méthode de Jacobi converge-t-elle pour la résolution d'un système linéaire dont la matrice est A ?

Exercice 57 (Méthode de Jacobi et relaxation)**.** *Suggestions en page 114, corrigé en page 122*

Soit $n \geq 1$. Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure de A et $-F$ la partie triangulaire supérieure de A, c'est-à-dire :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\ldots,n},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\ldots,n},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\ldots,n},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

Noter que $A = D - E - F$. Soit $b \in \mathbb{R}^n$. On cherche à calculer $x \in \mathbb{R}^n$ t.q. $Ax = b$. On suppose que D est définie positive (noter que A n'est pas forcément inversible). On s'intéresse ici à la méthode de Jacobi (par points), c'est–à–dire à la méthode itérative suivante :

Initialisation. $x^{(0)} \in \mathbb{R}^n$
Itérations. Pour $n \in \mathbb{N}$, $Dx^{(k+1)} = (E+F)x^{(k)} + b$.

On pose $J = D^{-1}(E+F)$.

1. Montrer, en donnant un exemple avec $n = 2$, que J peut ne pas être symétrique.

2. Montrer que J est diagonalisable dans $\mathbb{R}$ et, plus précisement, qu'il existe une base de $\mathbb{R}^n$, notée $\{f_1, \ldots, f_n\}$, et il existe $\{\mu_1, \ldots, \mu_n\} \subset \mathbb{R}$ t.q. $J\boldsymbol{f}_i = \mu_i \boldsymbol{f}_i$ pour tout $i \in \{1, \ldots, n\}$ et t.q. $D\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, n\}$.

   En ordonnant les valeurs propres de J, on a donc $\mu_1 \leq \ldots \leq \mu_n$, on conserve cette notation dans la suite.

3. Montrer que la trace de J est nulle et en déduire que $\mu_1 \leq 0$ et $\mu_n \geq 0$.

   On suppose maintenant que A et $2D - A$ sont symétriques définies positives et on pose $x = A^{-1}b$.

4. Montrer que la méthode de Jacobi (par points) converge (c'est-à-dire $x^{(k)} \to x$ quand $n \to \infty$). [Utiliser un théorème du cours.]

   On se propose maintenant d'améliorer la convergence de la méthode par une technique de relaxation. Soit $\omega > 0$, on considère la méthode suivante :

   Initialisation. $x^{(0)} \in \mathbb{R}^n$

   Itérations. Pour $n \in \mathbb{N}$, $D\tilde{x}^{(k+1)} = (E + F)x^{(k)} + b$, $x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}$.

5. Calculer les matrices M_ω (inversible) et N_ω telles que $M_\omega x^{(k+1)} = N_\omega x^{(k)} + b$ pour tout $n \in \mathbb{N}$, en fonction de ω, D et A. On note, dans la suite $J_\omega = (M_\omega)^{-1}N_\omega$.

6. On suppose dans cette question que $(2/\omega)D - A$ est symétrique définie positive. Montrer que la méthode converge (c'est-à-dire que $x^{(k)} \to x$ quand $n \to \infty$.)

7. Montrer que $(2/\omega)D - A$ est symétrique définie positive si et seulement si $\omega < 2/(1 - \mu_1)$.

8. Calculer les valeurs propres de J_ω en fonction de celles de J. En déduire, en fonction des μ_i, la valeur "optimale" de ω, c'est-à-dire la valeur de ω minimisant le rayon spectral de J_ω.

Exercice 58 (Méthodes de Jacobi et Gauss Seidel pour une matrice 3×3)**.** *Corrigé détaillé en page 125*

On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ et le vecteur $b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. Soit $x^{(0)}$ un vecteur de $\mathbb{R}^3$ donné.

1. *Méthode de Jacobi*

1.a Ecrire la méthode de Jacobi pour la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$, sous la forme $\boldsymbol{x}^{(k+1)} = B_J\boldsymbol{x}^{(k)} + \boldsymbol{c}_J$.

1.b Déterminer le noyau de B_J et en donner une base.

1.c Calculer le rayon spectral de B_J et en déduire que la méthode de Jacobi converge.

1.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\, x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii) x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

2. *Méthode de Gauss-Seidel.*

2.a Ecrire la méthode de Gauss-Seidel pour la résolution du système $A\boldsymbol{x} = b$, sous la forme $\boldsymbol{x}^{(k+1)} = B_{GS}\boldsymbol{x}^{(k)} + \boldsymbol{c}_{GS}$.

2.b Déterminer le noyau de B_{GS}.

2.c Calculer le rayon spectral de B_{GS} et en déduire que la méthode de Gauss-Seidel converge.

2.d Comparer les rayons spectraux de B_{GS} et B_J et vérifier ainsi un résultat du cours.

2.d Calculer $\boldsymbol{x}^{(1)}$ et $\boldsymbol{x}^{(2)}$ pour les choix suivants de $\boldsymbol{x}^{(0)}$:

$$(i)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Exercice 59 (Convergence en un nombre fini d'itérations)**.** *Corrigé détaillé en page 126.*

1 Soit α et β des réels. Soit $u^{(0)} \in \mathbb{R}$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite réelle définie par $u^{(k+1)} = \alpha u^{(k)} + \beta$.

1.a Donner les valeurs de α et β pour lesquelles la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge.

1.b On suppose que $\alpha \neq 0$, et que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer que s'il existe $K \in \mathbb{N}$ tel que $u_K = \overline{u}$, alors $u^{(k)} = \overline{u}$ pour tout $k \in \mathbb{N}$.

2 Soit $n > 1$, B une matrice réelle carrée d'ordre n et $b \in \mathbb{R}^n$. Soit $u^{(0)} \in \mathbb{R}^N$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite définie par $u^{(k+1)} = Bu^{(k)} + c$.

2.a Donner les conditions sur B et c pour que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite indépendante du choix initial $u_0 \in \mathbb{R}^N$.

2.b On suppose que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer qu'on peut avoir $u^{(1)} = \overline{u}$ avec $u^{(0)} \neq \overline{u}$.

Exercice 60 (SOR et Jacobi pour une matrice tridiagonale)**.**
Corrigé en page **??**

Soit $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n tridiagonale, c'est-à-dire telle que $a_{i,j} = 0$ si $|i-j| > 1$, et dont la matrice diagonale extraite $D = diag(a_{i,i})_{i=1,\dots,n}$ est inversible.
Soit B_ω la matrice d'itération de la méthode SOR associée à A. Montrer que λ est valeur propre de J si et seulement si ν_ω est valeur propre de B_ω, où $\nu_\omega = \mu_\omega^2$ et μ_ω vérifie $\mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0$.

En déduire que

$$\rho(B_\omega) = \max_{\lambda \text{ valeur propre de } J} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}.$$

Exercice 61 (Méthode de Jacobi pour des matrices particulières)**.** *Suggestions en page 114, corrigé en page 127*

On note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées d'ordre n à coefficients réels, et Id la matrice identité dans $\mathcal{M}_n(\mathbb{R})$. Soit $A = [a_{i,j}]_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$. On suppose que :

$$a_{i,j} \leq 0, \forall i, j = 1, \dots, n, i \neq j, \tag{1.122}$$

$$a_{i,i} > 0, \forall i = 1, \dots, n. \tag{1.123}$$

$$\sum_{i=1}^{n} a_{i,j} = 0, \forall j = 1, \dots, n. \tag{1.124}$$

Soit $\lambda \in \mathbb{R}_+^*$.

1. Pour $x \in \mathbb{R}^n$, on définit
$$\|x\|_A = \sum_{i=1}^{n} a_{i,i}|x_i|.$$
Montrer que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^n$.
2. Montrer que la matrice $\lambda\mathrm{Id} + A$ est inversible.
3. On considère le système linéaire suivant :
$$(\lambda\mathrm{Id} + A)u = b \tag{1.125}$$
Montrer que la méthode de Jacobi pour la recherche de la solution de ce système définit une suite $(u^{(k)})_{k\in n} \subset \mathbb{R}^n$.
4. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ vérifie :
$$\|u^{(k+1)} - u^{(k)}\|_A \leq (\frac{1}{1+\alpha})^k \|u^{(1)} - u^{(0)}\|_A,$$
où $\alpha = \min_{i=1,\dots,n} a_{i,i}$.

5. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ est de Cauchy, et en déduire qu'elle converge vers la solution du système (1.125).

Exercice 62 (Une méthode itérative particulière). *Corrigé en page 130.*

Soient $\alpha_1, \dots, \alpha_n$ des réels strictement positifs, et A la matrice $n \times n$ de coefficients $a_{i,j}$ définis par :

$$\begin{cases} a_{i,i} = 2 + \alpha_i \\ a_{i,i+1} = a_{i,i-1} = -1 \\ a_{i,j} = 0 \text{ pour tous les autres cas.} \end{cases}$$

Pour $\beta > 0$ on considère la méthode itérative $Px^{(k+1)} = Nx^{(k)} + b$ avec $A = P - N$ et $N = \text{diag}(\beta - \alpha_i)$ (c.à.d $\beta - \alpha_i$ pour les coefficients diagonaux, et 0 pour tous les autres).

1. Soit $\lambda \in \mathbb{C}$ une valeur propre de la matrice $P^{-1}N$; montrer qu'il existe un vecteur $x \in \mathbb{C}^n$ non nul tel que $Nx \cdot \overline{x} = \lambda Px \cdot \overline{x}$ (où $\overline{x}$ désigne le conjugué de x). En déduire que toutes les valeurs propres de la matrice $P^{-1}N$ sont réelles.

2. Montrer que le rayon spectral $\rho(P^{-1}N)$ de la matrice vérifie : $\rho(P^{-1}N) \leq \max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$

3. Déduire de la question 1. que si $\beta > \dfrac{\overline{\alpha}}{2}$, où $\overline{\alpha} = \max_{i=1,n} \alpha_i$, alors $\rho(P^{-1}N) < 1$, et donc que la méthode itérative converge.

4. Trouver le paramètre β minimisant $\max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$.
(On pourra d'abord montrer que pour tout $\beta > 0$, $|\beta - \alpha_i| \leq \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$ pour tout $i = 1, \dots, n$, avec $\underline{\alpha} = \min_{i=1,\dots,n} \alpha_i$ et $\bar{\alpha} = \max_{i=1,\dots,n} \alpha_i$ et en déduire que $\max_{i=1,n} |\beta - \alpha_i| = \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$) .

Exercice 63 (Méthode des directions alternées). *Corrigé en page 130.*

Soit $n \in \mathbb{N}$, $n \geq 1$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n symétrique inversible et $b \in \mathbb{R}^n$. On cherche à calculer $u \in \mathbb{R}^n$, solution du système linéaire suivant :

$$Au = b, \tag{1.126}$$

On suppose connues des matrices X et $Y \in \mathcal{M}_n(\mathbb{R})$, symétriques. Soit $\alpha \in \mathbb{R}_+^*$, choisi tel que $X + \alpha\text{Id}$ et $Y + \alpha\text{Id}$ soient définies positives (où Id désigne la matrice identité d'ordre n) et $X + Y + \alpha\text{Id} = A$.
Soit $u^{(0)} \in \mathbb{R}^n$, on propose la méthode itérative suivante pour résoudre (1.126) :

$$(X + \alpha Id)u^{(k+1/2)} = -Yu^{(k)} + b, \tag{1.127a}$$

$$(Y + \alpha Id)u^{(k+1)} = -Xu^{(k+1/2)} + b. \tag{1.127b}$$

1. Montrer que la méthode itérative (1.127) définit bien une suite $(u^{(k)})_{k\in\mathbb{N}}$ et que cette suite converge vers la solution u de (1.1) si et seulement si

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) < 1.$$

(On rappelle que pour toute matrice carrée d'ordre n, $\rho(M)$ désigne le rayon spectral de la matrice M.)

2. Montrer que si les matrices $(X + \frac{\alpha}{2}Id)$ et $(Y + \frac{\alpha}{2}Id)$ sont définies positives alors la méthode (1.127) converge. On pourra pour cela (mais ce n'est pas obligatoire) suivre la démarche suivante :

   (a) Montrer que

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) = \rho\big(X(X + \alpha Id)^{-1}Y(Y + \alpha Id)^{-1}\big).$$

(On pourra utiliser l'exercice 31 page 71).

(b) Montrer que

$$\rho\big(X(X+\alpha Id)^{-1}Y(Y+\alpha Id)^{-1}\big) \leq \rho\big(X(X+\alpha Id)^{-1}\big)\rho\big(Y(Y+\alpha Id)^{-1}\big).$$

(c) Montrer que $\rho\big(X(X+\alpha Id)^{-1}\big) < 1$ si et seulement si la matrice $(X + \frac{\alpha}{2} Id)$ est définie positive.

(d) Conclure.

3. Soit $f \in C([0,1]\times[0,1])$ et soit A la matrice carrée d'ordre $n = M \times M$ obtenue par discrétisation de l'équation $-\Delta u = f$ sur le carré $[0,1]\times[0,1]$ avec conditions aux limites de Dirichlet homogènes $u = 0$ sur $\partial\Omega$, par différences finies avec un pas uniforme $h = \frac{1}{M}$, et $\boldsymbol{b}$ le second membre associé.

(a) Donner l'expression de A et $\boldsymbol{b}$.

(b) Proposer des choix de X, Y et α pour lesquelles la méthode itérative (1.127) converge dans ce cas et qui justifient l'appellation “méthode des directions alternées" qui lui est donnée.

1.6 Valeurs propres et vecteurs propres

Les techniques de recherche des éléments propres, c.à.d. des valeurs et vecteurs propres (voir Définition 1.2 page 7) d'une matrice sont essentielles dans de nombreux domaines d'application, par exemple en dynamique des structures : la recherche des modes propres d'une structure peut s'avérer importante pour le dimensionnement de structures sous contraintes dynamiques ; elle est essentielle dans la compréhension des phénomènes acoustiques.
On peut se demander pourquoi on parle dans ce chapitre, intitulé "systèmes linéaires" du problème de recherche des valeurs propres : il s'agit en effet d'un problème non linéaire, les valeurs propres étant les solutions du polynôme caractéristique, qui est un polynôme de degré n, où n est la dimension de la matrice. Il n'est malheureusement pas possible de calculer numériquement les valeurs propres comme les racines du polynôme caractéristique, car cet algorithme est instable : une petite perturbation sur les coefficients du polynôme peut entraîner une erreur très grande sur les racines (voir par exemple le chapitre 5 du polycopié d'E. Hairer, en ligne sur le web (voir l'introduction de ce cours). De nombreux algorithmes ont été développés pour le calcul des valeurs propres et vecteurs propres. Ces méthodes sont en fait assez semblables aux méthodes de résolution de systèmes linéaires. Dans le cadre de ce cours, nous nous restreignons à deux méthodes très connues : la méthode de la puissance (et son adaptation de la puissance inverse), et la méthode dite QR.

1.6.1 Méthode de la puissance et de la puissance inverse

Pour expliquer l'algorithme de la puissance, commençons par un exemple simple. Prenons par exemple la matrice

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

dont les valeurs propres sont 1 et 3, et les vecteurs propres associés $\boldsymbol{f}^{(1)} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ et $\boldsymbol{f}^{(2)} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Partons de $\boldsymbol{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ et faisons tourner scilab en itérant les instructions suivantes :

```
-->x =  A * x ; x = x/norm(x)
```

ce qui correspond à la construction de la suite

$$\boldsymbol{x}^{(0)} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|},\ \boldsymbol{x}^{(1)} = \frac{A\boldsymbol{x}^{(0)}}{\|A\boldsymbol{x}^{(0)}\|}, \cdots, \boldsymbol{x}^{(k+1)} = \frac{A\boldsymbol{x}^{(k)}}{\|A\boldsymbol{x}^{(k)}\|} \tag{1.141}$$

où $\|\boldsymbol{x}\|$ désigne la norme euclidienne.
On obtient les résultats suivants :

0.8944272	0.7808688	0.7327935	0.7157819	0.7100107	0.7080761	0.7074300
- 0.4472136	- 0.6246950	- 0.6804511	- 0.6983239	- 0.7061361	- 0.7067834	- 0.706999

On voit clairement sur cet exemple que la suite $\boldsymbol{x}^{(k)}$ converge vers $\boldsymbol{f}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ lorsque $k \to +\infty$. Si maintenant on fait tourner Scilab en lui demandant de calculer ensuite le produit scalaire de $A\boldsymbol{x}$ avec $\boldsymbol{x}$:

```
-->x= A*x;x=x/norm(x);mu=(A*x)'*x
```

ce qui correspond au calcul de la suite $\mu_k = A\boldsymbol{x}^{(k)} \cdot \boldsymbol{x}^{(k)}, k \geq 0$, on obtient la suite :

$$2.8, 2.9756098, 2.9972603, 2.9996952, 2.9999661, ...$$

qui a tout l'air de converger vers 3 ! En fait on a le théoreème suivant, qui montre que dans un certain nombre de cas, on a effectivement convergence de l'algorithme vers la valeur propre dite dominante (celle qui correspond au rayon spectral).

Théorème 1.61 (Convergence de la méthode de la puissance)**.** *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. On note $\lambda_1, \cdots, \lambda_n$ les valeurs propres de A, $(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_n)$ une base orthonormée de trigonalisation de A telle que $A\boldsymbol{f}_n = \lambda_n \boldsymbol{f}_n$. On suppose que la valeur propre λ_n est dominante, c.à.d. que*

$$|\lambda_n| > |\lambda_{n-1}| \geq \cdots \geq |\lambda_1|,$$

et on suppose de plus que $\lambda_n \in \mathbb{R}$. Alors si $\boldsymbol{x} \notin Vect(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{n-1})$, la suite de vecteurs $\boldsymbol{x}_{2k}$ converge vers un vecteur unitaire qui est vecteur propre de A pour la valeur propre dominante λ_n.
De plus, si la norme choisie dans l'algorithme (1.141) *est la norme 2, alors la suite $(A\boldsymbol{x}_{2k}, \boldsymbol{x}_{2k})_{n\in\mathbb{N}}$ converge vers λ_n lorsque $k \to +\infty$.*

Démonstration. La démonstration de ce résultat fait l'objet de l'exercice 64 dans le cas plus simple où A est une matrice symétrique, et donc diagonalisable dans $\mathbb{R}$. □

La méthode de la puissance souffre de plusieurs inconvénients :

1. Elle ne permet de calculer que la plus grande valeur propre. Or très souvent, on veut pouvoir calculer la plus petite valeur propre.
2. De plus, elle ne peut converger que si cette valeur propre est simple.
3. Enfin, même dans le cas où elle est simple, si le rapport des des deux plus grandes valeurs propres est proche de 1, la méthode va converger trop lentement.

De manière assez miraculeuse, il existe un remède à chacun de ces maux :

1. Pour calculer plusieurs valeurs propres simultanément, on procède par blocs : on part de p vecteurs orthogonaux $\boldsymbol{x}_1^{(0)}, \ldots \boldsymbol{x}_p^{(0)}$ (au lieu d'un seul). Une itération de la méthode consiste alors à multiplier les p vecteurs par A et à les orthogonaliser par Gram-Schmidt. En répétant cette itération, on approche, si tout se passe bien, p valeurs propres et vecteurs propres de A, et la vitesse de convergence de la méthode est maintenant $\frac{\lambda_{n-p}}{\lambda_n}$.
2. Si l'on veut calculer la plus petite valeur propre, on applique la méthode de la puissance à A^{-1}. On a alors convergence (toujours si tout se passe bien) de $\frac{\|\boldsymbol{x}_{k+1}\|}{\|\boldsymbol{x}_k\|}$ vers $1/|\lambda_1|$. Bien sûr, la mise en oeuvre effective ne s'effectue pas avec l'inverse de A, mais en effectuant une décomposition LU de A qui permet ensuite la résolution du système linéaire $A\boldsymbol{x}_{k+1} = A\boldsymbol{x}^{(k)}$.
3. Enfin, pour accélérer la convergence de la méthode, on utilise une translation sur A, qui permet de se rapprocher de la valeur propre que l'on veut effectivement calculer. Voir à ce propos l'exercice 65.

1.6.2 Méthode QR

Toute matrice A peut se décomposer sous la forme $A = QR$, où Q est une matrice orthogonale et R une matrice triangulaire supérieure. Dans le cas où A est inversible, cette décomposition est unique. On a donc le théorème suivant :

Théorème 1.62 (Décomposition QR d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors il existe Q matrice orthogonale et R matrice triangulaire supérieure à coefficients diagonaux positifs ou nuls tels que $A = QR$. Si la matrice A est inversible, alors cette décomposition est unique.*

La démonstration est effectuée dans le cas inversible la question 1 de l'exercice 68. La décomposition QR d'une matrice A inversible s'obtient de manière très simple par la méthode de Gram-Schmidt, qui permet de construire

une base orthonormée $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ (les colonnes de la matrice Q), à partir de n vecteurs vecteurs indépendants $a_1, \ldots, a_n$ (les colonnes de la matrice A). On se reportera à l'exercice 66 pour un éventuel raffraichissement de mémoire sur Gram-Schmidt. Dans le cas où A n'est pas inversible (et même non carrée), la décomposition existe mais n'est pas unique. La démonstration dans le cadre général se trouve dans le livre de Ph. Ciarlet conseillé en début de ce cours.
L'algorithme QR pour la recherche des valeurs propres d'une matrice est extêmement simple : Si A est une matrice inversible, on pose $A_0 = A$, on effectue la décomposition QR de A : $A = A_0 = Q_0 R_0$ et on calcule $A_1 = R_0 Q_0$. Comme le produit de matrices n'est pas commutatif, les matrices A_0 et A_1 ne sont pas égales, mais en revanche elles sont semblables ; en effet, graâce à l'associativité du produit matriciel, on a :

$$A_1 = R_0 Q_0 = (Q_0^{-1} Q_0) R_0 Q_0 = Q_0^{-1} (Q_0 R_0) Q_0 = Q_0^{-1} A Q_0.$$

Les matrices A_0 et A_1 ont donc même valeurs propres.
On recommence alors l'opération : à l'itération n, on.effectue la décomposition QR de A_n : $A_n = Q_n R_n$ et on calcule $A_{n+1} = R_n Q_n$.
Par miracle, pour la plupart des matrices, les coefficients diagonaux de la matrice R_n tendent vers les valeurs propres de la matrice A, et les colonnes de la matrice Q_n vers les vecteurs propres associés. Notons que la convergence de l'algorithme QR est un problème ouvert. On siat démontrer que l'algorithme converge pour une large classe de matrices ; on pourra trouver par exemple dans les livres de Serre ou Hubbard-Hubert la démonstration sous une hypothèse assez technique et difficile à vérifier en pratique ; l'exercice 68 donne la démonstration (avec la même hypothèse technique) pour le cas plus simple d'une matrice symétrique définie positive.
Pour améliorer la convergence de l'algorithme QR, on utilise souvent la technique dite de "shift" (translation en français). A l'itération n, au lieu d'effectuer la décomposition QR de la matrice A_n, on travaille sur la matrice $A_n - bI$, où b est choisi proche de la plus grande valeur propre. En général on choisit le coefficient $b = a_{nn}^{(k)}$.
L'exercice 67 donne un exemple de l'application de la méthode QR avec shift.

1.6.3 Exercices

Exercice 64 (Méthode de la puissance)**.** *Suggestions en page 140, corrigé en page 140*

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Soit $\lambda_n \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_n| = \rho(A)$ et soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. On suppose que $-\lambda_n$ n'est pas une valeur propre de A et que $\boldsymbol{y}^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_n Id)$, ce qui revient à dire que lorsqu'on écrit le vecteur propre $\boldsymbol{y}^{(0)}$ dans la base des vecteurs propres, la composante sur le vecteur propre associé à λ_n est non nulle. On définit la suite $(\boldsymbol{y}^{(k)})_{n\in\mathbb{N}}$ par $\boldsymbol{y}^{(k+1)} = A\boldsymbol{y}^{(k)}$ pour $n \in \mathbb{N}$. Montrer que

   (a) $\dfrac{\boldsymbol{y}^{(k)}}{(\lambda_n)^k} \to \boldsymbol{y}$, quand $k \to \infty$, avec $\boldsymbol{y} \neq 0$ et $A\boldsymbol{y} = \lambda_n \boldsymbol{y}$.

   (b) $\dfrac{\|\boldsymbol{y}^{(k+1)}\|}{\|\boldsymbol{y}^{(k)}\|} \to \rho(A)$ quand $k \to \infty$.

   (c) $\frac{1}{\|\boldsymbol{y}^{2k}\|} \boldsymbol{y}^{2k} \to \boldsymbol{x}$ quand $k \to \infty$ avec $A\boldsymbol{x} = \lambda_n \boldsymbol{x}$ et $\|\boldsymbol{x}\| = 1$.

   Cette méthode de calcul de la plus grande valeur propre s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\boldsymbol{b} \in \mathbb{R}^n$. Pour calculer $\boldsymbol{y}$ t.q. $A\boldsymbol{y} = \boldsymbol{b}$, on considère un méthode itérative : on se donne un choix initial $\boldsymbol{y}^{(0)}$, et on construit la suite $\boldsymbol{y}^{(k)}$ telle que $\boldsymbol{y}^{(k+1)} = B\boldsymbol{y}^{(k)} + \boldsymbol{c}$ avec $\boldsymbol{c} = (Id - B)A^{-1}\boldsymbol{b}$, et on suppose B symétrique. On rappelle que si $\rho(B) < 1$, la suite $(\boldsymbol{y}^{(k)})_{n\in\mathbb{N}}$ tend vers $\boldsymbol{x}$. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|\boldsymbol{y}^{(k+1)} - \boldsymbol{x}\|}{\|\boldsymbol{y}^{(k)} - \boldsymbol{y}\|} \to \rho(B)$ quand $k \to \infty$ (ceci donne une estimation de la vitesse de convergence de la méthode itérative).

   (b) $\frac{\|\boldsymbol{y}^{(k+1)} - \boldsymbol{y}^{(k)}\|}{\|\boldsymbol{y}^{(k} - \boldsymbol{y}^{(k-1)}\|} \to \rho(B)$ quand $k \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 65 (Méthode de la puissance inverse avec shift)**.** *Suggestions en page 140. Corrigé en page 141.*

Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique et $\lambda_1, \ldots, \lambda_p$ ($p \leq n$) les valeurs propres de A. Soit $i \in \{1, \ldots, p\}$, on cherche à calculer λ_i. Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. On suppose que $\boldsymbol{x}^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_i Id)$. On suppose également connaître $\mu \in \mathbb{R}$ t.q. $0 < |\mu - \lambda_i| < |\mu - \lambda_j|$ pour tout $j \neq i$. On définit la suite $(\boldsymbol{x}^{(k)})_{n\in\mathbb{N}}$ par $(A - \mu Id)\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ pour $n \in \mathbb{N}$.

1. Vérifier que la construction de la suite revient à appliquer la méthode de la puissance à la matrice $A - \mu Id)^{-1}$.
2. Montrer que $\boldsymbol{x}^{(k)}(\lambda_i - \mu)^k \to \boldsymbol{x}$, quand $k \to \infty$, où $\boldsymbol{x}$ est un vecteur propre associé à la valeur propre λ_i, c.à.d. $\boldsymbol{x} \neq 0$ et $A\boldsymbol{x} = \lambda_i \boldsymbol{x}$.
3. Montrer que $\frac{\|\boldsymbol{x}^{(k+1)}\|}{\|\boldsymbol{x}^{(k)}\|} \to \frac{1}{|\mu - \lambda_i|}$ quand $k \to \infty$.

Exercice 66 (Orthogonalisation de Gram-Schmidt)**.** *Corrigé en page 141*

Soient $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de $\mathbb{R}^n$. On rappelle que la projection orthogonale $\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v})$ du vecteur $\boldsymbol{v}$ sur la droite vectorielle engendrée par $\boldsymbol{u}$ peut s'écrire de la manière suivante :

$$\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v}) = \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}} \boldsymbol{u},$$

où $\boldsymbol{u} \cdot \boldsymbol{v}$ désigne le produit scalaire des vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$. On note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$.

1. Soient $(\boldsymbol{a_1}, \ldots, \boldsymbol{a_n})$ une base de $\mathbb{R}^n$. On rappelle qu'à partir de cette base, on peut obtenir une base orthogonale $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ et une base orthonormale $(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)$ par le procédé de Gram-Schmidt qui s'écrit :

$$\begin{aligned}
&\boldsymbol{v}_1 = \boldsymbol{a_1}, && \boldsymbol{q}_1 = \frac{\boldsymbol{a_1}}{\|\boldsymbol{a_1}\|}\\
&\boldsymbol{v}_2 = \boldsymbol{a_2} - \mathrm{proj}_{\boldsymbol{v_1}}(\boldsymbol{a_2}), && \boldsymbol{q}_2 = \frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|}\\
&\boldsymbol{v}_3 = \boldsymbol{a_3} - \mathrm{proj}_{\boldsymbol{v_1}}(\boldsymbol{a_3}) - \mathrm{proj}_{\boldsymbol{v_2}}(\boldsymbol{a_3}), && \boldsymbol{q}_3 = \frac{\boldsymbol{v}_3}{\|\boldsymbol{v}_3\|}\\
&\boldsymbol{v}_4 = \boldsymbol{a_4} - \mathrm{proj}_{\boldsymbol{v_1}}(\boldsymbol{a_4}) - \mathrm{proj}_{\boldsymbol{v_2}}(\boldsymbol{a_4}) - \mathrm{proj}_{\boldsymbol{v_3}}(\boldsymbol{a_4}), && \boldsymbol{q}_4 = \frac{\boldsymbol{v}_4}{\|\boldsymbol{v}_4\|}\\
&\quad\vdots && \quad\vdots\\
&\boldsymbol{v}_k = \boldsymbol{a_k} - \sum_{\mathbf{j=1}}^{\mathbf{k-1}} \mathrm{proj}_{\boldsymbol{v_j}}(\boldsymbol{a_k}), && \boldsymbol{q}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}
\end{aligned}$$

On a donc

$$\boldsymbol{v}_k = \boldsymbol{a}_k - \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j, \qquad \boldsymbol{q}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}. \tag{1.142}$$

1. Montrer par récurrence que la famille $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ est une base orthogonale de $\mathbb{R}^n$.

2. Soient A la matrice carrée d'ordre n dont les colonnes sont les vecteurs $\boldsymbol{a}_j$ et Q la matrice carrée d'ordre N dont les colonnes sont les vecteurs $\boldsymbol{q}_j$ définis par le procédé de Gram-Schmidt (1.142), ce qu'on note :

$$A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_n \end{bmatrix}, \qquad Q = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \ldots & \boldsymbol{q}_n \end{bmatrix}.$$

Montrer que

$$\boldsymbol{a}_k = \|\boldsymbol{v}_k\| \boldsymbol{q}_k + \sum_{j=1}^{k-1} \frac{\boldsymbol{a}_k \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_j\|} \boldsymbol{q}_j.$$

En déduire que $A = QR$, où R est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3. Montrer que pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ inversible, on peut construire une matrice orthogonale Q (c.à. d. telle que $QQ^t = \mathrm{Id}$) et une matrice triangulaire supérieure R à coefficients diagonaux positifs telles que $A = QR$.

4. Donner la décomposition QR de $A = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}$.

5. On considère maintenant l'algorithme suivant (où l'on stocke la matrice Q orthogonale cherchée dans la matrice A de départ (qui est donc écrasée)

Algorithme 1.63 (Gram-Schmidt modifié)**.**

Pour $k = 1, \ldots, n$,

Calcul de la norme de $\boldsymbol{a}_k$

$r_{kk} := (\sum_{i=1}^n a_{ik}^2)^{\frac{1}{2}}$

Normalisation

Pour $\ell = 1, \ldots, n$

$a_{\ell k} := a_{\ell k}/r_{kk}$

Fin pour ℓ

Pour $j = k+1, \ldots, n$

Produit scalaire correspondant à $q_k \cdot a_j$

$r_{kj} := \sum_{i=1}^n a_{ik} a_{ij}$

On soustrait la projection de a_k *sur* q_j *sur tous les vecteurs de* A *après* k.

Pour $i = k+1, \ldots, n$,

$a_{ij} := a_{ij} - a_{ik} r_{kj}$

Fin pour i

Fin pour j

Montrer que la matrice A résultant de cet algorithme est identique à la matrice Q donnée par la méthode de Gram-Schmidt, et que la matrice R est celle de Gram-Schmidt. (Cet algorithme est celui qui est effectivement implanté, car il est plus stable que le calcul par le procédé de Gram-Schmidt original.)

Exercice 67 (Méthode QR avec shift)**.** *Soit* $A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & 0 \end{bmatrix}$

1. *Calculer les valeurs propres de la matrice* A.
2. *Effectuer la décomposition* QR *de la matrice* A.
3. *Calculer* $A_1 = RQ$ *et* $\tilde{A}_1 = RQ - bId$ *où* b *est le terme* a_{22}^1 *de la matrice* A_1
4. *Effectuer la décomposition* QR *de* A_1 *et* $\tilde{A}_1$*, et calculer les matrices* $A_2 = R_1 Q_1$ *et* $\tilde{A}_2 = \tilde{R}_1 \tilde{Q}_1$.

Exercice 68 (Méthode QR pour la recherche de valeurs propres)**.** *Corrigé en page 142*

Soit A une matrice inversible. Pour trouver les valeurs propres de A, on propose la méthode suivante, dite "méthode QR" : On pose $A_1 = A$ et on construit une matrice orthogonale Q_1 et une matrice triangulaire supérieure R_1 telles que $A_1 = Q_1 R_1$ (par exemple par l'algorithme de Gram-Schmidt). On pose alors $A_2 = R_1 Q_1$, qui est aussi une matrice inversible. On construit ensuite une matrice orthogonale Q_2 et une matrice triangulaire supérieure R_2 telles que $A_2 = Q_2 R_2$ et on pose $A_3 = R_3 Q_3$. On continue et on construit une suite de matrices A_k telles que :

$$A_1 = A = Q_1 R_1,\ R_1 Q_1 = A_2 = Q_2 R_2,\ \ldots,\ R_k Q_k = A_k = Q_{k+1} R_{k+1}. \tag{1.143}$$

Dans de nombreux cas, cette construction permet d'obenir les valeurs propres de la matrice A sur la diagonale des matrices A_k. Nous allons démontrer que ceci est vrai pour le cas particulier des matrices symétriques définies positives dont les valeurs propres sont simples (on peut le montrer pour une classe plus large de matrices).

On suppose à partir de maintenant que A est une matrice symétrique définie positive qui admet n valeurs propres (strictement positives) vérifiant $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. On a donc :

$$A = P\Lambda P^t, \text{ avec } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \text{ et } P \text{ est une matrice orthogonale.} \tag{1.144}$$

(La notation $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ désigne la matrice diagonale dont les termes diagonaux sont λ_1,... ,λ_n).

On suppose de plus que

$$P^t \text{ admet une décomposition } LU \text{ et que les coefficients diagonaux de } U \text{ sont strictement positifs.} \tag{1.145}$$

On va montrer que A_k tend vers $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

2. Soient Q_i et R_i les matrices orthogonales et triangulaires supérieures définies par (1.143).
2.1 Montrer que $A^2 = \tilde{Q}_2\tilde{R}_2$ avec $\tilde{Q}_k = Q_1Q_2$ et $\tilde{R}_k = R_2R_1$.

2.2 Montrer, par récurrence sur k, que

$$A^k = \tilde{Q}_k\tilde{R}_k, \tag{1.146}$$

avec

$$\tilde{Q}_k = Q_1Q_2 \ldots Q_{k-1}Q_k \text{ et } \tilde{R}_k = R_kR_{k-1} \ldots R_2R_1. \tag{1.147}$$

2.3 Justifier brièvement le fait que $\tilde{Q}_k$ est une matrice orthogonale et $\tilde{R}_k$ est une matrice triangulaire à coefficients diagonaux positifs.

3. Soit $M_k = \Lambda^k L \Lambda^{-k}$.
3.1 Montrer que $PM_k = \tilde{Q}_kT_k$ où $T_k = \tilde{R}_kU^{-1}\Lambda^{-k}$ est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3.2 Calculer les coefficients de M_k en fonction de ceux de L et des valeurs propres de A.

3.3 En déduire que M_k tend vers la matrice identité et que $\tilde{Q}_kT_k$ tend vers P lorsque $k \to +\infty$.

4. Soient $(B_k)_{k\in\mathbb{N}}$ et $(C_k)_{k\in\mathbb{N}}$ deux suites de matrices telles que les matrices B_k sont orthogonales et les matrices C_k triangulaires supérieures et de coefficients diagonaux positifs. On va montrer que si B_kC_k tend vers la matrice orthogonale B lorsque k tend vers l'infini alors B_k tend vers B et C_k tend vers l'identité lorsque k tend vers l'infini.
On suppose donc que B_kC_k tend vers la matrice orthogonale B. On note $b_1, b_2, \ldots, b_n$ les colonnes de la matrice B et $b_1^{(k)}, b_2^{(k)}, \ldots, b_n^{(k)}$ les colonnes de la matrice B_k, ou encore :

$$B = \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix}, \qquad B_k = \begin{bmatrix} b_1^{(k)} & b_2^{(k)} & \ldots & b_n^{(k)} \end{bmatrix}.$$

et on note $c_{i,j}^{(k)}$ les coefficients de C_k.

4.1 Montrer que la première colonne de B_kC_k est égale à $c_{1,1}^{(k)}b_1^{(k)}$. En déduire que $c_{1,1}^{(k)} \to 1$ et que $b_1^{(k)} \to b_1$.

4.2 Montrer que la seconde colonne de B_kC_k est égale à $c_{1,2}^{(k)}b_1^{(k)} + c_{2,2}^{(k)}b_2^{(k)}$. En déduire que $c_{1,2}^{(k)} \to 0$, puis que $c_{2,2}^{(k)} \to 1$ et que $b_2^{(k)} \to b_2$.

4.3 Montrer que lorsque $k \to +\infty$, on a $c_{i,j}^{(k)} \to 0$ si $i \neq j$, puis que $c_{i,i}^{(k)} \to 1$ et $b_i^{(k)} \to b_i$.

4.4 En déduire que B_k tend B et C_k tend vers l'identité lorsque k tend vers l'infini.

5. Déduire des questions 3 et 4 que $\tilde{Q}_k$ tend vers P et T_k tend vers Id lorsque $k \to +\infty$.

6. Montrer que $\tilde{R}_k(\tilde{R}_{k-1})^{-1} = T_k\Lambda T_{k-1}$. En déduire que R_k et A_k tendent vers Λ.

Université d'Aix-Marseille

3ème année, lic. Math. Géné.
Analyse numérique, semestre 5

Travaux pratique sous scilab n°2

Exercice 1 : Saisie et stockage d'une matrice

Plusieurs manières sont possibles pour saisir une matrice et stocker ses données.

On souhaite saisir la matrice de discrétisation du Laplacien qui est la matrice de taille $n \times n$ suivante :

$$\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & & \vdots \\ 0 & 0 & -1 & 2 & \ddots & 0 \\ \vdots & & & \ddots & \ddots & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix}$$

1. Ecrire un programme sous Scilab qui stocke la matrice du Laplacien dans la variable `A` à l'aide de boucles `for`.
   Après avoir fait un test avec n=5, mesurez le temps d'exécution du programme pour n=100 puis n=1000.
2. Ecrire un code qui réalise la même fonction à l'aide de la fonction `diag` (voir dans l'aide de Scilab).
3. Comparez les temps d'exécution pour les deux méthodes.

Indication : Pour évaluer le temps de calcul, on peut utiliser les fonctions `tic()` et `toc()`, sous la forme suivante :

```
... Lignes de code
tic(); //allume le chronomètre
code
temps_exec = toc()      //stoppe le chronomètre et stocke la valeur
//en secondes dans la variable temps_exec
... Suite du code
```

Exercice 2 : Décomposition LU

1. Ecrire une fonction scilab qui, à une matrice A de taille n, renvoie ses n mineurs principaux.

2. Pour les deux matrices qui suivent calculer les mineurs principaux, qu'en déduisez-vous ? Déterminer la décomposition LU de A avec la fonction pré-programmée dans Scilab (`lu`) et la décomposition de Choleski avec la fonction pré-programmée (`chol`) si cela est possible.

$$A_1 = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & & \vdots \\ 0 & 0 & -1 & 2 & \ddots & 0 \\ \vdots & & & \ddots & \ddots & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 1 & \ddots & \vdots & \vdots \\ -1 & & \ddots & \ddots & 0 & 1 \\ \vdots & & & -1 & 1 & 1 \\ -1 & \cdots & & -1 & -1 & 1 \end{pmatrix}$$

3. *Bonus :* Que remarquez-vous sur les profils des décompositions LU et de Choleski ?

Exercice 3

On considère la fonction suivante :

```
function [X,Y]=f(A)
    ncol=size(A,2);
    nlig=size(A,1);
    Y=A;
    X=eye(nlig,ncol);
    for j=1:ncol
      for i=j+1:nlig
            if abs(Y(j,j))<1e-13 then
                abort
            end
             X(i,j)=Y(i,j)/Y(j,j)
            Y(i,j:ncol)=Y(i,j:ncol)-Y(j,j:ncol)*Y(i,j)/Y(j,j);

        end
    end
    endfunction
```

Tester cette fonction sur la matrice A_1 de l'exercice précédent. Décrire l'action de cette fonction et comparer la vitesse d'exécution de cette fonction à celle de la fontion scilab correspondante.

Exercice 4 : Un exemple rigolo

L'objectif de cet exercice est de déterminer le nombre de faces d'un ballon de foot. Un ballon de foot est formé de faces de forme pentagonales ou hexagonales. On notera x le nombre de pentagones et y le nombre d'hexagones qui le constituent et f le nombre total de faces, a le nombre d'arêtes et s le nombre de sommets du ballon. On rappelle que ces nombres sont des entiers positifs.

Pour déterminer x et y, on écrit que

(a) chaque sommet appartient à exactement trois faces :

$$3s = 5x + 6y$$

(b) chaque arête est partagé par deux faces :

$$2a = 5x + 6y$$

(c) le nombre d'arêtes d'un polygone est égal à la somme du nombre de faces et du nombre de sommets moins 2 :

$$a = f + s - 2$$

(d) le nombre de faces est égal à la somme des nombres de pentagones et hexagones.

1. Montrer que le quintuplet $X = (x, y, f, a, s)$ vérifiant les propriétés (a)-(d) est solution d'un système linéaire $AX = b$, où A est une matrice et b un vecteur que l'on explicitera.
2. Résoudre avec scilab ce système linéaire et déterminer ses solutions entières. On pourra utiliser la commande (`rref`)de scilab.