Résolvante, stabilité et applications

Jean-François Coulombel

Résolvante, stabilité et applications

Jean-François Coulombel

CNRS, Université de Nantes, Laboratoire de Mathématiques Jean Leray (CNRS UMR6629)

2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France

Email : jean-francois.coulombel@univ-nantes.fr

31 octobre 2013

Résumé

On s'intéresse dans cet exposé au lien entre les estimations de résolvante d'une matrice et les estimations de stabilité sur les normes des puissances de cette matrice. L'accent est mis spécifiquement sur les matrices de rayon spectral inférieur à 1. On expose en détails un résultat de Boyd et Doyle [BD87] reliant l'estimation de la résolvante sur le cercle unité à la somme des normes des puissances de telles matrices. Deux applications de ce résultat sont traitées. L'une concerne la résolution itérative de grands systèmes linéaires ; l'autre concerne la stabilité de problèmes aux limites hyperboliques discrétisés.

Classification AMS : 15A45, 65F10, 65M12.
Mots-clés : matrices, résolvante, estimations de stabilité, grands systèmes linéaires, différences finies, problèmes aux limites hyperboliques.

On adopte dans tout cet exposé les notations suivantes : pour des vecteurs $X, Y \in \mathbb{C}^N$, on note $\langle X, Y \rangle := \sum_{i=1}^N \overline{X_i}\, Y_i$ leur produit scalaire hermitien. La norme correspondante est notée $|\cdot|$, sans préciser la dimension. La norme induite sur l'espace des matrices $\mathscr{M}_N(\mathbb{C})$ est notée $\|\cdot\|$. Pour $M \in \mathscr{M}_N(\mathbb{C})$, on note $\rho(M)$ le rayon spectral de M, et M^* la matrice transposée conjuguée de M de sorte que $\|M\| = \|M^*\| = \rho(M^*\,M)^{1/2}$. La matrice identité est notée I. Le groupe des matrices inversibles de taille N est noté $\mathrm{GL}_N(\mathbb{C})$. On définit enfin les ensembles suivants dans le plan complexe :

$$\mathscr{U} := \{\zeta \in \mathbb{C}, |\zeta| > 1\}\,, \quad \overline{\mathscr{U}} := \{\zeta \in \mathbb{C}, |\zeta| \geq 1\}\,,$$
$$\mathbb{D} := \{\zeta \in \mathbb{C}, |\zeta| < 1\}\,, \quad \mathbb{S}^1 := \{\zeta \in \mathbb{C}, |\zeta| = 1\}\,.$$

1 Introduction

De nombreux problèmes issus de la physique s'écrivent sous la forme d'équations d'évolution du premier ordre de la forme

$$\dot{u} = A\,u\,,$$

où le point désigne la dérivée par rapport à la variable de temps, et A est un opérateur linéaire, c'est-à-dire qu'on se donne un espace de Banach E et un sous-espace $D(A)$ dense dans E, A étant alors une application linéaire de $D(A)$ dans E. Si l'on voit, au moins formellement, l'équation d'évolution comme une équation différentielle à valeurs dans $D(A)$, ce qui serait le cas si A était une application linéaire continue de E dans E, on espère écrire la solution sous la forme $u(t) = \exp(t\,A)\,u_0$, où $u_0 \in D(A)$ est la condition initiale. La définition rigoureuse de tels groupes, ou semigroupes d'opérateurs fait l'objet d'une littérature abondante. On renvoie par exemple le lecteur à l'ouvrage [Paz83] pour une introduction à cette théorie.

Une fois qu'on a défini le semi-groupe, on peut s'intéresser aux propriétés des solutions comme par exemple à leur comportement asymptotique quand t tend vers $+\infty$. Le cas de la dimension finie est un résultat classique d'équations différentielles ; A est alors une matrice carrée, et le comportement de $\exp(t\,A)\,u_0$ lorsque t tend vers $+\infty$ dépend des propriétés spectrales de A. Si toutes les valeurs propres de A sont de partie réelle strictement négative, alors $\exp(t\,A)\,u_0$ converge exponentiellement vite vers 0. Si les valeurs propres de A sont de partie réelle négative et si de plus les valeurs propres imaginaires pures sont semi-simples (au sens où leurs multiplicités géométrique et algébrique coïncident), alors $\exp(t\,A)\,u_0$ est borné sur $\mathbb{R}^+$.

C'est à une version *quantitative* de ces derniers résultats que l'on s'intéresse ici. Plus précisément, pour montrer la convergence exponentielle vers 0 du semigroupe $\exp(t\,A)$, on peut commencer par réduire la matrice A sous forme de Jordan et traiter chaque bloc séparément. Cette approche s'avère efficace dans le cas d'une seule matrice A, mais elle se prête mal au cas où l'on cherche des bornes précises sur cette convergence exponentielle. Par exemple, si l'on part d'une équation aux dérivées partielles linéaire en d variables d'espace :

$$\dot{u} = \mathbf{A}\,u\,, \quad \mathbf{A} := \sum_{\alpha_1,\dots,\alpha_d=0}^{m} A_{\alpha_1,\dots,\alpha_d}\,\frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}}\cdots\frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}\,,$$

les A_α étant des matrices carrées, on peut appliquer la transformée de Fourier par rapport aux variables $x_1,\dots,x_d$ et se ramener à l'étude des équations différentielles :

$$\dot{v} = \mathbf{A}(\xi)\,v\,, \quad \mathbf{A}(\xi) := \sum_{\alpha_1,\dots,\alpha_d=0}^{m} (i\,\xi_1)^{\alpha_1}\dots(i\,\xi_d)^{\alpha_d}\,A_{\alpha_1,\dots,\alpha_d}\,.$$

Cette famille d'équations différentielles est paramétrée par les fréquences $\xi \in \mathbb{R}^d$. Malheureusement, la réduction de Jordan de $\mathbf{A}(\xi)$ se comporte souvent mal lorsque ξ varie, donc une information sur les valeurs propres de $\mathbf{A}(\xi)$ seulement s'avèrera souvent insuffisante pour obtenir des bornes sur le semigroupe $\exp(t\,\mathbf{A})$.

Le "bon outil" pour quantifier les informations spectrales et les convertir en bornes sur le semigroupe est la résolvante de la matrice (ou de l'opérateur) A. Dans la Section 2, on démontre deux résultats, propres à la dimension finie, qui relient une borne sur la résolvante d'une matrice à une borne sur les puissances de cette matrice. Par rapport aux problèmes d'équations différentielles mentionnés ci-dessus, ces résultats peuvent être vus comme des bornes sur le semigroupe $\exp(t\,A)$ en une suite de temps $n\,\Delta t$, où $n \in \mathbb{N}$ et $\Delta t > 0$ est un pas de temps fixé. Le Théorème 1 ci-dessous caractérise, en fonction du

comportement de leur résolvante, les familles de matrices dont les puissances sont uniformément bornées. Le Théorème 3 fournit une caractérisation similaire pour les familles de matrices dont les puissances sont uniformément sommables. Ces deux résultats fournissent des bornes qui dépendent de la taille des matrices et ne s'étendent donc pas à des espaces de dimension infinie. Les résultats correspondants sont d'ailleurs faux en dimension infinie. On renvoie néanmoins le lecteur à [HS10] pour des bornes de semigroupe en dimension infinie. Deux applications du Théorème 3 sont détaillées ensuite. Dans la Section 3, on s'intéresse à la résolution itérative de grands systèmes linéaires en tenant compte des erreurs d'arrondi qui, en pratique, interviennent nécessairement à chaque étape de l'itération. Le problème que l'on se pose est de quantifier de manière précise comment l'accumulation de ces erreurs d'arrondi entâche la convergence de l'itération. Dans la Section 4, on s'intéresse à la discrétisation de problèmes aux limites pour des équations hyperboliques. Après une brève introduction et quelques rappels, on démontre un résultat de stabilité pour des schémas décentrés amont qui ne requièrent pas de conditions aux limites numériques. Il s'agit bien sûr d'une situation très idéalisée mais qui se prête à une analyse simple et constitue une introduction à l'analyse de schémas plus généraux.

2 Famille de matrices de puissances uniformément sommables

Le but de cette section est de caractériser les sous-ensembles $\mathscr{F} \subset \mathscr{M}_N(\mathbb{C})$ pour lesquels la quantité suivante est finie :

$$\sup_{M\in\mathscr{F}} \sum_{k\geq 0} \|M^k\| \,.$$

Bien sûr, une condition nécessaire pour que cette quantité soit finie est que le rayon spectral de tout élément de $\mathscr{F}$ soit strictement inférieur à 1 (on utilise pour cela les relations $\rho(M)^k = \rho(M^k) \leq \|M^k\|$). On s'intéresse préalablement au cas plus simple des familles de matrices uniformément bornées pour lequel on rappelle le résultat fondamental de Kreiss [Kre62].

2.1 Rappels sur les familles de matrices de puissances bornées

On rappelle le Théorème suivant dont on va donner une démonstration complète.

Théorème 1 (Kreiss [Kre62])**.** *Soit $\mathscr{F} \subset \mathscr{M}_N(\mathbb{C})$ un ensemble de matrices. Alors sont équivalentes :*

(i) *Il existe une constante $C_1 > 0$ telle que pour tout élément $M \in \mathscr{F}$ et pour tout entier $k \in \mathbb{N}$, on a $\|M^k\| \leq C_1$.*

(ii) *Toute matrice $M \in \mathscr{F}$ a un rayon spectral inférieur ou égal à 1, et il existe une constante $C_2 > 0$ telle que pour tout élément $M \in \mathscr{F}$ et pour tout nombre complexe $z \in \mathscr{U}$, on a*

$$\|(z\,I - M)^{-1}\| \leq \frac{C_2}{|z| - 1} \,.$$

De plus, si la condition (i) *est vérifiée pour une constante* C_1*, on peut choisir* $C_2 = C_1$ *dans* (ii)*, et réciproquement, si la condition* (ii) *est vérifiée pour une constante* C_2*, on peut choisir* $C_1 = \mathrm{e}\, N\, C_2$ *dans* (i)*.*

Le Théorème 1 caractérise donc complètement les familles de matrices de puissances uniformément bornées par une certaine estimation de la résolvante en dehors du disque unité fermé. La démonstration originelle de [Kre62] établissait l'équivalence des conditions (i) et (ii). La quête de la meilleure constante fut l'objet de nombreux travaux qui trouvèrent leur épilogue dans [Spi91] (le résultat de [Spi91] ayant été conjecturé dans [LT84]). Le lecteur intéressé trouvera dans [SW97] une bibliographie complète ainsi que d'autres caractérisations et une discussion sur les opérateurs en dimension infinie. On rappelle maintenant la démonstration du Théorème 1 en guise de préliminaire au paragraphe suivant [1].

Démonstration du Théorème 1. Le fait que (i) implique (ii) est très aisé et repose sur le développement en série de la résolvante

$$(z\, I - M)^{-1} = \sum_{k \geq 0} \frac{M^k}{z^{k+1}}\,,$$

ce développement étant absolument convergent pour $z \in \mathscr{U}$ sous la condition (i). L'inégalité triangulaire permet de conclure.

On suppose maintenant que la condition (ii) est vérifiée, et on se donne $M \in \mathscr{F}$. Pour un entier $k \in \mathbb{N}$, un paramètre $r > 1$ à fixer ultérieurement, et des vecteurs $u, v \in \mathbb{C}^N$ de norme 1, on définit une fonction φ de classe $\mathscr{C}^\infty$ et $2\,\pi$-périodique par

$$\forall\, \theta \in \mathbb{R}\,, \quad \varphi(\theta) := \langle v, (r\, \mathrm{e}^{i\,\theta}\, I - M)^{-1}\, u \rangle\,.$$

Comme la résolvante de M est une fonction holomorphe sur le complémentaire du spectre de M, la formule de Cauchy donne

$$\langle v, M^k\, u \rangle = \frac{1}{2\, i\, \pi} \int_\gamma \zeta^k\, \langle v, (\zeta\, I - M)^{-1}\, u \rangle\, \mathrm{d}\zeta\,,$$

où γ est n'importe quel contour encerclant le spectre de M. En prenant pour γ le cercle de centre 0 et de rayon r (on rappelle le choix $r > 1$), on obtient

$$\langle v, M^k\, u \rangle = \frac{r^{k+1}}{2\,\pi} \int_0^{2\,\pi} \mathrm{e}^{i\,(k+1)\,\theta}\, \varphi(\theta)\, \mathrm{d}\theta = \frac{i\, r^{k+1}}{2\,\pi\,(k+1)} \int_0^{2\,\pi} \mathrm{e}^{i\,(k+1)\,\theta}\, \varphi'(\theta)\, \mathrm{d}\theta\,.$$

En particulier, l'inégalité de la moyenne donne

$$|\langle v, M^k\, u \rangle| \leq \frac{r^{k+1}}{2\,\pi\,(k+1)} \int_0^{2\,\pi} |\varphi'(\theta)|\, \mathrm{d}\theta\,. \tag{1}$$

Dans [Tad81, page 155], Tadmor montre que la fonction φ s'écrit sous la forme $\varphi(\theta) = P(r\, \mathrm{e}^{i\,\theta})/\Pi_M(r\, \mathrm{e}^{i\,\theta})$, où Π_M désigne le polynôme minimal de M, et P est un polynôme de

1. Le lecteur pourra aussi se référer à [TE05, chapitre 18] pour plus de commentaires historiques et un lien, peut-être inattendu, avec le problème des aiguilles de Buffon.

degré strictement inférieur à celui de Π_M. Pour être tout-à-fait complet, la formule de Taylor donne

$$(z\,I - M)^{-1} = \frac{1}{\Pi_M(z)} \sum_{j=1}^{\mathrm{d}^o\Pi_M - 1} \frac{1}{j!}\,\Pi_M^{(j)}(z)\,(M - z\,I)^{j-1}\,,$$

et il suffit d'insérer $z = r\,\mathrm{e}^{i\,\theta}$ pour obtenir l'écriture voulue de φ.

Les racines de Π_M étant les valeurs propres de M, et donc de module inférieur ou égal à 1, on vérifie que les racines de $\Pi_M(r\,\cdot)$ sont toutes incluses dans $\mathbb{D}$. On peut alors appliquer le résultat suivant :

Théorème 2 (Borwein et Erdélyi [BE96])**.** *Soit $f(z) = P(z)/Q(z)$ une fraction rationnelle complexe telle que*

$$Q(z) = \prod_{j=1}^{d}(z - a_j)\,, \quad a_1, \ldots, a_d \in \mathbb{D}\,,$$

et P est de degré inférieur ou égal à d. Alors on a

$$\forall\, z \in \mathbb{S}^1\,, \quad |f'(z)| \leq \|f\|_{L^\infty(\mathbb{S}^1)} \sum_{j=1}^{d} \frac{1 - |a_j|^2}{|z - a_j|^2}\,.$$

Dans notre cadre, le Théorème 2 donne l'estimation

$$\forall\, \theta \in \mathbb{R}\,, \quad |\varphi'(\theta)| \leq \|\varphi\|_{L^\infty(\mathbb{R})} \sum_{j=1}^{d} \frac{1 - |\lambda_j/r|^2}{|\mathrm{e}^{i\,\theta} - \lambda_j/r|^2}\,,$$

où d désigne le degré du polynôme minimal de M, et les λ_j sont des valeurs propres de M (certaines valeurs propres pouvant être répétées si elles sont racines multiples du polynôme minimal). On utilise cette estimation ponctuelle de φ' dans (1), et on obtient

$$|\langle v, M^k\,u\rangle| \leq \frac{r^{k+1}\,\|\varphi\|_{L^\infty(\mathbb{R})}}{k+1} \sum_{j=1}^{d} \frac{1}{2\,\pi} \int_0^{2\,\pi} \frac{1 - |\lambda_j/r|^2}{|\mathrm{e}^{i\,\theta} - \lambda_j/r|^2}\,\mathrm{d}\theta\,.$$

Chaque intégrale se calcule (on se ramène d'abord à λ_j réel puis on effectue le changement de variables classique $t = \tan(\theta/2)$), et on trouve

$$\frac{1}{2\,\pi} \int_0^{2\,\pi} \frac{1 - |\lambda_j/r|^2}{|\mathrm{e}^{i\,\theta} - \lambda_j/r|^2}\,\mathrm{d}\theta = 1\,.$$

Comme, par ailleurs, le degré d du polynôme minimal est inférieur ou égal à la taille N de la matrice (Cayley-Hamilton), on aboutit à l'estimation

$$|\langle v, M^k\,u\rangle| \leq N\,\frac{r^{k+1}\,\|\varphi\|_{L^\infty(\mathbb{R})}}{k+1}\,. \tag{2}$$

Il est temps d'utiliser la condition (ii), qui implique[2] $\|\varphi\|_{L^\infty(\mathbb{R})} \leq C_2/(r-1)$. On effectue alors le choix $1 + 1/(k+1)$ pour le paramètre r, et l'estimation (2) donne

$$|\langle v, M^k\, u\rangle| \leq N\, C_2 \left(1 + \frac{1}{k+1}\right)^{k+1} \leq \mathrm{e}\, N\, C_2\,.$$

Pour finir, on prend le supremum en v et u, et on obtient le résultat annoncé. □

La démonstration ci-dessus reprend essentiellement celle de [LT84]. La conclusion repose sur une estimation de la variation totale de φ sur $[0, 2\,\pi]$ en fonction de sa norme L^∞. L'estimation correspondante dans [LT84] manquait la constante optimale d'un facteur 2, et fut raffinée dans [Spi91]. On s'est appuyé ici sur l'estimation encore plus fine fournie par [BE96] (le Théorème 2 permet de contrôler ponctuellement φ' tandis que l'analyse dans [Spi91] ne donne qu'un contrôle dans L^1).

Même si on ne la reproduira pas ici, la démonstration du Théorème 2 est très accessible, et on invite le lecteur intéressé à consulter l'article [BE96] pour les détails qui y sont exposés de façon limpide.

Le Théorème 1 et ses nombreuses généralisations est un outil fondamental dans l'étude de la stabilité des schémas aux différences finies pour les équations d'évolution. La famille $\mathscr{F}$ est alors l'ensemble, paramétré par les fréquences, des matrices d'amplification du schéma numérique. Le lecteur pourra consulter les références [RM94, GKO95] pour avoir un aperçu de telles applications. On renvoie également à la Section 4.

2.2 Famille de matrices de puissances uniformément sommables

Le but de ce paragraphe est d'établir le résultat suivant :

Théorème 3 (Boyd et Doyle [BD87])**.** *Soit $\mathscr{F} \subset \mathscr{M}_N(\mathbb{C})$ un ensemble de matrices. Alors sont équivalentes :*

(i) *Il existe une constante $C_1 > 0$ telle que pour tout élément $M \in \mathscr{F}$ et pour tout entier $k \in \mathbb{N}$, on a $\sum_{k\geq 0} \|M^k\| \leq C_1$.*

(ii) *Toute matrice $M \in \mathscr{F}$ a un rayon spectral strictement inférieur à 1, et il existe une constante $C_2 > 0$ telle que pour tout élément $M \in \mathscr{F}$ et pour tout nombre complexe $z \in \overline{\mathscr{U}}$, on a*

$$\|(z\, I - M)^{-1}\| \leq C_2\,.$$

De plus, si la condition (i) *est vérifiée pour une constante C_1, on peut choisir $C_2 = C_1$ dans* (ii)*, et réciproquement, si la condition* (ii) *est vérifiée pour une constante C_2, on peut choisir $C_1 = 2\, N\, C_2$ dans* (i)*.*

On remarque que dans la condition (ii), on peut remplacer l'estimation uniforme de la résolvante sur $\overline{\mathscr{U}}$ par une estimation sur sa frontière $\mathbb{S}^1$. Ce n'est rien d'autre que le principe du maximum pour les fonctions holomorphes. En effet, si M est une matrice de rayon spectral strictement inférieur à 1, alors pour u et v vecteurs de norme 1, la fonction

$$\zeta \longmapsto \langle v, (I - \zeta\, M)^{-1}\, u\rangle\,,$$

2. On rappelle que les vecteurs u et v sont de norme 1.

est holomorphe sur $\mathbb{D}$, et continue sur son adhérence. On en déduit l'inégalité

$$\forall\, \zeta \in \mathbb{D}\,, \quad |\langle v, (I - \zeta\, M)^{-1}\, u\rangle| \leq \sup_{z\in\mathbb{S}^1} |\langle v, (I - z\, M)^{-1}\, u\rangle| \leq \sup_{z\in\mathbb{S}^1} \|(I - z\, M)^{-1}\|\,.$$

En prenant le supremum par rapport à v puis u, on obtient l'estimation raffinée

$$\forall\, \zeta \in \overline{\mathscr{U}}\,, \quad \|(\zeta\, I - M)^{-1}\| \leq \frac{1}{|\zeta|} \sup_{z\in\mathbb{S}^1} \|(z\, I - M)^{-1}\|\,.$$

Cela explique pourquoi, dans la démonstration ci-dessous, on utilisera uniquement l'estimation de la résolvante sur le cercle unité.

On commence par donner une démonstration du Théorème 3 inspirée des techniques du Théorème 1, mais qui ne fournira malheureusement pas la constante $2\, N\, C_2$ (la correction est d'un facteur logarithmique en C_2). La démonstration du Théorème 3, qu'on reprend de [LN91], sera détaillée ensuite.

Une démonstration aisée mais pas optimale du Théorème 3. Comme dans la démonstration du Théorème 1, le fait que (i) implique (ii) repose sur le développement en série de la résolvante

$$(z\, I - M)^{-1} = \sum_{k\geq 0} \frac{M^k}{z^{k+1}}\,,$$

ce développement étant absolument convergent pour $z \in \overline{\mathscr{U}}$ sous la condition (i). L'inégalité triangulaire permet encore de conclure.

On suppose désormais que la condition (ii) est satisfaite et on se donne une matrice $M \in \mathscr{F}$. On va montrer que (i) a lieu avec la constante [3] $C_1 = 2\, N\, C_2\, \ln(2\, C_2)$. Pour cela, on commence par étendre l'estimation de la résolvante dans une couronne à l'intérieur du disque unité [4].

Lemme 1. *Sous la condition* (ii)*, une matrice $M \in \mathscr{F}$ vérifie $1 - \rho(M) \geq 1/C_2$. De plus, si $\delta \geq 0$ vérifie $\delta\, C_2 < 1$, alors M n'a pas de valeur propre de module égal à $1 - \delta$, et on a l'estimation*

$$\sup_{z\in(1-\delta)\,\mathbb{S}^1} \|(z\, I - M)^{-1}\| \leq \frac{C_2}{1 - \delta\, C_2}\,.$$

Démonstration du Lemme 1. Supposons tout d'abord que le spectre de M n'est pas réduit à $\{0\}$. Soit λ une valeur propre réalisant le rayon spectral de M, et soit X un vecteur propre de norme 1 pour la valeur propre λ. Alors $\zeta := \lambda/|\lambda| \in \mathbb{S}^1$ vérifie

$$(\zeta\, I - M)\, X = \zeta\, (1 - |\lambda|)\, X\,,$$

c'est-à-dire (ζ n'étant pas valeur propre de M) :

$$X = \zeta\, (1 - \rho(M))\, (\zeta\, I - M)^{-1}\, X\,.$$

3. On pourra comparer avec [LN91, Théorème 2.3] où les auteurs montrent un résultat analogue avec la constante $C_1 = 2\, \mathrm{e}\, N\, C_2\, (1 + \ln(2\, C_2))$. On fait donc un petit peu mieux ici.

4. Ce genre d'arguments est assez usuel, voir par exemple [HS10, Lemme 1.2].

En passant à la norme, on obtient $1 \leq (1-\rho(M))\, C_2$, ce qui est l'inégalité cherchée. Dans le cas où la seule valeur propre de M est 0, on veut montrer $C_2 \geq 1$. Pour cela, on se donne X un vecteur de norme 1 dans le noyau de M. On écrit $X = (I-M)^{-1}\, X$ et on passe à la norme.

L'inégalité $1-\rho(M) \geq 1/C_2$ assure que pour tout $\delta \geq 0$ vérifiant $\delta\, C_2 < 1$, M n'a pas de valeur propre de module plus grand ou égal à $1-\delta$. Pour un tel paramètre δ, on se donne un nombre complexe $z \in \mathbb{C}$ de module $1-\delta$ (z est nécessairement non-nul), et on pose $\zeta := z/|z| \in \mathbb{S}^1$. On commence par remarquer que la norme de la matrice $(z-\zeta)\,(\zeta\, I - M)^{-1}$ vérifie

$$\|(z-\zeta)\,(\zeta\, I - M)^{-1}\| = |z-\zeta|\, \|(\zeta\, I - M)^{-1}\| \leq C_2\,(1-|z|) = \delta\, C_2 < 1\,.$$

Ainsi la matrice $I + (z-\zeta)\,(\zeta\, I - M)^{-1}$ est inversible, et son inverse est de norme plus petite ou égale à $1/(1-\delta\, C_2)$, voir [Ser10, Proposition 7.5] (c'est encore le développement en série de la résolvante). Or on a la relation

$$I + (z-\zeta)\,(\zeta\, I - M)^{-1} = (z\, I - M)\,(\zeta\, I - M)^{-1}\,,$$

et on a donc montré l'inégalité

$$\|(\zeta\, I - M)\,(z\, I - M)^{-1}\| \leq \frac{1}{1-\delta\, C_2}\,.$$

Pour finir, on écrit

$$(z\, I - M)^{-1} = (\zeta\, I - M)^{-1}\,(\zeta\, I - M)\,(z\, I - M)^{-1}\,,$$

et on utilise le fait que $\|\cdot\|$ est une norme matricielle pour conclure. □

On en vient maintenant à la démonstration (non-optimale) du Théorème 3, en supposant toujours que la condition (ii) est satisfaite. Pour $M \in \mathscr{F}$, $k \in \mathbb{N}$, $\delta > 0$ vérifiant $\delta\, C_2 < 1$, et des vecteurs $u, v \in \mathbb{C}^N$ de norme 1, on pose

$$\forall\, \theta \in \mathbb{R}\,, \quad \varphi(\theta) := \langle v, ((1-\delta)\, \mathrm{e}^{i\,\theta}\, I - M)^{-1}\, u\rangle\,.$$

On reprend la même démarche que dans la démonstration du Théorème 1, en choisissant cette fois comme contour d'intégration γ le cercle de centre 0 et de rayon $1-\delta$. Ce contour encercle bien le spectre de M d'après le Lemme 1. On obtient ainsi

$$\langle v, M^k\, u\rangle = \frac{i\,(1-\delta)^{k+1}}{2\,\pi\,(k+1)} \int_0^{2\,\pi} \mathrm{e}^{i\,(k+1)\,\theta}\, \varphi'(\theta)\, \mathrm{d}\theta\,.$$

L'inégalité de la moyenne puis les mêmes arguments[5] que dans la démonstration du Théorème 1 permettent d'aboutir à l'estimation (comparer avec l'inégalité (2)) :

$$|\langle v, M^k\, u\rangle| \leq N\, \frac{(1-\delta)^{k+1}\, \|\varphi\|_{L^\infty(\mathbb{R})}}{k+1}\,.$$

5. C'est-à-dire : l'écriture de φ en fraction rationnelle trigonométrique, l'estimation ponctuelle de φ' fournie par le Théorème 2, puis le calcul d'intégrales.

La norme L^∞ de φ se contrôle grâce au Lemme 1 et on obtient, après avoir pris le supremum par rapport aux vecteurs v et u :

$$\forall\, k \in \mathbb{N}\,, \quad \|M^k\| \leq \frac{N\, C_2}{1 - \delta\, C_2}\, \frac{(1-\delta)^{k+1}}{k+1}\,. \tag{3}$$

En prenant un paramètre δ indépendant de k, on somme par rapport à $k \in \mathbb{N}$ et on obtient

$$\sum_{k\geq 0} \|M^k\| \leq N\, C_2\, \frac{-\ln \delta}{1 - \delta\, C_2}\,.$$

On souhaiterait alors optimiser le majorant par rapport au paramètre δ, mais cela ne conduit pas à des formules très explicites[6]. Le choix $\delta := 1/(2\, C_2)$ conduit toutefois à l'estimation

$$\sum_{k\geq 0} \|M^k\| \leq 2\, N\, C_2\, \ln(2\, C_2)\,,$$

ce qui montre déjà que la famille $\mathscr{F}$ est de puissances uniformément sommables. □

On détaille maintenant la démonstration du Théorème 3, en faisant appel à des outils entièrement différents de ceux utilisés dans la démonstration du Théorème 1.

Démonstration du Théorème 3. On a vu le fait que (i) implique (ii) avec la même constante dans la démonstration ci-dessus. On se concentre donc sur le fait que (ii) implique (i). Pour cela, on commence par établir une première inégalité générale sur les matrices de rayon spectral strictement inférieur à 1.

Lemme 2. *Soit $M \in \mathscr{M}_N(\mathbb{C})$ une matrice de rayon spectral strictement inférieur à 1. Alors les formules*

$$P := \sum_{k\geq 0} M^k\, (M^k)^*\,, \quad Q := \sum_{k\geq 0} (M^k)^*\, M^k\,,$$

définissent des matrices hermitiennes définies positives. Les valeurs propres de la matrice $P\, Q$ sont des réels strictement positifs, dont les racines carrées sont appelées les valeurs singulières de Hankel de M. En notant ces dernières $\sigma_1, \ldots, \sigma_N$, on a

$$\sum_{k\geq 0} \|M^k\| \leq 2 \sum_{i=1}^{N} \sigma_i\,. \tag{4}$$

Démonstration du Lemme 2. Le Théorème de Householder, voir [Ser10, chapitre 7], assure que les normes $\|M^k\|$ décroissent géométriquement, ce qui montre la convergence absolue des séries définissant P et Q. Ces matrices sont clairement hermitiennes et définies positives (comme somme de l'identité et de matrices hermitiennes positives). On se rend compte que les valeurs propres de $P\, Q$ sont réelles et strictement positives en utilisant l'astuce standard

$$P\, Q = P^{1/2}\, (P^{1/2}\, Q\, P^{1/2})\, P^{-1/2}\,,$$

6. L'idéal serait bien sûr d'optimiser δ par rapport à k et C_2, puis de sommer par rapport à k, mais cela s'avère encore moins explicite. Il semble de toute façon impossible à ce stade d'obtenir une estimation finale qui soit linéaire en C_2.

qui montre que $P\,Q$ est semblable à la matrice hermitienne définie positive $P^{1/2}\,Q\,P^{1/2}$. Cela justifie l'existence des valeurs singulières de Hankel de M.

On s'intéresse maintenant à majorer la somme de la série des $\|M^k\|$. Pour cela, on se donne une matrice inversible T, à fixer ultérieurement. On définit

$$P_T := \sum_{k\geq 0} T^{-1}\,M^k\,(T^{-1}\,M^k)^* = T^{-1}\,P\,(T^{-1})^*\,,\quad Q_T := \sum_{k\geq 0}(M^k\,T)^*\,M^k\,T = T^*\,Q\,T\,, \tag{5}$$

et on pose $A := T^{-1}\,M\,T$ de sorte qu'on a la relation $M^k = T\,A^k\,T^{-1}$ pour tout entier k. Les définitions de P_T et Q_T se réécrivent donc

$$P_T = \sum_{k\geq 0} A^k\,T^{-1}\,(A^k\,T^{-1})^*\,,\quad Q_T = \sum_{k\geq 0}(T\,A^k)^*\,T\,A^k\,.$$

On majore

$$\begin{aligned}\sum_{k\geq 0}\|M^k\| &= \sum_{k\geq 0}\|M^{2\,k}\| + \sum_{k\geq 0}\|M^{2\,k+1}\| = \sum_{k\geq 0}\|T\,A^k\,A^k\,T^{-1}\| + \sum_{k\geq 0}\|T\,A^k\,A^{k+1}\,T^{-1}\| \\ &\leq \sum_{k\geq 0}\|T\,A^k\|\,\|A^k\,T^{-1}\| + \sum_{k\geq 0}\|T\,A^k\|\,\|A^{k+1}\,T^{-1}\|\,,\end{aligned}$$

puis on utilise l'inégalité de Cauchy-Schwarz pour majorer les deux dernières sommes par

$$\left(\sum_{k\geq 0}\|T\,A^k\|^2\right)^{1/2}\left(\sum_{k\geq 0}\|A^k\,T^{-1}\|^2\right)^{1/2} + \left(\sum_{k\geq 0}\|T\,A^k\|^2\right)^{1/2}\left(\sum_{k\geq 1}\|A^k\,T^{-1}\|^2\right)^{1/2}\,.$$

On aboutit ainsi à la première estimation

$$\sum_{k\geq 0}\|M^k\| \leq 2\,\left(\sum_{k\geq 0}\|T\,A^k\|^2\right)^{1/2}\left(\sum_{k\geq 0}\|A^k\,T^{-1}\|^2\right)^{1/2}\,. \tag{6}$$

A ce stade, on introduit la norme de Frobenius $\|\cdot\|_F$ d'une matrice :

$$\|B\|_F^2 := \sum_{i,j=1}^N |B_{i,j}|^2 = \text{Trace}\,(B^*\,B) = \text{Trace}\,(B\,B^*)\,.$$

L'inégalité de Cauchy-Schwarz (dans $\mathbb{C}^N$) permet de montrer l'inégalité $\|\cdot\| \leq \|\cdot\|_F$. En utilisant cette dernière inégalité dans (6), on obtient

$$\begin{aligned}\sum_{k\geq 0}\|M^k\| &\leq 2\,\left(\sum_{k\geq 0}\text{Trace}\,((T\,A^k)^*\,T\,A^k)\right)^{1/2}\left(\sum_{k\geq 0}\text{Trace}\,(A^k\,T^{-1}\,(A^k\,T^{-1})^*)\right)^{1/2} \\ &= 2\,(\text{Trace}\,Q_T)^{1/2}\,(\text{Trace}\,P_T)^{1/2}\,,\end{aligned} \tag{7}$$

où les matrices P_T, Q_T sont définies en (5).

L'inégalité (4) s'obtient à partir de (7) en choisissant judicieusement la matrice T. Précisément, la matrice Q (définie dans l'énoncé du Lemme 2) est hermitienne définie positive, et admet donc une décomposition de Cholesky $Q = R^* R$ avec R triangulaire supérieure à diagonale réelle strictement positive, voir [Ser10, chapitre 11.2]. Comme $R P R^*$ est hermitienne définie positive, on peut la diagonaliser au moyen d'une matrice unitaire :

$$R P R^* = U D^2 U^*, \quad U^* U = I, \quad D = \text{diag}\,(d_1, \dots, d_N),$$

où les d_i sont tous des nombres réels strictement positifs. En posant $T := (D^{-1/2} U^* R)^{-1}$, et en utilisant les relations (5), on calcule

$$P_T = T^{-1} P (T^{-1})^* = D^{-1/2} U^* R P R^* U D^{-1/2} = D^{-1/2} D^2 D^{-1/2} = D,$$

et

$$Q_T = T^* Q T = D^{1/2} U^* (R^{-1})^* R^* R R^{-1} U D^{1/2} = D^{1/2} D^{1/2} = D.$$

Au final, ce choix de la matrice inversible T donne $P_T = Q_T = D$, donc (7) se simplifie en

$$\sum_{k \geq 0} \|M^k\| \leq 2 \,\text{Trace}\, D.$$

Les relations (5) montrent que le produit $P_T Q_T$ est semblable à la matrice PQ. Pour notre choix de T, le produit $P_T Q_T$ est la matrice diagonale D^2 donc les coefficients diagonaux de D ne sont rien d'autre que les valeurs singulières de Hankel de M (l'ordre de numérotation n'importe pas ici car on en prend la somme). On a donc bien montré l'inégalité (4) et cela conclut la démonstration du Lemme 2. □

Pour finir la démonstration du Théorème 3, on va montrer la propriété (on garde les notations introduites au Lemme 2 pour les valeurs singulières de Hankel) :

$$\max_{i=1,\dots,N} \sigma_i \leq \sup_{z \in \mathbb{S}^1} \|(z I - M)^{-1}\|. \tag{8}$$

L'idée est de relier chacune de ces deux quantités à des normes d'opérateurs [7] sur l'espace ℓ^2. On commence par le membre de droite dans l'inégalité (8).

Lemme 3. *Soit $M \in \mathscr{M}_N(\mathbb{C})$ une matrice de rayon spectral strictement inférieur à 1. Alors l'opérateur*

$$G \quad : \quad (u_n)_{n \in \mathbb{Z}} \longmapsto \left(\sum_{k \leq n} M^{n-k} u_k \right)_{n \in \mathbb{Z}},$$

est borné sur $\ell^2(\mathbb{Z}; \mathbb{C}^N)$, et sa norme d'opérateur est égale à $\sup_{z \in \mathbb{S}^1} \|(z I - M)^{-1}\|$.

La norme d'une suite (u_n) de carré intégrable et à valeurs dans $\mathbb{C}^N$ est bien entendu définie par

$$\|(u_n)\|^2_{\ell^2(\mathbb{Z}; \mathbb{C}^N)} := \sum_{n \in \mathbb{Z}} |u_n|^2,$$

ce qui revient à sommer les carrés des normes ℓ^2 de chaque coordonnée de la suite (u_n).

7. Le lecteur trouvera dans [Nik02, Partie B] une étude approfondie des opérateurs de Hankel. Les résultats qu'on expose ici ne sont que des cas particuliers assez simples de résultats bien plus généraux.

Démonstration du Lemme 3. On commence par se donner une suite $(u_n)_{n\in\mathbb{Z}}$ à valeurs dans $\mathbb{C}^N$ ne comportant qu'un nombre fini de termes non-nuls (de telles suites sont évidemment denses dans ℓ^2). On définit le polynôme trigonométrique

$$\forall\, \theta \in \mathbb{R}\,, \quad u(\theta) := \sum_{n\in\mathbb{Z}} \mathrm{e}^{i\,n\,\theta}\, u_n\,,$$

ainsi que la fonction à valeurs matricielles

$$\forall\, \theta \in \mathbb{R}\,, \quad \mathbf{M}(\theta) := (I - \mathrm{e}^{i\,\theta}\, M)^{-1} = \sum_{k\geq 0} \mathrm{e}^{i\,k\,\theta}\, M^k\,.$$

Le Théorème de Parseval-Bessel (appliqué à chacune des coordonnées de u) donne

$$\frac{1}{2\,\pi} \int_0^{2\,\pi} |u(\theta)|^2\, \mathrm{d}\theta = \sum_{n\in\mathbb{Z}} |u_n|^2\,.$$

Calculons le développement en séries de Fourier de la fonction $\mathbf{M}\,u$. En multipliant les deux développements en séries de Fourier de $\mathbf{M}$ et u (on rappelle que u n'a qu'un nombre fini de coefficients de Fourier non-nuls donc la multiplication terme à terme des deux développements ne pose pas de problème), on obtient

$$\mathbf{M}(\theta)\, u(\theta) = \sum_{n\in\mathbb{Z}} \mathrm{e}^{i\,n\,\theta} \left(\sum_{p=-\infty}^{n} M^{n-p}\, u_p \right) = \sum_{n\in\mathbb{Z}} \mathrm{e}^{i\,n\,\theta}\, (G\,u)_n\,.$$

En appliquant de nouveau le Théorème de Parseval-Bessel, on obtient donc

$$\begin{aligned}\sum_{n\in\mathbb{Z}} |(G\,u)_n|^2 = \frac{1}{2\,\pi} \int_0^{2\,\pi} |\mathbf{M}(\theta)\, u(\theta)|^2\, \mathrm{d}\theta &\leq \left(\sup_{\theta\in\mathbb{R}} \|\mathbf{M}(\theta)\| \right)^2 \frac{1}{2\,\pi} \int_0^{2\,\pi} |u(\theta)|^2\, \mathrm{d}\theta \\ &= \left(\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\| \right)^2 \sum_{n\in\mathbb{Z}} |u_n|^2\,.\end{aligned}$$

Cette dernière inégalité montre que G est borné sur $\ell^2(\mathbb{Z};\mathbb{C}^N)$, et que sa norme est inférieure ou égale à la quantité $\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\|$. On montre désormais que cette quantité coïncide bien avec la norme d'opérateur de G. Pour cela, on se donne un argument $\theta_0 \in \mathbb{R}$ et un vecteur $X_0 \in \mathbb{C}^N$ de norme 1 tels que

$$\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\| = \|(I - \mathrm{e}^{i\,\theta_0}\, M)^{-1}\| = |(I - \mathrm{e}^{i\,\theta_0}\, M)^{-1}\, X_0|\,.$$

Pour tout entier k non-nul, on définit un polynôme trigonométrique u_k par

$$\forall\, \theta \in \mathbb{R}\,, \quad u_k(\theta) := \frac{1}{\sqrt{k}} \sum_{\nu=1}^{k} \mathrm{e}^{-i\,\nu(\theta-\theta_0)}\, X_0\,,$$

de sorte que la suite de fonctions (u_k) vérifie

$$\forall\, k \in \mathbb{N}\,, \quad \frac{1}{2\,\pi} \int_0^{2\,\pi} |u_k(\theta)|^2 \,\mathrm{d}\theta = 1\,,$$
$$\text{et} \quad \lim_{k\to+\infty} \frac{1}{2\,\pi} \int_0^{2\,\pi} |\mathbf{M}(\theta)\, u_k(\theta)|^2 \,\mathrm{d}\theta = |\mathbf{M}(\theta_0)\, X_0|^2 = \left(\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\|\right)^2 .$$

Le Théorème de Parseval-Bessel permet de conclure que la norme d'opérateur de G sur $\ell^2(\mathbb{Z};\mathbb{C}^N)$ est bien égale à $\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\|$. □

Lemme 4. *Soit $M \in \mathscr{M}_N(\mathbb{C})$ une matrice de rayon spectral strictement inférieur à 1. Alors l'opérateur*

$$H \quad : \quad (u_n)_{n\in\mathbb{N}} \longmapsto \left(\sum_{k\geq 0} M^{n+k}\, u_k\right)_{n\in\mathbb{N}} ,$$

est borné sur $\ell^2(\mathbb{N};\mathbb{C}^N)$ et compact. De plus sa norme d'opérateur est égale à la plus grande des valeurs singulières de Hankel de M (ces valeurs sont définies au Lemme 2).

Démonstration du Lemme 4. Pour montrer que H est borné et compact sur $\ell^2(\mathbb{N};\mathbb{C}^N)$, on écrit H comme la composée $H = H_2 \circ H_1$, avec

$$H_1 \quad : \quad (u_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N};\mathbb{C}^N) \longmapsto \sum_{k\geq 0} M^k\, u_k \in \mathbb{C}^N\,,$$

et

$$H_2 \quad : \quad v \in \mathbb{C}^N \longmapsto (M^n\, v)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N};\mathbb{C}^N)\,.$$

On vérifie sans peine que les opérateurs H_1 et H_2 sont bornés car la suite $(\|M^n\|)_{n\in\mathbb{N}}$ décroît géométriquement donc est de carré intégrable. Chacun des opérateurs H_1 et H_2 est compact car son espace d'arrivée ou de départ est de dimension finie, et donc H est bien un opérateur borné sur $\ell^2(\mathbb{N};\mathbb{C}^N)$ et compact (comme composée de deux opérateurs compacts).

Il n'est pas très difficile de voir que l'adjoint de H est l'opérateur

$$H^* \quad : \quad (v_n)_{n\in\mathbb{N}} \longmapsto \left(\sum_{k\geq 0} (M^*)^{n+k}\, v_k\right)_{n\in\mathbb{N}} .$$

Une fois que l'on sait que H est compact, on en déduit que $H^*\,H$ est également un opérateur compact. C'est de surcroît un opérateur autoadjoint positif, donc la norme d'opérateur de $H^*\,H$ est une valeur propre de $H^*\,H$, voir [Con90, Lemme 5.9]. On va calculer cette norme. On remarque tout d'abord que H est non-nul (appliquer H à la suite $(X, 0, \dots, 0, \dots)$, $X \in \mathbb{C}^N$ non-nul), ce qui implique que $H^*\,H$ a des valeurs propres strictement positives. Soit donc $\lambda > 0$ une valeur propre de $H^*\,H$, et $(v_n)_{n\in\mathbb{N}}$ un vecteur propre de norme 1 pour la valeur propre λ. On définit le vecteur

$$V := \sum_{k\geq 0} M^k\, v_k = H_1\, v\,,$$

de sorte que la suite $H\,v = H_2\,(H_1\,v)$ est donnée par

$$\forall\, n \in \mathbb{N}\,, \quad (H\,v)_n = M^n\,V\,.$$

Le vecteur V est nécessairement non-nul (car sinon tous les termes de la suite $H\,v$ seraient nuls, et on aurait $0 = H^*\,H\,v = \lambda\,v$). D'après l'expression de l'opérateur H^* donnée ci-dessus, on a

$$\forall\, n \in \mathbb{N}\,, \quad \lambda\,v_n = \sum_{j\geq 0} (M^*)^{j+n}\,(H\,v)_j = \sum_{j\geq 0} (M^*)^{j+n}\,M^j\,V = (M^*)^n\,Q\,V\,,$$

où la matrice Q est définie au Lemme 2. En multipliant l'égalité $\lambda\,v_n = (M^*)^n\,Q\,V$ par M^n puis en sommant par rapport à l'entier n, on trouve finalement $\lambda\,V = P\,Q\,V$, où P est également définie au Lemme 2. Comme le vecteur V est non-nul, λ est une valeur propre de $P\,Q$ et on aboutit à l'inégalité

$$\|H^*\,H\|_{\ell^2\to\ell^2} \leq \max_{i=1,\dots,N} \sigma_i^2\,. \tag{9}$$

Réciproquement, soit σ_i l'une des valeurs singulières de Hankel de M. Il existe donc un vecteur non-nul $X \in \mathbb{C}^N$ tel que $P\,Q\,X = \sigma_i^2\,X$. On définit un élément de $\ell^2(\mathbb{N};\mathbb{C}^N)$ par

$$\forall\, n \in \mathbb{N}\,, \quad v_n := (M^*)^n\,Q\,X\,.$$

La suite $(v_n)_{n\in\mathbb{N}}$ vérifie $H_1\,v = P\,Q\,X = \sigma_i^2\,X$, ce qui montre que v n'est pas la suite nulle (il est clair de toute façon que v_0 est non-nul). Par ailleurs, on a $(H\,v)_n = \sigma_i^2\,M^n\,X$ pour tout n, et on trouve donc

$$\forall\, n \in \mathbb{N}\,, \quad (H^*\,H\,v)_n = \sum_{k\geq 0} (M^*)^{n+k}\,(\sigma_i^2\,M^k\,X) = \sigma_i^2\,(M^*)^n\,Q\,X = \sigma_i^2\,v_n\,.$$

Les nombres σ_i^2 sont donc des valeurs propres de $H^*\,H$, ce qui montre qu'il y a en fait égalité entre les deux membres de l'inégalité (9). Comme, par ailleurs, on a $\|H\|^2_{\ell^2\to\ell^2} = \|H^*\,H\|_{\ell^2\to\ell^2}$, on a bien montré que la norme d'opérateur de H est égale à la plus grande valeur singulière de Hankel de M. (Le lecteur aura remarqué qu'on a en fait montré que les valeurs propres strictement positives de $H^*\,H$ sont exactement les valeurs singulières de Hankel de M.) □

Le Théorème 3 s'obtient en mettant bout-à-bout les Lemmes 2, 3 et 4. Pour une matrice $M \in \mathscr{F}$, l'inégalité (4) donne

$$\sum_{k\geq 0} \|M^k\| \leq 2\,N \max_{i=1,\dots,N} \sigma_i\,,$$

et il suffit donc de montrer l'inégalité (8) pour conclure. Les Lemmes 3 et 4 permettent de réécrire l'inégalité (8) comme $\|H\|_{\ell^2(\mathbb{N})\to\ell^2(\mathbb{N})} \leq \|G\|_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})}$. Cette dernière inégalité est très simple à obtenir. En effet, soit $(v_n)_{n\in\mathbb{N}}$ un élément de $\ell^2(\mathbb{N};\mathbb{C}^N)$. On pose

$$\forall\, j \in \mathbb{Z}\,, \quad u_j := \begin{cases} 0 & \text{si } j > 0\,, \\ v_{-j} & \text{si } j \leq 0\,, \end{cases}$$

et on remarque que pour tout $n \in \mathbb{N}$, on a $(G\,u)_n = (H\,v)_n$. On a donc

$$\|H\,v\|^2_{\ell^2(\mathbb{N})} = \sum_{n\geq 0} |(G\,u)_n|^2 \leq \|G\,u\|^2_{\ell^2(\mathbb{Z})} \leq \|G\|^2_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})}\,\|u\|^2_{\ell^2(\mathbb{Z})} = \|G\|^2_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})}\,\|v\|^2_{\ell^2(\mathbb{N})}\,,$$

ce qui montre que l'inégalité (8) a lieu, et conclut la démonstration du Théorème 3. □

On va voir dans les sections suivantes deux applications du Théorème 3 dans des domaines très distincts d'analyse numérique.

3 Résolution itérative de grands systèmes linéaires

3.1 Considérations générales

Les méthodes itératives de résolution d'un système linéaire

$$A\,x = b\,,\quad A \in \mathrm{GL}_N(\mathbb{C})\,,\quad b \in \mathbb{C}^N\,, \tag{10}$$

sont basées sur une décomposition de la matrice A en $A = B - C$. On écrit alors un schéma itératif de Picard (en supposant la matrice B inversible) :

$$x_{n+1} = B^{-1}\,(C\,x_n + b)\,,\quad n \in \mathbb{N}\,,\quad x_0 \in \mathbb{C}^N\,. \tag{11}$$

Les méthodes usuelles - par exemple celles de Jacobi, Gauss-Seidel, ou de sur-relaxation - sont exposées dans de nombreux ouvrages, voir notamment [All05, Cia82, Sch02, Ser10]. D'un point de vue général, le schéma itératif (11) converge, indépendamment de la donnée initiale x_0, vers l'unique solution de (10) si et seulement si $\rho(B^{-1}\,C) < 1$. Si l'on note x_∞ la limite de la suite $(x_n)_{n\in\mathbb{N}}$, c'est-à-dire l'unique vecteur dans $\mathbb{C}^N$ vérifiant $A\,x_\infty = b$, alors la suite des erreurs $e_n := x_n - x_\infty$ vérifie

$$\forall\, n \in \mathbb{N}\,,\quad e_n = (B^{-1}\,C)^n\,e_0\,.$$

En particulier, le rayon spectral $\rho(B^{-1}\,C)$ gouverne *asymptotiquement* le taux de convergence de la méthode itérative.

Pour une méthode itérative convergente, on sait qu'il existe une norme matricielle $\|\cdot\|_0$, subordonnée à une norme $|\cdot|_0$ sur $\mathbb{C}^N$, pour laquelle $\|B^{-1}\,C\|_0 < 1$, mais ce résultat a un intérêt pratique limité. En effet, la suite des erreurs $(e_n)_{n\in\mathbb{N}}$ vérifie l'estimation

$$\forall\, n \in \mathbb{N}\,,\quad |e_n|_0 \leq \|B^{-1}\,C\|_0^n\,|e_0|_0\,,$$

mais on est surtout intéressé par le comportement de la norme $|e_n|$, ou tout au moins d'une norme fixe aisément calculable. Il y a certes équivalence entre la norme $|\cdot|_0$ et la norme hermitienne $|\cdot|$ mais les constantes d'équivalence peuvent être très grandes. Il devient alors malaisé de prédire, pour un paramètre de tolérance $\varepsilon > 0$ donné, le rang à partir duquel on a $|e_n| \leq \varepsilon$ pour tout n et où on peut donc arrêter l'itération.

Si on ne considère que la norme subordonnée à la norme hermitienne, on obtient l'estimation

$$\forall\, n \in \mathbb{N}\,,\quad |e_n| \leq \|(B^{-1}\,C)^n\|\,|e_0|\,, \tag{12}$$

et on peut alors chercher à déterminer un critère d'arrêt pour la méthode itérative. Pour un paramètre de tolérance $\varepsilon > 0$ donné, on cherche un entier n au-delà duquel on a $\|(B^{-1}\,C)^n\|\,|e_0| \leq \varepsilon$. On voit bien à ce moment l'utilité des estimations démontrées à la partie précédente : le comportement des normes $\|(B^{-1}\,C)^n\|$ dépend très fortement du caractère "normal" de la matrice $B^{-1}\,C$, voir par exemple [TE05, chapitre 16]. Dans les estimations de la section précédente, on quantifie la "non-normalité" de la matrice $B^{-1}\,C$ en estimant la résolvante de cette matrice sur le cercle unité [8] et on obtient une estimation des puissances $\|(B^{-1}\,C)^n\|$ qui tient compte de cet éventuel défaut de normalité. Par exemple, l'estimation (3) donne (en choisissant le paramètre $\delta := 1/(2\,C_2)$ avec les notations de la Section 2) :

$$\|(B^{-1}\,C)^n\| \leq 2\,N\,\mathbf{C}\,\frac{(1-1/(2\,\mathbf{C}))^{n+1}}{n+1}\,, \quad \mathbf{C} := \sup_{z\in\mathbb{S}^1} \|(z\,I - B^{-1}\,C)^{-1}\|\,. \tag{13}$$

Il convient donc, au pire, d'arrêter la méthode itérative au rang n_ε vérifiant

$$\frac{(1-1/(2\,\mathbf{C}))^{n_\varepsilon+1}}{n_\varepsilon+1} \leq \frac{\varepsilon}{2\,N\,\mathbf{C}\,|e_0|}\,.$$

Il s'agit d'un critère d'arrêt *a priori*, c'est-à-dire qu'il ne requiert pas l'évaluation de la norme $|e_n|$ à chaque étape de l'itération (11).

3.2 Modélisation des erreurs d'arrondi

Quand on implémente la méthode itérative (11), il est bien sûr illusoire d'espérer résoudre exactement la suite de vecteurs $(x_n)_{n\in\mathbb{N}}$ vérifiant (11). Dans la pratique, chaque étape est entâchée d'une "erreur d'arrondi", c'est-à-dire qu'on remplace (11) par

$$x_{n+1} = B^{-1}\,(C\,x_n + b) + \varepsilon_n\,, \quad n \in \mathbb{N}\,, \quad x_0 \in \mathbb{C}^N\,. \tag{14}$$

où, pour tout entier n, ε_n est un "petit" vecteur. Comme cela est expliqué dans [All05, page 429], la question est de savoir si l'accumulation de ces erreurs d'arrondi n'entâche pas trop fortement le résultat final du calcul. Le Théorème 3 permet de modifier légèrement le résultat de [All05, Lemme 13.1.28] en ne considérant que la norme hermitienne et donc, là encore, de fournir un critère d'arrêt *a priori*. Plus précisément, on obtient le résultat suivant.

Lemme 5. *Soit $A \in \mathrm{GL}_N(\mathbb{C})$ une matrice admettant une décomposition $A = B - C$, avec $B \in \mathrm{GL}_N(\mathbb{C})$ et $B^{-1}\,C$ de rayon spectral strictement inférieur à 1. Alors l'itération (14) vérifie*

$$\forall\, n \in \mathbb{N}\,, \quad |x_n - x_\infty| \leq 2\,N\,\mathbf{C}\,\left(\frac{(1-1/(2\,\mathbf{C}))^{n+1}}{n+1}\,|x_0 - x_\infty| + \sup_{0\leq k\leq n-1} |\varepsilon_k|\right)\,,$$

où l'on a noté x_∞ l'unique solution du système linéaire $A\,x = b$ et $\mathbf{C}$ désigne la quantité $\sup_{z\in\mathbb{S}^1} \|(z\,I - B^{-1}\,C)^{-1}\|$.

8. Pour une matrice M normale, c'est-à-dire diagonalisable dans une base orthonormée de $\mathbb{C}^N$, on a $\sup_{z\in\mathbb{S}^1} \|(z\,I - M)^{-1}\| = 1/(1-\rho(M))$ et tout se ramène à l'estimation du rayon spectral de M.

Démonstration du Lemme 5. On reprend, presque mot pour mot, la démonstration du Lemme 13.1.28 de [All05]. L'itération (14) conduit à la relation

$$x_n - x_\infty = (B^{-1}\,C)^n\,(x_0 - x_\infty) + \sum_{k=0}^{n-1} (B^{-1}\,C)^k\,\varepsilon_{n-1-k}\,,$$

et l'inégalité trinagulaire donne

$$|x_n - x_\infty| \leq \|(B^{-1}\,C)^n\|\,|x_0 - x_\infty| + \left(\sum_{k=0}^{+\infty} \|(B^{-1}\,C)^k\|\right) \sup_{0\leq k\leq n-1} |\varepsilon_k|\,.$$

On applique alors l'estimation (13) pour borner la norme $\|(B^{-1}\,C)^n\|$ et le Théorème 3 pour borner la somme des normes $\|(B^{-1}\,C)^k\|$. □

Concrètement, si on dispose d'un algorithme fonctionnant avec une précision ε donnée, alors l'itération (14) a lieu avec des erreurs ε_k à chaque étape qui vérifient $|\varepsilon_k| \leq \varepsilon$ pour tout k. Si on souhaite aboutir à une solution approchée du système linéaire (10) avec une précision α donnée ($\alpha \gg \varepsilon$), on déduit du Lemme 5 que l'accumulation des erreurs numériques peut devenir un facteur limitant dès que la quantité $2\,N\,\mathbf{C}\,\varepsilon$ devient supérieur à α.

Voyons ce que cela donne sur le sempiternel exemple issu de la discrétisation du Laplacien avec conditions aux limites de Dirichlet sur l'intervalle $[0, 1]$. On s'intéresse dans ce cas à la matrice symétrique définie positive

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

La méthode itérative de Gauss-Seidel revient à décomposer A sous la forme $A = B - C$ avec

$$B = \begin{pmatrix} 2 & 0 & \cdots & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}. \tag{15}$$

Les valeurs propres de la matrice A sont les nombres (voir par exemple [Sch02, chapitre 11] pour le détail des calculs, qui sont vraiment élémentaires)

$$4\,\sin^2\left(\frac{j\,\pi}{2\,(N+1)}\right), \quad j = 1, \ldots, N,$$

et les valeurs propres de la matrice $B^{-1}\,C$ se calculent également aux moyens de résultats généraux sur les matrices tridiagonales (voir par exemple [Ser10, chapitre 12]). Ces valeurs propres sont, dans le cas où N est pair :

$$\mu_j := \cos^2\left(\frac{j\,\pi}{N+1}\right), \quad j = 1, \ldots, N/2,$$

avec multiplicité 1 chacune, et 0 avec multiplicité $N/2$. Le rayon spectral de $B^{-1}\,C$ est donc égal à $\mu_1 < 1$ et la méthode itérative (11) converge. Pour N grand, c'est-à-dire pour un grand nombre de points de discrétisation du Laplacien, on obtient

$$\frac{1}{1-\rho(B^{-1}\,C)} = \frac{(N+1)^2}{\pi^2} + O(1)\,.$$

Le noyau de $B^{-1}\,C$ coïncide avec celui de C, et donc est de dimension 1. Cela indique que $B^{-1}\,C$ est très loin d'être une matrice normale car elle possède un bloc de Jordan de taille $N/2$ (un très grand bloc donc). On peut donc craindre que l'estimation de la résolvante de $B^{-1}\,C$ sur le cercle unité diffère sensiblement de la quantité $(1-\rho(B^{-1}\,C))^{-1}$. Fort heureusement, il n'en est rien comme on le voit dans le résultat suivant.

Lemme 6. *Les matrices B et C définies par* (15) *vérifient*

$$\sup_{z\in\mathbb{S}^1} \|(z\,I - B^{-1}\,C)^{-1}\| \leq \frac{3}{2\,(1-\cos(\pi/(N+1)))} = \frac{3\,(N+1)^2}{\pi^2} + O(1)\,.$$

Démonstration du Lemme 6. Soit $z = \mathrm{e}^{i\,\theta} \in \mathbb{S}^1$ et soit $w := \mathrm{e}^{i\,\theta/2}$. On définit les deux matrices unitaires

$$Q(\theta) := \mathrm{diag}\left(w, w^2, \ldots, w^N\right), \quad Q(-\theta) := \mathrm{diag}\left(\overline{w}, \overline{w}^2, \ldots, \overline{w}^N\right).$$

Pour $Y \in \mathbb{C}^N$, on note X l'unique solution du système linéaire $(z\,I - B^{-1}\,C)\,X = Y$. Comme $B = 2\,I - C^T$, on voit que X satisfait

$$\left(z\,I - \frac{z}{2}\,C^T - \frac{1}{2}\,C\right)\,X = \left(I - \frac{1}{2}\,C^T\right)\,Y\,,$$

ou encore

$$Q(-\theta)\,\left(z\,I - \frac{z}{2}\,C^T - \frac{1}{2}\,C\right)\,Q(\theta)\,Q(-\theta)\,X = Q(-\theta)\,\left(I - \frac{1}{2}\,C^T\right)\,Y\,.$$

On calcule sans trop de peine

$$Q(-\theta)\,\left(z\,I - \frac{z}{2}\,C^T - \frac{1}{2}\,C\right)\,Q(\theta) = \frac{w}{2}\left(2\,(w-1)\,I + A\right),$$

et donc le vecteur X satisfait la relation

$$w\,Q(-\theta)\,X = \left(2\,(w-1)\,I + A\right)^{-1}\,Q(-\theta)\,(2\,I - C^T)\,Y\,.$$

En passant aux normes de part et d'autre de cette égalité, on obtient

$$|X| \leq \|\big(2\,(w-1)\,I + A\big)^{-1}\|\,\|2\,I - C^T\|\,|Y|\,. \tag{16}$$

Il est clair que C et C^T sont de norme 1, et donc $\|2\,I - C^T\| \leq 3$ par l'inégalité triangulaire. De plus, $2\,(w-1)\,I + A$ étant une matrice normale dont les valeurs propres sont

$$2\,\left(w - \cos\frac{j\,\pi}{N+1}\right)\,, \quad j = 1, \ldots, N,$$

on obtient

$$\|\big(2\,(w-1)\,I + A\big)^{-1}\| \leq \frac{1}{2\,(1-\cos(\pi/(N+1)))}\,.$$

Le résultat du Lemme 6 s'obtient en prenant dans (16) le supremum en Y puis en z. □

Pour l'itération de Gauss-Seidel de la matrice du Laplacien, on voit que le facteur $2\,N\,\mathbf{C}$ dans l'estimation du Lemme 5 se comporte grosso modo comme N^3. Si l'on considère une précision machine de l'ordre de $\varepsilon = 10^{-16}$ et si l'on désire atteindre une précision de l'ordre de 10^{-7}, il n'est pas clair qu'on puisse raffiner au-delà de 10^3 points sans que l'accumulation des erreurs d'arrondi entâche le calcul d'une trop grande imprécision. Néanmoins, la matrice $B^{-1}\,C$ n'est certes pas normale, mais on a vu au Lemme 6 que son défaut de normalité est indolore par rapport à l'estimation de sa résolvante. On peut donc légitimement espérer que le facteur N dans l'estimation du Lemme 5 soit superflu et que la somme des normes

$$\sum_{k\geq 0} \|(B^{-1}\,C)^k\|$$

soit exactement d'ordre N^2 quand N devient grand. Cela indiquerait qu'on pourrait prendre de l'ordre de 10^5 points de discrétisation et obtenir une précision d'ordre 10^{-6} dans la résolution du système (10) malgré l'accumulation des erreurs d'arrondi.

Le lecteur trouvera dans [TE05, chapitres 24 et 25] de plus amples informations sur l'importance de la non-normalité sur la convergence des méthodes itératives de résolution de grands systèmes linéaires. La référence [Hig02, chapitre 17] propose des inégalités semblables à celles discutées ici pour la prise en compte des erreurs d'arrondi. Passons maintenant à une application du Théorème 3 impliquant des matrices de taille fixe.

4 Problèmes aux limites hyperboliques discrets

4.1 Une brève introduction

Les systèmes d'équations aux dérivées partielles hyperboliques interviennent dans la modélisation de nombreux phénomènes physiques : la mécanique des fluides, l'électromagnétisme, l'élastodynamique etc. On renvoie par exemple à [Daf10] pour une présentation historique et de nombreuses références. On s'intéresse ici à un problème très simple qui

consiste en un système linéaire posé sur une demi-droite et qui ne nécessite pas de conditions au bord. Concrètement, on s'intéresse au problème

$$\begin{cases} \dfrac{\partial u}{\partial t} + A\,\dfrac{\partial u}{\partial x} = F(t,x)\,, & (t,x) \in \mathbb{R}^+ \times \mathbb{R}^+\,, \\ u(0,x) = f(x)\,, & x \in \mathbb{R}^+\,, \end{cases} \tag{17}$$

où $A \in \mathscr{M}_N(\mathbb{R})$ est une matrice diagonalisable dont toutes les valeurs propres sont strictement négatives [9]. On note $\lambda_1, \dots, \lambda_N$ les valeurs propres de A et $e_1, \dots, e_N$ une famille de vecteurs propres engendrant $\mathbb{R}^N$. On décompose les termes sources et la solution sur la base des e_i :

$$F(t,x) = \sum_{i=1}^{N} F_i(t,x)\,e_i\,, \quad f(x) = \sum_{i=1}^{N} f_i(x)\,e_i\,, \quad u(t,x) = \sum_{i=1}^{N} u_i(t,x)\,e_i\,,$$

ce qui permet d'écrire le système (17) sous la forme équivalente

$$\forall\, i = 1, \dots, N\,, \quad \begin{cases} \dfrac{\partial u_i}{\partial t} + \lambda_i\,\dfrac{\partial u_i}{\partial x} = F_i(t,x)\,, & (t,x) \in \mathbb{R}^+ \times \mathbb{R}^+\,, \\ u_i(0,x) = f_i(x)\,, & x \in \mathbb{R}^+\,. \end{cases}$$

La méthode dite des caractéristiques permet de résoudre chacune de ces équations en remarquant qu'une solution u_i doit nécessairement vérifier la relation

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[u_i(t, x + \lambda_i\,t)\big] = F_i(t, x + \lambda_i\,t)\,.$$

En intégrant, on obtient une formule explicite pour chacune des composantes u_i de la solution u :

$$\forall\,(t,x) \in \mathbb{R}^+ \times \mathbb{R}^+\,, \quad u(t,x) = \sum_{i=1}^{N} \left(f_i(x - \lambda_i\,t) + \int_0^t F_i(s, x - \lambda_i\,(t-s))\,\mathrm{d}s \right) e_i\,. \tag{18}$$

Remarquons que cette formule définit bien une solution car $\lambda_i < 0$ et donc on évalue bien les fonctions f_i, F_i en des points de leur domaine de définition. (Ce ne serait plus vrai dans le cas $\lambda_i > 0$ et il faudrait prescrire la trace de u_i en $x = 0$ pour pouvoir déterminer u_i dans tout le quart d'espace $\mathbb{R}^+ \times \mathbb{R}^+$.)

Le but que l'on se fixe maintenant est d'approcher, au moyen d'un schéma numérique, la solution de (17). Cela peut sembler inutile puisque l'on dispose de la formule (18) pour la solution de (17) mais il faut comprendre qu'une telle formule n'existe plus dès qu'on passe à des problèmes en dimension supérieure. Comprendre les schémas numériques en dimension un d'espace est donc une étape cruciale en vue du calcul de solutions inaccessibles d'un point de vue analytique.

Soient donc $\Delta x > 0$, respectivement $\Delta t > 0$, un pas d'espace, respectivement de temps, c'est-à-dire deux petits paramètres destinés *in fine* à tendre vers 0. On suppose

9. La diagonalisabilité de A est une condition nécessaire et suffisante (dite d'hyperbolicité) pour que le problème de Cauchy sur la droite $\mathbb{R}$ soit bien-posé dans $L^2(\mathbb{R})$.

que ces paramètres sont choisis de telle sorte que leur rapport $\lambda := \Delta t/\Delta x$ est constant. (Dans le jargon hyperbolicien, ce nombre est appelé paramètre de Courant, Friedrichs et Lewy.) On va approcher la solution u de (17) par une fonction en escaliers U constante sur chaque maille $[n\,\Delta t, (n+1)\,\Delta t[\,\times[j\,\Delta x, (j+1)\,\Delta x[$, $n, j \in \mathbb{N}$:

$$U(t,x) := U_j^n \quad \text{pour } (t,x) \in [n\,\Delta t, (n+1)\,\Delta t[\,\times[j\,\Delta x, (j+1)\,\Delta x[\,.$$

Supposons pour simplifier que la donnée initiale f dans (17) est nulle, de sorte qu'il est légitime d'initialiser le calcul des (U_j^n) en posant

$$\forall\, j \in \mathbb{N}\,, \quad U_j^0 := 0\,.$$

On cherche maintenant une formule de récurrence qui fait passer de la suite $(U_j^n)_{j\in\mathbb{N}}$ à la suite $(U_j^{n+1})_{j\in\mathbb{N}}$. Il existe de très nombreux moyens, tous pertinents par rapport à l'équation (17), pour calculer une telle approximation. On ne détaillera pas ici tel schéma ou tel autre, mais on adoptera plutôt un point de vue général en supposant que le passage de la suite $(U_j^n)_{j\in\mathbb{N}}$ à la suite $(U_j^{n+1})_{j\in\mathbb{N}}$ se fait au moyen d'une formule du type

$$U_j^{n+1} = \sum_{\ell=-r}^{p} A_\ell\, U_{j+\ell}^n + \Delta t\, F(n\,\Delta t, j\,\Delta x)\,, \tag{19}$$

où les entiers r, p et les matrices $A_{-r}, \ldots, A_p$ sont fixes. Ces dernières sont autorisées à dépendre du paramètre λ mais pas du pas de temps Δt (qu'on garde comme petit paramètre libre, Δx étant calculé par la relation $\Delta x = \Delta t/\lambda$). Les deux exemples les plus simples de tels schémas sont

$$\begin{aligned}
&\text{Le schéma de Lax-Friedrichs : } p = r = 1\,, \quad A_{\pm 1} := \frac{1}{2}\,(I \mp \lambda\, A)\,,\ A_0 := 0\,,\\
&\text{Le schéma décentré amont : } r = 0, p = 1\,, \quad A_0 := I + \lambda\, A\,,\ A_1 := -\lambda\, A\,,
\end{aligned}$$

le deuxième n'étant pertinent que lorsque les valeurs propres de A sont négatives, ce qui est le cas envisagé ici.

La formule de récurrence (19) n'a de sens que pour $j \geq r$ car la suite (U_j^n) est indéxée par $j \in \mathbb{N}$. Il faut donc prescrire autrement les valeurs de $U_0^{n+1}, \ldots, U_{r-1}^{n+1}$, et comme l'équation aux dérivées partielles qu'on cherche à approcher ne fournit pas aisément un moyen de prescrire la trace de u, on peut se contenter, dans un premier temps, de ne rien faire et imposer

$$U_0^{n+1} = \cdots = U_{r-1}^{n+1} := 0\,. \tag{20}$$

Partant de la condition initiale nulle ($U_j^0 = 0$ pour tout j), la question est maintenant de savoir si les formules (19), (20) conduisent à une bonne approximation U de la solution u de (17) lorsque le pas de temps Δt tend vers 0.

4.2 Le cas des schémas décentrés amont

On s'intéresse spécifiquement ici au cas $r = 0$, c'est-à-dire que la condition au bord numérique (20) disparait. On dit que le schéma est décentré "amont" car on va chercher l'information en "remontant" les caractéristiques (qui ici vont de la droite vers la

gauche donc on décentre vers la droite). Pour savoir si le schéma (19) fournit une bonne approximation de la solution u, on commence par chercher à vérifier que le schéma est stable ou, autrement dit, qu'il n'amplifie pas les éventuelles petites erreurs d'approximation des termes source. Dans le contexte des problèmes aux limites hyperboliques discrets, différentes notions de stabilité ont été proposées dans [Kre68] puis [GKS72]. On reprend ici la Définition 3.2 de [GKS72] :

Définition 1 (Stabilité [GKS72])**.** *Le schéma*

$$\begin{cases} U_j^{n+1} = \sum_{\ell=0}^{p} A_\ell \, U_{j+\ell}^n + \Delta t \, F_j^n \,, & j, n \in \mathbb{N} \,, \\ U_j^0 = 0 \,, & j \in \mathbb{N} \,, \end{cases} \tag{21}$$

est dit stable s'il existe une constante C_1 telle que pour tout $\gamma > 0$ et pour tout $\Delta t \in]0, 1]$, la solution (U_j^n) de (21) vérifie :

$$\frac{\gamma}{\gamma \, \Delta t + 1} \sum_{j,n \in \mathbb{N}} \Delta t \, \Delta x \, \mathrm{e}^{-2 \, \gamma \, n \, \Delta t} \, |U_j^n|^2 \leq C_1 \, \frac{\gamma \, \Delta t + 1}{\gamma} \sum_{j,n \in \mathbb{N}} \Delta t \, \Delta x \, \mathrm{e}^{-2 \, \gamma \, (n+1) \, \Delta t} \, |F_j^n|^2 \,. \tag{22}$$

Remarquons que le schéma (21) ne fait pas intervenir de conditions aux limites numériques, et on ne demande donc pas dans l'inégalité (22) de contrôler la trace de la suite (U_j^n) : on estime la solution dans le même espace que les données, ici la suite (F_j^n). La Définition 3.3 de [GKS72] traite le cas $r > 0$ avec des conditions aux limites inhomogènes, et requiert le contrôle de la trace (il s'agit donc d'une notion plus forte que celle considérée ici).

On reconnait dans (22) des normes L^2 en (t, x) de fonctions en escaliers, et il n'est donc pas étonnant qu'on puisse, via le théorème de Plancherel et une utilisation appropriée du théorème de Paley-Wiener, ramener l'étude de la stabilité de (21) à une propriété de résolubilité de l'équation résolvante associée :

$$W_j - \frac{1}{z} \sum_{\ell=0}^{p} A_\ell \, W_{j+\ell} = F_j \,, \quad j \geq 0 \,. \tag{23}$$

Plus précisément, on a le résultat suivant :

Théorème 4 (Gustafsson, Kreiss et Sundström [GKS72])**.** *Le schéma (21) est stable si et seulement s'il existe une constante $C_2 > 0$ telle que pour tout $z \in \mathscr{U}$ et pour tout $(F_j) \in \ell^2(\mathbb{N}; \mathbb{C}^N)$, il existe une unique solution $(W_j) \in \ell^2(\mathbb{N}; \mathbb{C}^N)$ de (23) et cette solution satisfait l'estimation :*

$$\frac{|z| - 1}{|z|} \sum_{j \geq 0} |W_j|^2 \leq C_2 \, \frac{|z|}{|z| - 1} \sum_{j \geq 0} |F_j|^2 \,.$$

Le lecteur intéressé trouvera dans [Cou13] une démonstration détaillée de ce résultat, y compris dans des situations plus générales que le cas traité ici. On se propose désormais de démontrer, au moyen des résultats de la Section 2, le résultat suivant :

Théorème 5. *Supposons que les matrices $A_0, \ldots, A_p$ vérifient :*

– *Il existe une constante* $C > 0$ *telle que pour tout* $\kappa \in \mathbb{S}^1$ *et pour tout* $n \in \mathbb{N}$*, on a*

$$\left\| \left(\sum_{\ell=0}^{p} \kappa^\ell \, A_\ell \right)^n \right\| \leq C \, .$$

– *Le rayon spectral de* A_0 *est strictement plus petit que* 1.
Alors le schéma (21) *est stable au sens de la définition 1.*

La première hypothèse du Théorème 5 assure la stabilité du schéma numérique (21) sur toute la droite réelle, c'est-à-dire si l'on remplace $j \in \mathbb{N}$ par $j \in \mathbb{Z}$. Le Théorème 1 assure que cette condition est équivalente à supposer une borne sur la résolvante de la matrice [10]

$$\sum_{\ell=0}^{p} \kappa^\ell \, A_\ell \, .$$

On utilisera d'ailleurs le Théorème 1 dans la démonstration du Lemme 8 ci-dessous. La seconde hypothèse peut sembler un peu plus technique. Elle revient à dire que le bord numérique est "non-caractéristique" pour l'équation résolvante.

Démonstration du Théorème 5. La seconde hypothèse du Théorème 5 va nous permettre de réécrire l'équation résolvante (23) sous la forme d'une récurrence du premier ordre. Plus précisément, on définit les matrices

$$\begin{aligned} \mathbb{A}_0(z) &:= I - \frac{1}{z} \, A_0 \, , \\ \forall \, \ell = 1, \dots, p \, , \quad \mathbb{A}_\ell(z) &:= -\frac{1}{z} \, A_\ell \, . \end{aligned}$$

Pour $z \in \mathscr{U}$, la matrice $\mathbb{A}_0(z)$ est inversible et on peut donc écrire (23) sous la forme équivalente

$$\forall \, j \in \mathbb{N} \, , \quad \begin{pmatrix} W_j \\ \vdots \\ W_{j+p-1} \end{pmatrix} = \mathbb{M}(z) \begin{pmatrix} W_{j+1} \\ \vdots \\ W_{j+p} \end{pmatrix} + \mathscr{F}_j(z) \, , \tag{24}$$

où l'on a introduit les notations

$$\forall \, z \in \mathscr{U} \, , \quad \mathbb{M}(z) := \begin{pmatrix} -\mathbb{A}_0(z)^{-1} \, \mathbb{A}_1(z) & \dots & \dots & -\mathbb{A}_0(z)^{-1} \, \mathbb{A}_p(z) \\ I & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix} , \tag{25}$$

$$\mathscr{F}_j(z) := \begin{pmatrix} \mathbb{A}_0(z)^{-1} \, F_j \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

On commence par établir une propriété fondamentale de la matrice $\mathbb{M}(z)$:

10. Que cette matrice ait un rayon spectral plus petit ou égal à 1 est une condition nécessaire, et parfois suffisante, de stabilité connue sous le nom de condition de von Neumann.

Lemme 7 (Kreiss [Kre68])**.** *Sous les hypothèses du Théorème 5, la matrice* $\mathbb{M}(z)$ *est de rayon spectral strictement plus petit que* 1 *pour tout* $z \in \mathscr{U}$ *.*

Démonstration du Lemme 7. Par définition des matrices $\mathbb{A}_0(z), \dots, \mathbb{A}_p(z)$, on voit que

$$\lim_{z\to\infty} \mathbb{A}_0(z) = I\,, \quad \lim_{z\to\infty} \mathbb{A}_\ell(z) = 0 \quad \text{si } \ell = 1, \dots, p\,.$$

Ainsi $\mathbb{M}(z)$ tend vers la matrice nilpotente

$$\begin{pmatrix} 0 & \dots & \dots & 0 \\ I & \ddots & & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix}$$

quand z tend vers l'infini. Par continuité de $\mathbb{M}$ et connexité de l'ensemble $\mathscr{U}$, il suffit donc de montrer que pour tout $z \in \mathscr{U}$, $\mathbb{M}(z)$ n'a aucune valeur propre sur $\mathbb{S}^1$. Raisonnons par l'absurde et supposons qu'il existe $z \in \mathscr{U}$ et $\kappa \in \mathbb{S}^1$ valeur propre de $\mathbb{M}(z)$. On se donne un vecteur propre $X \in \mathbb{C}^{N\,p}$, que l'on décompose sous la forme

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\,, \quad X_1, \dots, X_p \in \mathbb{C}^N\,.$$

Ces vecteurs vérifient les relations

$$\forall\, \ell = 1, \dots, p\,, \quad X_\ell = \kappa^{p-\ell}\, X_p\,, \quad \left(\sum_{\ell=0}^{p} \kappa^{p-\ell}\, \mathbb{A}_\ell(z) \right) X_p = 0\,.$$

Le vecteur X_p est nécessairement non-nul et on a

$$\left(\sum_{\ell=0}^{p} \overline{\kappa}^{\ell}\, A_\ell \right) X_p = z\, X_p\,,$$

ce qui est manifestement en contradiction avec la première hypothèse du Théorème 5 car z serait une valeur propre de module strictement plus grand que 1 de la matrice $\sum_{j=0}^{p} \overline{\kappa}^{\ell}\, A_\ell$. On a donc bien démontré le Lemme 7. □

Comme $\mathbb{M}(z)$ est de rayon spectral strictement plus petit que 1, la récurrence (24) possède une unique solution dans ℓ^2 et cette solution est donnée par la formule :

$$\forall\, j \in \mathbb{N}\,, \quad \begin{pmatrix} W_j \\ \vdots \\ W_{j+p-1} \end{pmatrix} = \sum_{\ell=0}^{+\infty} \mathbb{M}(z)^\ell\, \mathscr{F}_{j+\ell}(z)\,. \tag{26}$$

En effet, les normes $\|\mathbb{M}(z)^k\|$ décroissent géométriquement donc la formule ci-dessus définit bien une suite, et cette suite satisfait (24). L'appartenance à l'espace ℓ^2 est un cas particulier de l'inégalité de Young pour le produit de convolution. Plus précisément, on a

$$\begin{aligned}\sum_{j\geq 0}|W_j|^2 &\leq \sum_{j\geq 0}|W_j|^2+\cdots+|W_{j+p-1}|^2\\ &\leq \sum_{j\geq 0}\left(\sum_{\ell=0}^{+\infty}\|\mathbb{M}(z)^\ell\|\,|\mathscr{F}_{j+\ell}(z)|\right)^2\\ &\leq \sum_{j\geq 0}\left(\sum_{\ell=0}^{+\infty}\|\mathbb{M}(z)^\ell\|\right)\left(\sum_{\ell=0}^{+\infty}\|\mathbb{M}(z)^\ell\|\,|\mathscr{F}_{j+\ell}(z)|^2\right)\\ &\leq \left(\sum_{\ell=0}^{+\infty}\|\mathbb{M}(z)^\ell\|\right)^2\sum_{j\geq 0}|\mathscr{F}_j(z)|^2\,, \end{aligned}\tag{27}$$

où on a utilisé l'inégalité de Cauchy-Schwarz dans ℓ^2 puis le Théorème de Fubini.

Comme A_0 n'a pas de valeur propre dans $\overline{\mathscr{U}}$, on peut majorer uniformément la norme de $\mathbb{A}_0(z)^{-1}$ sur $\mathscr{U}$ (et même sur $\overline{\mathscr{U}}$). La définition (25) du vecteur $\mathscr{F}_j(z)$ permet de simplifier (27) en

$$\sum_{j\geq 0}|W_j|^2 \leq C\left(\sum_{\ell=0}^{+\infty}\|\mathbb{M}(z)^\ell\|\right)^2\sum_{j\geq 0}|F_j|^2\,,$$

où la constante C est indépendante de z. Pour appliquer le Théorème 4 et conclure la démonstration du Théorème 5, on voit qu'il suffit d'obtenir une estimation de la somme des normes $\|\mathbb{M}(z)^\ell\|$. C'est ici que le Théorème 3 rentre en jeu car on peut sans difficulté majeure estimer la résolvante de $\mathbb{M}(z)$ sur le cercle unité.

Lemme 8 (Gustafsson, Kreiss et Sundström [GKS72])**.** *Sous les hypothèses du Théorème 5, il existe une constante $C>0$ telle que pour tout $z\in\mathscr{U}$, la matrice $\mathbb{M}(z)$ vérifie*

$$\sup_{\kappa\in\mathbb{S}^1}\|(\mathbb{M}(z)-\kappa\,I)^{-1}\|\leq C\,\frac{|z|}{|z|-1}\,.$$

Comme expliqué ci-dessus, les résultats combinés du Lemme 8 et du Théorème 3 montrent que le schéma (21) est stable au sens de la définition 1. On s'attache donc pour finir à démontrer le Lemme 8.

Démonstration du Lemme 8. La première hypothèse du Théorème 5 indique que la famille de matrices $\{\sum_{j=0}^p\kappa^\ell\,A_\ell\,,\ \kappa\in\mathbb{S}^1\}$ est de puissances uniformément bornées. Le Théorème 1, dans son sens facile, implique qu'il existe une constante $C>0$, indépendante de $z\in\mathscr{U}$ et $\kappa\in\mathbb{S}^1$, telle que

$$\left\|\left(z\,I-\sum_{\ell=0}^{p}\kappa^\ell\,A_\ell\right)^{-1}\right\|\leq\frac{C}{|z|-1}\,.\tag{28}$$

Soient maintenant $z\in\mathscr{U}$, $\kappa\in\mathbb{S}^1$ et $b\in\mathbb{C}^{N\,p}$. On note $x\in\mathbb{C}^{N\,p}$ l'unique solution du système linéaire $(\mathbb{M}(z)-\kappa\,I)\,x=b$. Avec des notations évidentes pour la décomposition

par blocs des vecteurs de $C^{N\,p}$, on obtient les formules

$$\forall\, \ell = 1, \dots, p-1, \quad x_\ell = \kappa^{p-\ell}\, x_p + \sum_{j=\ell+1}^{p} \kappa^{j-\ell-1}\, b_j\,, \tag{29}$$

$$\left(I - \frac{\kappa^p}{z} \sum_{\ell=0}^{p} \overline{\kappa}^{\ell}\, A_\ell \right) x_p = -\widetilde{b}(\kappa, z)\,,$$

où le vecteur $\widetilde{b}(\kappa, z)$ est défini par

$$\widetilde{b}(\kappa, z) := \sum_{\ell=1}^{p-1} \mathbb{A}_\ell(z) \sum_{j=\ell+1}^{p} \kappa^{j-\ell-1}\, b_{\ell-j} + \mathbb{A}_0(z) \sum_{j=1}^{p} \kappa^{j-1}\, b_j\,.$$

Pour $z \in \mathscr{U}$ et $\kappa \in \mathbb{S}^1$, on a une estimation uniforme

$$|\widetilde{b}(\kappa, z)| \leq C_0\, |b|\,,$$

car les matrices $\mathbb{A}_\ell(z)$ sont bornées sur $\mathscr{U}$. L'estimation (28) fournit donc la borne

$$|x_p| \leq C\, \frac{|z|}{|z|-1}\, |b|\,,$$

et les autres composantes du vecteur x s'estiment ensuite par les relations (29). □

□

Un principe de base, donc très souvent vérifié, de l'analyse numérique est que la convergence d'une méthode est basée sur sa stabilité et sa consistance. Pour les problèmes aux limites hyperboliques discrets, la stabilité est toujours le point le plus difficile à démontrer. La convergence s'obtient en suivant les arguments de Gustafsson [Gus75] qu'on ne détaillera pas ici.

4.3 Pour aller plus loin

Les systèmes d'équations aux dérivées partielles hyperboliques pour lesquels les valeurs propres de A sont toutes négatives sont rares. Dans la plupart des problèmes issus de la physique, certaines valeurs propres de A sont positives et il faut adjoindre au système (17) un nombre adéquat de conditions au bord. Pour une dimension d'espace quelconque, la caractérisation des conditions aux limites conduisant à un problème aux limites

$$\begin{cases} \dfrac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j\, \dfrac{\partial u}{\partial x_j} = F(t,x)\,, & (t, x_1, \dots, x_{d-1}, x_d) \in \mathbb{R}^+ \times \mathbb{R}^{d-1} \times \mathbb{R}^+\,, \\ B\, u(t, x_1, \dots, x_{d-1}, 0) = g(t, x_1, \dots, x_{d-1})\,, & (t, x_1, \dots, x_{d-1}) \in \mathbb{R}^+ \times \mathbb{R}^{d-1}\,, \\ u(0,x) = f(x)\,, & x \in \mathbb{R}^{d-1} \times \mathbb{R}^+\,, \end{cases}$$

bien-posé - dans un cadre fonctionnel convenable - est (encore !) un résultat de Kreiss [Kre70], voir également Sakamoto [Sak70a, Sak70b]. Les résultats de Kreiss sont exposés de manière détaillée dans [BGS07, chapitre 4]. La démonstration des résultats analogues pour les équations discrétisées reste un problème encore largement ouvert. Même si le cas des problèmes unidimensionnels est globalement bien compris [11], les problèmes mul-

11. On renvoie par exemple à [Cou13] pour un état de l'art que l'auteur espère le plus complet possible.

tidimensionnels n'ont été abordés que par Michelson [Mic83] sous des hypothèses assez contraignantes. Il reste donc encore beaucoup à faire.

Références

[All05] G. Allaire. *Analyse numérique et optimisation*. Editions de l'Ecole Polytechnique, 2005.

[BD87] S. Boyd and J. Doyle. Comparison of peak and RMS gains for discrete-time systems. *Systems Control Lett.*, 9(1) :1–6, 1987.

[BE96] P. Borwein and T. Erdélyi. Sharp extensions of Bernstein's inequality to rational spaces. *Mathematika*, 43(2) :413–423 (1997), 1996.

[BGS07] S. Benzoni-Gavage and D. Serre. *Multidimensional hyperbolic partial differential equations*. Oxford Mathematical Monographs. Oxford University Press, 2007.

[Cia82] P. G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. Masson, 1982.

[Con90] J. B. Conway. *A course in functional analysis*. Graduate Texts in Mathematics. Springer-Verlag, 1990.

[Cou13] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. In *HCDTE Lecture Notes. Part I. Nonlinear Hyperbolic PDEs, Dispersive and Transport Equations*, pages 97–225. American Institute of Mathematical Sciences (AIMS), 2013.

[Daf10] C. M. Dafermos. *Hyperbolic conservation laws in continuum physics*, volume 325 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 2010.

[GKO95] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. Pure and Applied Mathematics. John Wiley & Sons Inc., 1995.

[GKS72] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26(119) :649–686, 1972.

[Gus75] B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.*, 29(130) :396–406, 1975.

[Hig02] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics (SIAM), second edition, 2002.

[HS10] B. Helffer and J. Sjöstrand. From resolvent bounds to semigroup bounds. Disponible sur `http://arxiv.org/abs/1001.4171`, 2010.

[Kre62] H.-O. Kreiss. Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren. *Nordisk Tidskr. Informations-Behandling*, 2 :153–181, 1962.

[Kre68] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22 :703–714, 1968.

[Kre70] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, 23 :277–298, 1970.

[LN91] C. Lubich and O. Nevanlinna. On resolvent conditions and stability estimates. *BIT*, 31(2) :293–313, 1991.

[LT84] R. J. LeVeque and L. N. Trefethen. On the resolvent condition in the Kreiss matrix theorem. *BIT*, 24(4) :584–591, 1984.

[Mic83] D. Michelson. Stability theory of difference approximations for multidimensional initial-boundary value problems. *Math. Comp.*, 40(161) :1–45, 1983.

[Nik02] N. K. Nikolski. *Operators, functions, and systems : an easy reading. Vol. 1.* Mathematical Surveys and Monographs. American Mathematical Society, 2002.

[Paz83] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Applied Mathematical Sciences. Springer-Verlag, 1983.

[RM94] R. D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*. Robert E. Krieger Publishing Co. Inc., 1994.

[Sak70a] R. Sakamoto. Mixed problems for hyperbolic equations. I. Energy inequalities. *J. Math. Kyoto Univ.*, 10 :349–373, 1970.

[Sak70b] R. Sakamoto. Mixed problems for hyperbolic equations. II. Existence theorems with zero initial datas and energy inequalities with initial datas. *J. Math. Kyoto Univ.*, 10 :403–417, 1970.

[Sch02] M. Schatzman. *Numerical analysis : a mathematical introduction*. Oxford University Press, 2002.

[Ser10] D. Serre. *Matrices*. Graduate Texts in Mathematics. Springer, 2010.

[Spi91] M. N. Spijker. On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem. *BIT*, 31(3) :551–555, 1991.

[SW97] J. C. Strikwerda and B. A. Wade. A survey of the Kreiss matrix theorem for power bounded families of matrices and its extensions. In *Linear operators (Warsaw, 1994)*, pages 339–360. Polish Acad. Sci., 1997.

[Tad81] E. Tadmor. The equivalence of L_2-stability, the resolvent condition, and strict H-stability. *Linear Algebra Appl.*, 41 :151–159, 1981.

[TE05] L. N. Trefethen and M. Embree. *Spectra and pseudospectra*. Princeton University Press, 2005. The behavior of nonnormal matrices and operators.