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 $\boldsymbol{x}_0 \in S_1$ 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 33 page 75. 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 $Q^* = Q^{-1}$ 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 telle que $|\lambda| = \rho(A)$.
On suppose tout d'abord que $\lambda \in \mathbb{R}$. Il existe alors un vecteur non nul de $\mathbb{R}^n$, noté $\boldsymbol{x}$, tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. 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 $|\lambda| \leq \|A\|$ et donc $\rho(A) \leq \|A\|$.

Si $\lambda \in \mathbb{C} \setminus \mathbb{R}$, la démonstration est un peu plus compliquée car la norme considérée est une norme dans $\mathbb{R}^n$ (et non dans $\mathbb{C}^n$). On montre tout d'abord que $\rho(A) < 1$ si $\|A\| < 1$.

En effet, Il existe $x \in \mathbb{C}^n$, $x \neq 0$, tel que $Ax = \lambda x$. En posant $x = y + iz$, avec $y, z \in \mathbb{R}^n$, on a donc pour tout $k \in \mathbb{N}$, $\lambda^k x = A^k x = A^k y + iA^k z$. Comme $\|A^k y\| \leq \|A\|^k\|y\|$ et $\|A^k z\| \leq \|A\|^k\|z\|$, on a, si $\|A\| < 1$, $A^k y \to 0$ et $A^k z \to 0$ (dans $\mathbb{R}^n$) quand $k \to +\infty$. On en déduit que $\lambda^k x \to 0$ dans $\mathbb{C}^n$. En choisissant une norme sur $\mathbb{C}^n$, notée $\|\cdot\|_a$, on a donc $|\lambda|^k\|x\|_a \to 0$ quand $k \to +\infty$, ce qui montre que $|\lambda| < 1$ et donc $\rho(A) < 1$.

Pour traiter le cas général (A quelconque dans $\mathcal{M}_n(\mathbb{R})$), il suffit de remarquer que la démonstration précédente donne, pour tout $\eta > 0$, $\rho(A/(\|A\| + \eta)) < 1$ (car $\|A/(\|A\| + \eta)\| < 1$). On a donc $\rho(A) < \|A\| + \eta$ pour tout $\eta > 0$, ce qui donne bien $\rho(A) \leq \|A\|$.

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

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

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

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

On en déduit que

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

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

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

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

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

on a donc, pour tout x dans $\mathbb{C}^n$, $x \neq 0$,

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

On en déduit que $\|A\|^2 \leq \rho(A)^2 + \eta C_T n^3$ et donc

$$\|A\| \leq \rho(A)\big(1 + \frac{\eta C_T n^3}{\rho(A)^2}\big)^{\frac{1}{2}} \leq \rho(A)(1 + \frac{\eta C_T n^3}{\rho(A)^2}).$$

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

■

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 $\|A\|_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon} \leq \mu^k \to 0$ 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 $n \to +\infty$. 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 $n \to +\infty$. Soit $\eta \in]l, 1[$ il existe donc n_0 tel que pour $n \geq n_0$, $\|A^{k_n}\|^{1/k_n} \leq \eta$. On en déduit que pour $n \geq n_0$, $\|A^{k_n}\| \leq \eta^{k_n}$, 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 38 page 76.
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 $\lambda_1, \dots, \lambda_n$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \dots, \lambda_n^{-1}$ et A^t a comme valeurs propres $\lambda_1, \dots, \lambda_n$.
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 42 page 77.
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 64). 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 :

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

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

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

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

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

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 69 et l'exercice 38 page 76) $(\mathrm{Id} + B)$ est inversible et $(\mathrm{Id} + B)^{-1} = \displaystyle\sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(\mathrm{Id} + B)^{-1}\| \leq \displaystyle\sum_{n=0}^{\infty}\|B\|^n = \frac{1}{1 - \|B\|} \leq \frac{1}{1 - \|A^{-1}\|\ \|\delta_A\|}$. 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 49 page 79) 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 12 page 20), et que

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

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 49 page 79) 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 51 page 79. 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.

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 $\mathbb{R}$