MA261 2005-2006

Introduction au calcul scientifique. Aspects algorithmiques

Patrick Ciarlet

L'auteur de ces notes de cours s'est parfois librement inspiré de l'ouvrage intitulé "Optimisation et algèbre linéaire", rédigé conjointement par Patrick Ciarlet et Pascal Joly entre 1998 et 2002, et édité par l'ENSTA.
Les Figures 4.1, 4.2, 5.1, 5.2, A.1, A.2, A.3, A.4, A.5 sont notamment extraites de cet ouvrage.

Table des matières

Chapitre 1

Considérations générales et modèles

Le but de ce cours est d'étudier des processus et cheminements qui permettent de résoudre des problèmes issus de la physique, de la mécanique, de la finance, etc. Que comprend un tel processus ? De façon générale, il est possible de le découper en quatre étapes :

1. *Identifier* la ou les quantité(s) à calculer.
   Modéliser le phénomène de nature physique (ou autre) associé : équation(s), inéquation(s), contrainte(s), etc.

2. Choisir une *méthode d'approximation* (ou *méthode de discrétisation*) lorsqu'on ne peut pas résoudre le problème exactement... En pratique, ça se passe toujours comme ça ! De plus, on est en général déjà amené à réaliser des *simplifications* pour construire le modèle à l'étape précédente.

3. Construire le *problème discrétisé* (ou *problème approché*.)
   Evaluer la *qualité* de la solution approchée par rapport à une solution exacte : on parle aussi de *précision* de la méthode de discrétisation.

4. *Résoudre* le problème discrétisé.
   Vérifier/visualiser/interpréter les résultats.

Dans les notes de cours qui sont proposées ci-après, nous allons explorer successivement tous ces aspects, de façon plus ou moins approfondie. Plus précisément, nous allons, dans la suite de ce chapitre, proposer quelques problèmes modèles élémentaires, issus pour la plupart de la physique. Il s'agit d'une présentation très intuitive, qui ne prétend pas respecter les *canons mathématiques* ! Pour de plus amples détails, ou pour des justifications précises, nous renvoyons le lecteur intéressé à [2], ainsi qu'aux autres cours de mathématiques appliquées, notamment en troisième année. Ensuite, nous allons les discrétiser, à l'aide de la *méthode des différences finies*[1]. Une fois le problème discrétisé construit, nous répondrons également, dans une certaine mesure, à la question de la qualité de l'approximation de la solution approchée, par comparaison à la quantité de départ à calculer. Ceci sera l'objet du chapitre 2. Ensuite, nous nous attacherons à la résolution des problèmes discrétisés, à l'aide de techniques et algorithmes issus de l'algèbre linéaire, aux chapitres 3, 4 et 5. Les algorithmes seront étudiés avec soin, ce qui nécessitera en particulier un certain nombre de rappels concernant les espaces vectoriels normés (de dimension finie), présentés en Annexe A et B. Enfin, les aspects liés à la vérification des résultats ainsi qu'à leur interprétation, seront abordés au cours de séances de Travaux Pratiques.

[1] Notons qu'il existe bien d'autres méthodes de discrétisation, comme par exemple les méthodes de type éléments finis ou volumes finis, les méthodes intégrales, les ondelettes, etc.

1.1 Problèmes statiques élémentaires

Par problèmes statiques, nous entendons problèmes à l'équilibre, par opposition aux problèmes qui dépendent du temps, évoqués à la prochaine section.

Commençons par un modèle élémentaire, celui du *fil pesant*. Soit donc un fil de longueur unité, fixé en ses deux extrémités. Le but est ici de calculer des déplacements verticaux, dont on suppose qu'ils sont "petits", lorsqu'il est soumis à la gravité. On note $\rho : x \mapsto \rho(x)$ la densité linéique de masse, et $u : x \mapsto u(x)$ le déplacement transversal. D'après les équations de l'élasticité linéaire, on sait que u vérifie

$$-u''(x) = f(x) \text{ sur }]0,1[, \quad \text{avec } f(x) = \rho(x)g. \tag{1.1}$$

Comme les extrémités du fil sont fixées, il est clair que le déplacement vertical est nul en celles-ci, ce que l'on exprime sous la forme

$$u(0) = u(1) = 0. \tag{1.2}$$

Enfin, le fait que l'énergie élastique de déformation soit bornée peut-être exprimé sous la forme

$$\int_0^1 u'(x)^2 \, dx < \infty. \tag{1.3}$$

On parle ici de **modèle monodimensionnel** ou de **modèle 1D**, puisque les équations (1.1)-(1.2) et l'inéquation (1.3) ne dépendent que de la variable x. Le **domaine de calcul** est dans ce cas l'intervalle *ouvert* $]0,1[$, et on rappelle que l'ensemble $\{0,1\}$ constitue sa **frontière**.

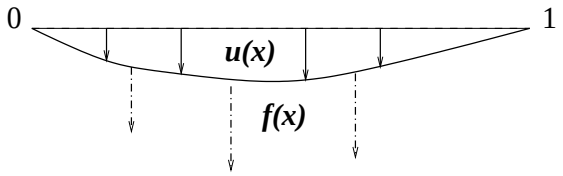


FIG. 1.1 – Modèles 1D : fil ou poutre

On peut introduire un second modèle 1D : celui de la *poutre*, appuyée en ses extrémités. Le but est encore une fois de déterminer des "petits" déplacements verticaux, lorsqu'elle est soumise à une charge transversale, égale à $f(x)\delta x$, où f est la densité de force par unité de longueur. D'après les équations de l'élasticité linéaire, on retrouve que le modèle est constitué par le même ensemble d'équations et d'inéquations (1.1)-(1.3) que précédemment.

Remarque 1.1.1 *L'équation (1.1) peut être écrite de façon équivalente*

$$-\frac{d^2u}{dx^2}(x) = f(x) \text{ pour } x \in]0,1[.$$

On peut également concevoir des modèles, provenant de problèmes statiques posés en dimension supérieure.

Commençons par celui de la *membrane élastique* à l'équilibre, dont la frontière est fixée. Encore une fois, on suppose que l'on veut déterminer des "petits" déplacements verticaux, lorsque cette membrane est soumise à une force verticale. Soit D le domaine *ouvert* occupé par la membrane au repos, que l'on suppose être inclus dans le plan Oxy (la membrane est donc horizontale au repos.) Cette fois, le déplacement vertical, toujours noté u, dépend de deux

variables, à savoir x et y, pour (x, y) parcourant le domaine de calcul D. Les forces agissant sur la membrane sont de la forme $\tau f(x, y)\delta S$, avec f la densité de force par unité de surface, et τ la tension de la membrane. D'après les équations de l'élasticité linéaire, u vérifie

$$-c(\tau)^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)(x, y) = f(x, y) \text{ pour } (x, y) \in D. \tag{1.4}$$

Ci-dessus, $c(\tau)$ est un nombre strictement positif, dépendant de τ.
Comme la frontière du domaine, ∂D, est fixée, on écrit

$$u(x, y) = 0, \text{ pour } (x, y) \in \partial D. \tag{1.5}$$

Enfin, le fait que l'énergie élastique de déformation soit bornée (ainsi que l'inégalité dite de Korn) permettent d'écrire

$$\int_D c(\tau)^2 \left(\left[\frac{\partial u}{\partial x} \right]^2 + \left[\frac{\partial u}{\partial y} \right]^2 \right)(x, y)\, dxdy < \infty. \tag{1.6}$$

On parle ici de **modèle bidimensionnel** ou de **modèle 2D**, puisque les équations (1.4)-(1.5)

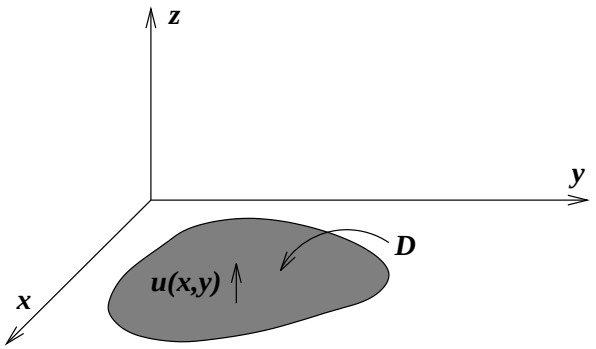


FIG. 1.2 – Modèle 2D : membrane élastique

et l'inéquation (1.6) dépendent du couple de variables (x, y).

Pour finir, rappelons le modèle associé à la *cavité électrostatique*, incluse dans $\mathbb{R}^3$. Le but est de déterminer le potentiel électrostatique autour d'un système de conducteurs parfaits disjoints, $C_1, \cdots, C_I$, eux-mêmes étant entourés d'un conducteur parfait C_0. Soit C le domaine de calcul *ouvert* inclus dans $\mathbb{R}^3$, constitué de la partie intérieure à ce dernier conducteur parfait, et privé de $\bar{C}_1, \cdots, \bar{C}_I$. Ici, C est la cavité électrostatique. On note $\partial C_1, \cdots, \partial C_I$ les frontières respectives de $C_1, \cdots, C_I$, ainsi que $\partial_{int} C_0$ la frontière interne de C_0 : par construction (cf. Figure 1.3), la frontière ∂C de notre cavité électrostatique est formée par l'union

$$\partial C = \partial C_1 \cup \cdots \cup \partial C_I \cup \partial_{int} C_0.$$

Si on considère que sur la frontière $\partial_{int} C_0$, on se trouve à l'équipotentielle nulle, le potentiel électrostatique est déterminé par la donnée de :

- ρ la densité de charge électrique par unité de volume ;
- V_k la valeur de l'équipotentielle sur la frontière ∂C_k, pour k variant de 1 à I.

On appelle u le potentiel électrostatique à calculer, c'est-à-dire $u(x, y, z)$, pour (x, y, z) parcourant C. D'après les équations de Maxwell, et plus précisément la relation de Coulomb, on sait que

$$-\epsilon_0 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)(x, y, z) = \rho(x, y, z) \text{ pour } (x, y, z) \in C. \tag{1.7}$$

Ci-dessus, ϵ_0 est la permittivité électrique du vide.
La frontière de chaque conducteur parfait étant équipotentielle, on écrit

$$u(x,y,z) = V_k,\ (x,y,z) \in \partial C_k,\ 1 \leq k \leq I,\ \text{et}\ V(x,y,z) = 0,\ (x,y,z) \in \partial_{int} C_0. \tag{1.8}$$

Enfin, le fait que l'énergie électrostatique soit bornée signifie que

$$\int_C \epsilon_0 \left(\left[\frac{\partial u}{\partial x}\right]^2 + \left[\frac{\partial u}{\partial y}\right]^2 + \left[\frac{\partial u}{\partial z}\right]^2 \right)(x,y,z)\, dxdydz < \infty. \tag{1.9}$$

On parle ici de **modèle tridimensionnel** ou de **modèle 3D**, puisque les équations (1.7)-(1.8) et l'inéquation (1.9) dépendent du triplet de variables (x,y,z).

FIG. 1.3 – Modèle 3D : cavité avec quatre conducteurs internes ($I = 4$)

Pour conclure cette section, nous proposons le récapitulatif ci-dessous, qui met en évidence un certain nombre de similitudes entre ces différents modèles statiques.

Remarque 1.1.2 *Bien évidemment, il existe beaucoup de modèles statiques, qui peuvent être complètement différents, par leur nature, de ceux présentés ici ! De fait, l'"unité" entre tous ces modèles permettra une modélisation* homogène*, comme on le verra au chapitre 2...*

Commençons par quelques définitions, utilisées fréquemment par la suite.

Définition 1.1.1 *On utilise le terme de* **condition aux limites** *pour qualifier les équations vérifiées par les inconnues du problème sur la frontière du domaine de calcul.*

Définition 1.1.2 *On appelle* **Laplacien dans** $\mathbb{R}^d$ $(d \geq 1)$ *l'opérateur scalaire défini par*

$$\Delta_d u = \sum_{k=1}^{k=d} \frac{\partial^2 u}{\partial {x_k}^2}.$$

Définition 1.1.3 *On appelle* **gradient dans** $\mathbb{R}^d$ $(d \geq 1)$ *l'opérateur vectoriel défini par*

$$\vec{\nabla}_d u = \sum_{k=1}^{k=d} \frac{\partial u}{\partial x_k} \vec{e}_k.$$

Pour récapituler :

- Problèmes posés dans des domaines (ouverts bornés) Ω_d de $\mathbb{R}^d$, $d = 1, 2, 3$, par rapport à la variable $\mathbf{x} = (x_1, \cdots, x_d)$:
  1. $d = 1$: fil, poutre ;
  2. $d = 2$: membrane ;
  3. $d = 3$: cavité.
- L'opérateur est identique, pour tous les problèmes. Il est composé de :
  (I) à l'intérieur, le Laplacien dans $\mathbb{R}^d$;
  (F) sur la frontière, une condition aux limites du type $u = 0$ (ou $u = constante$).
  De plus, même propriété pour l'intégrale du gradient, bornée pour tous les problèmes.
$$\int_{\Omega_d} \left| \vec{\nabla}_d u \right|^2 (\mathbf{x})\, d\mathbf{x} < \infty.$$
- Problèmes **linéaires** : si les problèmes avec les données f_1 et f_2 admettent pour solution u_1 et u_2, alors le problème avec pour donnée $\alpha_1 f_1 + \alpha_2 f_2$ admet pour solution $\alpha_1 u_1 + \alpha_2 u_2$.

1.2 Problèmes instationnaires élémentaires

Contrairement aux problèmes de la section précédente, nous nous intéressons ici à des modèles dépendant non seulement de la variable spatiale $\mathbf{x}$, mais aussi du temps t. Néanmoins, et pour éviter toute confusion, nous conservons la terminologie modèles 1D, 2D, 3D, par référence à la variable d'espace *uniquement*.

Commençons par un modèle 1D instationnaire, celui de la *corde vibrante*. Soit donc une corde de longueur unité, fixée en ses deux extrémités. Pour simplifier, nous négligeons la gravité, et nous supposons que la densité linéique de masse ρ est constante. Le but est encore une fois de calculer des "petits" déplacements verticaux, à partir d'une **configuration initiale**, connue à l'instant $t = 0$. On note $u : (x, t) \mapsto u(x, t)$ le déplacement transversal. D'après les équations de l'élasticité linéaire, on sait que u vérifie

$$\frac{\partial^2 u}{\partial t^2} - c(\tau)^2 \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } x \in]0, 1[\text{ et } t > 0. \tag{1.10}$$

(Ci-dessus, si τ est la tension de la corde, on a $c(\tau) = (\tau/\rho)^{1/2}$.)
Comme les extrémités du fil sont fixées, il est clair que le déplacement vertical est toujours nul en celles-ci, ce que l'on exprime sous la forme de conditions aux limites

$$u(0, t) = u(1, t) = 0 \text{ pour } t > 0. \tag{1.11}$$

Le but étant de déterminer les déplacements verticaux de la corde à partir d'une configuration initiale, celle-ci est donc connue : c'est une *donnée*. Pour cela, comme c'est une dérivée seconde en temps qui intervient dans le modèle, nous avons besoin de connaître *à la fois* sa position, ainsi que la dérivée partielle par rapport au temps[2]

$$u(x, 0) = u^0(x) \text{ et } \frac{\partial u}{\partial t}(x, 0) = u^1(x) \text{ pour } x \in]0, 1[. \tag{1.12}$$

La donnée dans l'équation (1.12) est le couple (u^0, u^1) : on parle de **conditions initiales**.
Enfin, le fait que l'énergie soit bornée peut-être exprimé sous la forme

$$\int_0^1 \left(\left[\frac{\partial u}{\partial t} \right]^2 + c(\tau)^2 \left[\frac{\partial u}{\partial x} \right]^2 \right) (x, t)\, dx < \infty \text{ pour } t > 0. \tag{1.13}$$

[2]Pour une justification intuitive, nous renvoyons le lecteur à la dernière section.

Le **domaine de calcul** est ici égal à $]0,1[\times]0,+\infty[$: c'est un *ouvert* de $\mathbb{R}^2$, par rapport au couple de variables (x,t). On peut aussi choisir de calculer la solution entre un instant initial $t=0$ et un instant final $T>0$, auquel cas il convient de remplacer la condition "pour $t>0$" par "pour $t \in]0,T[$" dans (1.10)-(1.13). Dans ce cas, le domaine de calcul est égal à $]0,1[\times]0,T[$.

Remarque 1.2.1 *Il convient de bien différencier "condition à un instant donné", de "condition aux limites en temps". Pour s'en convaincre, supposons que l'instant final T soit égal à un dans le modèle de la corde vibrante, de solution u_c. On l'a mis en équations sous la forme (1.10)-(1.13) pour (x,t) variant dans $]0,1[\times]0,1[$. Quelles valeurs de u_c connaît-on* a priori *?*

$$u_c(0,t),\ u_c(1,t) \text{ pour } t \in]0,1[\text{ (cf. (1.11))}, \text{ et } u_c(x,0) \text{ pour } x \in]0,1[\text{ (cf. (1.12))}.$$

Par contre, $x \mapsto u_c(x,1)$ est à déterminer !
Si on reprend le modèle 2D statique de la membrane élastique (1.4)-(1.6), de solution u_m, il est également posé dans le domaine $]0,1[\times]0,1[$, avec cette fois les variables (x,y). Quelles valeurs de u_m connaît-on a priori *? Celles imposées sur la frontière du domaine par (1.5), c'est-à-dire*

$$u_m(0,y),\ u_m(1,y) \text{ pour } y \in]0,1[, \text{ et } u_m(x,0),\ u_m(x,1) \text{ pour } x \in]0,1[.$$

Ainsi, $x \mapsto u_m(x,1)$ est connue !
Il y a donc là une différence fondamentale*, due à la nature des opérateurs associés à chaque modèle (adimensionnalisé) :*

$$-\frac{\partial^2 u_m}{\partial x^2} \underset{+}{\downarrow} \frac{\partial^2 u_m}{\partial y^2} \ \textit{vs.}\ -\frac{\partial^2 u_c}{\partial x^2} \underset{+}{\downarrow} \frac{\partial^2 u_c}{\partial t^2}.$$

Pour le modèle 2D instationnaire, reprenons celui de la *membrane*. Il s'agit de déterminer des "petits" déplacements verticaux, lorsque cette membrane est soumise à une force verticale qui dépend du temps ; soit $f(x,y,t)$ la densité de force par unité de surface. Le déplacement vertical est toujours noté u, et est lui aussi fonction de (x,y,t). Le couple (x,y) parcourt D et t est soit positif ($t>0$), soit compris entre zéro et T ($t \in]0,T[$), où l'instant final T est fixé. D'après les équations de l'élasticité linéaire, u vérifie

$$\left(\frac{\partial^2 u}{\partial t^2} - c(\tau)^2 \Delta_2 u\right)(x,y,t) = f(x,y,t) \text{ pour } (x,y) \in D \text{ et } t>0. \tag{1.14}$$

Comme la frontière ∂D est fixée, on écrit

$$u(x,y,t) = 0, \text{ pour } (x,y) \in \partial D \text{ et } t>0. \tag{1.15}$$

Les conditions initiales s'écrivent cette fois

$$u(x,y,0) = u^0(x,y) \text{ et } \frac{\partial u}{\partial t}(x,y,0) = u^1(x,y) \text{ pour } (x,y) \in D. \tag{1.16}$$

Enfin, on exprime le fait que l'énergie soit bornée par la relation

$$\int_D \left(\left[\frac{\partial u}{\partial t}\right]^2 + c(\tau)^2 \left|\vec{\nabla}_2 u\right|^2 \right)(x,y,t)\, dxdy < \infty \text{ pour } t>0. \tag{1.17}$$

Pour un modèle 3D instationnaire, intéressons à l'*acoustique* d'une salle S. Il s'agit de calculer la propagation du son engendrée par des hauts-parleurs, c'est-à-dire les *variations de pression* de l'air ambiant par rapport à une pression de référence P_{ref}. Précisément, c'est le déplacement de la membrane des hauts-parleurs, qui va générer ces variations.

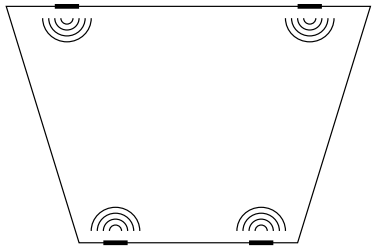


FIG. 1.4 – Modèle 3D instationnaire : acoustique avec quatre hauts-parleurs

On note ainsi p la variation de pression par rapport à la pression de référence, qui est "petite" par nature (nos tympans ne le supporteraient pas, dans le cas contraire !). C'est une fonction de quatre variables, (x, y, z, t) ; le couple (x, y, z) parcourt S et t est soit positif, soit dans $]0, T[$, où T est fixé. Le déplacement f des membranes dépend lui aussi des quatre variables, le triplet (x, y, z) parcourant dans ce cas la surface des membranes, appelée HP dans la suite.

Les équations vérifiées par p sont successivement

$$\left(\frac{\partial^2 p}{\partial t^2} - c^2 \Delta_3 p\right)(x, y, z, t) = 0 \text{ pour } (x, y, z) \in S \text{ et } t > 0. \tag{1.18}$$

(Pas de forces volumiques dans la salle ; c est la *vitesse du son* dans l'air.)
Pour ce qui concerne les conditions aux limites, elles sont de deux types. Sur les membranes, le flux de pression, noté $\partial p/\partial n$, est supposé proportionnel au déplacement f. Et, aux parois de la salle, $\partial S \setminus HP$, la pression est égale à P_{ref}, soit une variation nulle.

$$\left\{ \begin{array}{l} p(x, y, z, t) = 0, \text{ pour } (x, y, z) \in \partial S \setminus HP \\ \dfrac{\partial p}{\partial n}(x, y, z, t) = \alpha f, \text{ pour } (x, y, z) \in HP \end{array} \right. \text{ et } t > 0. \tag{1.19}$$

L'air étant au repos à l'instant initial, les conditions initiales s'écrivent cette fois

$$p(x, y, z, 0) = 0 \text{ et } \frac{\partial p}{\partial t}(x, y, z, 0) = 0 \text{ pour } (x, y, z) \in D. \tag{1.20}$$

Enfin, l'énergie acoustique est bornée, ce qui s'écrit

$$\int_S \left(\left[\frac{\partial p}{\partial t}\right]^2 + c^2 \left|\vec{\nabla}_3 p\right|^2 \right)(x, y, z, t)\, dxdy < \infty \text{ pour } t > 0. \tag{1.21}$$

Quels sont les points communs entre ces trois modèles instationnaires ?

- Problèmes posés dans des domaines $\Omega_d \times]0, T[$ de $\mathbb{R}^{d+1}$, $d = 1, 2, 3$, par rapport aux variables $(\mathbf{x}, t)$:
  1. $d = 1$: corde ;
  2. $d = 2$: membrane ;
  3. $d = 3$: acoustique.
- L'opérateur intérieur est identique, pour tous les problèmes :

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta_d u.$$

- Les conditions aux limites présentées sont de deux types :

une condition sur la valeur de l'inconnue à la frontière, *id est* $u = \cdots$;

une condition sur le flux de l'inconnue à travers la frontière, *id est* $\dfrac{\partial u}{\partial n} = \cdots$.

- Les conditions initiales sont toujours au nombre deux, pour l'opérateur intérieur présenté ci-dessus : l'une sur u, et l'autre sur $\partial_t u$, à un instant donné.
- Enfin, même propriété pour l'énergie, bornée pour tous les problèmes (voir la section 1.3 pour un résultat plus précis), pour tout t :

$$\int_{\Omega_d} \left(\left[\frac{\partial u}{\partial t} \right]^2 + c^2 \left| \vec{\nabla}_d u \right|^2 \right) (\mathbf{x}, t)\, d\mathbf{x} < \infty.$$

- Problèmes **linéaires** par rapport aux données f, u^0 et u^1.

Pour élargir la perspective, présentons brièvement un autre type de modèle 3D instationnaire, lui aussi basé sur le Laplacien : l'*équation de la chaleur*. Cette fois, on souhaite chauffer la salle S ! Il s'agit de calculer les *variations de température* de l'air ambiant par rapport à une température de référence T_{ref}. Plus précisément, supposons que ces variations soient le fait d'une source volumique de chaleur w. On note θ la variation de température, "petite" (variation de quelques dizaines de degrés au plus, essentiellement au voisinage de la source).
Les équations vérifiées par θ sont successivement

$$\left(\frac{\partial \theta}{\partial t} - \Delta_3 \theta \right) (x, y, z, t) = w(x, y, z, t) \text{ pour } (x, y, z) \in S \text{ et } t > 0. \tag{1.22}$$

Pour ce qui concerne les conditions aux limites, elles sont d'un seul type, si on suppose que les murs, sol et plafond restent à la température de référence[3]

$$\theta(x, y, z, t) = 0, \text{ pour } (x, y, z) \in \partial S \text{ et } t > 0. \tag{1.23}$$

Si la température est égale à T_{ref} à l'instant initial, la condition initiale s'écrit[4]

$$\theta(x, y, z, 0) = 0 \text{ pour } (x, y, z) \in S. \tag{1.24}$$

Enfin, l'énergie thermique est bornée (voir la section 1.3 pour un résultat plus précis), ce qui s'écrit

$$\int_S \theta^2(x, y, z, t)\, dxdydz + \int_0^t \int_S \left| \vec{\nabla}_3 \theta \right|^2 (x, y, z, s)\, dxdydzds < \infty \text{ pour } t > 0. \tag{1.25}$$

Remarque 1.2.2 *On imagine aisément le même type de problème dans une section 2D, pour arriver cette fois à une équation de la chaleur 2D.*

1.3 Classification et propriétés

Du point de vue de la terminologie, on peut classer les équations (celles qui sont définies dans l'intégralité du domaine de calcul) présentées précédemment en trois catégories. Notons tout d'abord qu'il s'agit d'**équations aux dérivées partielles** au sens où, si u est fonction de $x_1, \cdots, x_d$ (et de t), les dérivées partielles de la solution u par rapport à x_1, $\cdots$, x_d (et de t) apparaissent. Intéressons-nous maintenant à des modèles dépendant de trois variables, c'est-à-dire des modèles 3D statiques, ou 2D instationnaires. Si nous remplaçons *symboliquement* les dérivations partielles ∂_x, ∂_y, ∂_z, ∂_t respectivement par un facteur X, Y, Z, T, nous arrivons aux symboles (au signe près) :

[3] Si de l'énergie thermique s'échappait par une ouverture, on aurait sur celle-ci une condition aux limites du type $\lambda\theta + \dfrac{\partial \theta}{\partial n} = 0$, avec $\lambda > 0$.

[4] Pour ce type d'équations, la donnée de la répartition initiale de température suffit, puisque seule une dérivée partielle première par rapport à t intervient dans le modèle.

- $S_1 = X^2 + Y^2 + Z^2$ pour la cavité 3D statique ;
- $S_2 = X^2 + Y^2 - T^2$ pour la membrane 2D instationnaire ;
- $S_3 = X^2 + Y^2 - T$ pour l'équation de la chaleur 2D.

Si nous supposons que S_1, S_2 et S_3 sont des constantes (positive pour S_1), on a successivement l'équation d'une ellipse (ici une sphère), d'un hyperboloïde de révolution, et enfin d'un paraboloïde de révolution. C'est pourquoi, on parle :

- d'**EDP elliptique** pour qualifier les problèmes statiques de la section 1.1 ;
- d'**EDP hyperbolique** pour qualifier les problèmes instationnaires de la section 1.2, faisant intervenir une dérivée partielle seconde par rapport au temps (corde, membrane, acoustique) ;
- d'**EDP parabolique** pour qualifier les problèmes instationnaires de la section 1.2, faisant intervenir une dérivée partielle première par rapport au temps (équations de la chaleur).

Cette terminologie, outre son aspect pratique (classification des EDP en trois catégories), revêt une importance fondamentale, lorsqu'il s'agit d'étudier les propriétés de leurs solutions respectives. En effet, on peut vérifier, que deux solutions d'EDP d'une même classe possèdent un certain nombre de *propriétés identiques*. Donnons-en quelques exemples, d parcourant $\{1, 2, 3\}$. L'énoncé des résultats est parfois vague (mais c'est volontaire !), et le lecteur est invité à consulter [2] pour les résultats précis et les preuves qui les accompagnent.

Commençons par les problèmes elliptiques

$$-\Delta_d u = f \text{ sur } \Omega_d, \quad u = g \text{ sur } \partial\Omega_d. \tag{1.26}$$

(Ici, nous considérons que la condition aux limites sur la frontière $\partial\Omega_d$ peut être non-constante.)

Théorème 1.3.1 (Principe de positivité) *Supposons que f et g soient positives, respectivement sur Ω_d et $\partial\Omega_d$. Alors la solution u du problème (1.26) est positive sur Ω_d.*

Théorème 1.3.2 (Existence et unicité de la solution) *Le problème (1.26) admet une solution et une seule $\vec{x} \mapsto u(\vec{x})$ sur le domaine Ω_d, qui dépend continûment des données f et g.*

Remarque 1.3.1 *On note que si f et g sont nulles, l'unicité de la solution impose que $u = 0$.*

Poursuivons par les problèmes hyperboliques

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \Delta_d u = f \text{ sur } \Omega_d \times]0, T[, \quad u = 0 \text{ sur } \partial\Omega_d \times]0, T[, \\ u|_{t=0} = u^0, \ \dfrac{\partial u}{\partial t}|_{t=0} = u^1 \text{ sur } \Omega_d. \end{cases} \tag{1.27}$$

Dans ce cas, on peut établir une égalité d'énergie, en introduisant l'énergie associée à la solution u, somme d'une énergie cinétique et d'une énergie potentielle :

$$E(t) = \frac{1}{2} \int_{\Omega_d} \left(\left[\frac{\partial u}{\partial t} \right]^2 + c^2 \left| \vec{\nabla}_d u \right|^2 \right) (\mathbf{x}, t) \, d\mathbf{x}. \tag{1.28}$$

Théorème 1.3.3 (Egalité d'énergie) *A tout instant t ($0 < t < T$), l'énergie vérifie la relation*

$$E(t) = \int_0^t \int_{\Omega_d} (f \frac{\partial u}{\partial t})(\vec{x}, s) \, d\vec{x} ds + E(0). \tag{1.29}$$

Ainsi, lorsque la source volumique est nulle, alors l'énergie est conservée au cours du temps.

Corollaire 1.3.1 (Caractère conservatif) *Supposons que $f = 0$ dans (1.27). Pour t ($0 < t < T$), on a l'égalité*

$$E(t) = E(0). \tag{1.30}$$

Théorème 1.3.4 (Existence et unicité de la solution) *Le problème (1.27) admet une solution et une seule $(\vec{x}, t) \mapsto u(\vec{x}, t)$ sur le domaine $\Omega_d \times]0, T[$, qui dépend continûment des données f, u^0 et u^1.*

Finissons par les problèmes paraboliques

$$\frac{\partial u}{\partial t} - \Delta_d u = f \text{ sur } \Omega_d \times]0, T[, \quad u = 0 \text{ sur } \partial\Omega_d \times]0, T[, \quad u|_{t=0} = u^0 \text{ sur } \Omega_d. \tag{1.31}$$

(Ici, nous considérons que la condition initiale sur Ω_d peut être quelconque.)
Dans ce cas, on peut également établir une égalité d'énergie.

Théorème 1.3.5 (Egalité d'énergie) *A tout instant t ($0 < t < T$), la solution u du problème (1.31) vérifie la relation*

$$\frac{1}{2}\int_{\Omega_d} u^2(\vec{x}, t)\, d\vec{x} + \int_0^t \int_{\Omega_d} |\vec{\nabla}_d u|^2(\vec{x}, s)\, d\vec{x} ds = \int_0^t \int_{\Omega_d} (fu)(\vec{x}, s)\, d\vec{x} ds + \frac{1}{2}\int_{\Omega_d} (u^0)^2(\vec{x})\, d\vec{x}. \tag{1.32}$$

Si de plus la source volumique est nulle, alors la norme de la solution décroît au cours du temps.

Corollaire 1.3.2 (Caractère dissipatif) *Supposons que $f = 0$ dans (1.31). Pour t_1 et t_2 tels que $0 \leq t_1 < t_2 < T$ et $\displaystyle\int_{\Omega_d} u^2(\vec{x}, t_1)\, d\vec{x} \neq 0$, on a l'inégalité stricte*

$$\int_{\Omega_d} u^2(\vec{x}, t_2)\, d\vec{x} < \int_{\Omega_d} u^2(\vec{x}, t_1)\, d\vec{x}. \tag{1.33}$$

Théorème 1.3.6 (Décroissance exponentielle) *Supposons que $f = 0$ dans (1.31). Il existe une constante $\lambda > 0$* indépendante de T *telle que, pour tout t de $[0, T[$,*

$$\int_{\Omega_d} u^2(\vec{x}, t)\, d\vec{x} \leq \mathrm{e}^{-\lambda t} \int_{\Omega_d} (u^0)^2(\vec{x})\, d\vec{x}. \tag{1.34}$$

Théorème 1.3.7 (Existence et unicité de la solution) *Le problème (1.31) admet une solution et une seule $(\vec{x}, t) \mapsto u(\vec{x}, t)$ sur le domaine $\Omega_d \times]0, T[$, qui dépend continûment des données f et u^0.*

Définition 1.3.1 *Lorsqu'un problème admet une solution et une seule, qui de plus dépend continûment des données, on dit qu'il est* **bien posé**.

Pour conclure cette courte section sur la classification, notons qu'il existe bien d'autres modèles (avec ou sans Laplacien !), qui fassent partie d'une de ces trois classes. Examinons par exemple les équations de Maxwell dans la cavité 3D introduite à la section précédente, extérieure à un ensemble de conducteurs parfaits. Elles s'écrivent sous la forme bien connue suivante, composée des équations d'Ampère, de Faraday, de Coulomb et de l'absence de monopoles magnétiques libres :

$$\begin{aligned}
&\frac{\partial \vec{E}}{\partial t} - c^2 \vec{\mathrm{rot}}\, \vec{B} = -\frac{1}{\varepsilon_0}\vec{J} \text{ sur } C \times]0, T[,\\
&\frac{\partial \vec{B}}{\partial t} + \vec{\mathrm{rot}}\, \vec{E} = 0 \text{ sur } C \times]0, T[,\\
&\mathrm{div}\, \vec{E} = \frac{1}{\varepsilon_0}\rho \text{ sur } C \times]0, T[,\\
&\mathrm{div}\, \vec{B} = 0 \text{ sur } C \times]0, T[.
\end{aligned}$$

On rappelle que si μ_0 est la perméabilité magnétique du vide, alors c, la vitesse de la lumière, satisfait à $c^2\varepsilon_0\mu_0 = 1$. Les opérateurs différentiels rotationnel $\vec{\text{rot}}$ et divergence div sont définis par

$$\vec{\text{rot}}\,\vec{v} = (\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3})\vec{e}_1 + (\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1})\vec{e}_2 + (\frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2})\vec{e}_3, \quad \text{div}\,\vec{v} = \sum_{k=1}^{k=3} \frac{\partial v_k}{\partial x_k}.$$

Les données sont $\vec{J}$ et ρ, respectivement les densités de courant et de charge. Les inconnues sont $\vec{E}$ le champ électrique, et $\vec{B}$, l'induction magnétique : $(\vec{E}, \vec{B})$ est le **champ électromagnétique** à déterminer. Qui plus est, on suppose connue la valeur initiale du champ, puisque des dérivées partielles premières par rapport au temps interviennent dans notre modèle. On écrit :

$$\vec{E}|_{t=0} = \vec{E}^0, \ \vec{B}|_{t=0} = \vec{B}^0 \ \text{ sur } C.$$

Enfin, sur la frontière de la cavité, on a la relation

$$\vec{E} \times \vec{n} = 0, \ \text{ sur } \partial C \times]0, T[,$$

qui stipule que les composantes tangentielles du champ électrique $\vec{E}$ s'annulent (ci-dessus, $\vec{n}$ est un vecteur normal à la frontière de la cavité.)
Sous réserve que les densités de charge et de courant ρ et $\vec{J}$ vérifient la relation de conservation de la charge

$$\frac{\partial \rho}{\partial t} + \text{div}\,\vec{J} = 0,$$

ce modèle est *bien posé*. Par ailleurs, on peut prouver qu'il est de type *hyperbolique*.

1.4 Problèmes aux valeurs propres et problèmes stationnaires

Pour aborder cette dernière section, reprenons le modèle de la membrane 2D instationnaire (1.14)-(1.17), et posons-nous cette question :
Est-ce que la membrane est "stable", pour f dépendant du temps donnée ?
Ou bien, de façon équivalente :
Peut-on s'assurer que les déplacements verticaux restent "petits" ?

A cette question, on peut apporter deux types de réponses :
- L'une, *a posteriori*, consiste en la simulation numérique du comportement de la membrane, issue de la discrétisation de (1.14)-(1.17). Voir le chapitre 2 pour cette approche.
- L'autre, qui consiste en la décomposition du modèle (1.14)-(1.17) en une série de *problèmes aux valeurs propres*, puis en une étude *a priori*.

C'est cette seconde approche que nous allons utiliser ci-après. En effet, il est possible (consulter par exemple [2]) de décomposer la solution $u(x, y, t)$ de (1.14)-(1.17) selon

$$u(x, y, t) = \sum_{k \in \mathbb{N}} \alpha_k(t)\, u_k(x, y). \tag{1.35}$$

La somme est infinie[5], puisqu'elle porte sur tout nombre naturel k positif. Pour chaque $k \in \mathbb{N}$, u_k est solution d'un **problème statique 2D aux valeurs propres** : on doit déterminer $(x, y) \mapsto u_k(x, y)$ et $\lambda_k > 0$ tels que

$$-c^2\Delta_2 u_k = \lambda_k u_k \text{ sur } D, \quad u_k = 0 \text{ sur } \partial D. \tag{1.36}$$

[5]Par définition $u_k \neq 0$, $\forall k \in \mathbb{N}$.

Les éléments du couple (u_k, λ_k) sont respectivement appelés **mode propre** de la membrane, et **valeur propre** associée λ_k. Pour des raisons pratiques qui vont apparaître immédiatement, nous introduisons la **pulsation propre** ω_k, égale à $\sqrt{\lambda_k}$. On décompose également f, u^0 et u^1, sous la forme

$$f(x,y,t) = \sum_{k\in\mathbb{N}} f_k(t)\, u_k(x,y), \quad u^0(x,y) = \sum_{k\in\mathbb{N}} \alpha_k^0\, u_k(x,y),\ u^1(x,y) = \sum_{k\in\mathbb{N}} \alpha_k^1\, u_k(x,y). \quad (1.37)$$

On trouve donc

$$\sum_{k\in\mathbb{N}} \left(\alpha_k''(t) + \lambda_k\alpha_k(t) - f_k(t)\right) u_k(x,y) = 0 \text{ pour } (x,y,t) \in D\times]0,T[, \quad (1.38)$$

$$\sum_{k\in\mathbb{N}} \left(\alpha_k(0) - \alpha_k^0\right) u_k(x,y) = 0, \text{ pour } (x,y) \in D, \quad (1.39)$$

$$\sum_{k\in\mathbb{N}} \left(\alpha_k'(0) - \alpha_k^1\right) u_k(x,y) = 0 \text{ pour } (x,y) \in D. \quad (1.40)$$

Ainsi, une fois $(u_k)_{k\in\mathbb{N}}$ connus, il est équivalent de résoudre (1.14)-(1.17), ou la série ($k \in \mathbb{N}$) d'équations différentielles ordinaires (EDO) :

$$\alpha_k''(t) + \omega_k^2\alpha_k(t) = f_k(t), \text{pour } t \in]0,T[,\ \alpha_k(0) = \alpha_k^0,\ \alpha_k'(0) = \alpha_k^1. \quad (1.41)$$

Fixons k.
La *solution générale* ($f_k = 0$) de l'EDO correspondante est de la forme

$$\alpha_k^G(t) = A_k \sin(\omega_k t) + B_k \cos(\omega_k t).$$

A l'aide des *deux conditions initiales*, on peut déterminer les valeurs de A_k et B_k, pour arriver à

$$\alpha_k^G(t) = \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t), \text{pour } t \in]0,T[.$$

Remarque 1.4.1 *On comprend, à la suite de ce calcul, pourquoi* deux *conditions initiales sont nécessaires pour un problème faisant intervenir une dérivée partielle seconde par rapport au temps. De façon similaire, on trouverait qu'*une *condition initiale suffit, pour un problème comprenant uniquement une dérivée partielle première par rapport au temps.*

On peut facilement vérifier qu'une *solution particulière* ($\alpha_k^0 = \alpha_k^1 = 0$) de l'EDO est de la forme

$$\alpha_k^P(t) = \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds.$$

Ainsi, la solution complète de l'EDO s'écrit

$$\alpha_k(t) = \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t) + \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds, \text{pour } t \in]0,T[. \quad (1.42)$$

La solution de l'EDP instationnaire modélisant la membrane est finalement égale à

$$u(x,y,t) = \sum_{k\in\mathbb{N}} \left\{ \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t) + \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds \right\} u_k(x,y), \text{ pour } (x,y,t) \in D\times]0,T[. \quad (1.43)$$

Conclusion : les déplacements verticaux de la membrane sont donc "petits" si, et seulement si, tous les coefficients entre accolades (les $(\alpha_k^G + \alpha_k^P)(t)$) sont eux-mêmes "petits", pour t variant de 0 à T.
Or, l'amplitude des deux premiers termes – dont la somme est égale à $\alpha_k^G(t)$ – ne dépend que de α_k^0 et α_k^1/ω_k, et est par voie de conséquence indépendante de t : ils sont donc "petits" si, et seulement si, les données initiales le sont ! Ceci est intuitif...
Qu'en est-il du troisième et dernier terme, $\alpha_k^P(t)$? La réponse est moins immédiate... Considérons une "petite" sollicitation, prenant la forme simple

$$f(x,y,t) = f_l(t)u_l(x,y), \text{avec } f_l(t) = \mu_l \cos(\omega t),$$

pour une pulsation $\omega \geq 0$ et l donné. Le fait que la sollicitation soit "petite" revient à supposer que $\mu_l \neq 0$ est "petit". Que valent les coefficients $(\alpha_k^P)_{k\in\mathbb{N}}$? Si $k \neq l$, $\alpha_k^P = 0$. Par contre, si $k = l$, deux situations peuvent se produire. En effet, en calculant l'intégrale qui détermine ce coefficient, on arrive à deux résultats différents :

$$\text{si } \omega \neq \omega_l \ : \qquad \alpha_l^P(t) = \frac{\mu_l}{2\omega_l}\left\{\frac{1}{\omega-\omega_l} + \frac{1}{\omega+\omega_l}\right\}(\cos(\omega_l t) - \cos(\omega t))\,; \tag{1.44}$$

$$\text{si } \omega = \omega_l \ : \qquad \alpha_l^P(t) = \frac{\mu_l}{2\omega_l} t \sin(\omega_l t). \tag{1.45}$$

Ainsi, si $\omega = \omega_l$, l'amplitude des oscillations croît linéairement avec t – on parle de **résonance** – et il n'y a aucune chance que les déplacements verticaux restent "petits" dans ce cas. Si par contre ω est différent de ω_l, le terme de plus grande amplitude est

$$\frac{1}{\omega-\omega_l}\left[\frac{\mu_l}{2\omega_l}(\cos(\omega_l t) - \cos(\omega t))\right].$$

L'amplitude peut donc être "grande" si $|\omega - \omega_l|$ est "petit".

Bref, ce qui compte avant tout, pour une réponse *a priori*, c'est de déterminer les valeurs propres $(\lambda_k)_{k\in\mathbb{N}}$ ainsi que les modes propres associés, solution des *problèmes aux valeurs propres* (1.36), avec la condition

$$\int_C \left|\vec{\nabla}_2 u_k\right|^2 (x,y)\, dxdy < \infty.$$

A partir de là, on choisira la pulsation ω de la sollicitation f aussi "éloignée" que possible des $(\sqrt{\lambda_k})_{k\in\mathbb{N}}$. En particulier, il sera très intéressant de déterminer avec précision la plus petite valeur $\omega_{min} = \min_k \sqrt{\lambda_k}$, puisqu'un choix de ω dans $]0, \omega_{min}[$ permettra à coup sûr d'éviter les résonances.

Abordons brièvement, pour clore cette section, la notion de **problème stationnaire**. Pour l'obtenir, il suffit de considèrer un problème dépendant du temps t, pour lequel la dépendance par rapport à t est connue explicitement : oscillations forcées en $\sin(\nu t)$ ou $\cos(\nu t)$ avec $\nu \neq 0$, par exemple. On peut alors remplacer la dérivation par rapport au temps par une multiplication par $-\nu^2$, pour aboutir à un problème du type

$$-\nu^2 u - c^2 \Delta_2 u = g \text{ sur } D, \quad u = 0 \text{ sur } \partial D. \tag{1.46}$$

Il n'y a plus de condition initiale, puisque la résolution de (1.46) permet, par multiplication par le coefficient de dépendance en temps, de déterminer complètement la solution.

La différence avec le problème aux valeurs propres (1.36) tient en deux points :
– la valeur de ν est *connue*, alors qu'ω_k est une *inconnue* du problème (1.36) ;
– il y a un second membre g *a priori* non nul dans (1.46).

Chapitre 2

La méthode des différences finies

2.1 Introduction

Dans ce chapitre, nous présentons une méthode de discrétisation numérique. L'introduction de celle-ci nous permet de calculer des approximations des solutions d'EDP statiques. Dans la suite, on étudie plus en détail le calcul du déplacement transversal d'un fil ou d'une poutre, ainsi que du déplacement vertical d'une membrane élastique et du potentiel électrostatique. Comme on le verra, à ces trois problèmes correspondent des approximations différentes (bien que les modèlisations reposent toujours sur le Laplacien), puisque les modèles dont ils sont issus sont respectivement posés dans $\mathbb{R}$, $\mathbb{R}^2$ et $\mathbb{R}^3$. Néanmoins, la technique d'approximation retenue, appelée **méthode des différences finies**, reste la même pour ces trois problèmes. A la fin du chapitre, nous verrons comment cette méthode de discrétisation peut aussi être appliquée aux problèmes dépendant du temps.

2.2 Un problème monodimensionnel

Dans cette section, nous considérons le fil tendu entre ses extrémités, ou la poutre, appuyée en ses extrémités, situées en 0 et 1. On suppose qu'il ou elle est soumis à une force extérieure transverse (telle que son poids, dans le cas du fil pesant). On note $f : x \mapsto f(x)$ la densité linéique des forces appliquées, et $u : x \mapsto u(x)$ le déplacement transversal induit, que l'on cherche à approcher numériquement. Nous avons admis que les équations de l'élasticité linéaire monodimensionnelles (1D) normalisées permettent de modéliser correctement le phénomène. Rappelons-en la forme :

$$-u''(x) = f(x) \text{ sur }]0,1[, \quad u(0) = u(1) = 0. \tag{2.1}$$

Dans le cas particulier où $f \equiv 1$, la solution est égale à

$$u_0(x) = \frac{1}{2}x(1-x). \tag{2.2}$$

Ce résultat élémentaire sera utile par la suite...

On suppose que la solution u est de classe $\mathcal{C}^4([0,1])$ ou, ce qui est *équivalent*, que la donnée f est de classe $\mathcal{C}^2([0,1])$.

Pour déterminer une méthode d'approximation de l'équation aux dérivées partielles (2.1) (ça n'est pas la seule !), on utilise la

Proposition 2.2.1 *Soient $x \in]0,1[$ et h tel que $[x-h, x+h] \in [0,1]$. Alors*

$$\exists \theta \in]-1,1[\text{ tel que } -u''(x) = \frac{-u(x-h) + 2u(x) - u(x+h)}{h^2} + \frac{h^2}{12}u^{(4)}(x+\theta\, h). \tag{2.3}$$

Preuve : On utilise la formule de Taylor-Mac Laurin.

$$\exists \theta^- \in]-1,0[\text{ tel que } u(x-h) = u(x) - h\,u'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x+\theta^- h)$$
$$\exists \theta^+ \in]0,1[\text{ tel que } u(x+h) = u(x) + h\,u'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x+\theta^+ h).$$

On somme les deux égalités, pour trouver

$$-u''(x) = \frac{-u(x-h)+2u(x)-u(x+h)}{h^2} + \frac{h^2}{24}(u^{(4)}(x+\theta^- h) + u^{(4)}(x+\theta^+ h)).$$

Pour arriver à l'expression annoncée, il faut se souvenir du théorème des valeurs intermédiaires. Il permet, puisque $u^{(4)}$ est continue, de remplacer les deux termes en $u^{(4)}$ par $2u^{(4)}(x+\theta\,h)$, mais avec un paramètre θ appartenant à $[\theta^-,\theta^+]$, donc à $]-1,1[$ comme annoncé. ■

Remarque 2.2.1 *Le premier terme de (2.3) est une bonne approximation de $-u''(x)$, sous réserve que h est* petit. *En effet, comme on a la relation $-u^{(4)} = f''$, on sait que*

$$\left|\frac{h^2}{12}u^{(4)}(x+\theta\,h)\right| \leq \frac{C_{f,2}}{12}h^2, \text{ avec } C_{f,2} = \sup_{x\in[0,1]}|f''(x)|.$$

Ce résultat *simple* fournit une méthode de discrétisation et d'approximation de l'équation de départ (2.1) ; on parle souvent de **schéma numérique** de discrétisation. Le terme **différences finies** provient quant à lui de l'expression (2.3) : on remplace une dérivée, qui est par définition la *limite* d'un taux de variation, par un taux de variation, dont le dénominateur conserve une valeur finie *non nulle* (ici h^2 pour une dérivée seconde). En pratique, comment procède-t-on ?

Pour commencer, on choisit $N \in \mathbb{N}$, et on fixe $h = \dfrac{1}{N+1}$. Remarquons tout de suite que pour avoir une "bonne" approximation de $u''(x)$, il convient que h soit petit. Ceci signifie que N est un paramètre de discrétisation qui aura vocation à devenir "grand", lors de la réalisation des expériences numériques.

Nous allons construire une méthode qui permet d'approcher la valeur de u aux points $x_i = ih$, pour $i \in \{0,1,\cdots,N+1\}$, par des nombres, notés $(u_i)_{0\leq i\leq N+1}$. Puisque u est approchée en deux points consécutifs distants de h, on appelle h le **pas de discrétisation**.

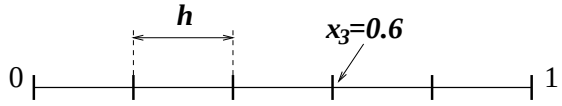


FIG. 2.1 – Le segment découpé ($N = 4$, $h = 0.2$)

Remarque 2.2.2 *Comme on sait que $u(0) = u(1) = 0$, on choisira toujours comme approximation $u_0 = u_{N+1} = 0$!*

On définit $f_i = f(x_i)$, pour $i \in \{1,\cdots,N\}$, et on considère l'ensemble des équations

$$\frac{-u_{i-1}+2u_i-u_{i+1}}{h^2} = f_i,\ 1 \leq i \leq N, \text{ avec } u_0 = u_{N+1} = 0. \tag{2.4}$$

Chaque équation faisant intervenir trois nombres parmi $(u_i)_i$, on parle de **schéma à trois points**.

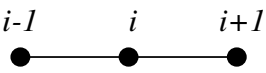


FIG. 2.2 – Le schéma aux différences finies à trois points

NB. Noter la similitude entre (2.4) d'une part, et (2.3) et (2.1) en $x = x_i$, d'autre part.

Si on appelle $\vec{u}$ (resp. $\vec{f}$) le vecteur de $\mathbb{R}^N$ de composantes $(u_i)_{1\leq i\leq N}$ (resp. $(f_i)_{1\leq i\leq N}$), on peut réécrire le système (2.4) sous la *forme vectorielle équivalente*

$$\mathbb{A}_1\,\vec{u} = \vec{f},\ \text{avec } \mathbb{A}_1 = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{2.5}$$

Par construction, la matrice $\mathbb{A}_1$ est
- tridiagonale, c'est-à-dire que tous les termes non nuls sont regroupés sur trois diagonales ;
- symétrique, puisque $(\mathbb{A}_1)_{i,j} = (\mathbb{A}_1)_{j,i}$, pour $1 \leq i, j \leq N$.

Il convient maintenant de vérifier qu'il existe une solution $\vec{u}$ unique de (2.5). Qui plus est, est-il possible de calculer et majorer l'erreur commise ? C'est l'objet des résultats ci-dessous. Tout d'abord, nous allons vérifier que la matrice $\mathbb{A}_1$ est inversible. Outre l'obtention de l'existence et de l'unicité de $\vec{u}$, ceci nous permettra de construire une formule explicite, exprimant l'erreur commise en fonction des données du problème. Par ailleurs, pour exploiter cette formule, c'est-à-dire pour majorer l'erreur, nous allons étudier les caractéristiques de l'inverse $\mathbb{A}_1^{-1}$.

Définition 2.2.1 *Un vecteur v de $\mathbb{R}^N$ est dit* **positif** *lorsque $v_i \geq 0$, $\forall 1 \leq i \leq N$.*
Une matrice A de $\mathbb{R}^{N\times N}$ est dite **positive** *lorsque $A_{i,j} \geq 0$, $\forall 1 \leq i, j \leq N$.*
Une matrice A de $\mathbb{R}^{N\times N}$ est dite **monotone** *lorsqu'elle est inversible, d'inverse positive.*

Avant de nous intéresser au cas particulier de la matrice issue du schéma à trois points, donnons une caractérisation simple des matrices monotones.

Proposition 2.2.2 *Une matrice A de $\mathbb{R}^{N\times N}$ est monotone si, et seulement si, on a l'inclusion*

$$\{v \in \mathbb{R}^N \ :\ Av \geq 0\} \subset \{v \in \mathbb{R}^N \ :\ v \geq 0\}. \tag{2.6}$$

Preuve : Supposons que A est monotone.
Soit v tel que $Av \geq 0$, alors $v = A^{-1}(Av)$ et, pour $1 \leq i \leq N$, $v_i = \sum_j (A^{-1})_{i,j}(Av)_j \geq 0$, puisque $(A^{-1})_{i,j}$ et $(Av)_j$ sont positifs par hypothèse. Ainsi $v \geq 0$, et l'inclusion (2.6) est vérifiée.

Réciproquement, si l'inclusion est satisfaite, montrons tout d'abord que A est inversible.
Soit donc v tel que $Av = 0$: on a $Av \geq 0$ et $A(-v) \geq 0$, ce qui implique $v \geq 0$ et $(-v) \geq 0$, *id est* $v = 0$, d'où l'inversibilité.
Sachant que A^{-1} existe, étudions sa positivité...
On note $(e_i)_i$ la base orthonormale canonique de $\mathbb{R}^N$. Alors les $f_i = A^{-1}e_i$, pour i variant de 1 à N, sont les vecteurs colonnes de A^{-1}. On a bien sûr $e_i = Af_i$, et l'inclusion (2.6) permet d'affirmer que f_i est positif, puisque e_i l'est. En d'autres termes, tous les éléments de A^{-1} sont positifs.
En conclusion, la matrice A est monotone. ∎

Proposition 2.2.3 *La matrice $\mathbb{A}_1$ correspondant à (2.5) est monotone.*

Preuve : Pour prouver que $\mathbb{A}_1$ est monotone, on reprend la proposition 2.2.2. Soit v tel que $\mathbb{A}_1 v \geq 0$, et $v_k = \min_{1\leq i\leq N} v_i$ (ou, de façon équivalente, $v_k \leq v_i$, $\forall i$). Le but est d'arriver à l'inégalité $v_k \geq 0$. On a

$$\left\{ \begin{array}{l} 2v_1 - v_2 \geq 0 \\ -v_{i-1} + 2v_i - v_{i+1} \geq 0, \quad 2 \leq i \leq N-1 \\ -v_{N-1} + 2v_N \geq 0 \end{array} \right. .$$

Si $v_k = v_1$, on trouve

$$v_k \geq v_2 - v_k \geq 0.$$

De même si $v_k = v_N$.

Si $k \in \{2, \cdots, N-1\}$, on trouve cette fois

$$(v_k - v_{k-1}) + (v_k - v_{k+1}) \geq 0.$$

Or, $v_k \leq v_{k-1}$ et, de même, $v_k \leq v_{k+1}$. On a donc $(v_k - v_{k-1}) + (v_k - v_{k+1}) \leq 0$, ce qui donne

$$v_{k-1} = v_{k+1} = v_k \ !$$

Par récurrence, on arrive facilement à $v_1 = \cdots = v_{k-1} = v_k = v_{k+1} = \cdots = v_N$. Et la première (ou la dernière) équation donne à nouveau $v_k \geq 0$. ∎

NB. Dans le corps de la preuve, on a obtenu l'égalité

$$\{v \in \mathbb{R}^N \ : \ \mathbb{A}_1 v \geq 0\} = \{v \in \mathbb{R}^N \ : \ v_i = \lambda, \ 1 \leq i \leq N, \ \lambda \in \mathbb{R}^+\}.$$

Exercice 2.2.1 *Déduire de la proposition précédente un* principe de positivité *pour le problème (2.5), homologue* discret *du principe général énoncé au théorème 1.3.1 (avec donnée nulle sur la frontière).*

Comme $\mathbb{A}_1$ est monotone, elle est en particulier inversible : c'est cette propriété que nous allons utiliser maintenant, pour déterminer l'erreur commise. Soit $\vec{e}$ le vecteur de $\mathbb{R}^N$ dont les composantes sont égales à $e_i = u_i - u(x_i)$, pour i variant de 1 à N. Comme pour $\vec{u}$, on adopte la convention $e_0 = e_{N+1} = 0$ (justifiée par le fait que $u(0) = u(1) = 0$, et $u_0 = u_{N+1} = 0\,!$). Sachant que u (resp. $\vec{u}$) est solution de l'équation (2.1) (resp. (2.5)), on a alors, d'après (2.3), pour i compris entre 1 et N :

$$\begin{aligned} (\mathbb{A}_1\vec{e})_i &= \frac{-e_{i-1} + 2e_i - e_{i+1}}{h^2} = (\mathbb{A}_1\vec{u})_i - \frac{-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})}{h^2} \\ &= f(x_i) - \frac{-u(x_i - h) + 2u(x_i) - u(x_i + h)}{h^2} = f(x_i) + u''(x_i) + \frac{h^2}{12} u^{(4)}(x_i + \theta_i\, h) \\ &= \frac{h^2}{12} u^{(4)}(x_i + \theta_i\, h) = \frac{h^2}{12} f''(x_i + \theta_i\, h), \text{ avec } \theta_i \in]-1, 1[. \end{aligned}$$

Dans l'esprit de la remarque 2.2.1, on introduit le vecteur $\vec{\varepsilon}$ de $\mathbb{R}^N$, dont les composantes sont égales à $\varepsilon_i = (\mathbb{A}_1\vec{e})_i = \dfrac{h^2}{12} f''(x_i + \theta_i\, h)$, pour i variant de 1 à N, et dont la norme est telle que $\|\vec{\varepsilon}\|_\infty \leq \dfrac{h^2}{12}\, C_{f,2}$ par construction. On en déduit alors l'expression de l'erreur commise

$$\vec{e} = \mathbb{A}_1^{-1}\vec{\varepsilon}. \tag{2.7}$$

Pour poursuivre, utilisons le fait que tous les éléments de $\mathbb{A}^{-1}$ sont positifs. En reprenant l'expression de l'erreur (2.7), on peut alors écrire

$$|e_i| = |\sum_j (\mathbb{A}_1^{-1})_{i,j}\varepsilon_j| \leq \sum_j (\mathbb{A}_1^{-1})_{i,j}|\varepsilon_j| \leq \sum_j (\mathbb{A}_1^{-1})_{i,j}\|\vec{\varepsilon}\|_\infty \leq \frac{h^2}{12}\, C_{f,2} \sum_j (\mathbb{A}_1^{-1})_{i,j}. \tag{2.8}$$

Pour finalement arriver à une majoration de l'erreur commise, il suffit de majorer $\sum_j (\mathbb{A}_1^{-1})_{i,j}$ dans (2.8), pour $1 \leq i \leq N$.

Proposition 2.2.4 *La somme des éléments de chaque ligne de $\mathbb{A}_1^{-1}$ est inférieure ou égale à 1/8.*

Preuve : On remarque que $\sum_j (\mathbb{A}_1^{-1})_{i,j} = \sum_j (\mathbb{A}_1^{-1})_{i,j}\delta_j$, sous réserve que $\delta_j = 1$, pour tout j. A quoi correspond le vecteur $\vec{\delta}$ ainsi construit ? On pose $\vec{u}_0 = \mathbb{A}_1^{-1}\vec{\delta}$, soit $\mathbb{A}_1\vec{u}_0 = \vec{\delta}$.
$\vec{\delta}$ joue le rôle d'un second membre de (2.5). Il correspond de fait à $f \equiv 1$, ce qui nous renvoie à la solution u_0 de (2.2). Or, dans ce cas *très particulier*,

$$-{u_0}''(x) = \frac{-u_0(x-h) + 2u_0(x) - u_0(x+h)}{h^2},\ \forall x \in]0,1[,\ \forall h \text{ t. q. } [x-h, x+h] \subset [0,1].$$

Ainsi, $\vec{u}_0$ tel que $(u_0)_i = u_0(x_i)$ vérifie

$$\mathbb{A}_1\vec{u}_0 = \vec{\delta},\ \text{soit } \vec{u}_0 = \mathbb{A}_1^{-1}\vec{\delta},\ \text{ou } \sum_j (\mathbb{A}_1^{-1})_{i,j} = u_0(x_i),\ 1 \leq i \leq N.$$

Et $\sup_{x\in[0,1]} u_0(x) = u_0(1/2) = 1/8$, ce qui permet de conclure. ∎

On a donc démontré le

Théorème 2.2.1 *Lorsque la solution u est de classe $\mathcal{C}^4([0,1])$, l'erreur est telle que*

$$\|\vec{e}\|_\infty \leq \frac{h^2}{96} \sup_{x\in[0,1]} |f''(x)|. \tag{2.9}$$

En conclusion, l'erreur "ponctuelle" tend **uniformément** vers 0 comme h^2.

NB. Lorsque h décroît, N croît en proportion inverse. Bref, l'erreur maximale décroît selon le carré de h, alors que le nombre d'inconnues croît en $1/h$...

Ceci étant, cette estimation dépend de la régularité de u (ou, ce qui revient au même dans le cas 1D, de celle de f). Que se passe-t-il si la solution u ou, ce qui est *équivalent*, la donnée f sont moins régulières ?

Théorème 2.2.2 *Lorsque la solution u est de classe $\mathcal{C}^2([0,1])$, l'erreur est telle que*

$$\lim_{h\to 0^+} \|\vec{e}\|_\infty = 0. \tag{2.10}$$

Preuve : On introduit à nouveau le vecteur $\vec{\varepsilon}$, de composantes $\varepsilon_i = (\mathbb{A}_1\vec{e})_i$. D'après les résultats portant sur la matrice $\mathbb{A}_1$ (propositions 2.2.3 et 2.2.4), il suffit de prouver que $\|\vec{\varepsilon}\|_\infty \to 0$, lorsque h tend vers 0, c'est-à-dire :

$$\forall \eta > 0,\ \exists h_\eta > 0,\ 0 < h < h_\eta \ \Rightarrow\ \|\vec{\varepsilon}\|_\infty < \eta.$$

(Bien évidemment, $\mathbb{A}_1$, $\vec{e}$ et $\vec{\varepsilon}$ dépendent de h, mais la dépendance est sous-entendue, notamment dans l'inégalité ci-dessus.)

Lorsque l'on sait simplement que f est continue, il n'est plus possible d'obtenir une expression des composantes $(\varepsilon_i)_i$ en fonction de la dérivée seconde f''... Mais tout n'est pas perdu ! Comme u est de classe $\mathcal{C}^2([0,1])$, on peut donc écrire, à l'aide de la formule de Taylor-Mac Laurin, les égalités

$$\exists \theta^- \in]-1,0[\text{ tel que } u(x-h) = u(x) - h\,u'(x) + \frac{h^2}{2} u''(x+\theta^- h)$$
$$\exists \theta^+ \in]0,1[\text{ tel que } u(x+h) = u(x) + h\,u'(x) + \frac{h^2}{2} u''(x+\theta^+ h).$$

On en déduit que, pour i variant de 1 à N,

$$\varepsilon_i = f(x_i) + \frac{1}{2}\left(u''(x_i+\theta^- h) + u''(x_i+\theta^+ h)\right) = f(x_i) - \frac{1}{2}\left(f(x_i+\theta^- h) + f(x_i+\theta^+ h)\right).$$

Or, f étant continue sur le segment $[0,1]$, elle est uniformément continue. Ce que l'on peut exprimer mathématiquement sous la forme :

$$\forall \eta > 0,\ \exists h_\eta > 0,\ \forall x, y \in [0,1],\ |x-y| < h_\eta \ \Rightarrow\ |f(x) - f(y)| < \eta.$$

Or, si $h \in]0, h_\eta[$, on a $|x_i - (x_i + \theta^\pm h)| < h_\eta$, d'où $|\varepsilon_i| < \eta$, pour tout i : par voie de conséquence, $\|\vec{\varepsilon}\|_\infty < \eta$. On en conclut finalement que, pour tout h dans $]0, h_\eta[$, on a l'inégalité

$$\|\vec{e}\|_\infty < \frac{\eta}{8}.$$

■

Sur cet exemple simple, on constate donc que la méthode des différences finies convergera *a priori* d'autant mieux que la solution du problème initial est régulière. Il est à noter, et c'est un point très important, que l'on retrouve effectivement ce type de comportement lorsque l'on réalise des expériences numériques...

Pour résoudre le système linéaire (2.5), nous allons établir une autre propriété concernant la matrice $\mathbb{A}_1$, qui nous permettra d'utiliser les algorithmes étudiés aux chapitres 3 et 4 :

- méthode de Cholesky (chapitre 3) ;
- méthodes de Jacobi ou de Gauss-Seidel, puisque $\mathbb{A}_1$ est tridiagonale (chapitre 4).

Définition 2.2.2 *Une matrice A de $\mathbb{R}^{N\times N}$ est dite* **définie-positive** *lorsque $(v, Av) > 0$, $\forall v \in \mathbb{R}^N$.*

Notons tout de suite le résultat général ci-dessous.

Proposition 2.2.5 *Toute matrice définie-positive A de $\mathbb{R}^{N\times N}$ est inversible.*

Preuve : En effet, $Av = 0 \Longrightarrow (v, Av) = 0 \Longrightarrow v = 0$. ■

Proposition 2.2.6 *La matrice $\mathbb{A}_1$ de (2.5) est symétrique définie-positive.*

Preuve : Nous avons déjà remarqué que la matrice $\mathbb{A}_1$ est symétrique.
Il reste à vérifier qu'elle est définie-positive. On applique la définition, en formant le produit scalaire $h^2(v, \mathbb{A}_1 v)$, pour un vecteur v de $\mathbb{R}^N$ quelconque :

$$\begin{aligned} h^2(v,\mathbb{A}_1 v) &= h^2 \sum_{i=1}^{N} v_i (\mathbb{A}_1 v)_i = v_1(2v_1 - v_2) + \sum_{i=2}^{N-1} v_i(-v_{i-1} + 2v_i - v_{i+1}) + v_N(-v_{N-1} + v_N) \\ &= 2\sum_{i=1}^{N} {v_i}^2 - 2\sum_{i=1}^{N-1} v_i v_{i+1} = {v_1}^2 + {v_N}^2 + \sum_{i=1}^{N-1} \left({v_i}^2 - 2v_i v_{i+1} + {v_{i+1}}^2\right) \\ &= {v_1}^2 + {v_N}^2 + \sum_{i=1}^{N-1} (v_i - v_{i+1})^2 . \end{aligned}$$

Ainsi, $(v, \mathbb{A}_1 v) \geq 0$. De plus,$(v, \mathbb{A}_1 v) = 0$ entraîne que $v_1 = v_N = 0$, et $v_i = v_{i+1}$, pour $i = 1, \cdots, N-1$. On en déduit par récurrence que $v_i = 0$, pour $i = 2, \cdots, N-1$, et donc $v = 0$. ■

2.3 Un problème bidimensionnel

Dans cette section, on va proposer une méthode de calcul des (petits) déplacements de la membrane élastique, soumise à des forces verticales. Dans $\mathbb{R}^2$, on considère une membrane carrée, $D =]0,1[^2$, soumise à une force verticale, de densité surfacique $f(x,y)$, et fixée en sa **frontière** ∂D. On note $u : (x,y) \mapsto u(x,y)$ le déplacement transversal induit par l'application de la force, dont on rappelle qu'il est solution du problème bidimensionnel (2D) *normalisé*

$$-\Delta_2 u = f \text{ sur } D, \quad u = 0 \text{ sur } \partial D. \tag{2.11}$$

Pour discrétiser le problème à l'aide de la méthode des différences finies, on s'inspire très fortement de la méthode employée pour le problème 1D (2.1). En effet, on remarque qu'en 1D on a les égalités $-u'' = -\dfrac{d^2}{dx^2}u = -\Delta_1 u$, *id est* (2.1) est un Laplacien 1D à résoudre !

Si la solution u est de classe $\mathcal{C}^4([0,1]^2)$, on peut écrire l'équivalent de la proposition 2.2.1.

NB. Malheureusement, et contrairement au cas 1D, on n'a plus l'équivalence entre u de classe $\mathcal{C}^4([0,1]^2)$ et f de classe $\mathcal{C}^2([0,1]^2)$...

Proposition 2.3.1 *Soient* $(x,y) \in]0,1[^2$ *et* (h_1, h_2) *tels que* $[x-h_1, x+h_1] \in [0,1]$, *et* $[y-h_2, y+h_2] \in [0,1]$. *Alors*

$\exists (\theta_1, \theta_2) \in]-1,1[^2$ *tels que*

$$-\frac{\partial^2 u}{\partial x^2}(x,y) = \frac{-u(x-h_1,y) + 2u(x,y) - u(x+h_1,y)}{{h_1}^2} + \frac{{h_1}^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\theta_1 h_1, y)$$
$$-\frac{\partial^2 u}{\partial y^2}(x,y) = \frac{-u(x,y-h_2) + 2u(x,y) - u(x,y+h_2)}{{h_2}^2} + \frac{{h_2}^2}{12}\frac{\partial^4 u}{\partial y^4}(x, y+\theta_2 h_2).$$

On en déduit que

$$-\Delta_2 u(x,y) = \frac{-u(x-h_1,y) + 2u(x,y) - u(x+h_1,y)}{{h_1}^2} + \frac{-u(x,y-h_2) + 2u(x,y) - u(x,y+h_2)}{{h_2}^2}$$
$$+\frac{{h_1}^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\theta_1 h_1, y) + \frac{{h_2}^2}{12}\frac{\partial^4 u}{\partial y^4}(x, y+\theta_2 h_2). \tag{2.12}$$

Remarque 2.3.1 *Les deux premiers termes de (2.12) sont une bonne approximation de* $-\Delta_2 u(x,y)$, *sous réserve que* h_1 *et* h_2 *soient* petits. *Le reste est en effet borné par*

$$\frac{1}{12}\left({h_1}^2 + {h_2}^2\right) C_{u,4}, \text{ avec } C_{u,4} = \max\left(\sup_{(x,y)\in[0,1]^2} |\frac{\partial^4 u}{\partial x^4}(x,y)|, \sup_{(x,y)\in[0,1]^2} |\frac{\partial^4 u}{\partial y^4}(x,y)| \right).$$

Dans la suite, on va considérer un pas de discrétisation *identique*[1] selon la direction des x, et celle des y, c'est-à-dire $h = h_1 = h_2$. Chacun des intervalles $[0,1]$ est découpé en $n+1$ intervalles égaux de longueur $h = 1/(n+1)$. Puis on calcule les nombres $(u_{i,j})_{0\leq i,j\leq n+1}$, qui sont les *valeurs approchées* de la solution u aux points d'abscisse $x_i = ih$ et d'ordonnée $y_j = jh$. On note $(f_{i,j})_{1\leq i,j\leq n}$ les valeurs $f_{i,j} = f(x_i, y_j)$, pour i et j variant de 1 à n.

[1] En pratique, il est tout à fait possible de considérer des pas différents selon les directions x et y, voire variables dans chaque direction...

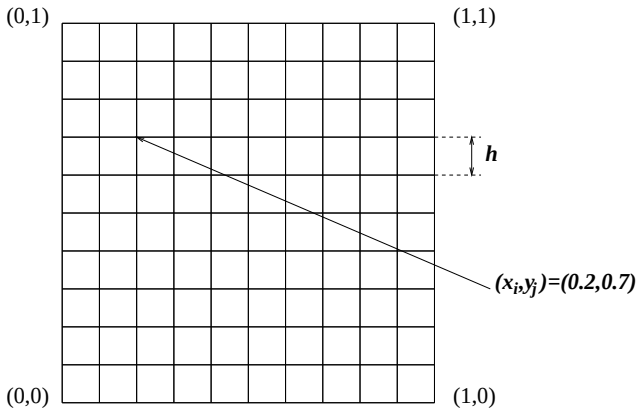


FIG. 2.3 – D et les points de discrétisation ($n = 9$, $h = 0.1$)

Le Laplacien 2D est *approché* par une combinaison linéaire de valeurs $u_{i,j}$, selon le **schéma à cinq points**

$$-\Delta_2 u(x_i, y_j) \approx \frac{-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2}, \quad 1 \leq i, j \leq n. \tag{2.13}$$

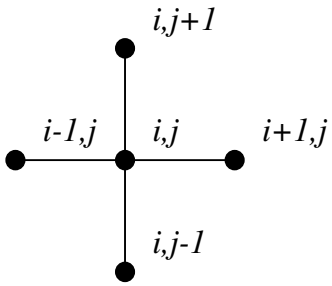


FIG. 2.4 – Le schéma aux différences finies à cinq points

Le problème (2.11) est donc approché de la manière suivante : on remplace la recherche de la fonction u, par la recherche des n^2 valeurs $u_{i,j} \in \mathbb{R}$ qui vérifient les relations

$$\frac{-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f_{i,j}, \quad 1 \leq i, j \leq n. \tag{2.14}$$

Les valeurs de u sur la frontière ∂D, ici $u(0, \cdot)$ et $u(1, \cdot)$, ainsi que $u(\cdot, 0)$ et $u(\cdot, 1)$, sont connues (et égales à zéro). Il en est donc de même pour $u_{0,j}$, $u_{n+1,j}$, $u_{i,0}$ et $u_{i,n+1}$, pour i et j variant de 1 à n. Il reste donc au total $N = n^2$ valeurs à calculer.

On les regroupe n par n, ainsi que les $(f_{i,j})_{i,j}$, en opérant l'*identification* $u_{\cdot,j} = (u_{i,j})_{1 \leq i \leq n}$. Le bloc $u_{\cdot,j}$ appartient à $\mathbb{R}^n$, avec

$$u_{\cdot,j} = \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{n,j} \end{pmatrix}.$$

Il en est de même pour $f_{\cdot,j} \in \mathbb{R}^n$. Ensuite, on pose

$$\vec{u} = \begin{pmatrix} u_{\cdot,1} \\ \vdots \\ u_{\cdot,n} \end{pmatrix} \in \mathbb{R}^N \text{ et } \vec{f} = \begin{pmatrix} f_{\cdot,1} \\ \vdots \\ f_{\cdot,n} \end{pmatrix} \in \mathbb{R}^N.$$

Ainsi, on a *numéroté* les inconnues ligne par ligne, dans le sens croissant, pour les indices i (au sein d'une ligne) et j (numéro de ligne). Les inconnues $(u_{i,j})_{1\leq i,j\leq n}$ sont solutions du système linéaire formé des N relations (2.14). Ce système linéaire peut être écrit sous la forme

$$\mathbb{A}_2\vec{u} = \vec{f}, \tag{2.15}$$

avec $\mathbb{A}_2$ une matrice de $\mathbb{R}^{N\times N}$. Si l'on s'intéresse à sa structure interne, on vérifie facilement que l'on peut écrire

$$\mathbb{A}_2 = \frac{1}{h^2}\begin{pmatrix} \mathbb{B}_1 & T & & & \\ T & \mathbb{B}_1 & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & \mathbb{B}_1 & T \\ & & & T & \mathbb{B}_1 \end{pmatrix}. \tag{2.16}$$

Ci-dessus, les blocs autres que $\mathbb{B}_1$ et T sont nuls et, par ailleurs, $T = -I_n$, avec I_n la matrice identité d'ordre n, et $\mathbb{B}_1 \in \mathbb{R}^{n\times n}$ est la matrice tridiagonale définie par

$$\mathbb{B}_1 = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} = 2I_n + h^2\mathbb{A}_1.$$

La matrice $\mathbb{A}_2$ est donc pentadiagonale par points (i. e. avec tous les éléments non nuls regroupés sur cinq diagonales), et tridiagonale par blocs, *lorsque* la numérotation est celle indiquée ci-dessus : ligne par ligne (j croissant), et i croissant au sein d'une ligne.

Récapitulons. Si on note avec un seul indice les composantes de $\vec{u}$, c'est-à-dire $(u_I)_{1\leq I\leq N}$, on a les correspondances :

$$\text{composante } _I \equiv \ _{i,j} \quad \Longleftrightarrow \quad I = i + (j-1)n \text{ ou } \begin{cases} i = (I-1) \mod n + 1 \\ j = \lfloor (I-1)/n \rfloor + 1 \end{cases}. \tag{2.17}$$

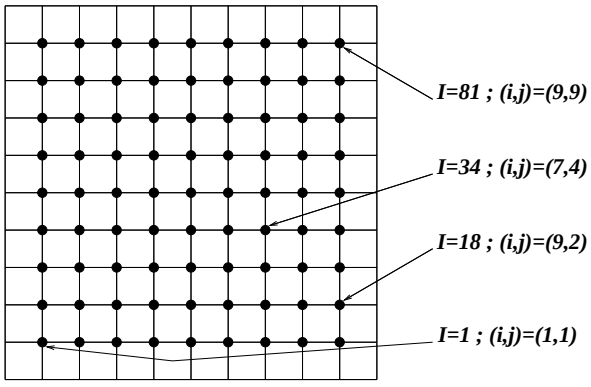


FIG. 2.5 – Les deux numérotations : $I \in \{1,\cdots,81\}$ et $i,j \in \{1,\cdots,9\}$

Nous allons maintenant étudier les propriétés de la matrice $\mathbb{A}_2$. Nous allons d'abord établir qu'elle est monotone, puis qu'elle est symétrique définie-positive, comme $\mathbb{A}_1$.

Théorème 2.3.1 *La matrice $\mathbb{A}_2$ des systèmes linéaires (2.15) et (2.16) est monotone.*

Preuve : Soit donc $v \in \mathbb{R}^N$ tel que $\mathbb{A}_2 v \geq 0$. Composante par composante (avec le double indiçage $_{i,j}$), ceci signifie

$$4v_{1,1} - v_{2,1} - v_{1,2} \geq 0 \tag{2.18}$$
$$4v_{n,1} - v_{n-1,1} - v_{n,2} \geq 0 \tag{2.19}$$
$$4v_{1,n} - v_{1,n-1} - v_{2,n} \geq 0 \tag{2.20}$$
$$4v_{n,n} - v_{n,n-1} - v_{n-1,n} \geq 0 \tag{2.21}$$
$$4v_{i,1} - v_{i-1,1} - v_{i+1,1} - v_{i,2} \geq 0 \quad 2 \leq i \leq n-1 \tag{2.22}$$
$$4v_{1,j} - v_{1,j-1} - v_{2,j} - v_{1,j+1} \geq 0 \quad 2 \leq j \leq n-1 \tag{2.23}$$
$$4v_{n,j} - v_{n,j-1} - v_{n-1,j} - v_{n,j+1} \geq 0 \quad 2 \leq j \leq n-1 \tag{2.24}$$
$$4v_{i,n} - v_{i,n-1} - v_{i-1,n} - v_{i+1,n} \geq 0 \quad 2 \leq i \leq n-1 \tag{2.25}$$
$$4v_{i,j} - v_{i,j-1} - v_{i-1,j} - v_{i+1,j} - v_{i,j+1} \geq 0 \quad 2 \leq i,j \leq n-1. \tag{2.26}$$

Ci-dessus, on a isolé des lignes de la matrice $\mathbb{A}_2$ correspondant respectivement

1. En (2.18)-(2.21) :
   aux *coins* de la grille, d'indices (i,j) parmi $\{(1,1),(n,1),(1,n),(n,n)\}$.
2. En (2.22)-(2.25) :
   au *bord* de la grille, coins exclus, d'indices (i,j) avec $i \in \{1,n\}$ ou (exclusif) $j \in \{1,n\}$.
3. En (2.26) :
   aux *points internes* de la grille, d'indices (i,j) avec $i,j \in \{2,\cdots,n-1\}$.

Soit maintenant $v_{min} = \min_{i,j} v_{i,j}$. On veut prouver que $v_{min} \geq 0$, pour pouvoir conclure grâce à la proposition 2.2.2. Comme la grille comprend des points aux coins, sur le bord (coins exclus) et intérieurs, on traite les trois cas correspondants.

1. Si v_{min} correspond à un coin (par exemple $v_{min} = v_{1,1}$), (2.18) fournit l'inégalité

   $$2v_{min} \geq (v_{2,1} - v_{min}) + (v_{1,2} - v_{min}) \geq 0.$$

2. Si v_{min} correspond à un point du bord différent d'un coin (par exemple $v_{min} = v_{i,1}$, $i \in \{2,\cdots,n-1\}$ donné), (2.22) fournit l'inégalité

   $$v_{min} \geq (v_{i-1,1} - v_{min}) + (v_{i+1,1} - v_{min}) + (v_{i,2} - v_{min}) \geq 0.$$

3. Si v_{min} correspond à un point intérieur, (2.26) fournit l'inégalité

   $$(v_{min} - v_{i,j-1}) + (v_{min} - v_{i-1,j}) + (v_{min} - v_{i+1,j}) + (v_{min} - v_{i,j+1}) \geq 0.$$

   Pour conclure dans le cas 3., il faut se ramener aux cas 1. ou 2. Or, d'après la définition de v_{min}, chacun des quatre termes entre parenthèses est négatif ou nul. Il sont donc tous nuls, c'est-à-dire

   $$v_{min} = v_{i,j-1} = v_{i-1,j} = v_{i+1,j} = v_{i,j+1}.$$

   Comme dans le cas 1D (cf. la preuve de la proposition 2.2.3), on peut raisonner par récurrence (sur i et sur j), pour trouver finalement que $v_{i,j} = v_{min}$ en tous les points, coins *exceptés*. Mais ceci est suffisant pour conclure, car les points du bord (hors coins) sont compris, et le cas 2. s'applique.

■

Pour établir le caractère défini-positif de $\mathbb{A}_2$, nous allons reprendre l'écriture par blocs $n \times n$ de celle-ci (2.16).

Proposition 2.3.2 *La matrice $\mathbb{A}_2$ du système linéaire (2.15) est symétrique définie-positive.*

Preuve : Pour commencer, la matrice $\mathbb{A}_2$ est symétrique par construction.
On forme, pour $v \in \mathbb{R}^N$, $h^2(v, \mathbb{A}_2 v)$:

$$\begin{aligned}
h^2(v,\mathbb{A}_2 v) &= \left(\begin{pmatrix} \mathbb{B}_1 & -I_n & & & \\ -I_n & \mathbb{B}_1 & -I_n & & \\ & \ddots & \ddots & \ddots & \\ & & -I_n & \mathbb{B}_1 & -I_n \\ & & & -I_n & \mathbb{B}_1 \end{pmatrix} \begin{pmatrix} v_{\cdot,1} \\ \vdots \\ v_{\cdot,n} \end{pmatrix}, \begin{pmatrix} v_{\cdot,1} \\ \vdots \\ v_{\cdot,n} \end{pmatrix} \right) \\
&= (\mathbb{B}_1 v_{\cdot,1} - v_{\cdot,2}, v_{\cdot,1})_n + \sum_{2\le j\le n-1} (-v_{\cdot,j-1} + \mathbb{B}_1 v_{\cdot,j} - v_{\cdot,j+1}, v_{\cdot,j})_n + (-v_{\cdot,n-1} + \mathbb{B}_1 v_{\cdot,n}, v_{\cdot,n})_n \\
&= \sum_{1\le j\le n} (\mathbb{B}_1 v_{\cdot,j}, v_{\cdot,j})_n - 2 \sum_{1\le j\le n-1} (v_{\cdot,j}, v_{\cdot,j+1})_n \\
&= \sum_{1\le j\le n} \big(2\|v_{\cdot,j}\|_n^2 + h^2(\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n\big) - 2 \sum_{1\le j\le n-1} (v_{\cdot,j}, v_{\cdot,j+1})_n \\
&= \|v_{\cdot,1}\|_n^2 + \|v_{\cdot,n}\|_n^2 + h^2 \sum_{1\le j\le n} (\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n + \sum_{1\le j\le n-1} \|v_{\cdot,j} - v_{\cdot,j+1}\|_n^2.
\end{aligned}$$

Ci-dessus, on a noté $(\cdot,\cdot)_n$ le produit scalaire usuel de $\mathbb{R}^n$, et $\|\cdot\|_n$ la norme associée. Comme $\mathbb{A}_1$ est définie-positive, on en déduit immédiatement que $h^2(v, \mathbb{A}_2 v)$ est positif ou nul, puisque c'est une somme de termes positifs. Si on suppose que $h^2(v, \mathbb{A}_2 v) = 0$, on trouve $\sum_{1\le j\le n} (\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n = 0$, ce qui implique $v_{\cdot,j} = 0$ pour $1 \le j \le n$, soit finalement $v = 0$. ∎

Malheureusement, on ne peut pas aller plus loin, c'est-à-dire estimer la précision de l'approximation obtenue grâce au schéma à cinq points, en continuant à suivre la méthode pour le cas 1D[2]. Cependant, à l'aide d'autres techniques relativement lourdes à mettre en œuvre (cf. [6]), on peut malgré tout prouver le

Théorème 2.3.2 *Lorsque la solution u est de classe $\mathcal{C}^4([0,1]^2)$, l'erreur est telle que*

$$\|\vec{e}\|_\infty \le C\, h^2 (C_{u,4} + h\, C_{u,3}), \text{ avec } C_{u,3} = \max \left(\sup_{(x,y)\in[0,1]^2} |\frac{\partial^3 u}{\partial x^3}(x,y)|, \sup_{(x,y)\in[0,1]^2} |\frac{\partial^3 u}{\partial y^3}(x,y)| \right),$$

où C est une constante qui est indépendante de u et de h.

Bref, *"tout est bien qui finit bien"* (librement traduit de [13]...).

Notons au passage que comme $\mathbb{A}_2$ est symétrique définie-positive, on peut utiliser, pour résoudre (2.15), la méthode de Cholesky (chapitre 3), ou bien les méthodes de Jacobi ou de Gauss-Seidel, puisque $\mathbb{A}_2$ est tridiagonale par blocs $n \times n$ (chapitre 4).

Pour conclure cette section, considérons un problème 2D plus général, au sens où la condition aux limites, c'est-à-dire la valeur prise par la solution sur la frontière du domaine, notée ci-dessous g (c'est une *seconde donnée*, après f), n'est plus nulle. On doit alors résoudre

$$-\Delta_2 u = f \text{ sur } D, \quad u = g \text{ sur } \partial D. \tag{2.27}$$

Si l'on construit une approximation à l'aide des différences finies, on démontre cette fois encore que les inconnues $(u_{i,j})_{1\le i,j\le n}$ sont solutions d'un système linéaire formé de N relations,

[2] En effet, si $f \equiv 1$, quelle est la solution du problème (2.11), quelle est sa régularité ?

qui s'écrit

$$\mathbb{A}_2\vec{u} = \vec{f} + \frac{1}{h^2}\begin{pmatrix} g(\cdot,0) \\ \cdot \\ \cdot \\ \cdot \\ g(\cdot,1) \end{pmatrix}. \tag{2.28}$$

Le dernier vecteur regroupe l'ensemble des valeurs prises par u sur la frontière ∂D, qui sont connues (car égales à la valeur prise par g au même point) : ce vecteur est donc une *donnée* du problème et, à ce titre, il se trouve à droite du signe égal. Par ailleurs, il faut que i ou j soit égal à 1 ou n, pour que sa composante $_{i,j}$ soit le cas échéant non nulle ; en effet, dans ce cas, cf. les figures 2.3 et 2.4, un des points voisins au moins se trouve sur la frontière ∂D. Au contraire, si i et j appartiennent tous les deux à $\{2, \cdots, n-1\}$, tous les voisins se trouvent dans D, et sa composante $_{i,j}$ est nécessairement nulle.

Exercice 2.3.1 *Etablir un* principe de positivité *pour le problème (2.28), homologue* discret *du principe général énoncé au théorème 1.3.1.*

2.4 Un problème tridimensionnel

Dans cette section, on propose une méthode de calcul numérique du potentiel électrostatique, toujours basée sur les différences finies, sous la forme d'un exercice. Précisons tout d'abord le modèle. Dans $\mathbb{R}^3$, on considère une cavité cubique, $C =]0,1[^3$, dans laquelle on a fait le vide. On suppose qu'elle est entourée d'un conducteur parfait, et que le potentiel électrostatique, sur sa frontière ∂C, est nul. Enfin, on place des charges dans la cavité. On rappelle que si ρ est la densité de charge et ϵ_0 la permittivité électrique, le potentiel électrostatique u généré par les charges dans la cavité est solution de l'équation dite de Coulomb, qui s'écrit

$$-\Delta_3 u = \frac{\rho}{\epsilon_0} \text{ sur } C, \quad u = 0 \text{ sur } \partial C. \tag{2.29}$$

Nous proposons pour finir les deux exercices ci-dessous.

Exercice 2.4.1 *Suggérer une généralisation de l'approche du problème 2D au cas 3D, pour résoudre le problème du potentiel électrostatique (2.29).*
1. Quel sera a priori *le nombre de points du schéma aux différences finies, dans ce cas ? Le construire, en le justifiant.*
2. Examiner en détail les propriétés de la matrice $\mathbb{A}_3$ ainsi obtenue. Etablir successivement que :

- *$\mathbb{A}_3$ est monotone ;*
- *$\mathbb{A}_3$ est symétrique ;*
- *$\mathbb{A}_3$ est définie-positive.*

3. En déduire un principe de positivité discret associé à $\mathbb{A}_3$.

Exercice 2.4.2 *Suggérer une généralisation au cas d-dimensionnel ($d \geq 4$), pour résoudre le problème ci-dessous, posé dans $\Omega_d =]0,1[^d$:*

$$-\Delta_d u = f \text{ sur } \Omega_d, \quad u = g \text{ sur } \partial\Omega_d, \qquad \text{où } \Delta_d = \sum_{k=1}^{d} \frac{\partial^2}{\partial {x_k}^2}. \tag{2.30}$$

1. Quel sera a priori *le nombre de points du schéma aux différences finies, dans ce cas ? Le construire, en le justifiant.*
2. Examiner en détail les propriétés de la matrice $\mathbb{A}_d$ ainsi obtenue. Etablir successivement que :

- $\mathbb{A}_d$ *est monotone ;*
- $\mathbb{A}_d$ *est symétrique ;*
- $\mathbb{A}_d$ *est définie-positive.*

(On pourra raisonner par récurrence.)
3. En déduire un principe de positivité discret associé à $\mathbb{A}_d$.

2.5 Problèmes dépendant du temps

Dans cette section, nous considérons tout d'abord un problème instationnaire 1d, à savoir celui de la corde vibrante, fixée en ses deux extrémités. On souhaite calculer numériquement les déplacements verticaux, à partir d'une position initiale connue (à $t = 0$), jusqu'à un instant final $T > 0$. Si on suppose que la corde est de longueur L, on a vu que u vérifie

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } x \in]0, L[\text{ et } t \in]0, T[, \tag{2.31}$$

$$u(0,t) = u(L,t) = 0 \text{ pour } t \in]0, T[, \tag{2.32}$$

$$u(x,0) = u^0(x) \text{ et } \frac{\partial u}{\partial t}(x,0) = u^1(x) \text{ pour } x \in]0, L[. \tag{2.33}$$

La solution dépend de deux variables, l'une spatiale x, et l'autre temporelle t, et l'équation (2.31) met en jeu deux dérivées partielles secondes, l'une par rapport à x, et l'autre par rapport à t. Il est donc naturel de considérer une approximation par différences finies, pour chacune des deux dérivées. Par ailleurs, la seconde condition initiale comprend une dérivée première par rapport au temps, que nous allons également approcher par différences finies. Enfin, pour avoir accès aux valeurs en $(x,t) \in \{(0,0),(L,0),(0,T),(L,T)\}$, nous allons supposer que d'une part (2.32) est valable pour $t = T$, et d'autre part que la première partie de (2.33) est valable en $x = 0$ et en $x = L$, avec $u^0(0) = u^0(L) = 0$.
Dans la suite, on introduit un pas de discrétisation spatial, à savoir $h_x = L/(n_x + 1)$, et un pas de discrétisation temporel *a priori* différent, $h_t = T/(n_t + 1)$. L'intervalle $[0, L]$ est donc découpé en $n_x + 1$ intervalles, et $[0, T]$ est découpé en $n_t + 1$ intervalles. Pour bien différencier les discrétisations en espace et en temps, on note les *valeurs approchées* de la solution u aux "points" d'abscisse $x_i = ih_x$ et d'ordonnée $t_m = mh_t$ par $(u_i^m)_{0\le i\le n_x+1, 0\le m\le n_t+1}$. On arrive donc aux approximations :

$$\frac{\partial^2 u}{\partial t^2}(x_i,t_m) \approx \frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{h_t^2},\ 1 \le i \le n_x,\ 1 \le m \le n_t, \tag{2.34}$$

$$\frac{\partial^2 u}{\partial x^2}(x_i,t_m) \approx \frac{u_{i+1}^m - 2u_i^m + u_{i-1}^m}{h_x^2},\ 1 \le i \le n_x,\ 1 \le m \le n_t, \tag{2.35}$$

$$\frac{\partial u}{\partial t}(x_i,0) \approx \frac{u_i^1 - u_i^0}{h_t},\ 1 \le i \le n_x. \tag{2.36}$$

A partir de là, il est aisé de construire l'ensemble des équations vérifiées par les inconnues $(u_i^m)_{0\le i\le n_x+1, 0\le m\le n_t+1}$. Commençons par l'approximation de (2.31) aux "points" *intérieurs* de discrétisation $(x_i,t_m)_{1\le i\le n_x, 1\le m\le n_t}$, c'est-à-dire dans l'*ouvert* $]0, L[\times]0, T[$. Ainsi, on trouve :

$$\frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{h_t^2} - c^2 \frac{u_{i+1}^m - 2u_i^m + u_{i-1}^m}{h_x^2} = 0,\ 1 \le i \le n_x,\ 1 \le m \le n_t. \tag{2.37}$$

Les conditions aux limites (2.32) permettent de déterminer les valeurs extrémales *en espace*, c'est-à-dire pour $i = 0$ et $i = n_x + 1$, ce qui correspond aux extrémités du domaine de calcul spatial $x_0 = 0$ et $x_{n_x+1} = L$. Par identification, on en déduit

$$u_0^m = u_{n_x+1}^m = 0,\ 1 \le m \le n_t + 1. \tag{2.38}$$

Enfin, les deux conditions initiales permettent de calculer les valeurs approchées aux deux premiers instant, à savoir en t_0 et t_1, par l'intermédiaire de

$$u_i^0 = u^0(x_i),\ 0 \leq i \leq n_x + 1, \tag{2.39}$$

$$\frac{u_i^1 - u_i^0}{h_t} = u^1(x_i),\ 1 \leq i \leq n_x. \tag{2.40}$$

Si on regroupe (2.37-2.40), on aboutit après réorganisation à

$$m = 0\ : \quad \left\{ \begin{array}{l} u_i^0 = u^0(x_i),\ \text{pour } 1 \leq i \leq n_x \\ u_0^0 = u_{n_x+1}^0 = 0 \end{array} \right. ; \tag{2.41}$$

$$m = 1\ : \quad \left\{ \begin{array}{l} u_i^1 = u^0(x_i) + h_t u^1(x_i),\ \text{pour } 1 \leq i \leq n_x \\ u_0^1 = u_{n_x+1}^1 = 0 \end{array} \right. ; \tag{2.42}$$

Pour $m = 1, \cdots, n_t$

$$\left\{ \begin{array}{l} u_i^{m+1} = \dfrac{c^2 h_t^2}{h_x^2}(u_{i+1}^m - 2u_i^m + u_{i-1}^m) + 2u_i^m - u_i^{m-1},\ \text{pour } 1 \leq i \leq n_x \\ u_0^{m+1} = u_{n_x+1}^{m+1} = 0 \end{array} \right. . \tag{2.43}$$

Que constate-t-on ?
Tout d'abord, on calcule les valeurs de la solution approchée en *incrémentant en temps*, en commençant par l'instant $t_0 = 0$, en poursuivant par t_1, puis t_2, $\cdots$, jusqu'à l'instant final $t_{n_t+1} = T$. La valeur finale est bien une inconnue du problème, comme suggéré par la remarque 1.2.1 sur la solution exacte.
Par ailleurs, si on s'intéresse à la dépendance entre les données, on constate que u_i^{m+1} dépend de $(u_{i_1}^m)_{i-1 \leq i_1 \leq i+1}$ et de u_i^{m-1}. Ces valeurs dépendent elles-mêmes de $(u_{i_2}^{m-1})_{i-2 \leq i_2 \leq i+2}$, de $(u_{i_1}^{m-2})_{i-1 \leq i_1 \leq i+1}$ et de u_i^{m-3}, et ainsi de suite... Le schéma ci-dessous résume la situation, où l'on a représenté le **cône de dépendance discret** de la solution calculée.

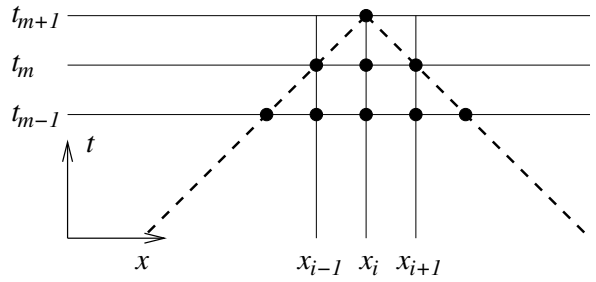


FIG. 2.6 – Cône de dépendance discret

D'un point de vue algorithmique, il est intéressant d'introduire la suite de vecteurs $(\vec{v}^m)_{0 \leq m \leq n_t+1}$ de $\mathbb{R}^{n_x}$, contenant les approximations spatiales successives sur $]0, L[$, calculées à l'aide de (2.41-2.43) : $(\vec{v}^m)_i = u_i^m$, pour $1 \leq i \leq n_x$, et $0 \leq m \leq n_t + 1$. Comme dans le cas statique, il n'est pas nécessaire de stocker explicitement les valeurs aux extrémités $x = 0$ et $x = L$, qui sont connues à tout instant. On peut alors réécrire (2.43) sous la *forme vectorielle équivalente*, pour $1 \leq m \leq n_t$:

$$\vec{v}^{m+1} = (2I_{n_x} + (ch_t)^2 \mathbb{A}_1')\vec{v}^m - \vec{v}^{m-1},\ \mathbb{A}_1' = \frac{1}{(h_x)^2} \begin{pmatrix} 2 & -1 & \dots & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & \dots & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n_x \times n_x}. \tag{2.44}$$

De cette façon, on retrouve bien une forme vectorielle de type 1d, semblable à la formulation vectorielle statique 1d (2.5). Par contre, on obtient un **schéma explicite**, au sens où il n'est pas nécessaire de résoudre un système linéaire pour déterminer la solution approchée. Enfin, les matrices $\mathbb{A}_1$ et $\mathbb{A}_1'$ diffèrent par un coefficient multiplicatif...

Dans la suite, nous n'étudions pas dans les détails la convergence de la solution discrète (notée u_{app}) vers la solution exacte u. Nous nous contentons, pour fixer les idées, d'une définition, avec une mesure de l'erreur par l'intermédiaire d'une norme *abstraite* ($\|\cdot\|$.)

Définition 2.5.1 *Un schéma est dit* **convergent** *pour la norme $\|\cdot\|$ si, et seulement si, pour toute donnée initiale (u^0, u^1), on a la propriété*

$$\|u_{app} - u\| \to 0, \textit{ lorsque } h_x, h_t \to 0,$$

le rapport h_t/h_x étant fixé.

NB. Dans la définition ci-dessus, on remarque que les pas de discrétisation h_x et h_t sont liés entre eux.

Cette étude de la convergence repose usuellement (cf. [2]) sur deux ingrédients : la **consistance**, et la **stabilité** du schéma. Nous allons uniquement évoquer la notion de stabilité, à l'aide d'un calcul simple. Si on suppose que le domaine spatial est égal à $\mathbb{R}$, alors on peut vérifier que la solution exacte est de la forme

$$u(x,t) = \frac{1}{2}(u^0(x+ct) + u^0(x-ct)) + \frac{1}{2}(v^1(x+ct) - v^1(x-ct)),\ v^1(t) = \int_0^t u^1(s)ds. \quad (2.45)$$

Remarque 2.5.1 *La vitesse de propagation de l'information, associée à (2.45), est égale à $\pm c$. Les ondes se propagent donc à la vitesse c (en module) sur la corde. Cette propriété reste valable pour les problèmes hyperboliques en dimension 2 ou 3 d'espace (cas 2d, 3d instationnaires hyperboliques.)*

En particulier, la valeur de u au "point" (x,t) dépend de
- la valeur de u^0 en $x \pm ct$;
- la valeur de v^1 en $x \pm ct$, c'est-à-dire la valeur de u^1 sur $[x-ct, x+ct]$.

On peut donc, à l'instar du cas discret, définir le **cône de dépendance** de la solution exacte.

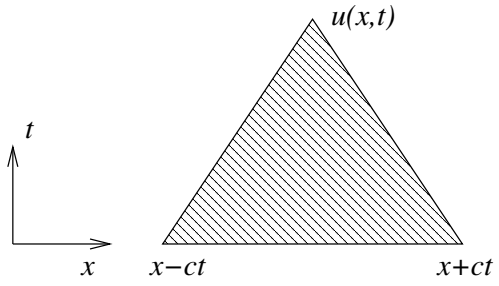


FIG. 2.7 – Cône de dépendance

Si maintenant on revient à la question de la stabilité, il est clair qu'une *condition nécessaire* de stabilité du schéma est que le cône de dépendance discret contienne *suffisamment* d'informations : en d'autres termes, il doit contenir le cône de convergence associé à la solution exacte !

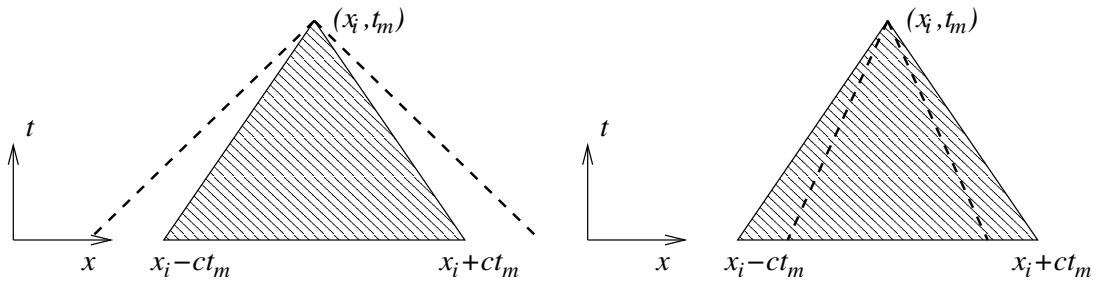


FIG. 2.8 – Stabilité ou instabilité

Si l'on compare les pentes, à savoir h_t/h_x pour le schéma discret et $1/c$ pour la solution exacte, la condition nécessaire de stabilité du schéma explicite se résume à

$$ch_t \leq h_x. \tag{2.46}$$

On parle habituellement de **condition CFL**, pour Courant, Lax et Friedrichs.

Il est possible (cf. [2]) de prouver que la condition CFL (2.46) est en fait une *condition nécessaire et suffisante* de stabilité du schéma explicite. Pour s'affranchir d'une condition de stabilité, on peut introduire des **schémas implicites** (voir les séances de Travaux Pratiques.) Bien sûr, ce type d'approximation à l'aide des différences finies, résultant en des schémas explicites ou implicites, est utilisable *a priori* pour les problèmes instationnaires 2d et 3d, hyperboliques ou paraboliques, présentés au chapitre 1.

Pour conclure cette section, nous abordons brièvement la discrétisation des problèmes aux valeurs propres, puis celle des problèmes stationnaires. Reprenons le cas de la membrane élastique $\Omega_2 =]0,1[^2$.

Le problème aux valeurs propres s'écrit : trouver les couples modes propres[3] - valeurs propres $(x,y) \mapsto u_k(x,y)$ et $\lambda_k > 0$ tels que

$$-c^2\Delta_2 u_k = \lambda_k u_k \text{ sur } \Omega_2, \quad u_k = 0 \text{ sur } \partial\Omega_2. \tag{2.47}$$

(Ci-dessus, k est un entier naturel non nul par convention.)
Une fois le pas de discrétisation h fixé ($h = 1/(n+1)$, $N = n^2$), on en déduit l'approximation par différences finies : trouver les couples modes propres discrets - valeurs propres discrètes $\vec{u}_l \in \mathbb{R}^N$ et λ_l tels que

$$c^2\mathbb{A}_2\vec{u}_l = \lambda_l\vec{u}_l. \tag{2.48}$$

Comme $\mathbb{A}_2$ est définie-positive (proposition 2.3.2), on en déduit immédiatement que toute valeur propre discrète λ_l est strictement positive. En effet :

$$\lambda_l\|u_l\|_2^2 = \lambda_l(u_l,u_l) = c^2(\mathbb{A}_2u_l,u_l) > 0.$$

Des méthodes de calcul des valeurs propres et vecteurs propres d'une matrice sont présentées au chapitre 5. Ceci répond à la question de la résolution du problème discret. Comment peut-on vérifier que celui-ci est une bonne approximation du problème initial ? La question étant complexe, nous nous contentons de mettre en avant la problématique associée... Les valeurs propres discrètes étant en nombre fini d'après un résultat classique (proposition A.2.1), il est clair qu'on ne peut espérer approcher tous les couples $(u_k,\lambda_k)_{k\in\mathbb{N}}$, pour h (et donc N) fixés ! Néanmoins,

[3] On rappelle qu'un mode propre n'est pas identiquement nul.

le nombre de modes propres discrets croît comme N, et tend bien vers l'infini lorsque $h \to 0$... Ensuite, concernant la convergence des modes discrets vers les modes du problème initial, nous sommes confrontés à deux difficultés : identifier vers quel mode propre une suite de modes propres discrets (indexée par N) converge, et s'assurer qu'à la limite $N \to +\infty$, on atteint tous les modes propres.

Nous finissons par les problèmes stationnaires du type (pour $\nu > 0$ donné) : trouver $(x, y) \mapsto u(x, y)$ tel que

$$-\nu^2 u - c^2 \Delta_2 u = g \text{ sur } \Omega_2, \quad u = 0 \text{ sur } \partial\Omega_2. \tag{2.49}$$

Le pas h étant fixé ($h = 1/(n+1)$, $N = n^2$), l'approximation par différences finies consiste en le système linéaire

$$\mathbb{A}_{2,\nu}\vec{u} = \vec{u}, \text{ avec } \mathbb{A}_{2,\nu} = c^2\mathbb{A}_2 - \nu^2 I_N. \tag{2.50}$$

On retrouve là un système linéaire à résoudre, dont la matrice $\mathbb{A}_{2,\nu}$ est symétrique. Malheureusement, lorsque ν est suffisamment grand, cette matrice n'est plus défnie-positive. Il est donc nécessaire d'utiliser une méthode de type factorisation de Gauss (chapitre 3), pour laquelle des problèmes de stabilité numérique peuvent se produirent. Quant à la question de la convergence, elle est également complexe, étant étroitement reliée à celle de la convergence du problème aux valeurs propres.

Chapitre 3

Les méthodes directes

3.1 Introduction

Le calcul de la solution d'un système linéaire d'ordre 20 par les formules de Cramer est impossible à réaliser sur un ordinateur, car il requiert environ 10^{21} opérations élémentaires $(+, -, *, /)$ [1].

En conséquence, il est nécessaire de déterminer d'autres méthodes, qui soient utilisables en pratique, par exemple pour résoudre les systèmes linéaires du chapitre 2.
On montre pour commencer que si la matrice d'un système linéaire est diagonale ou de forme triangulaire, ceci apporte une simplification importante dans le calcul explicite de la solution de ce système. Comment utiliser cette particularité pour traiter le cas général ? Une première approche conduit à la méthode dite d'élimination, une seconde à la méthode de factorisation. Ces méthodes sont décrites dans les paragraphes suivants. Les algorithmes classiques qui en découlent sont appelés **méthodes directes** ; les méthodes de Gauss, Crout et Cholesky font partie de cet ensemble d'algorithmes, leur étude fait l'objet de ce chapitre. Le dernier paragraphe constitue une introduction à l'utilisation pratique de ces méthodes pour la résolution de systèmes linéaires à matrice creuse, tels que ceux obtenus à partir de la discrétisation par différences finies (cf. chapitre 2.)

3.2 Systèmes linéaires simples à résoudre

3.2.1 Système linéaire à matrice diagonale

Définition 3.2.1 *On appelle matrice* **diagonale** *une matrice $D \in \mathbb{R}^{n\times n}$ (D comme Diagonale), telle que $D_{i,j} = 0$ pour tout couple (i, j) tel que $i \neq j$.*

Proposition 3.2.1 *Le déterminant d'une matrice diagonale est égal au produit des coefficients diagonaux :*

$$dét\,(D) = \prod_{i=1}^{n} D_{i,i}.$$

[1] En effet, pour calculer un déterminant d'ordre n il faut faire la somme de $n!$ produits de n facteurs, soit au total $n \times n!$ opérations :

$$\text{dét}\,(A) = \sum_{\sigma\in\mathcal{S}(n)} \varepsilon(\sigma) A_{1,\sigma(1)}\, A_{2,\sigma(2)} \dots A_{n,\sigma(n)}.$$

Le coût calcul des n composantes du vecteur x, solution du système linéaire $Ax = b$ par les formules de Cramer, est donc de $n(n + 1) \times n!$ opérations. Ainsi pour la résolution d'un système linéaire d'ordre 10 (respectivement 20), environ 4.10^8 opérations (resp. 10^{21} opérations) sont nécessaires pour calculer la solution par les formules de Cramer.

Proposition 3.2.2 *Soit $D \in \mathbb{R}^{n\times n}$ une matrice diagonale inversible, la solution du système linéaire $Dy = b$ est obtenue par les formules :*

pour $i = 1, \ldots, n$ **faire**

$y_i = b_i / D_{i,i}$.

fin

Coût calcul : dans le cas particulier d'un système linéaire à matrice diagonale, la Proposition 3.2.1 montre que le calcul du déterminant requiert n multiplications. D'après la Proposition 3.2.2, pour un système linéaire à matrice diagonale, $Dy = b$, le coût calcul des n composantes du vecteur y est également de n opérations, ce qui est extrêmement raisonnable !

3.2.2 Système linéaire à matrice triangulaire

Définition 3.2.2 *On appelle matrice* **triangulaire inférieure** *une matrice $L \in \mathbb{R}^{n\times n}$ (L comme Lower), telle que $L_{i,j} = 0$ pour tout couple (i, j) tel que $i < j$.*
On appelle matrice **triangulaire supérieure** *une matrice $U \in \mathbb{R}^{n\times n}$ (U comme Upper), telle que $U_{i,j} = 0$ pour tout couple (i, j) tel que $j < i$.*

$$L = \begin{bmatrix} \mathrm{x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{bmatrix} \quad \text{et} \quad U = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{x} \end{bmatrix}$$

Proposition 3.2.3 *Le déterminant d'une matrice triangulaire (supérieure ou inférieure) est égal au produit des coefficients diagonaux :*

$$dét\,(T) = \prod_{i=1}^{n} T_{i,i}.$$

Proposition 3.2.4 *Soit $L \in \mathbb{R}^{n\times n}$ une matrice triangulaire inférieure inversible, la solution du système linéaire $Ly = b$ est obtenue par les formules :*

pour $i = 1, \ldots, n$ **faire**

$y_i = [b_i - \sum_{j<i} L_{i,j} y_j] / L_{i,i}$.

fin

Soit $U \in \mathbb{R}^{n\times n}$ une matrice triangulaire supérieure inversible, la solution du système linéaire $Ux = y$ est obtenue par les formules :

pour $i = n, \ldots, 1$ **faire**

$x_i = [y_i - \sum_{j>i} U_{i,j} x_j] / U_{i,i}$.

fin

Pour une matrice triangulaire supérieure, ces relations montrent que l'on peut calculer le vecteur y de proche en proche, en commençant par la dernière composante y_n ; on dit alors que l'on résout le système linéaire **en remontant**. Pour le système linéaire à matrice triangulaire inférieure, on calcule également le vecteur x de proche en proche, en commençant par la première composante x_1 ; on dit que l'on résout le système linéaire **en descendant**.

Remarque 3.2.1 *Noter que l'hypothèse d'inversibilité entraîne que les coefficients diagonaux de deux matrices sont tous différents de zéro, puisque*

$$dét\,(L) = \prod_{i=1}^{n} L_{i,i} \quad et \quad dét\,(U) = \prod_{i=1}^{n} U_{i,i}.$$

Coût calcul : dans le cas particulier d'un système linéaire à matrice triangulaire T, la Proposition 3.2.3 montre que le calcul du déterminant de T nécessite n multiplications. D'après les formules de la Proposition 3.2.4, pour un système linéaire à matrice triangulaire inférieure $Ly = b$, le calcul de la composante y_i requiert :

- pour $i = 1$: 1 division ;
- pour $i > 1$: 1 soustraction, $(i-1)$ multiplications, $(i-2)$ additions, 1 division.

Soit un total de $(2i-1)$ opérations élémentaires $(+, -, *, /)$. Le coût calcul des n composantes du vecteur y, est donc égal à

$$\sum_{i=1}^{i=n} (2i-1) = 2\frac{n(n+1)}{2} - n = n^2 \text{ opérations},$$

ce qui reste raisonnable ! On obtient le même coût pour un système linéaire à matrice triangulaire supérieure $Ux = y$.

3.2.3 Conclusion

Ces propriétés sur les matrices diagonales ou triangulaires nous conduisent naturellement à construire d'autres approches de la résolution des systèmes linéaires que celle basée sur les formules de Cramer : on essaiera de se ramener au cas de matrices diagonales ou triangulaires.

3.3 Partition des matrices en blocs

Les notions précédentes concernent une approche des matrices coefficient par coefficient ; il est souvent utile d'effectuer une partition de la matrice A en $p \times p$ blocs, pour écrire formellement

$$\begin{bmatrix} [A]_{1,1} & [A]_{1,2} & \dots & [A]_{1,p} \\ [A]_{2,1} & [A]_{2,2} & \ddots & [A]_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ [A]_{p,1} & [A]_{p,2} & \dots & [A]_{p,p} \end{bmatrix}.$$

Dans cette écriture les $[A]_{i,j}$ sont p^2 matrices de $\mathbb{R}^{m_i \times n_j}$ appelées **blocs**, dont seuls les p blocs diagonaux sont nécessairement carrés ($m_i = n_i$, $1 \leq i \leq p$). Les définitions précédentes se généralisent naturellement suivant :

Définition 3.3.1 *On appelle matrice* **triangulaire inférieure par blocs** *une matrice $L \in \mathbb{R}^{n \times n}$ telle que $[L]_{i,j} = 0 \in \mathbb{R}^{m_i \times n_j}$ pour tout couple (i,j) tel que $i < j$;*

On appelle matrice **triangulaire supérieure par blocs** *une matrice $U \in \mathbb{R}^{n \times n}$ telle que $[U]_{i,j} = 0$ pour tout couple (i,j) tel que $j < i$.*

$$L = \begin{bmatrix} [L]_{1,1} & & & & & \\ [L]_{2,1} & [L]_{2,2} & & & & \\ \ddots & \ddots & \ddots & & & \\ \ddots & \ddots & \ddots & \ddots & & \\ [L]_{p,1} & \ddots & \ddots & [L]_{p,p-1} & [L]_{p,p} \end{bmatrix} \quad et \quad U = \begin{bmatrix} [U]_{1,1} & [U]_{1,2} & \ddots & \ddots & [U]_{1,p} \\ & [U]_{2,2} & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \\ & & & & [U]_{p,p} \end{bmatrix}.$$

Enfin on appelle matrice **diagonale par blocs** *une matrice $D \in \mathbb{R}^{n\times n}$ telle que $[D]_{i,j} = 0$ pour tout couple (i,j) tel que $i \neq j$.*

$$D = \begin{bmatrix} [D]_{1,1} & & & & \\ & [D]_{2,2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & [D]_{p,p} \end{bmatrix}.$$

La plupart des résultats démontrés pour les matrices se généralisent au cas des matrices définies par blocs, en prenant la précaution dans l'écriture des formules de noter que les termes qui interviennent sont des matrices ; ceci nécessite en particulier une adaptation spécifique aux règles de calcul algébrique puisque si le produit de deux blocs est bien un bloc, cette opération n'est pas commutative !

3.4 Exercices sur les matrices triangulaires

On établira les résultats suivants :

Exercice 3.4.1 *Montrer que le produit de deux matrices triangulaires supérieures (resp. inférieures) inversibles $T', T'' \in \mathbb{R}^{n\times n}$ est une matrice triangulaire supérieure (resp. inférieure) inversible de $\mathbb{R}^{n\times n}$.*

Attention ! le produit d'une matrice triangulaire inférieure par une matrice triangulaire supérieure est une matrice à structure **quelconque**. Il en est de même du produit d'une matrice triangulaire supérieure par une matrice triangulaire inférieure.

Exercice 3.4.2 *Montrer que la matrice inverse d'une matrice triangulaire inférieure (resp. supérieure) inversible est une matrice triangulaire inférieure (resp. supérieure) inversible.*

Exercice 3.4.3 *Montrer que le déterminant d'une matrice triangulaire par blocs (supérieure ou inférieure), ou encore d'une matrice diagonale par blocs, est égal au produit des déterminants des blocs diagonaux.*

Exercice 3.4.4 *Soit $L \in \mathbb{R}^{n\times n}$ une matrice triangulaire inférieure par blocs inversible, montrer que la solution du système linéaire $Ly = b$ est obtenue par les formules :*

pour $i = 1, \dots p$ **faire**

$$[y]_i = [L]_{i,i}^{-1} \left([b]_i - \sum_{j<i} [L]_{i,j}[y]_j \right).$$

fin

Soit $U \in \mathbb{R}^{n\times n}$ une matrice triangulaire supérieure par blocs inversible, montrer que la solution du système linéaire $Ux = y$ est obtenue par les formules :

pour $i = p, \dots 1$ **faire**

$$[x]_i = [U]_{i,i}^{-1} \left([y]_i - \sum_{j>i} [U]_{i,j}[x]_j \right).$$

fin

3.5 Déterminant d'une matrice carrée

La Proposition 3.2.3 montre que le déterminant d'une matrice triangulaire est très facile à calculer. Cette propriété peut être exploitée pour le calcul du déterminant d'une matrice carrée quelconque $A \in \mathbb{R}^{n\times n}$. Si on suppose que l'on peut écrire la matrice A sous la forme d'un produit $A = LU$ dans lequel L est un matrice triangulaire inférieure à diagonale unité et U un matrice triangulaire supérieure, on peut alors écrire

$$\text{dét}\,(A) = \text{dét}\,(L)\ \text{dét}\,(U) = \text{dét}\,(U) = \prod_{i=1}^{n} U_{i,i}.$$

Coût calcul : On montre dans la suite de ce chapitre (voir la Proposition 3.10.1) que le calcul des matrices L et U à partir de la matrice $A \in \mathbb{R}^{n\times n}$ nécessite $O(n^3)$ opérations. On a vu que le calcul de dét (U) peut être effectué en $O(n)$ opérations. Finalement le déterminant de la matrice A d'ordre n peut être calculé en $O(n^3)$ opérations, ce qui est beaucoup plus favorable que la méthode de Cramer...

3.6 La méthode d'élimination

On considère le système linéaire (dont la matrice A est inversible)

$$\begin{pmatrix} A_{1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{i,1} & \cdot & \cdot & A_{i,i} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{n,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_i \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_i \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$$

dans lequel on *suppose* $A_{1,1} \neq 0$; à l'aide de la première ligne de ce système, on exprime la composante x_1 en fonction des autres :

$$x_1 = [b_1 - \sum_{2\leq j\leq n} A_{1,j}x_j]/A_{1,1}$$

en reportant cette identité dans la ligne i du système linéaire ($2 \leq i \leq n$), on obtient

$$A_{i,1}x_1 + \sum_{j=2}^{n} A_{i,j}x_j = b_i, \text{ soit } \sum_{j=2}^{n} \left(A_{i,j} - A_{i,1}A_{1,j}/A_{1,1}\right) x_j = b_i - A_{i,1}/A_{1,1}b_1.$$

On introduit alors les notations $A^{(0)} = A$ et $b^{(0)} = b$, puis on définit la matrice $A^{(1)}$ et le vecteur $b^{(1)}$ suivant :

- pour la première ligne :
$A^{(1)}_{1,j} = A^{(0)}_{1,j}$, $1 \leq j \leq n$, et $b^{(1)}_1 = b^{(0)}_1$;
- pour les $n-1$ dernières lignes :
$A^{(1)}_{i,j} = A^{(0)}_{i,j} - A^{(0)}_{i,1} \times A^{(0)}_{1,j}/A^{(0)}_{1,1}$, et $b^{(1)}_i = b^{(0)}_i - A^{(0)}_{i,1} \times b^{(0)}_1/A^{(0)}_{1,1}$, $2 \leq i \leq n$, $1 \leq j \leq n$.

On obtient un système linéaire équivalent au précédent, au sens où les deux systèmes admettent la même solution. Noter que la première colonne de la matrice $A^{(1)}$ est **nulle** par construction à l'exception du coefficient $A^{(1)}_{1,1}$.

Si $A^{(1)}_{2,2} \neq 0$, on peut réitérer le procédé en éliminant cette fois l'inconnue x_2 des $n-2$ lignes $i = 3, 4, \ldots, n$, et ainsi de suite... On génère de cette façon une suite de matrices et de seconds membres par l'algorithme :

1) **initialisation :**

$A^{(0)} = A \in \mathbb{R}^{n \times n}$.

$b^{(0)} = b \in \mathbb{R}^n$.

2) **itérations : pour** $k = 1, 2, \ldots n-1$ **faire**

(1) élimination de l'inconnue x_k

$$A^{(k)}_{i,j} = A^{(k-1)}_{i,j} \qquad 1 \leq i \leq k, \quad 1 \leq j \leq n$$

$$A^{(k)}_{i,j} = A^{(k-1)}_{i,j} - A^{(k-1)}_{i,k} \times A^{(k-1)}_{k,j}/A^{(k-1)}_{k,k} \qquad k < i \leq n, \quad 1 \leq j \leq n$$

(2) modification du second membre

$$b^{(k)}_i = b^{(k-1)}_i \qquad 1 \leq i \leq k$$

$$b^{(k)}_i = b^{(k-1)}_i - A^{(k-1)}_{i,k} \times b^{(k-1)}_k/A^{(k-1)}_{k,k} \qquad k < i \leq n$$

fin

Noter que les coefficients $A^{(k)}_{i,j}$ pour $k < i \leq n$ et $1 \leq j < k$ définis par ces formules, sont nuls par construction. En effet, pour de tels couples (i,j), on sait que d'une part $A^{(k-1)}_{i,j} = 0$, et d'autre part que $A^{(k-1)}_{k,j} = 0$ d'après l'itération précédente. En d'autres termes, on peut écrire, pour $k = 1, \cdots, n-1$:

$$A^{(k)} = \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & \mathcal{S}^{(k)} \end{pmatrix},$$

avec $[U]_{1,1} \in \mathbb{R}^{k \times k}$ une matrice *triangulaire supérieure*, $[U]_{1,2} \in \mathbb{R}^{k \times (n-k)}$, et $\mathcal{S}^{(k)} \in \mathbb{R}^{(n-k) \times (n-k)}$. On appelle $\mathcal{S}^{(k)}$ le **complément de Schur** .

Après $n-1$ itérations de cet algorithme (en supposant que les différents coefficients $A^{(k)}_{k,k}$ sont non nuls pour chaque k) la matrice $A^{(n-1)}$ obtenue est une matrice triangulaire supérieure et le système linéaire

$$A^{(n-1)}x = b^{(n-1)}$$

peut être résolu à l'aide de la Proposition 3.2.4. De plus, il a la même solution que le système initial $Ax = b$, puisque tous les systèmes linéaires $A^{(k)}x = b^{(k)}$ sont équivalents entre eux, pour $k = 1, \cdots, n-1$.

On introduit ensuite la matrice *triangulaire inférieure* $L^{(k)}$ de rang n, identique à la matrice I_n, à l'exception de la colonne k :

$$L^{(k)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 0 & 0 & 1 \end{pmatrix},$$

avec $L^{(k)}_{i,k} = -A^{(k-1)}_{i,k}/A^{(k-1)}_{k,k}$ pour $i > k$. Par construction, on a la relation :

$$\text{dét}\,(L^{(k)}) = 1.$$

De plus, on vérifie facilement que pour tout $k < n$, $A^{(k)} = L^{(k)}A^{(k-1)}$ et $b^{(k)} = L^{(k)}b^{(k-1)}$.

Finalement, en posant $U = A^{(n-1)}$ et $\tilde{L} = L^{(n-1)} \dots L^{(1)}$, on peut écrire

$$U = \tilde{L}A \quad \text{et} \quad Ux = \tilde{L}b, \tag{3.1}$$

où U et $\tilde{L}$ sont des matrices triangulaires inversibles. Le calcul de la solution x par les formules de **remontée** est alors immédiat.

La seule question qui se pose alors est de savoir si on peut toujours calculer cette matrice $A^{(n-1)}$ par les formules précédentes :

il faut pour cela que $A^{(k-1)}_{k,k} \neq 0$ pour $k = 1, 2, \dots, n-1$.

Si en cours de calcul, on rencontre un coefficient diagonal $A^{(k-1)}_{k,k}$ nul, on peut procéder de la façon suivante : on recherche dans la colonne k de la matrice $A^{(k-1)}$ un coefficient $A^{(k-1)}_{i_k,k}$ non nul pour $i_k > k$, et *s'il en existe un*, on échange alors les lignes i_k et k de la matrice. Cette modification revient à multiplier à gauche la matrice courante $A^{(k-1)}$ par une matrice de permutation $P(i_k, k)$ qui amène le coefficient $A^{(k-1)}_{i_k,k}$ sur la diagonale[2]

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} A^{(k-1)}_{1,1} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & A^{(k-1)}_{2,2} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & A^{(k-1)}_{3,3} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & A^{(k-1)}_{k,j} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & A^{(k-1)}_{i_k,k} & \mathrm{x} & \mathrm{x} & A^{(k-1)}_{i_k,j} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

[2]Notons que pour toute matrice de permutation $P(i_k, k)$, on a la relation $P(i_k, k)P(i_k, k) = I_n$. De plus, pour $k \geq 2$, on peut toujours écrire $P(i_k, k)$ sous la forme par blocs

$$P(i_k, k) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & P \end{pmatrix},$$

puisque $i_k > k$.

soit

$$P(i_k,k)A^{(k-1)} = \begin{pmatrix} A_{1,1}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & A_{2,2}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & A_{3,3}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & A_{i_k,k}^{(k-1)} & \text{x} & \text{x} & A_{i_k,j}^{(k-1)} & \text{x} \\ 0 & 0 & 0 & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & A_{k,j}^{(k-1)} & \text{x} \\ 0 & 0 & 0 & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \end{pmatrix}$$

Dans la factorisation en cours, cette multiplication n'affecte pas les lignes d'indice inférieur à k déjà calculées, mais seulement les lignes i et k de la matrice $A^{(k)}$; noter que les composantes $b_{i_k}^{(k-1)}$ et $b_k^{(k-1)}$ doivent aussi être échangées pour que le nouveau système linéaire soit équivalent au précédent.

Que se passe-t-il si à l'étape k (k fixé) *tous les coefficients* $A_{i,k}^{(k-1)}$ de la colonne k sont nuls ? Cela veut dire que l'on a obtenu une matrice de la forme

$$\begin{pmatrix} A_{1,1}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & A_{2,2}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & A_{3,3}^{(k-1)} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & A_{k,j}^{(k-1)} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & A_{i,j}^{(k-1)} & \text{x} \\ 0 & 0 & 0 & \mathbf{0} & \text{x} & \text{x} & \text{x} & \text{x} \end{pmatrix}$$

et la matrice $A^{(k-1)}$ est donc au plus de rang $n-1$, ce qui est contraire à l'hypothèse A inversible car la relation

$$A^{(k-1)} = L^{(k-1)} \dots L^{(1)} A$$

entraîne bien sûr

$$\text{dét}\,(A^{(k-1)}) = \text{dét}\,(A) \neq 0.$$

Exercice 3.6.1 *Montrer que pour tout k l'inverse $L^{-(k)}$ de la matrice $L^{(k)}$ est une matrice* triangulaire inférieure *de rang n, identique à I_n, à l'exception de la colonne k, avec, pour $i > k$, $L_{i,k}^{-(k)} = A_{i,k}^{(k-1)}/A_{k,k}^{(k-1)}$. En d'autres termes, on a la relation*

$$L^{-(k)} = 2I_n - L^{(k)}.$$

3.7 La méthode de factorisation

Dans le paragraphe précédent on a obtenu pour une matrice A donnée, les matrices triangulaires inversibles U et $\tilde{L}$. D'après (3.1), à l'aide du résultat de l'exercice 3.6.1, on définit

$$L = \tilde{L}^{-1} = L^{-(1)} \dots L^{-(n-1)}.$$

La matrice L est triangulaire inférieure à diagonale unité, qui vérifie la relation

$$A = L\,U.$$

Cette relation est appelée **factorisation de Gauss** de la matrice A. A partir de cette factorisation la solution du système linéaire $Ax = b$ est obtenue en deux étapes :

- la **descente**, qui consiste à calculer le vecteur y solution de $Ly = b$;
- la **remontée**, dans laquelle on calcule le vecteur x solution de $Ux = y$.

Au cours des calculs, on peut être avoir à effectuer une certain nombre de permutations de lignes pour amener des coefficients non nuls sur la diagonale. Si on doit effectuer une permutation à l'itération k, on factorise $P(i_k, k)A^{(k-1)}$ sous la forme

$$A^{(k)} = L^{(k)}P(i_k, k)A^{(k-1)}.$$

Au terme des itérations, on a alors factorisé la matrice A de la façon suivante[3]

$$U = L^{(n-1)}P(i_{n-1}, n-1)\cdots L^{(k)}P(i_k, k)\cdots L^{(1)}P(i_1, 1)A.$$

On a donc un produit "mêlé" de matrices du type $L^{(k)}$ et de matrices de permutation $P(i_p, p)$. Fort heureusement, on peut prouver le résultat suivant

Proposition 3.7.1 *Soit k et p tels que $k < p$. Alors il existe une matrice $L^{(k,p)}$ possédant la même structure que $L^{(k)}$ et telle que*

$$P(i_p, p)L^{(k)} = L^{(k,p)}P(i_p, p). \tag{3.2}$$

Preuve : Il suffit de considérer la matrice $L^{(k,p)}$ définie par

$$L^{(k,p)}_{i_p,k} = L^{(k,p)}_{p,k},\ L^{(k,p)}_{p,k} = L^{(k,p)}_{i_p,k},\ L^{(k,p)}_{i,j} = L^{(k,p)}_{i,j} \text{ sinon.}$$

■

A partir de ce résultat, on écrit tout simplement

$$\begin{aligned}
U &= L^{(n-1)}P(i_{n-1}, n-1)L^{(n-2)}P(i_{n-2}, n-2)\cdots A \\
&= L^{(n-1)}\bar{L}^{(n-2)}P(i_{n-1}, n-1)P(i_{n-2}, n-2)L^{(n-3)}P(i_{n-3}, n-3)\cdots A \\
&= L^{(n-1)}\bar{L}^{(n-2)}\bar{L}^{(n-3)}P(i_{n-1}, n-1)P(i_{n-2}, n-2)P(i_{n-3}, n-3)\cdots A \\
&\ \ \vdots \\
&= L^{(n-1)}\bar{L}^{(n-2)}\cdots\bar{L}^{(1)}P(i_{n-1}, n-1)\cdots P(i_1, 1)A.
\end{aligned}$$

Ci-dessus, on a bien sûr, pour $1 \le k \le n-2$,

$$\bar{L}^{(k)} = P(i_{n-1}, n-1)\cdots P(i_{k+1}, k+1)L^{(k)}P(i_{k+1}, k+1)\cdots P(i_{n-1}, n-1),$$

ce qui définit des matrices triangulaires inférieures possédant la même structure que $L^{(k)}$, cf. la Proposition 3.7.1.

On aboutit pour finir à $P\,A = L\,U$, où L est triangulaire inférieure, U triangulaire supérieure, et P est un produit de matrices de permutation, $P = P(i_{n-1}, n-1)\ \cdots\ P(i_1, 1)$. Notons que $P^{-1} = P(i_1, 1)\ \cdots\ P(i_{n-1}, n-1) = P^T$. On a ainsi démontré le résultat suivant

Théorème 3.7.1 *Si la matrice A est inversible, alors il existe une matrice de permutation P telle que*

$$P\ A = L\ U,$$

avec L matrice triangulaire inférieure à diagonale unité, U matrice triangulaire supérieure.

[3] Avec la convention $P(i_k, k) = I_n$ lorsque on n'effectue pas de permutation à l'étape k...

3.8 Stabilité numérique et stratégies de pivotage

On voit bien que cette écriture n'est pas unique puisqu'à chaque échange, on peut avoir le choix entre plusieurs lignes pour effectuer la permutation. On peut alors introduire un critère supplémentaire pour déterminer la ligne à permuter, par exemple un critère de stabilité numérique. Supposons que le coefficient courant $A_{k,k}^{(k-1)}$ soit petit, de l'ordre de ε, alors les formules de calcul

$$A_{i,j}^{(k)} = A_{i,j}^{(k-1)} - \frac{1}{\varepsilon}\, A_{i,k}^{(k-1)}\; A_{k,j}^{(k-1)}$$

montrent que dans le complément de Schur $\mathcal{S}^{(k)}$ résultant, le second terme est dominant, c'est-à-dire que l'on a

$$\mathcal{S}^{(k)} \approx -\frac{1}{\varepsilon} u \cdot v^T$$

les vecteurs u et v représentant respectivement la colonne k et la ligne k de la matrice $A^{(k-1)}$.

Exercice 3.8.1 *Montrer que pour tout $u, v \in \mathbb{R}^n$, tels que $u \neq 0$ et $v \neq 0$, la matrice $u \cdot v^T \in \mathbb{R}^{n\times n}$ est de rang 1.*

Dans ce cas, la matrice $\mathcal{S}^{(k)} \in \mathbb{R}^{(n-k)\times(n-k)}$ est *quasi-singulière* dès que $n-1 > k$! Autrement dit, le choix d'un petit coefficient diagonal peut conduire à une *instabilité numérique* de la factorisation.

Le choix de ce coefficient, appelé **pivot** doit donc être effectué avec la plus grande attention, et cela introduit naturellement deux variantes de la factorisation de Gauss :

- la factorisation avec **pivot partiel** revient à rechercher à chaque étape de l'algorithme le plus grand coefficient en valeur absolue parmi les $A_{i,k}^{(k-1)}$ pour $i \geq k$. La permutation de lignes associée à ce choix correspond à une multiplication à gauche de la matrice A par une matrice de permutation P.
- la factorisation avec **pivot total** revient à rechercher à chaque étape de l'algorithme le plus grand coefficient en valeur absolue parmi tous les $A_{i,j}^{(k-1)}$ pour $i \geq k$ et $j \geq k$. Si on choisit un coefficient $A_{i_k,j_k}^{(k-1)}$ en dehors de la colonne k, il faut ajouter à la permutation de lignes $P(i_k, k)$ une permutation de colonnes qui correspond à une multiplication à droite de la matrice A par une matrice de permutation $Q(j_k, k)$ (avec $j_k > k$). En fin de factorisation, on a obtenu
$$P\,A\,Q = L\,U.$$
- Si enfin on effectue une recherche avec **pivot total** sur les *coefficients diagonaux* $A_{i,i}^{(k-1)}$ uniquement (pour $i \geq k$), on a alors une permutation de ligne et une permutation de colonne identiques, i. e. $Q(i_k, k) = P(i_k, k)$. En fin de factorisation, on arrive à
$$P\,A\,P^T = L\,U.$$
Cette propriété sera utilisée pour factoriser des matrices symétriques.

Remarque 3.8.1 *En reprenant les notations précédentes, on voit qu'une permutation de lignes $P(i_k, k)$ ou de colonnes $Q(j_k, k)$ de la matrice $\mathcal{S}^{(k)}$ à l'étape k, ne modifie pas les blocs déjà calculés $[L]_{1,1}$ et $[U]_{1,1}$:*

$$P(i_k,k)\,A\,Q(j_k,k) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & P \end{pmatrix} \begin{pmatrix} [L]_{1,1} & 0 \\ [L]_{2,1} & I_{n-k+1} \end{pmatrix} \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & \mathcal{S}^{(k)} \end{pmatrix} \begin{pmatrix} I_{k-1} & 0 \\ 0 & Q \end{pmatrix}$$

soit encore

$$P(i_k,k)\,A\,Q(j_k,k) = \begin{pmatrix} [L]_{1,1} & 0 \\ P[L]_{2,1} & I_{n-k+1} \end{pmatrix} \begin{pmatrix} [U]_{1,1} & [U]_{1,2}Q \\ 0 & P\mathcal{S}^{(k)}Q \end{pmatrix}$$
$$= \begin{pmatrix} [L]_{1,1} & 0 \\ P[L]_{2,1} & P[L]_{2,2} \end{pmatrix} \begin{pmatrix} [U]_{1,1} & [U]_{1,2}Q \\ 0 & [U]_{2,2}Q \end{pmatrix}.$$

On énonce pour finir un résultat d'unicité.

Proposition 3.8.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible, pour des matrices de permutation P et Q données, la factorisation $P\,A\,Q = L\,U$ est unique.*

Preuve : Cela est évident par construction des matrices L et U, leurs coefficients étant déterminés de manière unique par les formules de l'algorithme. Mais on peut aussi démontrer ce résultat par une méthode qui sera utile par la suite. Supposons qu'il existe des matrices triangulaires inférieures L et L', et triangulaires supérieures U et U' qui vérifient $P\,A\,Q = L\,U = L'\,U'$. Par voie de conséquence,

$$(L')^{-1}L = U'U^{-1}(= M).$$

Or, $(L')^{-1}L$ (resp. $U'U^{-1}$) est une matrice triangulaire inférieure (resp. triangulaire supérieure). Ainsi M est nécessairement une matrice diagonale. De plus, L et L' étant à diagonale unité, il en est de même pour $(L')^{-1}L$. Finalement, $M = I_n$. ■

3.9 Les méthodes directes

Dans la suite on appellera **méthode directe** de résolution d'un système linéaire $Ax = b$ tout algorithme qui calcule la solution x en un nombre d'opérations déterminé *a priori*. Il s'agit ici de faire la distinction avec les méthodes itératives, étudiées au chapitre 4, pour lesquelles le nombre d'opérations dépend du nombre d'itérations de la méthode, nombre qu'il est impossible de connaître à l'avance car il est lié au choix de la solution initiale $x^0 \in \mathbb{R}^n$ relativement au second membre b.

Les méthodes d'élimination et de factorisation sont à ce titre des méthodes directes, car il est évident que le nombre d'opérations nécessaires au calcul des matrices L et U est fini - ce nombre d'opérations est calculé précisément plus loin - Par ailleurs le coût d'une descente et d'une remontée est égal à $2n^2$ opérations.

Noter que ce nombre d'opérations dépend des propriétés de la matrice A, car on peut tirer parti d'une éventuelle symétrie par exemple. On distingue ainsi plusieurs types de factorisation :

- La méthode de Gauss $A = LU$, avec L matrice triangulaire inférieure à diagonale unité, et U matrice triangulaire supérieure. A doit être inversible.
- La méthode de Crout $A = LDL^T$, avec L matrice triangulaire inférieure à diagonale unité, et D matrice diagonale. A doit être symétrique inversible.
- La méthode de Cholesky $A = LL^T$, avec L matrice triangulaire inférieure. A doit être symétrique définie-positive.

Dans ce qui suit, on reprend l'étude de la factorisation de la matrice A dans une formulation plus générale, en supposant uniquement que cette matrice est inversible.

3.10 Algorithme de factorisation de Gauss

Maintenant que l'existence des matrices L et U est établie, on vérifie que l'on peut calculer leurs coefficients directement par identification, à partir des relations

$$\forall\ i,j \quad 1 \le i,j \le n \qquad A_{i,j} = \sum_k L_{i,k}\,U_{k,j}.$$

On procède en calculant pour un indice k donné, tous les coefficients $L_{\cdot,k}$ de la colonne k de la matrice L, puis tous les coefficients $U_{k,\cdot}$ de la ligne k de la matrice U. Ce processus peut être représenté par les schémas suivants, dans lesquels les coefficients $\cdot$ sont supposés connus, les coefficients x sont inconnus et le coefficient $\bullet$ est en cours de calcul à l'aide des coefficients connus $\circ$.

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{i,j} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & \\ \cdot & 1 & & & & & & \\ \cdot & \cdot & 1 & & & & & \\ \cdot & \cdot & \cdot & 1 & & & & \\ \cdot & \cdot & \cdot & \cdot & 1 & & & \\ \circ & \circ & \circ & \circ & \bullet & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \mathrm{x} & \mathrm{x} & 1 & \\ \cdot & \cdot & \cdot & \cdot & \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \circ & \cdot & \cdot & \cdot \\ & \cdot & \cdot & \cdot & \circ & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \circ & \cdot & \cdot & \cdot \\ & & & \cdot & \circ & \cdot & \cdot & \cdot \\ & & & & \circ & \cdot & \cdot & \cdot \\ & & & & & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ & & & & & & \mathrm{x} & \mathrm{x} \\ & & & & & & & \mathrm{x} \end{pmatrix}$$

Calcul d'un coefficient de L.

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_{i,j} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & \\ \cdot & 1 & & & & & & \\ \cdot & \cdot & 1 & & & & & \\ \cdot & \cdot & \cdot & 1 & & & & \\ \cdot & \cdot & \cdot & \cdot & 1 & & & \\ \circ & \circ & \circ & \circ & \circ & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \mathrm{x} & \mathrm{x} & 1 & \\ \cdot & \cdot & \cdot & \cdot & \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \circ & \cdot \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \circ & \cdot \\ & & \cdot & \cdot & \cdot & \cdot & \circ & \cdot \\ & & & \cdot & \cdot & \cdot & \circ & \cdot \\ & & & & \cdot & \cdot & \circ & \cdot \\ & & & & & \cdot & \bullet & \mathrm{x} \\ & & & & & & \mathrm{x} & \mathrm{x} \\ & & & & & & & \mathrm{x} \end{pmatrix}$$

Calcul d'un coefficient de U.

En résumé, pour une matrice A donnée, les coefficients des matrices L et U sont calculés (à une permutation de lignes et de colonnes près) par les formules

pour $k = 1, \dots n-1$ **faire**

$L_{k,k} = 1$

$U_{k,k} = A_{k,k} - \sum_{j<k} L_{k,j} U_{j,k}$

pour $i = k+1, \dots n$ **faire**

$L_{i,k} = [A_{i,k} - \sum_{j<k} L_{i,j} U_{j,k}] \ / \ U_{k,k}$

fin

pour $i = k+1, \dots n$ **faire**

$U_{k,i} = A_{k,i} - \sum_{j<k} L_{k,j} U_{j,i}$

fin

fin

Exercice 3.10.1 *Les formules précédentes calculent les coefficients de la matrice L colonne par colonne, et ceux de la matrice U ligne par ligne. Montrer que l'algorithme suivant définit les*

mêmes matrices, bien que les coefficients de la matrice L' soient calculés ligne par ligne et ceux de la matrice U' colonne par colonne.

pour $k = 1, \dots n-1$ **faire**

  pour $i = 1, \dots k-1$ **faire**

    $L'_{k,i} = [A_{k,i} - \sum_{j<i} L'_{k,j} U'_{j,i}] \ / \ U'_{i,i}$

  fin

  pour $i = 1, \dots k-1$ **faire**

    $U'_{i,k} = A_{i,k} - \sum_{j<i} L'_{i,j} U'_{j,k}$

  fin

  $L'_{k,k} = 1$

  $U'_{k,k} = A_{k,k} - \sum_{j<k} L'_{k,j} U'_{j,k}$

fin

On étudie pour finir le **coût calcul** associé à la résolution d'un système linéaire, à l'aide de la méthode de Gauss.

Proposition 3.10.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible, et $b \in \mathbb{R}^n$ un vecteur. Le nombre d'opérations élémentaires $(+, -, *, /)$ nécessaires au calcul de la solution du système linéaire $Ax = b$ par la méthode de Gauss est de l'ordre de $2n^3/3$.*

Preuve : Traitons d'abord le coût de la factorisation de Gauss par les formules précédentes. On peut choisir de calculer les n colonnes de L, d'après l'algorithme précédent. Soit donc k variant de 1 à n. Pour déterminer la colonne k de L, il faut calculer un coefficient diagonal et $n-k$ coefficients non-diagonaux. Or, $L_{k,k} = 1$, et chaque coefficient $(L_{i,k})_{i>k}$ requiert $2k-1$ opérations élémentaires $(+, -, *, /)$. Le nombre total d'opérations est donc égal à

$$\sum_{k=1}^{k=n} (2k-1)(n-k) = 2n \sum_{k=1}^{k=n} k - 2 \sum_{k=1}^{k=n} k^2 + O(n^2) = n^3 - \frac{2}{3} n^3 + O(n^2) = \frac{1}{3} n^3 + O(n^2),$$

soit un nombre total d'opérations d'environ $n^3/3$ pour déterminer L. Bien sûr, une étude de l'algorithme de calcul des lignes de U fournit un nombre total d'opérations lui aussi égal à

$$\frac{1}{3} n^3 + O(n^2).$$

Maintenant que les matrices L et U sont calculées, la résolution du systéme linéaire peut s'effectuer par une descente-remontée, dont le coût est égal à $2n^2$ opérations.

Le coût total de la résolution du système linéaire $Ax = b$ est donc de l'ordre de $2n^3/3$ opérations élémentaires $(+, -, *, /)$. ■

Pour $n = 20$, on doit donc effectuer environ 5.400 opérations.

A noter que la partie la plus coûteuse de l'algorithme est la factorisation $A = LU$. En conséquence lorsque l'on a plusieurs (par exemple 10) systèmes linéaires à résoudre avec la même matrice A, on ne calcule les matrices L et U qu'une seule fois (voir par exemple le paragraphe 5.3) et le coût total de la résolution de ces 10 systèmes linéaires reste de l'ordre de $2n^3/3$ opérations.

3.11 Factorisation de Gauss-Jordan. Factorisation de Crout

Dans la factorisation précédente $A = LU$, on a imposé le choix d'une matrice L triangulaire inférieure à diagonale unité. Si on note D la matrice diagonale formée à partir des coefficients diagonaux de U : $D_{i,i} = U_{i,i}$, alors

$$A = L\,U = L\,D\,\tilde{U}.$$

Cette factorisation est appelée factorisation de Gauss-Jordan, dans laquelle la matrice $\tilde{U}$ est triangulaire supérieure à diagonale unité. Les coefficients des matrices D et $\tilde{U}$ sont déterminés à partir des coefficients de U par les relations

pour $k = 1, \dots n$ **faire**
  $D_{k,k} = U_{k,k}$
  pour $i = 1, \dots k$ **faire**
    $\tilde{U}_{k,i} = U_{i,k} \;/\; U_{k,k} = U_{i,k} \;/\; D_{k,k}$.
  fin
fin

Cette écriture s'avère utile dans le cas où la matrice A est symétrique. En effet, lorsqu'on emploie une stratégie de pivot total sur les éléments diagonaux ($Q = P^T$), on a alors

Proposition 3.11.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible et symétrique, à une matrice de permutation P près, la factorisation $P\,A\,P^T = L\,D\,L^T$ est unique.*

Preuve : On a par construction $(P\,A\,P^T)^T = P\,A^T\,P^T = P\,A\,P^T$, d'où

$$L\,D\,\tilde{U} = \tilde{U}^T\,D\,L^T,$$

et en utilisant la Proposition 3.8.1 sur l'unicité de la factorisation de Gauss, on trouve bien

$$L = \tilde{U}^T.$$

■

On obtient ainsi la **factorisation de Crout** pour une matrice symétrique :

$$A = L\,D\,L^T.$$

Par identification, les coefficients de L et D sont calculés suivant

pour $k = 1, \dots n$ **faire**
  $L_{k,k} = 1 \quad$ et $\quad D_{k,k} = A_{k,k} - \displaystyle\sum_{j<k} L_{k,j}^2$
  pour $i = k+1, \dots n$ **faire**
    $L_{i,k} = [A_{i,k} - \displaystyle\sum_{j<k} L_{i,j}L_{k,j}] \;/\; D_{k,k}$.
  fin
fin

La résolution du système linéaire $Ax = b$ s'effectue alors en trois étapes : le calcul du vecteur z solution de $Lz = b$, puis du vecteur y par $Dy = z$ et enfin du vecteur x par $L^T x = y$. Le coût calcul de ces trois résolutions est identique à celui de l'étape correspondante de la méthode de Gauss puisque la matrice L est à diagonale unité.

On s'est limité au cas $Q = P^T$. Ceci signifie que dans la stratégie du pivot total la recherche du meilleur pivot est limité aux coefficients diagonaux de la matrice en cours de calcul, pour conserver la symétrie. Cette contrainte peut être levée dans le cadre de la factorisation par blocs, qui sera étudiée plus loin.

Pour finir, comme il n'est plus besoin de calculer la matrice triangulaire supérieure U $(= L^T)$, on déduit de la Proposition 3.10.1 le résultat sur le **coût calcul**.

Proposition 3.11.2 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique inversible, et $b \in \mathbb{R}^n$ un vecteur. Le nombre d'opérations élémentaires $(+, -, *, /)$ nécessaires au calcul de la solution du système linéaire $Ax = b$ par la méthode de Crout est de l'ordre de $n^3/3$.*

3.12 Factorisation de Cholesky

Supposons maintenant que $A \in \mathbb{R}^{n\times n}$ soit une matrice symétrique définie-positive, il résulte de la Proposition 3.11.1 que l'on peut écrire $A = P^T L\, D\, L^T P$, puisque $P\, P^T = I_n$. De plus, par définition, on a la propriété

$$\forall x \in \mathbb{R}^n,\ x \neq 0 \qquad 0 < (x, Ax) = (x, P^T LDL^T Px) = (L^T Px, D\, L^T Px).$$

Pour tout $y = L^T Px \neq 0$, on a donc $(y, Dy) > 0$: la matrice diagonale D est définie-positive, et pour tout $1 \leq i \leq n$, $D_{i,i} > 0$, ce qui permet de définir la matrice diagonale $D^{1/2}$ par $D^{1/2}_{i,i} = \sqrt{D_{i,i}}$. En posant alors $\mathcal{L} = LD^{1/2}$, on obtient finalement

$$P\ A\ P^T = \mathcal{L}\mathcal{L}^T.$$

Proposition 3.12.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, il existe une matrice $\mathcal{L}$ triangulaire inférieure, telle que la factorisation*

$$PAP^T = \mathcal{L}\mathcal{L}^T$$

est unique à une matrice de permutation P près.

En utilisant les formules générales, les coefficients de $\mathcal{L}$ sont calculés colonne par colonne par les relations

pour $k = 1, \ldots n$ **faire**

$$\mathcal{L}_{k,k} = [A_{k,k} - \sum_{j<k} \mathcal{L}^2_{k,j}]^{1/2}$$

pour $i = k+1, \ldots n$ **faire**

$$\mathcal{L}_{i,k} = [A_{i,k} - \sum_{j<k} \mathcal{L}_{k,j}\mathcal{L}_{i,j}] \ /\ \mathcal{L}_{k,k}.$$

fin

fin

Exercice 3.12.1 *Montrer que l'on peut aussi calculer la matrice $\mathcal{L}$ ligne par ligne suivant*

pour $k = 1, \dots n$ **faire**

  pour $i = 1, \dots k-1$ **faire**

    $\mathcal{L}_{k,i} = [A_{k,i} - \sum_{j<i} \mathcal{L}_{k,j}\mathcal{L}_{i,j}] \ / \ \mathcal{L}_{i,i}$

  fin

    $\mathcal{L}_{k,k} = [A_{k,k} - \sum_{j<k} \mathcal{L}_{k,j}^2]^{1/2}$.

fin

Une conséquence importante de la Proposition 3.12.1, dans le cas où la matrice A est symétrique définie-positive, est que l'on n'a pas besoin de vérifier que les termes

$$A_{k,k} - \sum_{j<k} \mathcal{L}_{k,j}^2$$

sont tous strictement positifs, pour tout k. L'existence des coefficients diagonaux $\mathcal{L}_{k,k}$ étant assurée par cette Proposition, il suffit de vérifier par identification que l'on peut les calculer de cette manière.

Noter que cette propriété n'est pas satisfaite si A est supposée seulement symétrique : dans la factorisation de Crout certains coefficients diagonaux $D_{i,i}$ peuvent en effet être négatifs.

Une autre conséquence remarquable de la Proposition 3.8.1 peut être obtenue en faisant intervenir à nouveau la matrice complément de Schur $\mathcal{S}$:

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$$

avec $A_{1,1} = A_{1,1}^T \in \mathbb{R}^{n_1 \times n_1}$, $A_{2,1} \in \mathbb{R}^{n_2 \times n_1}$, $A_{1,2} = A_{2,1}^T \in \mathbb{R}^{n_1 \times n_2}$, $A_{2,2} = A_{2,2}^T \in \mathbb{R}^{n_2 \times n_2}$, $n = n_1 + n_2$. On suppose que l'on connait une factorisation de Cholesky du bloc : $A_{1,1} = \mathcal{L}_{1,1}\mathcal{L}_{1,1}^T$, avec $\mathcal{L}_{1,1} \in \mathbb{R}^{n_1 \times n_1}$ matrice triangulaire inférieure, alors (cf. Remarque 3.8.1)

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} = \begin{pmatrix} \mathcal{L}_{1,1} & 0 \\ \mathcal{L}_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} I_{n_1} & 0 \\ 0 & \mathcal{S} \end{pmatrix} \times \begin{pmatrix} \mathcal{L}_{1,1}^T & \mathcal{L}_{2,1}^T \\ 0 & I_{n_2} \end{pmatrix}.$$

Proposition 3.12.2 *Soit $A \in \mathbb{R}^{n \times n}$ une matrice symétrique définie-positive, pour tout couple (n_1, n_2) avec $n_1 + n_2 = n$, le complément de Schur $\mathcal{S} \in \mathbb{R}^{n_2 \times n_2}$ est une matrice symétrique définie-positive.*

Preuve : Par construction la matrice $\mathcal{S} = A_{2,2} - \mathcal{L}_{2,1}\mathcal{L}_{2,1}^T$ est symétrique ; il suffit alors de prendre des vecteurs $x \in \mathbb{R}^n$ dont les n_1 premières composantes sont nulles pour obtenir

$$\begin{aligned}
(x, Ax) &= \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix}^T \begin{pmatrix} A_{1,1} & A_{2,1}^T \\ A_{2,1} & A_{2,2} \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix}^T \begin{pmatrix} \mathcal{L}_{1,1} & 0 \\ \mathcal{L}_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} I_{n_1} & 0 \\ 0 & \mathcal{S} \end{pmatrix} \times \begin{pmatrix} \mathcal{L}_{1,1}^T & \mathcal{L}_{2,1}^T \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \\
&= (\tilde{x}, \mathcal{S}\tilde{x})
\end{aligned}$$

ainsi $\forall \tilde{x} \in \mathbb{R}^{n_2}, \tilde{x} \neq 0 \quad (\tilde{x}, \mathcal{S}\tilde{x}) > 0.$ ■

En conséquence, on en déduit qu'il est toujours possible de réaliser la factorisation (de Cholesky) d'une matrice symétrique définie-positive *sans permutation*, ce qui est *faux* pour la factorisation (de Gauss) d'une matrice inversible quelconque, et qui reste *faux* pour la factorisation (de Crout) d'une matrice symétrique inversible.

Corollaire 3.12.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie- positive. Il existe une matrice $\mathcal{L}$ triangulaire inférieure unique, telle que*

$$A = \mathcal{L}\mathcal{L}^T.$$

Preuve : Pour réaliser la factorisation sans permutation, on raisonne par récurrence sur k, qui varie de 1 à $n-1$, et l'on montre que $A_{k,k}^{(k-1)} = (A^{(k-1)}e_k, e_k) > 0$.
Avec la convention $\mathcal{S}^{(0)} = A$, ceci revient à vérifier que $(\mathcal{S}^{(k-1)}e_k, e_k) > 0$.

Pour cela, il suffit de montrer que $\mathcal{S}^{(k-1)}$ est symétrique définie-positive, par récurrence sur k. Or, $\mathcal{S}^{(0)} = A$ est symétrique définie-positive et, d'après la Proposition précédente, l'assertion "$\mathcal{S}^{(k-1)}$ symétrique définie-positive entraîne $\mathcal{S}^{(k)}$ symétrique définie-positive" est vraie. ■

Et, pour finir, une évaluation du **coût calcul**.

Proposition 3.12.3 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, et $b \in \mathbb{R}^n$ un vecteur. Le nombre d'opérations élémentaires $(+, -, *, /)$ nécessaires au calcul de la solution du système linéaire $Ax = b$ par la méthode de Cholesky est de l'ordre de $n^3/3$.*

3.13 Factorisation par blocs

Dans ce qui précède les matrices triangulaires qui définissent les différentes factorisations ont été calculées coefficient par coefficient, on peut à ce titre les qualifier de **factorisations ponctuelles** ; cependant si on effectue une une partition des matrices A, L et U en $p \times p$ blocs, on obtient formellement

$$\begin{bmatrix} [A]_{1,1} & [A]_{1,2} & \dots & [A]_{1,p} \\ [A]_{2,1} & [A]_{2,2} & \ddots & [A]_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ [A]_{p,1} & [A]_{p,2} & \dots & [A]_{p,p} \end{bmatrix} = \begin{bmatrix} [L]_{1,1} & 0 & \dots & 0 \\ [L]_{2,1} & [L]_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ [L]_{p,1} & [L]_{p,2} & \dots & [L]_{p,p} \end{bmatrix} \times \begin{bmatrix} [U]_{1,1} & [U]_{1,2} & \dots & [U]_{1,p} \\ 0 & [U]_{2,2} & \ddots & [U]_{2,p} \\ \dots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & [U]_{p,p} \end{bmatrix}$$

Dans cette écriture seuls les blocs diagonaux sont nécessairement carrés. Par identification on obtient la **factorisation par blocs** de Gauss

pour $k = 1, \dots p$ **faire** (*factorisation ponctuelle* du bloc $[A]_{k,k}$)

$[L]_{k,k}\ [U]_{k,k} = [A]_{k,k} - \sum_{j<k} [L]_{k,j}[U]_{j,k}$

pour $i = k+1, \dots p$ **faire**

$[L]_{i,k}[U]_{k,k} = [A]_{i,k} - \sum_{j<k} [L]_{i,j}[U]_{j,k}$

fin

pour $i = k+1, \dots p$ **faire**

$[L]_{k,k}[U]_{k,i} = [A]_{k,i} - \sum_{j<k} [L]_{k,j}[U]_{j,i}$

fin

fin

dans cette écriture les produits $[L]_{i,j}[U]_{j,k}$ sont des produits de matrices, et les blocs $[L]_{i,k}$ et $[U]_{k,i}$ sont obtenus par résolution de systèmes linéaires à matrice triangulaire supérieure, et dont les seconds membres sont des matrices $[B]$ et $[B']$:

$$[L]_{i,k}[U]_{k,k} = [B] \text{ soit } [U]^T_{k,k}[L]^T_{i,k} = [B]^T \ ; \quad [L]_{k,k}[U]_{k,i} = [B'].$$

Comme pour la factorisation de Gauss ponctuelle, il peut être nécessaire d'effectuer des permutations de lignes ou de colonnes afin de placer des blocs réguliers sur la diagonale.

Mentionnons également la factorisation de Crout par blocs

pour $k = 1, \dots p$ **faire** (*factorisation ponctuelle* du bloc $[A]_{k,k}$)

$[L]_{k,k}[D]_{k,k}[L]^T_{k,k} = [A]_{k,k} - \sum_{j<k} [L]_{k,j}[L]^T_{k,j}$

pour $i = k+1, \dots p$ **faire**

$[D]_{k,k}[L]_{i,k} = [A]_{i,k} - \sum_{j<k} [L]_{i,j}[L]^T_{k,j}$

fin

fin

et enfin la factorisation de Cholesky par blocs

pour $k = 1, \dots p$ **faire** (*factorisation ponctuelle* du bloc $[A]_{k,k}$)

$[\mathcal{L}]_{k,k}[\mathcal{L}]_{k,k}^T = [A]_{k,k} - \sum_{j<k} [\mathcal{L}]_{k,j}[\mathcal{L}]_{k,j}^T$

pour $i = k+1, \dots p$ **faire**

$[\mathcal{L}]_{k,k}[\mathcal{L}]_{i,k}^T = [A]_{i,k} - \sum_{j<k} [\mathcal{L}]_{i,j}[\mathcal{L}]_{k,j}^T$

fin

fin

Cette élégante formulation est récursive puisque l'on peut y remplacer à nouveau les factorisations ponctuelles par des factorisations par blocs... Cette approche est intéressante pour un calcul concret sur ordinateur dans le cas de matrices géantes qui ne tiennent pas en mémoire (on les découpe alors en blocs suffisamment petits) ou le plus souvent pour des matrices qui ne sont effectivement connues que sous la forme d'une partition.

Une autre application de cette formulation est la recherche d'un pivot extra-diagonal pour des matrices symétriques dont la factorisation pose des problèmes numériques, comme les matrices non définies. En effet dans le cas d'une stratégie de pivot total par points, il faut, pour *préserver la symétrie*, limiter la recherche du coefficient de plus grande valeur absolue aux seuls termes diagonaux. Cette procédure peut s'avérer inefficace si les coefficients diagonaux sont trop petits. On applique alors la technique des **pivots jumeaux** : supposons qu'à l'étape k de la factorisation, le pivot idéal $\mathcal{S}_{k,k'} = \mathcal{S}_{k',k}$ soit en position extra-diagonale

$$\mathcal{S} = \begin{pmatrix} \mathcal{S}_{k,k} & \cdot & \cdot & \mathcal{S}_{k,k'} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathcal{S}_{k',k} & \cdot & \cdot & \mathcal{S}_{k',k'} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

alors à l'aide d'une permutation symétrique des lignes et des colonnes, on commence par écrire

$$P\mathcal{S}P^T = \begin{pmatrix} \mathcal{S}_{k,k} & \mathcal{S}_{k,k'} & \cdot & \cdot & \cdot \\ \mathcal{S}_{k',k} & \mathcal{S}_{k',k'} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

puis on effectue une factorisation par blocs symétrique de la matrice $\mathcal{S}$ avec un premier **bloc diagonal** 2×2, et les autres blocs diagonaux de la partition de rang 1. On assure ainsi la stabilité numérique de la factorisation, tout en conservant la symétrie.

3.14 Profil et conservation du profil

Une propriété importante des matrices que l'on rencontre fréquemment en calcul scientifique est leur caractère creux : pour les matrices qui proviennent de l'approximation de la solution d'une équation aux dérivées partielles par la méthode des différences finies ou la méthode des éléments finis, le nombre moyen de coefficients non nuls par ligne est petit (voir le paragraphe B.6 et les Travaux Pratiques) et cela quel que soit le nombre total d'inconnues n. L'exploitation

de cette propriété apporte une économie considérable tant sur le plan du temps calcul, que de la place mémoire. Examinons un exemple :

$$A = \begin{bmatrix} \bullet & \bullet & 0 & \bullet & 0 & 0 & 0 & \bullet \\ \bullet & \bullet & \bullet & 0 & 0 & \bullet & 0 & \bullet \\ 0 & \bullet & \bullet & \bullet & 0 & \bullet & 0 & \bullet \\ \bullet & 0 & \bullet & \bullet & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bullet & \bullet & 0 & 0 \\ 0 & \bullet & \bullet & 0 & \bullet & \bullet & 0 & \bullet \\ 0 & 0 & 0 & 0 & 0 & 0 & \bullet & 0 \\ \bullet & \bullet & \bullet & 0 & 0 & \bullet & 0 & \bullet \end{bmatrix}$$

Les coefficients non nuls de cette matrice sont représentés par le symbole $\bullet$ et les autres par 0. On remarque tout de suite que la matrice A ci-dessus est à *structure symétrique*, puisque les $\bullet$ sont placés symétriquement par rapport à la diagonale. En d'autres termes,

$$A_{i,j} \neq 0 \Longleftrightarrow A_{j,i} \neq 0. \tag{3.3}$$

On voudrait dans la suite que cette propriété (3.3) soit satisfaite (notamment pour simplifier les notations !). A cette fin, on *symétrisera* la structure de la matrice, en ajoutant un *couple* de symbole $\bullet$ non-diagonaux, dès lors que *l'un au moins* des $A_{i,j}$ et $A_{j,i}$ correspondants est non nul.

Pour chaque ligne k de la matrice A, on peut définir $il(k)$ le plus petit indice de colonne l tel que $A_{k,l} \neq 0$; on définit de même pour chaque colonne k, $ic(k)$ le plus petit indice de ligne l, tel que $A_{l,k} \neq 0$. On note que pour pouvoir définir $il(k)$ et $ic(k)$ pour tout k, une condition *suffisante*[4] est l'inversibilité de A, ce qui tombe bien, puisque on s'intéresse à la factorisation en vue de la résolution de systèmes linéaires !
Comme (3.3) est satisfaite (éventuellement par symétrisation de la structure), on a automatiquement $il(k) = ic(k)$ pour tout k, ce qui permet de s'affranchir des indices $ic(k)$ dans la suite.

On introduit la propriété suivante :

$$il(k) \leq k, \ \forall k. \tag{3.4}$$

On a le résultat ci-dessous.

Proposition 3.14.1 *Si les factorisations de Gauss et Crout se font sans permutation, alors la propriété (3.4) est vraie.*

Preuve : Supposons qu'il existe k tel que $il(k) > k$. Dans ce cas, on a par définition $A_{k,1} = A_{k,2} = A_{k,k} = 0$. Ecrivons que $A = LU$ avec L triangulaire inférieure et U triangulaire supérieure, sur la $k^{ème}$ ligne

$$\begin{aligned} 0 &= A_{k,1} = L_{k,1}U_{1,1}, \\ 0 &= A_{k,2} = L_{k,1}U_{1,2} + L_{k,2}U_{2,2}, \\ &\vdots \\ 0 &= A_{k,k-1} = L_{k,1}U_{1,k-1} + L_{k,2}U_{2,k-1} + \cdots + L_{k,k-1}U_{k-1,k-1} \\ 0 &= A_{k,k} = L_{k,1}U_{1,k} + L_{k,2}U_{2,k} + \cdots + L_{k,k-1}U_{k-1,k} + L_{k,k}U_{k,k}. \end{aligned}$$

On déduit de la première équation que $L_{k,1} = 0$, puisque U est inversible. De la deuxième équation, on déduit alors que $L_{k,2} = 0$, et ainsi de suite jusqu'à la $(k-1)^{ème}$ équation, qui

[4] En effet, lorsque A est inversible, ses lignes et ses colonnes sont *toutes* non nulles.

fournit le résultat $L_{k,k-1} = 0$. En passant enfin à la $k^{ème}$ équation, on arrive à $L_{k,k}U_{k,k} = 0$, ce qui contredit l'hypothèse L et U inversibles. ■

En d'autres termes, la propriété (3.4) est *nécessaire* pour pouvoir factoriser une matrice sans permutation, mais elle n'est pas *suffisante...*

Exercice 3.14.1 *Soit A la matrice symétrique et inversible*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 3 & 3 \end{bmatrix}.$$

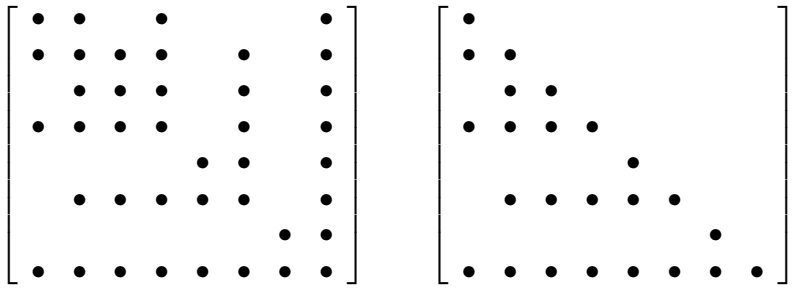
Vérifier que A ne peut être factorisée sans permutation, bien que la propriété (3.4) soit satisfaite.

Sous réserve que les propriétés (3.3) et (3.4) soient satisfaites, on introduit les ensembles

$$Pl(A) = \{(k,l), 1 \leq k \leq n, il(k) \leq l \leq k\} \text{ et } Pc(A) = \{(l,k), 1 \leq k \leq n, il(k) \leq l \leq k\}.$$

Définition 3.14.1 *on appelle* **profil** *de la matrice A*
- *l'ensemble $Pr(A) = Pl(A)$ si A est symétrique*
- *l'ensemble $Pr(A) = Pl(A) \cup Pc(A)$ sinon.*

Reprenons l'exemple ci-dessus et décrivons l'ensemble $Pr(A)$ correspondant.



$Pr(A)$: à gauche si A vérifie uniquement (3.3), à droite si A est de plus symétrique.

Remarque 3.14.1 *1) Par définition, $(i,j) \in Pr(A)$ n'entraîne pas que $A_{i,j} \neq 0$. En d'autres termes il peut exister des coefficients de A nuls à l'intérieur du profil ! Il faut donc distinguer l'ensemble $Pr(A)$ de l'ensemble*

$$\mathcal{S}q(A) = \{(k,l), 1 \leq l \leq k \leq n, A_{k,l} \neq 0\}$$

qui est appelé le **squelette** *de la matrice A.*

2) On a défini le **profil ponctuel**, *il est évident que l'on peut associer à toute partition de la matrice A un* **profil par blocs**.

Proposition 3.14.2 *Les factorisations de Gauss et Crout, lorsqu'elles se font sans permutation, ainsi que la factorisation de Cholesky, conservent le profil.*
En d'autres termes, $(i,j) \notin Pr(A)$ entraîne $L_{i,j} = U_{i,j} = 0$.

Preuve : On raisonne pour la factorisation de Gauss $LU = A$. On vérifie tout d'abord que L conserve le profil en reprenant la preuve de la Proposition 3.14.1, i. e. en écrivant que, pour la ligne k, $A_{i,k} = 0$ tant que $i < il(k)$. Ensuite, pour U, on écrit cette fois des égalités pour la colonne k de A, sachant que $A_{k,j} = 0$ tant que $j < il(k)$.
Cette propriété est commune aux trois factorisations, la structure des calculs étant identique. ■

Remarque 3.14.2 *ce résultat est valable pour le profil par points comme pour le profil par blocs.*

Remarque 3.14.3 *Pour les matrices creuses, on définit la largeur de bande de la ligne k par la relation $lb(k) = k - il(k) + 1$. La largeur de bande moyenne d'une matrice $A \in \mathbb{R}^{n\times n}$ est donc $l = [\sum_{k=1}^{n} l(k)]/n$. Pour les matrices creuses la largeur de bande moyenne l est en général petite devant n. Si on a $lb(k) \sim l$ pour tout k, on vérifie que le coût calcul de la factorisation est de l'ordre de $O(n\, l^2)$.*

Chapitre 4

Les méthodes itératives

4.1 Introduction

On appelle **méthode itérative** de résolution d'un système linéaire $Ax = b$, tout algorithme qui construit à partir d'une estimation initiale x^0 une suite de vecteurs $\{x^k\}_{k\in\mathbb{N}}$ destinée à converger vers la solution x du système. Les méthodes présentées dans ce chapitre sont associées à la notion de décomposition régulière d'une matrice, qui nécessite quelques rappels sur l'analyse numérique matricielle.

4.2 Décomposition régulière

Définition 4.2.1 *on appelle* **décomposition régulière** *d'une matrice $A \in \mathbb{R}^{n\times n}$ la donnée de deux matrices $M, N \in \mathbb{R}^{n\times n}$ telles que*
(i) $A = M - N$;
(ii) M est inversible.

On associe à toute décomposition régulière la méthode itérative

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Mx^{k+1} = Nx^k + b \\ \textbf{fin} \end{array}\right. \tag{4.1}$$

On voit que si cette méthode converge vers un vecteur x, celui–ci vérifie néssairement la relation

$$Mx = Nx + b \quad \text{soit encore} \quad Ax = b.$$

Si on appelle $r^k = b - Ax^k$ le **vecteur résidu**, cette méthode peut aussi s'écrire sous la forme

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + M^{-1}r^k \\ \textbf{fin} \end{array}\right. \tag{4.2}$$

On note si que $M = A$ la méthode itérative converge en une seule itération quel que soit le vecteur initial : $x^1 = x^0 + A^{-1}(b - Ax^0) = A^{-1}b$, mais il s'agit là d'une méthode directe qui nécessite la résolution du système linéaire $Az = r$. Dans la pratique on recherche donc des matrices M pour lesquelles cette résolution n'est pas trop coûteuse. L'objet de ce chapitre est l'étude générale de ce type de méthode et de leur convergence vers x solution du système linéaire suivant le choix de la matrice M.

On introduit le **vecteur erreur** $e^k = x - x^k$, alors

$$\left.\begin{array}{c} Mx = Nx + b \\ Mx^{k+1} = Nx^k + b \end{array}\right\} \Longrightarrow e^{k+1} = M^{-1}Ne^k.$$

Proposition 4.2.1 *La méthode itérative converge si et seulement si*

$$\rho(M^{-1}N) < 1.$$

Preuve : Par construction le vecteur e^k vérifie $e^{k+1} = M^{-1}Ne^k = \left[M^{-1}N\right]^{k+1} e^0$. Une condition nécessaire et suffisante pour que e^{k+1} tende vers 0 quel que soit le vecteur initial x^0 est que $\rho(M^{-1}N) < 1$, d'après le Théorème B.5.1. ∎

Ainsi, pour définir une méthode itérative convergente il faut donc choisir la matrice M de façon que

(i) $\rho(M^{-1}N) < 1$,

(ii) la résolution du système linéaire $Mx^{k+1} = Nx^k + b$ ne soit pas trop coûteuse car il faudra l'effectuer à chaque itération !

Théorème 4.2.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive et soit une décomposition régulière $A = M - N$. Si la matrice $M^T + N$ est définie-positive alors*

$$\rho(M^{-1}N) < 1.$$

Preuve : Puisque A est symétrique par construction $M^T + N = M^T + M - A$ est aussi symétrique. On va utiliser la Proposition B.4.1, qui permet d'affirmer que $\rho(M^{-1}N) \leq \|M^{-1}N\|$ pour toute norme matricielle $\|\cdot\|$. On choisit la norme vectorielle induite par la norme $\|v\|_A = \sqrt{(v, Av)}$ qui est bien une norme vectorielle puisque A est symétrique définie-positive. Dans ce cas,

$$\|M^{-1}N\|_A = \max_{v\neq 0} \frac{\|M^{-1}Nv\|_A}{\|v\|_A} = \max_{v\neq 0} \frac{\|v - M^{-1}Av\|_A}{\|v\|_A},$$

puisque $N = M - A$. Soit maintenant $v \in \mathbb{R}^n$ tel que $\|v\|_A = 1$, et $w = M^{-1}Av$; alors

$$\begin{aligned}
\|M^{-1}Nv\|_A^2 &= \|v - M^{-1}Av\|_A^2 = \|v - w\|_A^2 \\
&= (v - w, A(v - w)) \\
&= (v, Av) - (w, Av) - (v, Aw) + (w, Aw) \\
&= \|v\|_A^2 - (w, Mw) - (A^T v, w) + (w, Aw) \\
&= \|v\|_A^2 - (w, Mw) - (Av, w) + (w, Aw) \\
&= \|v\|_A^2 - (w, Mw) - (Mw, w) + (w, Aw) \\
&= \|v\|_A^2 - (w, Mw) - (w, M^T w) + (w, Aw) \\
&= \|v\|_A^2 - (w, (M^T + N)w) \\
&\leq \|v\|_A^2 - \lambda_{min}(M^T + N)\|w\|_2^2.
\end{aligned}$$

Ici $\lambda_{min}(M^T + N) > 0$ car la matrice $M^T + N$ est définie- positive par hypothèse. Il reste à minorer $\|w\|_2$ en fonction de $\|v\|_2$:

$$\begin{aligned} Av &= Mw \\ \Longrightarrow (Av, v) &= (Mw, v) \\ \lambda_{min}(A)\|v\|_2^2 \leq \|v\|_A^2 &\leq \|M\|_2\|w\|_2\|v\|_2 \\ \Longrightarrow \lambda_{min}(A)\|v\|_A^2 &\leq \|M\|_2^2\|w\|_2^2, \end{aligned}$$

avec encore $\lambda_{min}(A) > 0$ car A est définie-positive. Finalement

$$\|M^{-1}Nv\|_A^2 \leq [1 - \frac{\lambda_{min}(M^T + N)\lambda_{min}(A)}{\|M\|_2^2}]\|v\|_A^2$$

et on conclut que

$$\rho(M^{-1}N) \leq \|M^{-1}N\|_A < 1.$$

■

Remarque 4.2.1 *Ce résultat est encore valable si la matrice A n'est plus symétrique, mais reste définie-positive [10].*

Il est raisonnable de rechercher la meilleure convergence possible : pour cela il faut savoir comparer les vitesses de convergence de deux méthodes itératives associées aux décompositions régulières $A = M_1 - N_1 = M_2 - N_2$; si $\rho(M_1^{-1}N_1) < \rho(M_2^{-1}N_2) < 1$, alors la première méthode converge plus vite que la seconde.

On définit de manière plus précise la vitesse de convergence d'une méthode itérative comme la quantité

$$R(M^{-1}N) = -\log \rho(M^{-1}N)$$

qui est d'autant plus grande que le rayon spectral de la matrice $M^{-1}N$ est petit.

4.3 Itérations par points – Itérations par blocs

On présente maintenant quelques décompositions régulières classiques, en écrivant la matrice A sous la forme $A = D - E - F$ où $D, E, F \in \mathbb{R}^{n\times n}$ sont les matrices définies par

$$\begin{aligned} D_{i,j} = A_{i,i} \text{ si } i = j \text{ et } D_{i,j} = 0 \text{ si } i \neq j \\ E_{i,j} = -A_{i,j} \text{ si } i > j \text{ et } E_{i,j} = 0 \text{ si } i \leq j \\ F_{i,j} = -A_{i,j} \text{ si } i < j \text{ et } F_{i,j} = 0 \text{ si } i \geq j. \end{aligned}$$

D est une matrice diagonale,
E est donc une matrice triangulaire inférieure stricte,
F une matrice triangulaire supérieure stricte.

Cette écriture est dite **ponctuelle** puisque les indices i et j varient de 1 à n, elle se généralise à l'écriture **par blocs** en utilisant un découpage (ou partition) par blocs de la matrice A en p^2 blocs (cf. le chapitre 3) :

$$\begin{aligned} [D]_{k,l} = [A]_{k,l} \text{ si } k = l \text{ et } [D]_{k,l} = [0] \text{ si } k \neq l \\ [E]_{k,l} = -[A]_{k,l} \text{ si } k > l \text{ et } [E]_{k,l} = [0] \text{ si } k \leq l \\ [F]_{k,l} = -[A]_{k,l} \text{ si } k < l \text{ et } [F]_{k,l} = [0] \text{ si } k \geq l, \end{aligned}$$

où cette fois les indices k et l varient de 1 à p nombre de blocs de la partition. Dans ce formalisme les blocs extra-diagonaux peuvent être rectangulaires, mais les blocs diagonaux sont nécessairement carrés (cf. le chapitre 3). Cette écriture permet de distinguer les méthodes itératives **par points** des méthodes **par blocs**. Noter qu'une méthode par points constitue le cas extrême de la méthode par blocs dans lequel chaque bloc est réduit à un seul coefficient de la matrice A !

4.4 Critère de convergence

Au paragraphe B.6, on a introduit la notion de convergence d'un algorithme à la précision ε après k itérations, par la majoration (ici, on a ajouté la norme du résidu initial)

$$\|r^k\| \leq \varepsilon \|r^0\|, \tag{4.3}$$

pour une norme vectorielle $\|\cdot\|$ à déterminer. D'après la relation $e^k = (M^{-1}N)^k e^0$ et la Proposition B.4.2, on déduit qu'il existe une norme $\|\cdot\|$ telle que

$$\|e^k\| \leq (\rho(M^{-1}N))^k \|e^0\|, \text{ soit } \|A^{-1}r^k\| \leq (\rho(M^{-1}N))^k \|A^{-1}r^0\|,$$

puisque $r^k = b - Ax^k = -Ae^k$. Si on introduit le **nombre de conditionnement** de A, dans la norme $\|\cdot\|$, défini par :

$$\kappa(A) = \|A^{-1}\| \, \|A\|,$$

on en déduit

$$\|r^k\| = \|AA^{-1}r^k\| \leq \|A\|(\rho(M^{-1}N))^k \|A^{-1}r^0\| \leq \kappa(A)(\rho(M^{-1}N))^k \|r^0\|. \tag{4.4}$$

En comparant (4.3) et (4.4), on *estime* le nombre d'itérations K nécessaires pour vérifier le critère de convergence par la formule

$$\varepsilon = \kappa(A)(\rho(M^{-1}N))^k, \text{ soit } K = \frac{\log(\kappa(A)/\varepsilon)}{R(M^{-1}N)}. \tag{4.5}$$

4.5 Méthode de Jacobi

C'est la méthode itérative la plus ancienne : à partir d'une estimation x^k de la solution, on calcule le nouvel itéré x^{k+1} composante par composante en écrivant que chaque composante du résidu est nulle :

$$r_i^{k+1} = b_i - \sum_{j \neq i} A_{i,j} x_j^k - A_{i,i} x_i^{k+1} = 0.$$

Soit encore

$$A_{i,i} x_i^{k+1} = b_i - \sum_{j \neq i} A_{i,j} x_j^k.$$

Cette méthode correspond au choix $M = D$ et $N = E + F$ et les itérations s'écrivent

initialisation

$x^0 \in \mathbb{R}^n$

itérations : pour $k = 0, 1, \ldots,$ **faire**

$Dx^{k+1} = (E + F)x^k + b$

fin (4.6)

La matrice d'itération associée est notée

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

4.6 Méthode de Gauss-Seidel

Dans la formule précédente, on peut prendre en compte les nouvelles valeurs des composantes de x^{k+1} au fur et à mesure de leur calcul, en commençant par la première ($i = 1$), puis la deuxième ($i = 2$), etc. On obtient alors la relation

$$r_i^{k+1} = b_i - \sum_{j<i} A_{i,j} x_j^{k+1} - A_{i,i} x_i^{k+1} - \sum_{j>i} A_{i,j} x_j^k = 0.$$

Cette méthode correspond au choix $M = D - E$ et $N = F$, d'où les itérations

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (D - E)x^{k+1} = Fx^k + b \\ \textbf{fin} \end{array} \right. \tag{4.7}$$

La matrice d'itération associée est notée

$$G = (D - E)^{-1} F.$$

Dans ce cas, l'*ordre de numérotation* des inconnues à une influence sur l'algorithme, contrairement à l'algorithme de la méthode de Jacobi.

4.7 Méthode de relaxation

On introduit un paramètre réel $\omega \neq 0$ et on écrit que chaque composante x_i^{k+1} est une combinaison de la valeur connue x_i^k et de celle fournie par la méthode de Gauss-Seidel $\tilde{x}_i^{k+1}$:

$$\begin{aligned} & x_i^{k+1} = (1 - \omega) x_i^k + \omega \tilde{x}_i^{k+1} \\ \Longrightarrow\ & A_{i,i} x_i^{k+1} = (1 - \omega) A_{i,i} x_i^k + \omega \left(b_i - \sum_{j<i} A_{i,j} x_j^{k+1} - \sum_{j>i} A_{i,j} x_j^k \right). \end{aligned}$$

Ce qui revient à prendre

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

Les itérations s'écrivent

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (D - \omega E)x^{k+1} = \left(\omega F + (1 - \omega) D\right) x^k + \omega b \\ \textbf{fin} \end{array} \right. \tag{4.8}$$

La matrice d'itération est notée

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

On remarque que pour $\omega = 1$ on retrouve bien la méthode de Gauss-Seidel, i. e. $L_1 = G$.

Pour cette matrice, on peut écrire les relations

$$|\text{dét}(L_\omega)| = \prod_i |\lambda_i(L_\omega)| \leq \rho(L_\omega)^n,$$

ainsi que

$$\text{dét}(L_\omega) = \text{dét}\left((D)^{-1}\right) (1-\omega)^n \text{dét}(D) = (1-\omega)^n,$$

car les matrices E et F sont triangulaires à diagonale nulle ; on obtient finalement l'encadrement

$$|1-\omega| \leq |\text{dét}(L_\omega)|^{1/n} \leq \rho(L_\omega).$$

Ainsi la méthode diverge pour $\omega \notin]0,2[$. La méthode est dite

- de sous–relaxation quand $0 < \omega < 1$;
- de Gauss-Seidel quand $\omega = 1$;
- de sur–relaxation quand $1 < \omega < 2$.

En général on prend $\omega \in]1,2[$ et la méthode s'appelle en anglais Successive Over Relaxation (S.O.R.). On peut démontrer le théorème :

Théorème 4.7.1 *[Ostrowski–Reich] Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, la méthode de relaxation converge si et seulement si $\omega \in]0,2[$.*

Preuve : La condition est nécessaire puisque pour $\omega \notin]0,2[$ la méthode diverge.

Pour démontrer que la condition est suffisante, on vérifie que la matrice $M^T + N = \dfrac{2-\omega}{\omega} D$ est symétrique définie-positive quand $\omega \in]0,2[$ et A est symétrique définie-positive. Le résultat s'en déduit par application du Théorème 4.2.1. ∎

4.8 Matrices tridiagonales par blocs

Les matrices tridiagonales par blocs sont très courantes dans le cadre de l'approximation de solution d'équations différentielles par les méthodes des différences finies ou des éléments finis. On considère dans ce paragraphe les matrices dont la structure est de la forme

$$A = \begin{bmatrix} [D]_1 & -[F]_1 & & & \\ -[E]_1 & [D]_2 & -[F]_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -[E]_{p-2} & [D]_{p-1} & -[F]_{p-1} \\ & & & -[E]_{p-1} & [D]_p \end{bmatrix}$$

dans laquelle seuls les blocs diagonaux $[D]_j \in \mathbb{R}^{q_j \times q_j}$ sont a priori carrés.

Proposition 4.8.1 *Si la matrice A est tridiagonale par blocs, alors $\rho(G) = \rho(J)^2$.*

Preuve : On commence par définir pour tout $\lambda \neq 0$ la matrice tridiagonale par blocs

$$C(\lambda) = \begin{bmatrix} [D']_1 & -\lambda [F']_1 & & & \\ -\dfrac{1}{\lambda}[E']_1 & [D']_2 & -\lambda [F']_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -\dfrac{1}{\lambda}[E']_{p-2} & [D']_{p-1} & -\lambda [F']_{p-1} \\ & & & -\dfrac{1}{\lambda}[E']_{p-1} & [D']_p \end{bmatrix};$$

avec $[D']_j \in \mathbb{R}^{q_j \times q_j}$. On introduit ensuite la matrice

$$Q(\lambda) = \begin{bmatrix} \lambda I_{q_1} & & & & \\ & \lambda^2 I_{q_2} & & & \\ & & \ddots & & \\ & & & \lambda^{p-1} I_{q_{p-1}} & \\ & & & & \lambda^p I_{q_p} \end{bmatrix}.$$

Alors pour tout $\lambda \neq 0$, la matrice $C(\lambda)$ est semblable à $C(1)$ car

$$C(\lambda) = Q(1/\lambda)C(1)Q(\lambda) = Q^{-1}(\lambda)C(\lambda)Q(\lambda).$$

Examinons maintenant les valeurs propres de la matrice de Jacobi : par définition ce sont les racines du polynôme caractéristique

$$p_J(\lambda) = \text{dét}\left(D^{-1}(E+F) - \lambda I_n\right) = \text{dét}(\lambda D - E - F)/\text{dét}(-D)\,;$$

de même les valeurs propres de la matrice de Gauss-Seidel sont les racines du polynôme

$$p_G(\lambda) = \text{dét}\left((D-E)^{-1}F - \lambda I_n\right) = \text{dét}(\lambda D - \lambda E - F)/\text{dét}(E-D).$$

Noter que $\text{dét}(E-D) = \text{dét}(-D)$ et que

$$\begin{aligned} p_G(\lambda^2) &= \text{dét}(\lambda^2 D - \lambda^2 E - F)/\text{dét}(-D) \\ &= \text{dét}\ Q(\lambda)\ \text{dét}(\lambda^2 D - \lambda E - \lambda F)\ \text{dét}\ Q(1/\lambda)/\text{dét}(-D) \\ &= \lambda^n \text{dét}(\lambda D - E - F)/\text{dét}(-D) = \lambda^n p_J(\lambda) \end{aligned}$$

Ainsi lorsque λ est racine de p_J, λ^2 est racine de p_G, et réciproquement quand $\lambda \neq 0$. En fait ce calcul n'est correct que si $\lambda \neq 0$, puisqu'il fait intervenir la matrice $Q(1/\lambda)$. Noter que si $\lambda = 0$ est valeur propre de G, cela n'intervient pas dans le résultat car on étudie $\rho(G) = \max|\lambda|$. ■

Remarque 4.8.1 *Dans ce calcul on a défini la matrice $C(\lambda)$ à partir de la matrice A, de la manière suivante :*

$$\left[D'\right]_j = \lambda^2 \left[D\right]_j,\ \textit{puis}\ \left[E'\right]_j = -\lambda^2 \left[E\right]_j\ \textit{ et }\ \left[F'\right]_j = \left[F\right]_j\,;$$

c'est-à-dire que

$$C(\lambda) = \lambda^2 D - \lambda^2 E - F = Q^{-1}(\lambda)(\lambda^2 D - \lambda E - \lambda F)Q(\lambda).$$

Ce résultat montre que la méthode de Gauss–Seidel converge (ou diverge !) deux fois plus vite que la méthode de Jacobi pour les matrices tridiagonales par blocs.

Théorème 4.8.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice tridiagonale par blocs, telle que les valeurs propres de J soient toutes réelles, alors les méthodes de Jacobi, de Gauss-Seidel et de relaxation avec $\omega \in]0,2[$, divergent ou convergent simultanément.*

Comme pour la Proposition 4.8.1 on écrit

$$p_{L_\omega}(\lambda) = \text{dét}\left((\frac{1}{\omega}D - E)^{-1}(\frac{1-\omega}{\omega}D + F) - \lambda I_n\right)$$

d'où, en utilisant la relation $\det(E - \frac{1}{\omega}D) = \omega^{-n}\det(-D)$,

$$\begin{aligned}
p_{L_\omega}(\lambda) &= \det(\frac{\lambda+\omega-1}{\omega}D - \lambda E - F)\ \omega^n/\det(-D)\\
p_{L_\omega}(\lambda^2) &= \det(\frac{\lambda^2+\omega-1}{\omega}D - \lambda^2 E - F)\ \omega^n/\det(-D)\\
&= \det\ Q^{-1}(\lambda)\ \det(\frac{\lambda^2+\omega-1}{\omega}D - \lambda E - \lambda F)\ \det\ Q(\lambda)\ \omega^n/\det(-D)\\
&= \lambda^n \det(\frac{\lambda^2+\omega-1}{\lambda\omega}D - E - F)\ \omega^n/\det(-D)\\
&= \lambda^n\omega^n\ p_J(\frac{\lambda^2+\omega-1}{\lambda\omega}).
\end{aligned}$$

NB. *dans ce qui suit* $\zeta^{1/2}$ *représente une racine carrée complexe de* ζ.

Si λ est valeur propre non nulle de L_ω alors $\mu = \dfrac{\lambda+\omega-1}{\lambda^{1/2}\omega}$ est valeur propre de J.
Réciproquement si μ est valeur propre de J, alors les racines $\lambda_\pm$ de l'équation

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

sont valeurs propres de L_ω. Cette équation se met encore sous la forme

$$\lambda^2 + \lambda\left(2(\omega-1) - \mu^2\omega^2\right) + (\omega-1)^2 = 0$$

et on en déduit

$$\lambda_\pm = \frac{1}{2}(\mu^2\omega^2 - 2\omega + 2) \pm \frac{\mu\omega}{2}(\mu^2\omega^2 - 4\omega + 4)^{1/2}.$$

On suppose dans la suite que les valeurs propres μ de J sont réelles. Pour connaitre la valeur de $\rho(L_\omega)$, il faut étudier les variations de $\lambda_\pm$ en fonction de ω.

Exercice 4.8.1 *Etudier les variations de* $\lambda_\pm$ *en fonction de* ω.

On se contente de reproduire la courbe représentative de ces variations (figure 4.1)

Théorème 4.8.2 *Soit* $A \in \mathbb{R}^{n\times n}$ *une matrice symétrique définie-positive tridiagonale par blocs, alors les méthodes de Jacobi, de Gauss-Seidel et de relaxation pour* $\omega \in]0,2[$ *convergent.*
De plus, il existe une valeur optimale du paramètre ω

$$\omega_{opt} = \frac{2}{1+\sqrt{1-\rho(J)^2}}$$

telle que

$$\rho(L_{\omega_{opt}}) = \min_{\omega\in]0,2[}\rho(L_\omega) < \rho(G) = \rho(J)^2 < \rho(J) < 1.$$

Preuve : Pour appliquer le Théorème 4.8.1 il suffit de vérifier que les valeurs propres μ de J sont réelles :

$$\begin{aligned}
Jv = D^{-1}(E+F)v = \mu v &\Longrightarrow (E+F)v = \mu Dv\\
\text{soit encore } Av = (1-\mu)Dv &\Longrightarrow (v,Av) = (1-\mu)(v,Dv).
\end{aligned}$$

Si A est symétrique définie-positive alors nécessairement D l'est aussi, ainsi (v,Av) et (v,Dv) sont des réels positifs, et μ valeur propre de J est réelle et plus petite que 1. Alors d'après le Théorème 4.8.1 les méthodes de Jacobi, de Gauss-Seidel et de relaxation pour $0 < \omega < 2$ convergent. ■

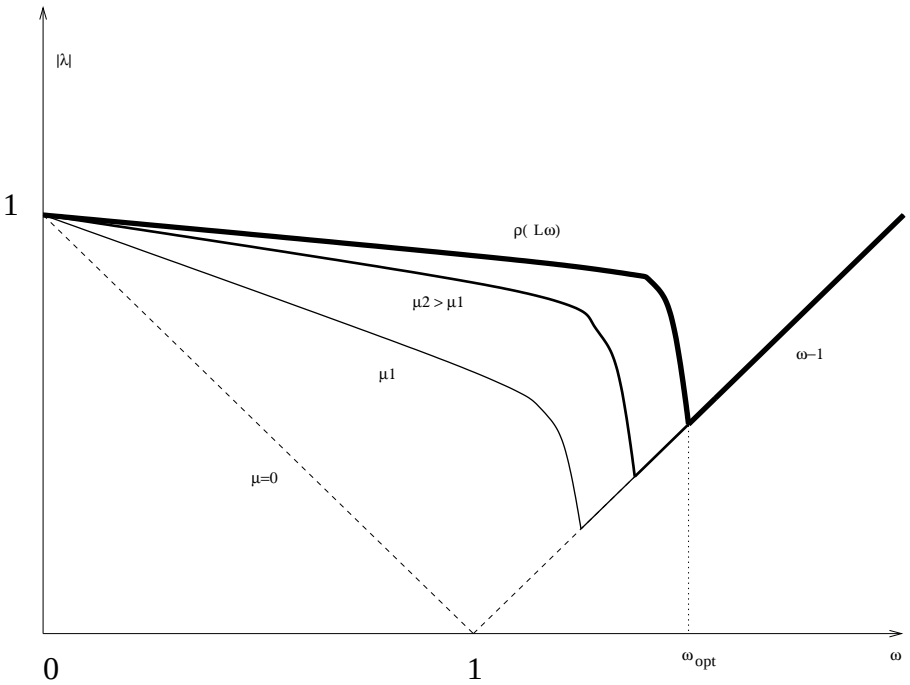


FIG. 4.1 – Les variations du rayon spectral $\rho(L_\omega)$

Remarque 4.8.2 *si on ne connait pas exactement l'expression de la valeur optimale ω_{opt} l'étude des variations de $\rho(L_\omega)$ montre qu'il vaut mieux l'approcher par valeurs supérieures puisque la dérivée $\dfrac{\partial\rho(L_\omega)}{\partial\omega}$ vaut 1 quand $\omega \to \omega_{opt+}$ mais tend vers $-\infty$ quand $\omega \to \omega_{opt-}$!*

4.9 Méthode de Jacobi relaxée

On peut définir une méthode de Jacobi relaxée :

$$\begin{aligned} x_i^{k+1} &= (1-\omega)x_i^k + \omega\tilde{x}_i^{k+1} \\ \Longrightarrow A_{i,i}x_i^{k+1} &= (1-\omega)A_{i,i}x_i^k + \omega[b_i - \sum_{i\neq j} A_{i,j}x_j^k]. \end{aligned}$$

ce qui revient à prendre $M = \dfrac{1}{\omega}D$ et $N = \dfrac{1-\omega}{\omega}D + E + F$, les itérations s'écrivent

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Dx^{k+1} = (\omega E + \omega F + (1-\omega)D)\, x^k + \omega b \\ \textbf{fin} \end{array} \right. \tag{4.9}$$

On note

$$J_\omega = D^{-1}\left((1-\omega)D + \omega E + \omega F\right) = (1-\omega)I_n + \omega J$$

la matrice d'itération associée, en remarquant que pour $\omega = 1$ on retrouve bien la méthode de Jacobi, i. e. $J_1 = J$.

Proposition 4.9.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive tridiagonale par blocs, alors la méthode de Jacobi relaxée converge si et seulement si*

$$0 < \omega < \frac{2}{1+\rho(J)}.$$

Preuve : Les valeurs propres de la matrice A étant réelles positives, celles de la matrice

$$J_\omega = (1-\omega)I_n + \omega(I_n - D^{-1}A) = I_n - \omega D^{-1}A$$

sont réelles pour tout ω. En particulier pour $\omega = 1$, on retrouve la matrice de Jacobi J dont les valeurs propres μ_i sont rangées par ordre décroissant

$$\mu_n \leq \mu_{n-1} \leq \ldots \leq \mu_2 \leq \mu_1 < 1$$

Les hypothèses du Théorème 4.8.2 étant satisfaites, la méthode de Jacobi converge et en conséquence $\rho(J) = \max|\mu_1|, |\mu_n| < 1$.

La matrice J_ω a pour valeurs propres

$$\mu_i(\omega) = 1 - \omega + \omega\mu_i$$

il faut donc étudier les variations de

$$\rho(J_\omega) = \max_i |1 - \omega + \omega\mu_i|$$

Il faut maintenant revenir sur une propriété des valeurs propres de J quand A est une matrice tridiagonale par blocs : si μ est valeur propre de J, alors $-\mu$ l'est aussi ! Pour cela on reprend l'argument utilisé dans la démonstration de la Proposition 4.8.1 :

$$p_J(\lambda) = \text{dét}\left(D^{-1}(E+F) - \lambda I_n\right) = \text{dét}(\lambda D - E - F)/\text{dét}(-D)$$

et on remarque que

$$\begin{aligned} p_J(-\lambda) &= \text{dét}(-\lambda D - E - F)/\text{dét}(-D) \\ &= \text{dét}\left(Q(1/-1)\right)\text{dét}(-\lambda D + E + F)\text{dét}\left(Q(-1)\right)/\text{dét}(-D) \\ &= (-1)^n\text{dét}(\lambda D - E - F)/\text{dét}(-D) \\ &= (-1)^n p_J(\lambda). \end{aligned}$$

On en déduit que $\rho(J) = \mu_1 = -\mu_n$, et la représentation graphique montre que $\rho(J_\omega) < 1$ si $\omega < 2/(1+\rho(J))$, ce qui termine la démonstration. ■

La représentation graphique montre également que la valeur optimale du paramètre est

$$\omega_{opt} = \frac{2}{2 - (\mu_1 + \mu_n)}.$$

Soit encore $\omega_{opt} = 1$ puisque $\mu_1 + \mu_n = 0$, il est donc inutile de relaxer la méthode de Jacobi pour les matrices tridiagonale par blocs !

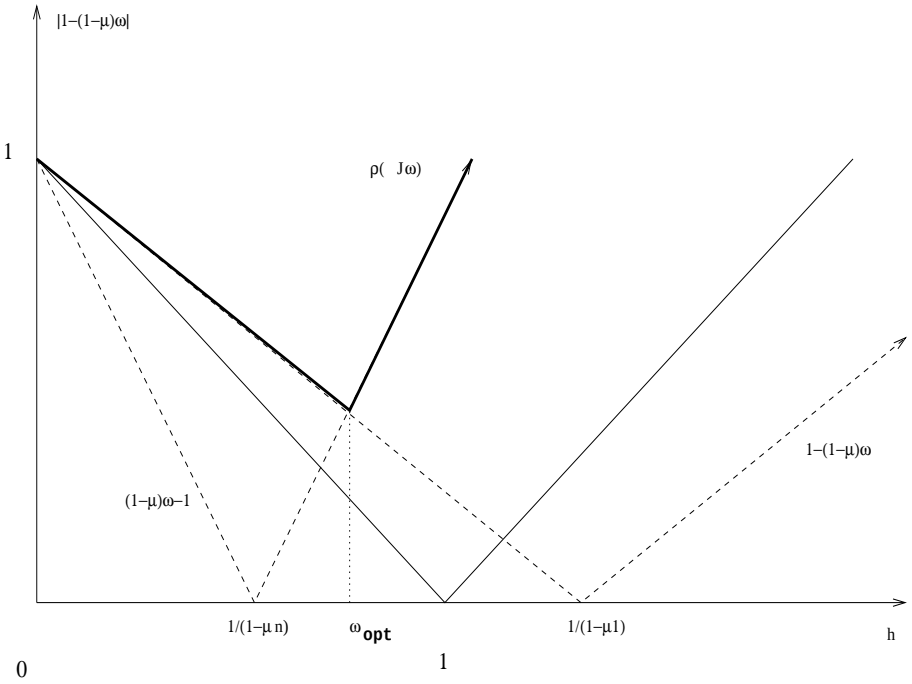


FIG. 4.2 – Les variations du rayon spectral $\rho(J)$

4.10 Méthode de Richardson

Une méthode itérative très utilisée en optimisation (cf. [4]) est la méthode de gradient à pas constant, qui calcule le nouvel itéré sous la forme

$$x^{k+1} = x^k + \alpha r^k, \quad \alpha \neq 0.$$

Le résidu r^k est le gradient d'une fonctionnelle que l'on cherche à minimiser et le paramètre α est le pas de descente. Lorsque A est une matrice symétrique définie-positive, on peut en effet résoudre le problème $Ax = b$ en minimisant la fonctionnelle

$$v \mapsto \frac{1}{2}(v, Av) - (b, v)$$

sur $\mathbb{R}^n$. Cette méthode correspond à la décomposition

$$M = \frac{1}{\alpha} I_n \quad \text{et} \quad N = \frac{1}{\alpha} I_n - A,$$

régulière pour tout $\alpha \neq 0$; on l'appelle méthode de Richardson du premier ordre à pas constant, et on l'écrit sous la forme

$$\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \alpha r^k \\ \textbf{fin} \end{array} \tag{4.10}$$

Proposition 4.10.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, la méthode de Richardson converge si et seulement si*

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

Preuve : On note $0 < \lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_1$ les valeurs propres de A, les valeurs propres de

$$R_\alpha = M^{-1}N = I_n - \alpha A$$

sont les $1 - \alpha\lambda_j$. Alors $\rho(R_\alpha) = |1 - \alpha\lambda_1|$; comme pour la méthode de Jacobi relaxée la convergence n'a lieu que si

$$0 < \alpha < \frac{2}{\lambda_1} = \frac{2}{\rho(A)}$$

et la valeur optimale est

$$\alpha = \frac{2}{(\lambda_1 + \lambda_n)}.$$

■

4.11 Méthode de Richardson à pas variable

En général on dispose de peu d'informations sur le spectre des matrices que l'on traite, et il est difficile de donner au paramètre α une valeur qui assure la convergence. Une variante de la méthode de Richardson fournit une solution à ce problème : on modifie le paramètre α à chaque itération, en lui donnant la valeur qui minimise la norme du résidu $r^{k+1} = r^k - \alpha A r^k$. La méthode de Richardson à pas variable s'écrit

$$\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \alpha^k r^k \\ \textbf{fin} \end{array} \tag{4.11}$$

Proposition 4.11.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, si*

$$\alpha^k = \frac{\|r^k\|_2^2}{\|r^k\|_A^2}$$

alors la méthode de Richardson à pas variable converge.

Preuve : Par construction

$$r^{k+1} = b - Ax^{k+1} = r^k - \alpha^k A r^k.$$

Pour toute norme vectorielle $\|\cdot\|_C$ avec C matrice symétrique définie-positive, on peut écrire

$$\begin{aligned} \|r^{k+1}\|_C^2 &= \left(r^{k+1}, Cr^{k+1}\right) = \left(r^k - \alpha^k A r^k, C(r^k - \alpha^k A r^k)\right) \\ &= \|r^k\|_C^2 - 2\alpha^k (Ar^k, Cr^k) + (\alpha^k)^2 (Ar^k, CAr^k). \end{aligned}$$

Cette expression atteint son minimum

$$\|r^{k+1}\|_C^2 = \|r^k\|_C^2 - \frac{|(CAr^k, r^k)|^2}{(Ar^k, CAr^k)}$$

quand

$$\alpha^k = \frac{(CAr^k, r^k)}{(Ar^k, CAr^k)}.$$

Lorsque la matrice A est symétrique définie-positive un choix simple pour le calcul de α^k est de prendre $C = A^{-1}$, alors

$$\alpha^k = \frac{\|r^k\|_2^2}{\|r^k\|_A^2} \quad \text{et} \quad \|r^{k+1}\|_{A^{-1}}^2 = \|r^k\|_{A^{-1}}^2 - \frac{\|r^k\|_2^4}{\|r^k\|_A^2}.$$

Exercice 4.11.1 *Montrer que pour toute matrice A, symétrique définie-positive*

$$\forall v \in \mathbb{R}^n \quad \frac{\|v\|_A^2}{\|v\|_2^2} \times \frac{\|v\|_{A^{-1}}^2}{\|v\|_2^2} \leq \kappa_2(A)$$

avec $\kappa_2(A) = \dfrac{\lambda_1}{\lambda_n} = \|A\|_2 \, \|A^{-1}\|_2$ nombre de conditionnement de A dans la norme $\|\cdot\|_2$.

En utilisant ce résultat, on obtient pour $v = r^k$ la majoration

$$\begin{aligned}\|r^{k+1}\|_{A^{-1}}^2 &= \|r^k\|_{A^{-1}}^2 \left[1 - \frac{\|r^k\|_2^4}{\|r^k\|_A^2 \, \|r^k\|_{A^{-1}}^2}\right] \\ &\leq \|r^k\|_{A^{-1}}^2 \, [1 - 1/\kappa_2(A)]\end{aligned}$$

Puisque par définition $\kappa_2(A) \geq 1$, on voit donc que $\|r^{k+1}\|_{A^{-1}}$ tend vers 0 quand k tend vers $+\infty$, et la méthode converge. ∎

4.12 Matrices à diagonale dominante

Il existe une catégorie de matrices importante dans l'histoire de l'étude des méthodes itératives : les matrices à diagonale dominante.

Définition 4.12.1 *une matrice $A \in \mathbb{R}^{n\times n}$ est dite à* **diagonale dominante** *si et seulement si*

$$\forall i, \quad 1 \leq i \leq n \qquad \sum_{j\neq i} |A_{i,j}| \leq |A_{i,i}|.$$

Une matrice $A \in \mathbb{R}^{n\times n}$ est dite à **diagonale strictement dominante** *si et seulement si*

$$\forall i, \quad 1 \leq i \leq n \qquad \sum_{j\neq i} |A_{i,j}| < |A_{i,i}|.$$

Exercice 4.12.1 *Montrer qu'une matrice à diagonale strictement dominante est inversible.*

Proposition 4.12.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice à diagonale strictement dominante, alors les méthodes de Jacobi et Gauss-Seidel par point convergent.*

Preuve : Par définition de la matrice de Jacobi, $J = D^{-1}(E+F)$ et

$$\|J\|_\infty = \max_i \sum_{j\neq i} \frac{|A_{i,j}|}{|A_{i,i}|}.$$

Donc si A est une matrice à diagonale strictement dominante

$$\rho(J) \leq \|J\|_\infty < 1.$$

Soit maintenant λ une valeur propre de la matrice de Gauss-Seidel $G = (D-E)^{-1}F$:

$$(D-E)^{-1}Fv = \lambda v \quad \Longleftrightarrow \quad Fv = \lambda(D-E)v,$$

et soit i la composante telle que $|v_i| = \max_j |v_j|$, alors on écrit

$$\begin{aligned} \lambda A_{i,i} v_i &= \lambda \sum_{j<i} A_{i,j} v_j - \sum_{j>i} A_{i,j} v_j \\ \Longrightarrow |\lambda||A_{i,i}| &\leq |\lambda| \sum_{j<i} |A_{i,j}| + \sum_{j>i} |A_{i,j}|. \\ \Longrightarrow |\lambda| \left(|A_{i,i}| - \sum_{j<i} |A_{i,j}| \right) &\leq \sum_{j>i} |A_{i,j}|. \end{aligned}$$

Par ailleurs on sait que

$$\begin{aligned} |A_{i,i}| - \sum_{j<i} |A_{i,j}| &> \sum_{j>i} |A_{i,j}| \geq 0, \\ \Longrightarrow |\lambda| &\leq \frac{\displaystyle\sum_{j>i} |A_{i,j}|}{\displaystyle |A_{i,i}| - \sum_{j<i} |A_{i,j}|} < 1; \end{aligned}$$

ainsi $\rho(G) < 1$ et la méthode converge. ∎

4.13 Méthode de relaxation symétrique (S.S.O.R.)

L'ordre des inconnues a-t-il une influence sur la convergence de la méthode ? Comme on l'a déjà remarqué, cette question a un sens car la numérotation des inconnues joue un rôle effectif dans la définition des méthodes de Gauss-Seidel et de relaxation : chaque composante x_i^{k+1} du nouvel itéré x_j^{k+1} est définie à partir des composantes d'indice inférieur x_j^{k+1} pour $j < i$ (sur ce sujet voir par exemple Adams et Jordan [1]).

Pour éviter les problèmes liés à la numérotation des inconnues quand on n'a pas d'information utile à exploiter, il est donc préférable de *symétriser* les itérations de S.O.R. en inversant l'ordre des calculs à chaque itération : on effectue une itération dans l'ordre croissant des inconnues et l'itération suivante dans l'ordre décroissant.

On obtient ainsi la méthode de **sur-relaxation symétrique**, Symmetric Successive Over Relaxation (S.S.O.R. en abrégé), qui s'écrit

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (D - \omega E)x^{k+1/2} = (\omega F + (1-\omega)D)\, x^k + \omega b \\ \quad (D - \omega F)x^{k+1} = (\omega E + (1-\omega)D)\, x^{k+1/2} + \omega b \\ \textbf{fin} \end{array} \right. \tag{4.12}$$

On note

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

la matrice d'itération associée. L'étude directe des valeurs propres de cette matrice est assez compliquée, mais on peut néanmoins vérifier le résultat suivant

Proposition 4.13.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, la méthode S.S.O.R. converge si et seulement si $\omega \in]0, 2[$.*

Remarque 4.13.1 *on peut encore montrer qu'il existe une valeur optimale du paramètre ω (voir Young [14]), mais on ne sait pas en donner une expression analytique dans le cas général. Pour certains systèmes linéaires, la valeur*

$$\omega'_{opt} = \frac{2}{1 + \sqrt{2\left(1 - \rho(J)\right)^2}}$$

est une bonne valeur pour laquelle S.S.O.R. converge environ deux fois plus vite que S.O.R. avec ω_{opt}, mais comme chaque itération de S.S.O.R. coûte environ deux fois plus cher qu'une itération de S.O.R., on ne gagne pas beaucoup ! Le véritable intérêt de cette méthode est de diminuer l'influence de la numérotation des inconnues sur la convergence.

4.14 Etude d'un exemple simple

Comparons maintenant les différentes méthodes présentées dans ce chapitre sur le problème du fil pesant fixé en ses extrémités, discrétisé par différences finies (voir le chapitre 2). On rappelle que le système linéaire résultant s'écrit

$$\mathbb{A}_1 x = b, \text{ avec } \mathbb{A}_1 = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & \ldots & -1 & 2 \end{pmatrix}, \tag{4.13}$$

où $h = \dfrac{1}{n+1}$ est le pas de discrétisation, avec x et b deux vecteurs de $\mathbb{R}^n$, et $\mathbb{A}_1 \in \mathbb{R}^{n\times n}$.

Exercice 4.14.1 *Montrer que les valeurs propres de la matrice* $\mathbb{A}_1$ *sont les*

$$\lambda^k = \frac{2}{h^2}\left[1-\cos(k\pi h)\right] \quad 1 \leq k \leq n,$$

associées aux vecteurs propres φ_h^k

$$\varphi^k = \left[\, sin(k\pi h), \quad sin(2k\pi h), \quad \ldots, \quad sin((n-1)k\pi h), \quad sin(nk\pi h)\,\right]^T.$$

On résout le système linéaire (4.13) par les méthodes itératives décrites dans ce chapitre :
– la matrice de la méthode de Jacobi est

$$J_1 = D_1^{-1}(E_1 + E_1^T) = I_n - D_1^{-1}\mathbb{A}_1.$$

Cette matrice a les mêmes vecteurs propres que $\mathbb{A}_1$ et ses valeurs propres sont les

$$\mu_h^k = 1 - \frac{h^2}{2}\lambda_h^k = \cos(k\pi h) = 1 - 2\sin^2(k\pi h/2) \quad 1 \leq k \leq n.$$

Ainsi pour h petit,

$$\rho(J_1) = 1 - 2\sin^2(\pi h/2) \approx 1 - \frac{\pi^2 h^2}{2}$$

et la vitesse de convergence de la méthode de Jacobi est donnée par la formule

$$R(J_1) = -\log\rho[J_1] \approx \pi^2 h^2/2\ ;$$

elle diminue quand le nombre d'inconnues augmente !

– il en est de même pour la méthode de Gauss-Seidel ; en appliquant la Proposition 4.8.1 à la matrice $\mathbb{A}_1$ qui est tridiagonale, on obtient pour la matrice

$$G_1 = (D_1 - E_1)^{-1}E_1^T$$

$$\rho(G_1) = \rho(J_1)^2 = \left[1 - 2\sin^2(\pi h/2)\right]^2 \approx 1 - \pi^2 h^2$$

et la vitesse de convergence de la méthode de Gauss-Seidel est

$$R(G_1) = -\log\rho[G_1] \approx \pi^2 h^2.$$

La méthode de Gauss-Seidel converge deux fois plus vite que la méthode de Jacobi, mais sa vitesse de convergence diminue aussi quand le nombre d'inconnues augmente !

– pour la méthode de relaxation

$$L_{1,\omega} = (D_1 - \omega E_1)^{-1}\left((1-\omega)D_1 + \omega E_1^T\right),$$

on utilise le résultat du Théorème 4.8.1 avec la valeur optimale ω_{opt}

$$\rho(L_{1,\omega_{opt}}) = \omega_{opt} - 1 = \frac{2}{1+\sqrt{1-\rho(J_1)^2}} - 1 = \frac{1-\sqrt{1-\rho(J_1)^2}}{1+\sqrt{1-\rho(J_1)^2}}$$

quand h tend vers 0,

$$1 - \rho(J_1)^2 = 1 - \rho(G_1) \approx 4\sin^2(\pi h/2) \approx \pi^2 h^2$$

$$\rho(L_{1,\omega_{opt}}) \approx \frac{1-\pi h}{1+\pi h} \approx 1 - 2\pi h.$$

Finalement la vitesse de convergence de la méthode de relaxation est

$$R(L_{1,\omega_{opt}}) = -\log \rho[L_{1,\omega_{opt}}] \approx 2\pi h.$$

– pour la méthode S.S.O.R. avec le paramètre "optimal" ω'_{opt}

$$\rho(S_{1,\omega'_{opt}}) = \omega'_{opt} - 1 = \frac{2}{1+\sqrt{2\left(1-\rho(J_1)\right)^2}} - 1 = \frac{1-\sqrt{2\left(1-\rho(J_1)\right)^2}}{1+\sqrt{2\left(1-\rho(J_1)\right)^2}};$$

quand h tend vers 0, la vitesse de convergence de la méthode S.S.O.R. vérifie

$$R(S_{1,\omega'_{opt}}) = -\log \rho[S_{1,\omega'_{opt}}] \approx 2\sqrt{2}\pi h = \sqrt{2}R(L_{1,\omega_{opt}}).$$

– pour la méthode de Richardson à pas constant

$$\lambda_1 = \frac{2}{h^2}\left[1+\cos(\pi h)\right] \quad \text{et} \quad \lambda_n = \frac{2}{h^2}\left[1-\cos(\pi h)\right]$$

d'où

$$\alpha_{opt} = \frac{2}{\lambda_1+\lambda_n} = \frac{h^2}{2} \quad \text{et} \quad \rho(R_{1,\alpha_{opt}}) = 1-\alpha_{opt}\lambda_n = \cos \pi h$$

soit pour la vitesse de convergence de la méthode de Richardson à pas constant

$$R(R_{1,\alpha_{opt}}) = -\log \cos \pi h \approx \pi^2 h^2/2.$$

Exercice 4.14.2 *Montrer que pour la matrice $\mathbb{A}_1$, le nombre d'opérations à effectuer à chaque itération est de l'ordre du nombre d'inconnues, quelle que soit la méthode itérative choisie.*

Remarque 4.14.1 *Cette propriété est vraie pour toutes les matrices creuses.*

On peut donc calculer le nombre total d'opérations pour obtenir la solution du système linéaire (c'est-à-dire le temps calcul) à partir du nombre d'itérations, *estimé* par la formule (4.5). Le tableau suivant rassemble toutes ces estimations en ordre de grandeur, au facteur $\log(\kappa(\mathbb{A}_1)/\varepsilon)$ près, qui ne dépend pas de la méthode itérative retenue.

Méthode	$R(B)$	Nb. itérations	Nb. opérations
Jacobi	$\pi^2h^2/2$	$2n^2$	$O(n^3)$
Gauss-Seidel	π^2h^2	n^2	$O(n^3)$
S.O.R. ω_{opt} (*)	$2\pi h$	$2\pi n$	$O(n^2)$ (*)
S.S.O.R. ω'_{opt} (*)	$2\sqrt{2}\pi h$	$\sqrt{2}\pi n$	$O(n^2)$ (*)
Richardson avec α_{opt}	$\pi^2h^2/2$	$2n^2$	$O(n^3)$

Sur cet exemple, les méthodes les moins coûteuses sont donc les méthodes S.O.R. et S.S.O.R. (*). Ces résultats sont néanmoins à prendre avec *précaution*. En effet, on connait explicitement les valeurs propres des matrices utilisées, et il s'agit donc d'un *cas très particulier*. Il n'existe pas de résultat comparable dans le cas général, puisque les valeurs optimales ω_{opt} et ω'_{opt} sont souvent inaccessibles. En pratique, il convient donc de comparer *toutes* les méthodes dont on dispose pour résoudre un système linéaire.

4.15 Itérations par points ou par blocs ?

Revenons sur la notion de méthodes itératives par points et par blocs. Ces méthodes ont été définies à partir de la structure des matrices : quand la matrice est tridiagonale par blocs, il est naturel d'utiliser une décomposition $A = M - N$ qui tienne compte de cette propriété. Comme il est presque toujours possible de définir une méthode *par points* là où on a défini une méthode *par blocs* - il suffit qu'il n'y ait pas de coefficient diagonal nul - on a tendance à penser intuitivement que la méthode *par blocs* convergera alors plus vite.
Examinons deux exemples. Pour commencer,

$$A' = \begin{bmatrix} 1 & 0 & -1/4 & 1/4 \\ 0 & 1 & -1/4 & 1/4 \\ -1/4 & -1/4 & 1 & 0 \\ 1/4 & 1/4 & 0 & 1 \end{bmatrix}$$

pour laquelle le rayon spectral de la matrice de Gauss-Seidel par points est $\rho(G'_p) = 0.25$. Si on définit la décomposition

$$D' = \begin{bmatrix} 1 & 0 & -1/4 & \vdots & 0 \\ 0 & 1 & -1/4 & \vdots & 0 \\ -1/4 & -1/4 & 1 & \vdots & 0 \\ \dots & \dots & \dots & \cdot & \dots \\ 0 & 0 & 0 & \vdots & 1 \end{bmatrix} \quad E' = \begin{bmatrix} 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ \dots & \dots & \dots & \cdot & \dots \\ -1/4 & -1/4 & 0 & \vdots & 0 \end{bmatrix}.$$

alors le rayon spectral de la matrice de Gauss-Seidel par blocs

$$G'_b = (D' - E')^{-1} E'^T$$

vaut $\rho(G'_b) = 0.1429$; il est bien plus petit que $\rho(G'_p)$.

Cette propriété n'est pas toujours satisfaite comme le montre le second exemple, tiré de [14] :

$$A'' = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix}$$

est symétrique définie-positive et on définit la décomposition

$$D'' = \begin{bmatrix} 5 & 2 & \vdots & 0 \\ 2 & 5 & \vdots & 0 \\ \dots & \dots & \cdot & \dots \\ 0 & 0 & \vdots & 5 \end{bmatrix} \quad E'' = \begin{bmatrix} 0 & 0 & \vdots & 0 \\ 0 & 0 & \vdots & 0 \\ \dots & \dots & \dots & \dots \\ -2 & -3 & \vdots & 0 \end{bmatrix}.$$

Le rayon spectral de la matrice de Gauss-Seidel par blocs $\rho(G''_b) = 0.3905$ est plus grand que le rayon spectral de la matrice de Gauss-Seidel par points $\rho(G''_p) = 0.3098$!

Chapitre 5

Méthode de la puissance itérée

5.1 Introduction

Le but de ce chapitre est de construire des algorithmes de calcul effectif des valeurs propres et vecteurs propres d'une matrice. Dans ce chapitre, est présentée la méthode de la puissance itérée, ainsi que les méthodes dérivées : la puissance itérée avec translation, avec déflation, et la puissance itérée inverse.

5.2 Etude d'un exemple

Avant de présenter en détail la méthode de la puissance itérée, examinons sur un exemple concret l'effet de la multiplication répétée d'un vecteur par une matrice. Ce type de problème rentre dans la catégorie de l'étude de la stabilité des systèmes dynamiques, voir [8].

Soit à résoudre le problème suivant : *Les statistiques du marché du travail pour les étudiants de l'enseignement supérieur montrent que chaque mois un étudiant (diplômé !) sur deux est pris en stage, et que un stagiaire sur quatre est embauché. Vers quel état évolue la population des étudiants diplômés ?*

Pour le mois k, on appelle e^k le nombre d'étudiants qui ne sont ni en stage, ni embauchés, s^k le nombre de stagiaires et E^k le nombre d'étudiants embauchés. Le problème est résumé par la relation linéaire :

$$\begin{bmatrix} e^{k+1} \\ s^{k+1} \\ E^{k+1} \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 3/4 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} e^k \\ s^k \\ E^k \end{bmatrix}.$$

La matrice de cette relation ayant ses valeurs propres distinctes est diagonalisable, d'après la Corollaire A.2.1 :

$$A = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 3/4 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}^{-1}.$$

Après k mois l'état de la population est donné par la formule

$$\begin{bmatrix} e^k \\ s^k \\ E^k \end{bmatrix} = A^k \begin{bmatrix} e^0 \\ s^0 \\ E^0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 1 \end{bmatrix}^k \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} e^0 \\ s^0 \\ E^0 \end{bmatrix}.$$

Quand k tend vers l'infini, on tend vers l'état stable

$$\begin{bmatrix} e^\infty \\ s^\infty \\ E^\infty \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} e^0 \\ s^0 \\ E^0 \end{bmatrix}$$

dans lequel tout le monde est embauché, quelle que soit la situation initiale ! Si on examine cette dernière relation, on constate lorsque k tend vers l'infini les trois vecteurs colonnes de la matrice A^k tendent vers le vecteur propre associé à la valeur propre de plus grand module : $|\lambda| = 1$.

Cet exemple nous incite à considérer l'algorithme suivant, faisant intervenir les puissances successives d'une matrice A de $\mathbb{C}^{n\times n}$ pour calculer une valeur propre de plus grand module et un vecteur propre associé. Si on note $\|\cdot\|$ une norme quelconque de $\mathbb{C}^n$, soit l'algorithme :

1) **initialisation :**

$v_0 \in \mathbb{C}^n$ tel que $\|v_0\| = 1$

2) **itérations : pour** $k = 1, 2, \ldots$ **faire**

$v_k = Av_{k-1}/\|Av_{k-1}\|$

fin

Par construction, on a pour tout $k \geq 1$, $\|v_k\| = 1$. On impose cette propriété pour éviter que la norme de ce vecteur tende vers l'infini... A partir des relations de cet algorithme et de la décomposition spectrale de la matrice A établie au paragraphe A.6, on vérifie que

$$v_0 = \sum_{i=1}^{d} P_i v_0, \quad A^k v_0 = \sum_{i=1}^{d} (\lambda_i P_i + D_i)^k P_i v_0, \ k \geq 1.$$

$$\text{Donc } v_k = \frac{1}{\alpha_k} \sum_{i=1}^{d} (\lambda_i P_i + D_i)^k P_i v_0, \text{ avec } \alpha_k = \|\sum_{i=1}^{d} (\lambda_i P_i + D_i)^k P_i v_0\|.$$

Théorème 5.2.1 *Soit $A \in \mathbb{C}^{n\times n}$. On suppose qu'il n'existe qu'une seule valeur propre λ_1 de plus grand module et que cette valeur propre est semi–simple. Soit v_0 un choix initial possédant une composante non nulle sur le sous–espace M_1 ($P_1 v_0 \neq 0$). Alors, en appelant r_1 (resp. θ_1) le module (resp. l'argument) de λ_1 ($\lambda_1 = r_1 e^{\imath\theta_1}$), on peut démontrer successivement que*

(i) $\lim_{k\to\infty} (e^{-\imath k\theta_1} v_k) = \dfrac{P_1 v_0}{\|P_1 v_0\|}$;

(ii) $\lim_{k\to\infty} \|Av_k\| = r_1$;

(iii) Soit j telle que $(P_1 v_0)_j \neq 0$: $\lim_{k\to\infty} \dfrac{v_{k+1}^j}{v_k^j} = e^{\imath\theta_1}$, *avec v^j la $j^{\text{ème}}$ composante de v.*

Remarque 5.2.1 *Dans le cas où $\lambda_1 \in \mathbb{R}_\star^+$, (ii) signifie que $\lim_{k\to\infty} \|Av_k\| = \lambda_1$. C'est toujours le cas d'une matrice hermitienne (resp . symétrique) définie-positive de $\mathbb{C}^{n\times n}$ (resp. $\mathbb{R}^{n\times n}$).*

Preuve : Puisque λ_1 est supposée semi–simple, $D_1 = [0]$ et on écrit

$$v_k = \frac{1}{\alpha_k} \left[\lambda_1^k P_1 v_0 + \sum_{i=2}^{d} (\lambda_i P_i + D_i)^k P_i v_0 \right] = \frac{\lambda_1^k}{\alpha_k} \left[P_1 v_0 + e_k \right], \text{ avec } e_k = \sum_{i=2}^{d} \frac{1}{\lambda_1^k} (\lambda_i P_i + D_i)^k P_i v_0.$$

Le rayon spectral de la matrice

$$Q = \sum_{i=2}^{d} \frac{1}{\lambda_1}(\lambda_i P_i + D_i)P_i,$$

égal à $|\lambda_2|/|\lambda_1|$, est strictement plus petit que 1 par hypothèse. Ainsi la suite de vecteurs $(e_k)_k$ tend vers 0 quand $k \to +\infty$, d'après le Théorème B.5.1. Par ailleurs,

$$\frac{\alpha_k}{r_1^k} = \frac{1}{r_1^k}\|\lambda_1^k(P_1v_0 + e_k)\| = \|P_1v_0 + e_k\| \to \|P_1v_0\|.$$

Notons que, d'après la Proposition A.6.1, puisque par hypothèse $P_1v_0 \neq 0$, P_1v_0 est un vecteur propre associé à λ_1. On introduit maintenant les vecteurs auxiliaires $w_k = \mathrm{e}^{-\imath k\theta_1}\, v_k$. On trouve

$$w_k = \mathrm{e}^{-\imath k\theta_1}\frac{\lambda_1^k}{\alpha_k}(P_1v_0 + e_k) = \frac{r_1^k}{\alpha_k}(P_1v_0 + e_k) \to \frac{P_1v_0}{\|P_1v_0\|}, \text{ c'est-à-dire (i).}$$

Pour prouver (ii), on remarque que

$$Av_k = A\left(\mathrm{e}^{\imath k\theta_1}w_k\right) = \mathrm{e}^{\imath k\theta_1}Aw_k, \text{ d'où } \|Av_k\| = \|Aw_k\|.$$

Comme $(w_k)_k$ est une suite convergente d'après (i), il en est de même pour $(Aw_k)_k$, et

$$\|Av_k\| = \|Aw_k\| \to \frac{\|AP_1v_0\|}{\|P_1v_0\|} = r_1, \text{ c'est-à-dire (ii).}$$

Pour prouver (iii), nous considérons une coordonnée j telle que $(P_1v_0)_j \neq 0$, ou $(e_j, P_1v_0) \neq 0$, avec $(e_j)_j$ la base canonique de $\mathbb{C}^n$.
On a la relation $v_k^j = (e_j, v_k) = \mathrm{e}^{\imath k\theta_1}(e_j, w_k)$. D'après (i), (e_j, w_k) tend vers $(P_1v_0)_j/\|P_1v_0\|$ qui est non nul par hypothèse. Ainsi, il existe k_0 tel que, pour tout $k \geq k_0$, $(e_j, w_k) \neq 0$, et donc $v_k^j \neq 0$. Par ailleurs,

$$v_{k+1}^j = (e_j, v_{k+1}) = \frac{(e_j, Av_k)}{\|Av_k\|} = \mathrm{e}^{\imath k\theta_1}\frac{(e_j, Aw_k)}{\|Aw_k\|}.$$

Pour $k \geq k_0$, on a donc

$$\frac{v_{k+1}^j}{v_k^j} = \frac{1}{\|Aw_k\|}\frac{(e_j, Aw_k)}{(e_j, w_k)} \to \frac{\lambda_1}{r_1} = \mathrm{e}^{\imath\theta_1}, \text{ c'est-à-dire (iii).}$$

■

On note que

1) l'algorithme fournit une valeur propre et un vecteur propre associé. En effet, d'une part (ii-iii) fournissent λ_1 et, d'autre part, (i) fournit un vecteur propre de M_1, puisque P_1v_0 appartient toujours à ce sous-espace propre.

2) la vitesse de convergence de l'algorithme est lié au rapport $\rho_{1,2} = |\lambda_2|/|\lambda_1|$, où λ_2 est la deuxième valeur propre de plus grand module. De fait, la vitesse de convergence est liée à la façon dont $(e_k)_k$ tend vers 0, ce qui dépend du rayon spectral $\rho(Q)$, qui est lui-même inférieur ou égal à $\rho_{1,2}$ (cf. la discussion du paragraphe 4.4.)

5.3 Méthode de la puissance inverse itérée

Si on suppose que la matrice $A \in \mathbb{C}^{n\times n}$ est inversible, alors 0 n'est pas valeur propre. Rangeons les valeurs propres par ordre de module décroissant

$$Spe(A) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_2, \lambda_1\}$$

alors

$$Spe(A^{-1}) = \{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_{n-1}}, \frac{1}{\lambda_n}\}$$

et les vecteurs propres de A sont aussi vecteurs propres de A^{-1} : $Au = \lambda u \Longleftrightarrow A^{-1}u = \frac{1}{\lambda}u$. Donc si on veut calculer la valeur propre de A de plus petit module λ_n, on applique l'algorithme de la puissance itérée à la matrice inverse A^{-1} : c'est la méthode de la puissance itérée inverse :

1) **initialisation :**

$v_0 \in \mathbb{C}^n$ tel que $\|v_0\| = 1$

2) **itérations : pour** $k = 1, 2, \ldots$ **faire**

$v_k = A^{-1}v_{k-1}/\|A^{-1}v_{k-1}\|$

fin

Cet algorithme fournit la valeur propre de plus grand module de A^{-1} : soit $\frac{1}{\lambda_n}$. La vitesse de convergence est liée, cette fois, au rapport $\rho' = |\lambda_n|/|\lambda_{n-1}|$.

Dans la pratique pour calculer v_k, on effectue une factorisation de la matrice A par la méthode de Cholesky si A est symétrique définie-positive (resp. par la méthode de Gauss si A est quelconque), et on résout le système linéaire $L\,L^T v_k = v_{k-1}$ (respectivement $L\,U v_k = v_{k-1}$).

5.4 Technique de translation

Le problème qui se pose maintenant est comment obtenir les autres valeurs propres, une fois que l'on a calculé les valeurs propres extrêmes ? Une réponse est fournie par la technique de **translation** (**shift** en anglais), qui consiste à rechercher les valeurs propres de la matrice $A - \sigma I$. Si le spectre de A est

$$Spe(A) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_2, \lambda_1\}$$

le spectre de la matrice $\tilde{A} = A - \sigma I$ est

$$Spe(\tilde{A}) = \{\lambda_n - \sigma, \lambda_{n-1} - \sigma, \ldots, \lambda_2 - \sigma, \lambda_1 - \sigma\}.$$

Un choix judicieux de σ, tel que $|\lambda_1 - \sigma| < |\lambda_j - \sigma|$, permet à la méthode de la puissance itérée appliquée à $\tilde{A}$ de converger vers une valeur propre $\lambda_j - \sigma$, avec $\lambda_j \neq \lambda_1$.

Il faut être prudent dans le choix de σ car on n'obtient pas obligatoirement les valeurs propres dans l'ordre des modules décroissants par cette technique. Par exemple si $Spe(A) = \{-2, 3, 5\}$, la méthode de la puissance itérée appliquée à A converge vers $\lambda_1 = 5$, si on l'applique à la matrice $A - 2I$, elle converge vers -4 car $Spe(A - 2I) = \{-4, 1, 3\}$; on a donc calculé la valeur propre $\lambda_3 = -4 + 2 = -2$ et non $\lambda_2 = 3$!

Sur le graphique 5.1, on voit que la méthode de translation ne permet pas d'atteindre la valeur propre $\lambda_2 = 3$, car pour toute valeur du paramètre σ, la courbe représentant les variations de

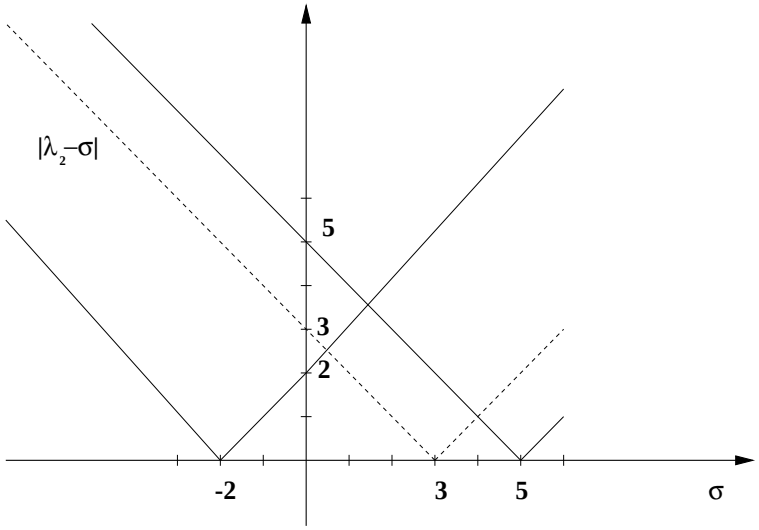


FIG. 5.1 – Les variations de $|\lambda - \sigma|$

$|\lambda_2 - \sigma|$ est toujours comprise entre les courbes $|\lambda_1 - \sigma|$ et $|\lambda_3 - \sigma|$. Pour obtenir λ_2, il faut travailler sur le spectre de A^{-1}, comme le montre la figure 5.2. Dans ce cas, on applique la technique de translation à l'algorithme de la puissance itérée inverse, en factorisant la matrice $\tilde{A} = A - \sigma I$ pour le calcul des itérés successifs... Si λ est la valeur propre la plus proche de σ, alors $\dfrac{1}{\lambda - \sigma}$ est la valeur propre de plus grand module de $(A - \sigma I)^{-1}$. La convergence est liée cette fois au rapport

$$\frac{\dfrac{1}{|\lambda' - \sigma|}}{\dfrac{1}{|\lambda - \sigma|}} = \frac{|\lambda - \sigma|}{|\lambda' - \sigma|},$$

avec λ' telle que $|\lambda' - \sigma|$ est le deuxième plus petit module. Ce rapport peut être très petit si σ est proche de λ (et assez éloigné de λ'). La convergence de la méthode est donc très rapide (quelques itérations) si on dispose d'une bonne estimation de λ.

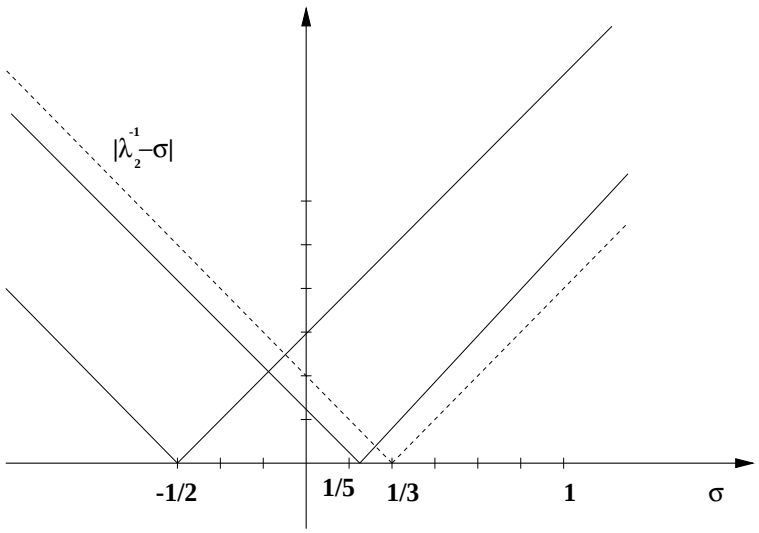


FIG. 5.2 – Les variations de $|\lambda^{-1} - \sigma|$

Cette méthode est donc utilisée comme accélération de la méthode de la puissance itérée

inverse, mais aussi pour le calcul des vecteurs propres lorsque l'on a obtenu une estimation des valeurs propres par un autre algorithme. On voit par ailleurs qu'il n'est pas nécessaire d'avoir une estimation fine de ces valeurs propres puisque la méthode de la puissance itérée inverse fournit des valeurs plus précises.

Remarque 5.4.1 *quand la valeur σ est proche de la valeur exacte de λ, la matrice $A - \sigma I$ est presque singulière ; ce phénomène pourrait introduire des problèmes numériques au cours de la factorisation de cette matrice, mais Parlett a montré que les calculs restaient stables et qu'on pouvait utiliser cette méthode sans modification [11].*

5.5 Technique de déflation

Une autre façon de calculer différentes valeurs propres d'une matrice par la méthode de la puissance itérée, consiste à retirer les valeurs propres du spectre de A de la manière suivante appelée technique de **déflation**.

On suppose connue une valeur propre λ_j de la matrice A et un vecteur propre associé u_j, on définit alors la matrice

$$\tilde{A} = A - \sigma u_j \cdot v^*$$

où σ est un paramètre complexe et $v^* \in \mathbb{C}^{1\times n}$ un vecteur ligne tel que $v^* u_j = 1$.

Théorème 5.5.1 *Soit une matrice $A \in \mathbb{C}^{n\times n}$* diagonalisable *possédant d valeurs propres distinctes, de spectre*

$$Spe(A) = \{\lambda_d, \lambda_{d-1}, \ldots, \lambda_2, \lambda_1\}.$$

On suppose que la valeur propre λ_j est simple. *Alors la matrice $\tilde{A}$ a pour spectre*

$$Spe(\tilde{A}) = \{\lambda_d, \lambda_{d-1}, \ldots, \lambda_j - \sigma, \ldots, \lambda_2, \lambda_1\},$$

avec, le cas échéant, $\lambda_j - \sigma \in \{\lambda_i,\ i \neq j\}$.

Preuve : On sait que les valeurs propres sont associées indifféremment à des vecteurs propres à gauche, ou à droite, cf. la Proposition A.2.5. Soit donc, pour $\lambda_i \neq \lambda_j$, $w_i^* \in \mathbb{C}^{1\times n}$ un vecteur propre à gauche de A. D'après la Proposition A.2.6, comme λ_i est distincte de λ_j, on a $w_i^* u_j = 0$:

$$w_i^* \tilde{A} = w_i^* A - \sigma (w_i^* u_j) v^* = w_i^* A = \lambda_i w_i^*.$$

Ainsi λ_i est valeur propre de $\tilde{A}$, et w_i^* vecteur propre à gauche de A est aussi vecteur propre à gauche de $\tilde{A}$. Ceci est valable pour tout vecteur propre à gauche associé à λ_i : les ordres de multiplicité de $\tilde{A}$ et A vérifient donc

$$m_i(\tilde{A}) \geq m_i(A),\ \text{pour } i \neq j,\ \text{d'où } \sum_{i\neq j} m_i(\tilde{A}) \geq n-1.$$

D'autre part

$$\tilde{A} u_j = A u_j - \sigma u_j (v^* u_j) = A u_j - \sigma u_j = (\lambda_j - \sigma) u_j.$$

Donc u_j est un vecteur propre associé à la valeur propre $\lambda_j - \sigma$ de $\tilde{A}$. Deux cas peuvent se présenter :

- $\lambda_j - \sigma \notin \{\lambda_i,\ i \neq j\}$: son ordre de multiplicité est de 1 pour $\tilde{A}$, et on a bien retrouvé toutes les valeurs propres de $\tilde{A}$ (avec le même ordre de multiplicité que pour A.)
- $\lambda_j - \sigma = \lambda_i$, pour $i \neq j$. On note que, d'après la Proposition A.4.6, il existe m_i vecteurs propres à gauche indépendants $(w_{i,k}^*)_{1\leq k\leq m_i}$ associés à λ_i tels que $w_{i,k}^* u_j = 0$ (ou, par transconjugaison, $(w_{i,k}, u_j) = 0$.) Ainsi la famille $(w_{i,1}, \cdots, w_{i,m_i}, u_j)$ est libre, et $m_i(\tilde{A}) = m_i(A) + 1$. On a également retrouvé toutes les valeurs propres de $\tilde{A}$.

■

Quels sont les autres vecteurs propres à droite de la matrice $\tilde{A}$? On les cherche sous la forme $\tilde{u}_i = u_i - \gamma_i u_j$ pour $i \neq j$:

$$\tilde{A}\tilde{u}_i = (A - \sigma u_j \cdot v^*)(u_i - \gamma_i u_j) = \lambda_i u_i - (\gamma_i(\lambda_j - \sigma) + \sigma v^* u_i)\, u_j.$$

Pour que $\tilde{u}_i$ soit vecteur propre de $\tilde{A}$ associé à λ_i, il faut et il suffit que

$$\lambda_i u_i - (\gamma_i(\lambda_j - \sigma) + \sigma v^* u_i)\, u_j = \lambda_i\, (u_i - \gamma_i u_j)$$

soit encore

$$\gamma_i(\lambda_j - \lambda_i - \sigma) = \sigma v^* u_i.$$

Finalement on a l'alternative

a) $\sigma \neq \lambda_j - \lambda_i \Longrightarrow \gamma_i = \dfrac{\sigma v^* u_i}{\lambda_j - \lambda_i - \sigma}$ et $u_i - \gamma_i u_j$ est aussi vecteur propre ;

b) $\sigma = \lambda_j - \lambda_i \Longrightarrow \lambda_i = \lambda_j - \sigma$ est alors valeur propre multiple de $\tilde{A}$
et u_j est le seul vecteur propre connu .

Remarque 5.5.1 *1) le choix du vecteur v^* du théorème ne pose pas de difficulté a priori, on peut par exemple prendre v^* égal à w_j^*, vecteur propre à gauche associé à λ_j. Ce choix conduit à $\gamma_i = 0$, car dans ce cas on a automatiquement $v^* u_i = 0$.*

2) dans la pratique, il n'est pas nécessaire de calculer la matrice $\tilde{A}$, car dans l'algorithme de la puissance itérée, il suffit de calculer le produit $\tilde{A}v_k = Av_k - \sigma u_j(v^* v_k)$.

Annexe A

Valeurs propres et vecteurs propres

A.1 Introduction

Dans ce chapitre, on donne quelques résultats théoriques fondamentaux qui sont utilisés pour construire des algorithmes de calcul des éléments propres des matrices. La présentation de la forme de Jordan, puis de la décomposition spectrale d'une matrice permet d'établir une distinction entre matrices diagonalisables et matrices défectives.

A.2 Rappels

Avant de commencer l'étude des propriétés spectrales des matrices il faut rappeler quelques notions utiles : tout d'abord il est nécessaire de se placer dans le corps $\mathbb{C}$ des nombres complexes, car les valeurs propres et vecteurs propres d'une matrice à coefficients réels peuvent être complexes. A l'exception de certains cas particuliers, tous les calculs présentés dans ce chapitre sont donc effectués avec des nombres complexes.

Soit $B = \{b_1, b_2, \ldots, b_n\}$ une base de $\mathbb{C}^n$. On peut écrire tout vecteur x sous la forme $x = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n$, avec $(x_i)_{i=1,n}$ les n coordonnées de x dans la base B. On peut alors définir

$$(u, v) = u^* v = \sum_{i=1}^{n} \overline{u}_i \; v_i$$

le produit scalaire complexe de $\mathbb{C}^n$, $\overline{x}$ désignant le complexe conjugué de x.

On associe à toute matrice $A \in \mathbb{C}^{n \times m}$, la matrice "transconjuguée", notée A^*, définie par

$$A^*_{i,j} = \overline{A}_{j,i} \quad 1 \leq i \leq n, \quad 1 \leq j \leq m.$$

On dit que cette matrice est l'**adjointe** de A pour le produit scalaire complexe, puisque

$$\forall u, v \in \mathbb{C}^n \qquad (Au, v) = (u, A^* v).$$

Définition A.2.1 – *On appelle* **valeur propre** *d'une matrice A, toute racine complexe λ_i, ou $\lambda_i(A)$, du* **polynôme caractéristique** $p(\lambda) = dét(A - \lambda I)$. *A ce titre on associe à chaque valeur propre sa* **multiplicité algébrique** m_i, *qui est l'ordre de multiplicité de λ_i en tant que racine de p. Si on note d le nombre de racines distinctes de p, on peut donc écrire*

$$p(\lambda) = \prod_{i=1}^{d} (\lambda - \lambda_i)^{m_i}.$$

– *On définit aussi la* **valeur propre** λ_i *et un* **vecteur propre** *associé* u_i *comme un couple* (λ_i, u_i) *solution du problème* $Au_i = \lambda_i u_i$, avec $u_i \neq 0$, *ce qui peut encore s'exprimer par la relation d'appartenance* $u_i \in \mathrm{Ker}\,(A - \lambda_i I) \setminus \{0\}$. *On introduit donc naturellement la notion de* **multiplicité géométrique** *de* λ_i *par* $g_i = dim\,(\mathrm{Ker}\,(A - \lambda_i I))$, *pour* i *compris entre 1 et* d.

Proposition A.2.1 *Les multiplicités* $(m_i)_{i=1,d}$ *et* $(g_i)_{i=1,d}$ *vérifient les relations :*

(i) $\displaystyle\sum_{i=1}^{d} m_i = n$;

(ii) $g_i \leq m_i$, *pour* $i \in \{1, \cdots, d\}$;

(iii) $\displaystyle\sum_{i=1}^{d} g_i \leq n$.

Preuve : Comme tout polynôme à coefficients complexes est scindé dans $\mathbb{C}$, on en déduit immédiatement la relation (i).

Pour prouver (ii), on introduit, pour i fixé, $B_i = (e_1, \cdots, e_{g_i})$ une base de $\mathrm{Ker}\,(A - \lambda_i I)$. On la complète en une base B' de $\mathbb{C}^n$. La matrice A est alors semblable à la matrice par blocs ci-dessous,

$$A' = \begin{bmatrix} \lambda_i I_{g_i} & X \\ 0 & Y \end{bmatrix},$$

qui représente l'application linéaire asssociée, exprimée cette fois dans la base B'. On a alors

$$p(\lambda) = \mathrm{dét}\,(A' - \lambda I_n) = (\lambda_i - \lambda)^{g_i}\,\mathrm{dét}\,(Y - \lambda I_{n-g_i}).$$

Ainsi λ_i est racine de p d'ordre au moins g_i, ce qui prouve (ii).

Pour finir, (iii) est une conséquence immédiate des deux points précédents. ■

Avant de détailler plus avant la présentation, nous rappelons le résultat bien connu

Proposition A.2.2 *Soient* $(v_k)_k$ d' *vecteurs propres de* A, *associés à des valeurs propres deux à deux distinctes. Alors* $(v_k)_k$ *est une famille libre de* $\mathbb{C}^n$.

Preuve : Raisonnons par récurrence sur d'.
Pour $d' = 1$, on note que, par définition des vecteurs propres, $v_1 \neq 0$. Ainsi, (v_1) est bien une famille libre.
Supposons le résultat vrai $d' - 1$. Considérons une famille $(v_k)_k$ de d' vecteurs propres de A, associés à des valeurs propres deux à deux distinctes. Si la famille $(v_k)_k$ était liée, on pourrait par exemple écrire

$$v_1 = \sum_{k=2}^{d'} \alpha_k v_k.$$

Or, par application de A (resp. par multiplication par λ_1), on trouve

$$\lambda_1 v_1 = \sum_{k=2}^{d'} \lambda_k \alpha_k v_k \text{ (resp. } \lambda_1 v_1 = \sum_{k=2}^{d'} \lambda_1 \alpha_k v_k).$$

Par différence, on en déduit que

$$\sum_{k=2}^{d'} (\lambda_k - \lambda_1) \alpha_k v_k = 0.$$

Si on applique l'hypothèse de récurrence, on trouve $(\lambda_k - \lambda_1)\alpha_k = 0$, pour $k \in \{2, \cdots, d'\}$: Comme $\lambda_k \neq \lambda_1$, on a en fait $\alpha_k = 0$, pour $k \in \{2, \cdots, d'\}$, ce qui entraîne $v_1 = 0$ et aboutit à une contradiction. La famille $(v_k)_k$ est donc libre. ∎

Définition A.2.2 *Une matrice A de $\mathbb{C}^{n\times n}$ est dite* **diagonalisable** *lorsqu'elle est semblable à une matrice diagonale.*

On commence par le résultat ci-dessous.

Proposition A.2.3 *Une matrice $A \in \mathbb{C}^{n\times n}$ est diagonalisable si, et seulement si, il existe une base de $\mathbb{C}^n$ formée de vecteurs propres de A.*

Preuve : Si A est diagonalisable, on peut écrire $A = U\Lambda U^{-1}$, avec U inversible et Λ diagonale,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

On note $(u_i)_{1\leq i\leq n}$ les vecteurs colonnes de U. Comme U est inversible, ils sont linéairement indépendants : ils forment donc une base de $\mathbb{C}^n$. Qui plus est, si on écrit $AU = U\Lambda$ colonne par colonne, on trouve $Au_i = \lambda_i u_i$, pour i variant de 1 à n.

Réciproquement s'il existe n vecteurs propres $u_1, u_2, \cdots, u_n$ linéairement indépendants, alors la matrice

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \in \mathbb{C}^{n\times n},$$

est inversible. Des relations $Au_i = \lambda_i u_i$, pour $1 \leq i \leq n$, on tire successivement $AU = U\Lambda$, puis $A = U\Lambda U^{-1}$. ∎

Proposition A.2.4 *Une matrice A de $\mathbb{C}^{n\times n}$ est diagonalisable si, et seulement si,*

$$\sum_{i=1}^{d} g_i = n.$$

Preuve : Par définition, si A est diagonalisable, alors elle est semblable à la matrice par blocs

$$A' = \begin{bmatrix} \lambda_1 I_{g_1} & 0 & \dots & 0 \\ 0 & \lambda_2 I_{g_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_d I_{g_d} \end{bmatrix},$$

écrite dans la base de vecteurs propres $(e_1^1, \cdots, e_{g_1}^1, e_1^2, \cdots, e_{g_2}^2, \cdots, e_1^d, \cdots, e_{g_d}^d)$. On constate alors que les racines du polynôme caractéristique p sont $\lambda_1, \cdots, \lambda_d$ et que, de plus, λ_i est exactement racine d'ordre g_i. Dans ce cas, $\sum_i g_i = n$.

Réciproquement, soit $(e_i^1, \cdots, e_{g_i}^1)$ une base de $\mathrm{Ker}\,(A - \lambda_i I)$, pour chaque i. Par hypothèse, $(e_k)_k = (e_1^1, \cdots, e_{g_1}^1, e_1^2, \cdots, e_{g_2}^2, \cdots, e_1^d, \cdots, e_{g_d}^d)$ est une famille à n éléments de $\mathbb{C}^n$. Pour prouver que c'est une base, il suffit de vérifier qu'elle est libre. Soit donc $(\alpha_k)_k$ tels que

$$\sum_{k=1}^{n} \alpha_k e_k = 0.$$

On regroupe alors les éléments de chaque $\mathrm{Ker}\,(A-\lambda_i I)$, pour trouver $\sum_{i=1}^{d} v_i = 0$, avec

$$v_i = \sum_{m=1}^{g_i} \alpha_m^i e_m^i \in \mathrm{Ker}\,(A-\lambda_i I).$$

D'après la Proposition A.2.2, chaque $v_i = 0$. Finalement, comme $(e_i^1, \cdots, e_{g_i}^1)$ est une base de $\mathrm{Ker}\,(A-\lambda_i I)$, on a de plus $\alpha_m^i = 0$, pour $m \in \{1, \cdots, g_i\}$. En conclusion, tous les $(\alpha_k)_k$ sont nuls, et $(e_k)_k$ est une base de $\mathbb{C}^n$ formée de vecteurs propres de A. ■

Corollaire A.2.1 *Si toutes les valeurs propres d'une matrice sont distinctes, elle est diagonalisable.*

Preuve : Dans ce cas, on a $g_i = 1$, pour i variant de 1 à n. Leur somme vaut donc n. ■

Définition A.2.3 *Une matrice A de $\mathbb{C}^{n\times n}$ est dite* **défective** *lorsqu'elle n'est pas diagonalisable.*

On introduit ensuite les notions suivantes :

Définition A.2.4 *L'ensemble des valeurs propres d'une matrice A s'appelle le* **spectre** *de A.*
- *λ_i est dite valeur propre* **simple** *si, et seulement si, $m_i = 1$; sinon λ_i est valeur propre* **multiple**.
- *λ_i valeur propre* **multiple**, *est dite* **semi-simple** *si, et seulement si, $m_i = g_i > 1$; sinon λ_i est valeur propre* **défective** *(on a alors $m_i > g_i$).*

Remarque A.2.1 *D'après ce que l'on a vu ci-dessous, une matrice A admet au moins une valeur propre défective si, et seulement si, elle est défective.*
Par ailleurs, seule une valeur propre multiple peut être défective, puisque pour une valeur propre simple λ_i, on a $m_i = g_i = 1$!

Enfin pour en terminer avec les définitions, rappelons encore que

Définition A.2.5 *On dit que $v \in \mathbb{C}^{1\times n} \setminus \{0\}$ est* **vecteur propre à gauche** *de la matrice A, si et seulement si il existe $\mu \in \mathbb{C}$ tel que $v^*A = \mu v^*$. Par cohérence les vecteurs propres usuels sont appelés* **vecteurs propres à droite**.

Un vecteur propre à gauche est un *vecteur ligne* de $\mathbb{C}^{1\times n}$. Un vecteur propre à droite est un *vecteur colonne* de $\mathbb{C}^{n\times 1}$, que l'on identifie à $\mathbb{C}^n$. On ne distingue pas valeur propre à droite de valeur propre à gauche. De fait, ces notions coïncident.

Proposition A.2.5 *A chaque valeur propre correspond un vecteur propre à droite, et un vecteur propre à gauche.*

Preuve : 1) Pour commencer, on a la série d'équivalences sur les valeurs propres (à droite)

$$\lambda_i \text{ v. p. de } A \Longleftrightarrow \text{dét}\,(A-\lambda_i I_n) = 0 \Longleftrightarrow \text{dét}\,(A-\lambda_i I_n)^* = 0 \Longleftrightarrow \text{dét}\,(A^*-\overline{\lambda_i} I_n) = 0 \Longleftrightarrow \overline{\lambda_i} \text{ v. p. de } A^*.$$

2) Ensuite, on vérifie par transposition et passage au complexe conjugué que, pour $u \in \mathbb{C}^n$, $u \neq 0$, $Au = \mu u$ équivaut à $vA^* = \overline{\mu} v$, avec $v = u^* = (\overline{u}_1, \cdots, \overline{u}_n) \in \mathbb{C}^{1\times n} \setminus \{0\}$: λ_i est valeur propre à droite de A si, et seulement si, $\overline{\lambda_i}$ est valeur propre à gauche de A^*.
3) On conclut que

$$\lambda_i \text{ v. p. à gauche de } A \overset{2)}{\Longleftrightarrow} \overline{\lambda_i} \text{ v. p. à droite de } A^* \overset{1)}{\Longleftrightarrow} \lambda_i = \overline{\overline{\lambda_i}} \text{ v. p. à droite de } A.$$

■

Proposition A.2.6 *Soit u_i un vecteur propre à droite de la matrice $A \in \mathbb{C}^{n\times n}$: $Au_i = \lambda_i u_i$, et soit v_j un vecteur propre à gauche de la matrice A : $v_j^* A = \lambda_j v_j^*$. Si $\lambda_i \neq \lambda_j$, alors $v_j^* u_i = 0$.*

Preuve : On note que, comme $v_j^* \in \mathbb{C}^{1\times n}$ et $u_i \in \mathbb{C}^{n\times 1}$, leur produit $v_j^* u_i$ appartient à $\mathbb{C}$. Soient λ_i et u_i tels que $Au_i = \lambda_i u_i$, λ_j et v_j tels que $v_j^* A = \lambda_j v_j^*$, alors

$$\left.\begin{array}{r} Au_i = \lambda_i u_i \Longrightarrow v_j^* A u_i = \lambda_i v_j^* u_i \\ A^* v_j = \overline{\lambda}_j v_j \Longrightarrow u_i^* A^* v_j = \overline{\lambda}_j u_i^* v_j \end{array}\right\} \Longrightarrow \lambda_i v_j^* u_i = v_j^* A u_i = (u_i^* A^* v_j)^* = \lambda_j v_j^* u_i,$$

soit finalement $(\lambda_i - \lambda_j) v_j^* u_i = 0$. ∎

Pour conclure ce paragraphe, considérons les deux matrices suivantes :

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{et} \quad A_2 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix};$$

elles ne différent que par le dernier coefficient diagonal, mais ont des propriétés spectrales distinctes

$$\mathrm{Spe}(A_1) = \{1, 2, 3\} \quad \text{et} \quad \mathrm{Spe}(A_2) = \{1, 2\}.$$

La matrice A_1 ayant ses valeurs propres réelles distinctes, ses vecteurs propres forment une base de $\mathbb{R}^3$. A_1 est donc *diagonalisable* :

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}^{-1}.$$

La matrice A_2 est *défective* : la valeur propre $\lambda(A_2) = 2$ a une multiplicité algébrique $m = 2$ car le polynôme caractéristique est divisible par $(\lambda-2)^2$. Sa multiplicité géométrique est $g = 1$: en effet tout vecteur propre $v = (v_1, v_2, v_3)^T$ associé à la valeur propre $\lambda = 2$ vérifie nécessairement les relations

$$\begin{aligned} v_1 + 2v_2 - 4v_3 &= 2v_1 \\ 2v_2 + 2v_3 &= 2v_2 \\ 2v_3 &= 2v_3 \end{aligned}$$

on en déduit que $v_3 = 0$ et $v_1 = 2v_2$; le sous–espace propre relatif à la valeur propre $\lambda = 2$ est donc engendré par le vecteur $v = (2, 1, 0)^T$. La matrice A_2 est défective, et on peut seulement écrire

$$A_2 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1/2 \end{pmatrix}^{-1}.$$

Cette différence peut s'avérer très importante dans la pratique, en particulier lorsque l'on doit évaluer A_1^k et A_2^k.

A.3 Localisation des valeurs propres

Théorème A.3.1 *[Gerschgorin–Hadamard] Le spectre de la matrice $A \in \mathbb{C}^{n\times n}$ est contenu dans l'ensemble D réunion des disques D_i du plan complexe définis par*

$$D_i = \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

Preuve : Soit λ une valeur propre de A et u un vecteur propre associé, et soit $|u_i| = \max_j |u_j|$, alors $|u_i| \neq 0$ et

$$\sum_j A_{i,j}\ u_j = \lambda u_i \Longleftrightarrow \lambda - A_{i,i} = \sum_{j\neq i} A_{i,j}\frac{u_j}{u_i} \Longleftrightarrow |\lambda - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|.$$

Ainsi, λ appartient au disque D_i de rayon $\sum_{j\neq i}|A_{i,j}|$ centré en $A_{i,i}$. A toute valeur propre λ, on peut ainsi associer un disque D_i, et le spectre de la matrice A est donc contenu dans l'ensemble

$$\mathrm{D} = \cup_{i=1}^n \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

■

Ce résultat permet de localiser rapidement les valeurs propres d'une matrice dans le plan complexe ; on vérifie par exemple sur la figure A.1 que les valeurs propres de la matrice S (représentées par des croix) sont situées à l'intérieur de deux disques

$$S = \begin{pmatrix} 0.5000 & 0.1667 & 0.0417 & 0.0083 \\ 0.1667 & 0.0417 & 0.0083 & 0.0014 \\ 0.0417 & 0.0083 & 0.0014 & 0.0002 \\ 0.0083 & 0.0014 & 0.0002 & 0.0000 \end{pmatrix}$$

$$\begin{aligned} \lambda_1(S) &= \ \ 0.5575 \text{ dans le disque de centre } (0.5000, 0.) \quad \text{de rayon } 0.2167 \\ \lambda_2(S) &= -0.0146 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \\ \lambda_3(S) &= \ \ 0.0001 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \\ \lambda_4(S) &= \ \ 0.0000 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \end{aligned}$$

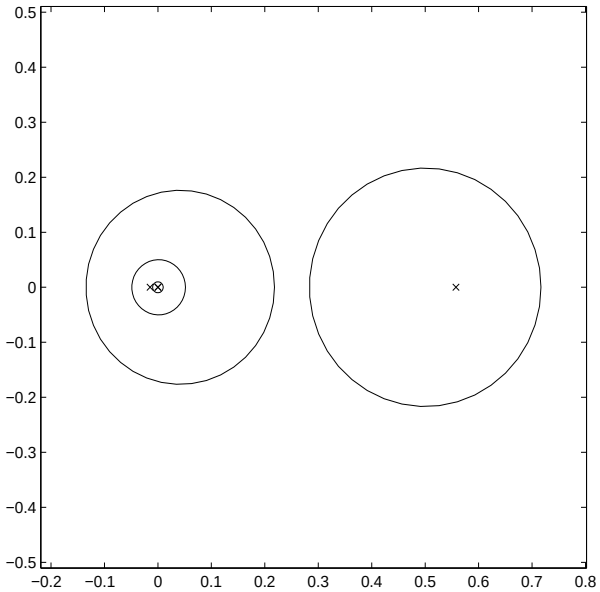


FIG. A.1 – Le spectre de la matrice S

Dans le cas d'une matrice non symétrique N, on peut appliquer le Théorème A.3.1 aux matrices N et N^T qui ont les mêmes valeurs propres, mais des disques de Gerschgorin différents, ce qui permet une localisation plus précise du spectre de N :

$$N = \begin{pmatrix} 0.5000 & -0.1000 & 0.0417 & 0.0083 \\ 0.1667 & 0.0417 & 0.0083 & 0.0014 \\ 0.0417 & 0.0000 & 0.0014 & 0.0002 \\ 0.0083 & 0.0014 & 0.0000 & 0.0100 \end{pmatrix}.$$

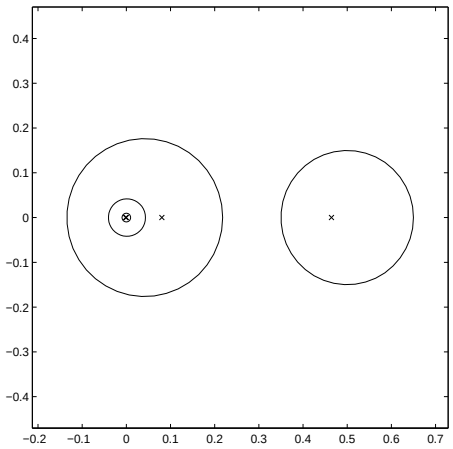


FIG. A.2 – Le spectre de N

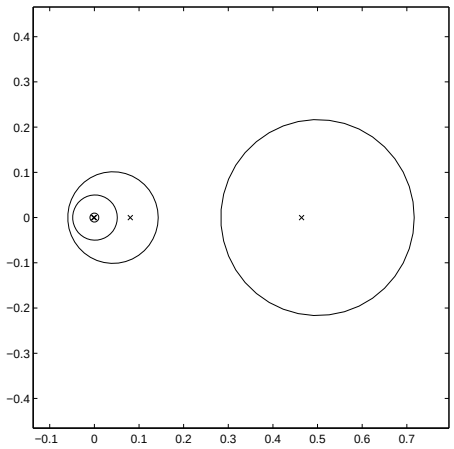


FIG. A.3 – Le spectre de N^T

Il existe de nombreuses configurations. L'exemple de la matrice circulante C illustre un cas particulier qui fait l'objet de l'exercice A.3.1

$$C = \begin{pmatrix} 1. & 2. & 3. & 4. \\ 4. & 1. & 2. & 3. \\ 3. & 4. & 1. & 2. \\ 2. & 3. & 4. & 1. \end{pmatrix}.$$

Pour cette matrice l'ensemble D est contenu tout entier dans un seul disque (voir figure A.4), avec une valeur propre sur la frontière

$$\begin{array}{lll} \lambda_1(C) = & 10 & \text{sur le cercle de centre } (1.,0.) \quad \text{de rayon } 9 \\ \lambda_2(C) = & -2+2\imath & \text{dans le disque de centre } (1.,0.) \quad \text{de rayon } 9 \\ \lambda_3(C) = & -2-2\imath & \text{dans le disque de centre } (1.,0.) \quad \text{de rayon } 9 \\ \lambda_4(C) = & -2 & \text{dans le disque de centre } (1.,0.) \quad \text{de rayon } 9 \end{array}.$$

Exercice A.3.1 *Soit λ une valeur propre de la matrice irréductible $A \in \mathbb{C}^{n\times n}$. Montrer que si λ appartient à la frontière de l'ensemble D, alors tous les cercles de Gerschgorin passent par λ.*

En conséquence tout point de la frontière de D, intersection de cercles de Gerschgorin, et tel il existe au moins un cercle qui ne le contienne pas, ne peut correspondre à une valeur propre de la matrice.

Enfin il existe une variante du Théorème de Gerschgorin-Hadamard, qui permet de mieux localiser les valeurs propres :

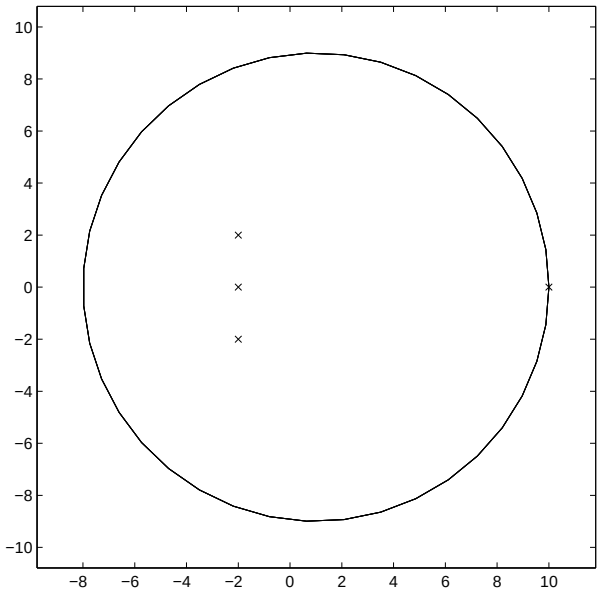


FIG. A.4 – Le spectre de C

Théorème A.3.2 *Soit une matrice $A \in \mathbb{C}^{n\times n}$ et soient les n disques D_i du plan complexe définis par*

$$D_i = \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

S'il existe p disques D_i formant un ensemble connexe E, sans intersection avec les $n-p$ disques restant, alors E contient exactement p valeurs propres de la matrice A.

Preuve : On écrit A sous la forme $A = Dia + R$ où Dia est la partie diagonale de A, et on définit les n rayons

$$r_i = \sum_{j\neq i} |A_{i,j}| = \sum_{j} |R_{i,j}| \qquad 1 \leq i \leq n,$$

ainsi que les n disques

$$D_i = \{z \in \mathbb{C},\ |z - Dia_{i,i}| \leq r_i\}.$$

Puis pour tout $\varepsilon \geq 0$ on définit de manière cohérente la matrice $A(\varepsilon) = Dia + \varepsilon R$, et les n disques associés

$$D_i(\varepsilon) = \{z \in \mathbb{C},\ |z - Dia_{i,i}| \leq \varepsilon\ r_i\}.$$

Par définition $A(0) = Dia$, $A(1) = A$ et pour tout i et tout ε inférieur à 1, $D_i(\varepsilon) \subset D_i(1) = D_i$. Par application du Théorème A.3.1, on sait que le spectre de $A(\varepsilon)$ est contenu dans l'ensemble $\cup_{i=1}^{n} D_i(\varepsilon)$ pour tout ε. Sans nuire à la généralité on suppose que l'ensemble connexe E est constitué par l'union des p premiers disques : $E = \cup_{i=1}^{p} D_i$. On définit alors l'ensemble $E(\varepsilon) = \cup_{i=1}^{p} D_i(\varepsilon)$. L'hypothèse

$$\forall j > p \quad D_j \cap E = \emptyset$$

entraine

$$\forall j > p, \forall \varepsilon \leq 1 \quad D_j(\varepsilon) \cap E(\varepsilon) = \emptyset.$$

Pour $\varepsilon = 0$, chaque disque $D_i(0)$ est réduit à un point et

$$E(0) = \cup_{i=1}^{p} D_i(0) = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$$

quand ε tend vers 1, $E(\varepsilon) \subset E$ contient toujours exactement p valeurs propres : $\lambda_1, \lambda_2, \ldots, \lambda_p$, les autres valeurs propres restant dans leurs disques. Cette configuration reste vraie à la limite, puisque les valeurs propres de $A(\varepsilon)$ dépendent continûment de ε. ■

On peut voir sur la figure A.5 le cas d'une matrice A (proche de S)

$$A = \begin{pmatrix} 0.5000 & 0.1267 & 0.0417 & 0.0083 \\ 0.1267 & 1.0417 & 0.0083 & 0.0014 \\ 0.0417 & 0.0083 & 0.2514 & 0.0002 \\ 0.0083 & 0.0014 & 0.0002 & 0.0100 \end{pmatrix},$$

pour laquelle tous les disques de Gerschgorin sont disjoints :

$\lambda_1 = 1.0702$ dans le disque de centre $(1.0417, 0.)$ de rayon 0.1364

$\lambda_2 = 0.4786$ dans le disque de centre $(0.5000, 0.)$ de rayon 0.1767

$\lambda_3 = 0.2444$ dans le disque de centre $(0.2514, 0.)$ de rayon 0.0502

$\lambda_4 = 0.0099$ dans le disque de centre $(0.0100, 0.)$ de rayon 0.0099

.

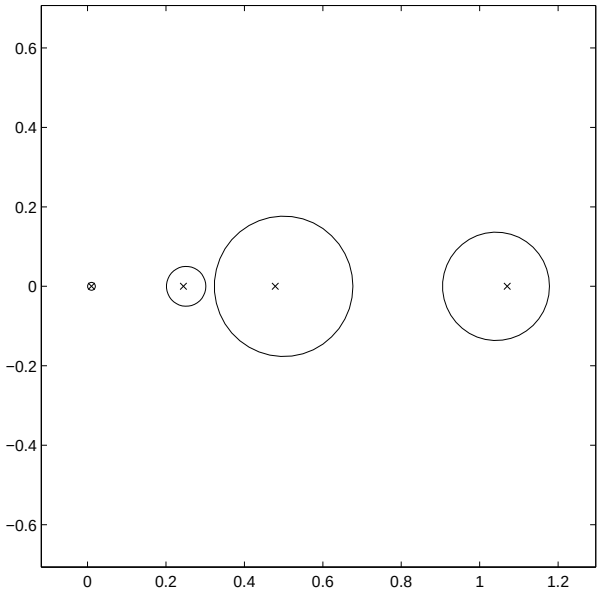


FIG. A.5 – Le spectre de A

A.4 Matrices diagonalisables

On se place dans un espace vectoriel sur $\mathbb{C}$. Les mêmes résultats restent valables dans un espace vectoriel sur $\mathbb{R}$, sous réserve que l'on remplace partout

- complexe par réelle ($A \in \mathbb{R}^{n \times n}$),
- hermitienne par symétrique ($A^T = A$),
- normale par normale ($A^T A = A A^T$),
- unitaire par orthogonale ($O O^T = O^T O = I_n$).

Les résultats s'appliquent donc en particulier aux matrices issues de la discrétisation par différences finies du Laplacien (voir le chapitre 2).

Définition A.4.1 *Une matrice $A \in \mathbb{C}^{n\times n}$ est dite* **hermitienne** *lorsque $A^* = A$.*
Une matrice $A \in \mathbb{C}^{n\times n}$ est dite **normale** *lorsque $A^*A = AA^*$.*
Une matrice $U \in \mathbb{C}^{n\times n}$ est dite **unitaire** *lorsque $UU^* = U^*U = I_n$.*

Proposition A.4.1 *Les valeurs propres d'une matrice hermitienne sont réelles.*

Preuve : En effet, pour un couple vecteur-valeur propres (u, λ), on a la suite d'égalités,

$$\lambda(u,u) = (u, Au) \overset{A^*=A}{=} (Au, u) = (\lambda u, u) = \overline{\lambda}(u,u).$$

On en déduit que $\lambda \in \mathbb{R}$. ■

Proposition A.4.2 *Les vecteurs propres d'une matrice hermitienne correspondant à des valeurs propres distinctes sont orthogonaux.*

Preuve : En effet, pour deux couples vecteur-valeur propres (u, λ) et (v, μ), on peut écrire,

$$\left.\begin{array}{l}(Au,v) = (\lambda u, v) = \overline{\lambda}(u,v) \overset{\overline{\lambda}=\lambda}{=} \lambda(u,v) \\ (Au,v) \overset{A^*=A}{=} (u, Av) = (u, \mu v) = \mu(u,v)\end{array}\right\} \Longrightarrow (\lambda-\mu)(u,v) = 0 \Longrightarrow (u,v) = 0 \text{ si } \lambda \neq \mu.$$

■

Proposition A.4.3 *Toute matrice hermitienne est diagonalisable. Qui plus est, on peut choisir ses vecteurs propres de sorte qu'ils forment une base orthonormale.*
En d'autres termes, il existe Q unitaire et D diagonale telles que $A = QDQ^*$.

Cette propriété découle d'un résultat plus général

Proposition A.4.4 *[Forme de Schur] Soit $A \in \mathbb{C}^{n\times n}$ il existe une matrice unitaire Q telle que $T = Q^*AQ$ soit une matrice triangulaire supérieure avec pour éléments diagonaux les valeurs propres de la matrice A.*

Preuve : La démontration est effectuée par récurrence : la propriété est évidente à l'ordre $n = 1$. Supposons la vraie jusqu'à l'ordre $n-1$ inclus. Soit $A \in \mathbb{C}^{n\times n}$ et λ une valeur propre de A, u un vecteur propre associé de norme 1 ; d'après le théorème de la base incomplète, il existe une matrice $U \in \mathbb{C}^{n-1\times n-1}$ unitaire ($U^*U = UU^* = I$) telle que la matrice $[u, U] \in \mathbb{C}^{n\times n}$ soit aussi unitaire, car on peut toujours construire une base orthogonale de $\mathbb{C}^n$ dont $u \neq 0$ soit le premier vecteur de base.

Ainsi par construction $U^*u = 0$ et $A[u, U] = [\lambda u, AU]$, soit encore

$$[u,U]^*\ A\ [u,U] = \begin{bmatrix} u^* \\ U^* \end{bmatrix}\ [\lambda u, A\ U] = \begin{bmatrix} \lambda & u^*AU \\ 0 & U^*AU \end{bmatrix}.$$

Comme U^*AU est de rang $n-1$, on peut lui appliquer l'hypothèse de récurrence : il existe $\tilde{Q} \in \mathbb{C}^{n-1\times n-1}$ unitaire telle que $\tilde{Q}^*U^*AU\tilde{Q} = \tilde{T}$; alors

$$[u,U\tilde{Q}]^*\ A\ [u,U\tilde{Q}] = \begin{bmatrix} u^* \\ \tilde{Q}^*U^* \end{bmatrix}\ [\lambda u, A\ U\tilde{Q}] = \begin{bmatrix} \lambda & u^*AU\tilde{Q} \\ 0 & \tilde{T} \end{bmatrix} = T.$$

Enfin puisque $Q = [u, U\tilde{Q}]$ est unitaire, les matrices A et T sont semblables, et possèdent de ce fait les mêmes valeurs propres. Plus précisément, si μ est une valeur propre de U^*AU associée au vecteur propre $v \in \mathbb{C}^{n-1}$, μ est aussi une valeur propre de A puisque

$$U^*AU\ v = \mu v \quad \Longrightarrow A(Uv) = \mu(Uv).$$

On obtient donc la propriété à l'ordre n avec $Q = [u, U]$, et dans cette écriture les termes diagonaux sont bien les valeurs propres de A.

Les vecteurs colonnes de la matrice Q sont appelés **vecteurs de Schur** ; ils vérifient la relation $AQ = QT$. ∎

Dans le cas où la matrice A est hermitienne, la matrice triangulaire supérieure $T = Q^*AQ$ est aussi hermitienne, car elle vérifie $T^* = Q^*A^*Q = Q^*AQ$. Elle est donc diagonale, à coefficients réels, ce qui démontre la Proposition A.4.3.

De plus dans ce cas particulier, la relation $AQ = QT$ avec T diagonale, montre que les vecteurs de Schur sont les vecteurs propres de la matrice hermitienne A. La matrice Q dont les colonnes sont les vecteurs propres de A, est unitaire par construction. On en déduit que les vecteurs propres de A forment une base orthogonale de $\mathbb{C}^n$. Ce résultat étant vrai, que les valeurs propres soient distinctes ou non, constitue donc une extension de la Proposition A.4.2.

Proposition A.4.5 *Une matrice $A \in \mathbb{C}^{n\times n}$ est normale si, et seulement si, elle est diagonalisable dans une base orthonormale.*

Preuve : Supposons d'abord que A soit diagonalisable dans une base orthonormale : il existe Q unitaire et D diagonale telles que $A = QDQ^*$. On a alors la suite d'égalités

$$AA^* = QDQ^*QD^*Q^* \overset{Q^*Q=I_n}{=} QDD^*Q^* \overset{D^*D=DD^*}{=} QD^*DQ^* \overset{Q^*Q=I_n}{=} QD^*Q^*QDQ^* = A^*A.$$

Réciproquement, soit A une matrice normale. On écrit $A = QTQ^*$, avec Q matrice unitaire et T matrice triangulaire supérieure. De l'égalité $A^*A = AA^*$ on tire $T^*T = TT^*$. Mais la matrice T étant triangulaire supérieure, on peut écrire pour tout i

$$(TT^*)_{i,i} = \sum_{j\geq i} |T_{i,j}|^2 = (T^*T)_{i,i} = \sum_{j\leq i} |T_{j,i}|^2.$$

Pour $i = 1$ on trouve donc que

$$(TT^*)_{1,1} = \sum_{j\geq 1} |T_{1,j}|^2 = (T^*T)_{1,1} = |T_{1,1}|^2,$$

ce qui entraine que tous les coefficients extra-diagonaux $T_{1,j}$ sont nuls. En appliquant le même raisonnement à la ligne $i = 2$, on trouve

$$(TT^*)_{2,2} = \sum_{j\geq 2} |T_{2,j}|^2 = (T^*T)_{2,2} = |T_{1,2}|^2 + |T_{2,2}|^2 = |T_{2,2}|^2,$$

puisque $T_{1,2} = 0$. Tous les coefficients extra-diagonaux $T_{2,j}$ sont nuls... En réitérant ce procédé, on montre que la matrice T est diagonale. ∎

Remarque A.4.1 *Des relations $AQ = QD$ et $A^*Q = QD^*$ on déduit que si A est normale, alors les matrices A et A^* admettent la même base de vecteurs propres, qui sont les vecteurs colonnes de la matrice Q.*

Il s'agit bien d'une *généralisation* des résultats concernant les matrices hermitiennes, car la matrice A définie par ($\imath^2 = -1$)

$$A = \begin{bmatrix} \imath & 0 & \dots & 0 \\ 0 & \imath & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \imath \end{bmatrix}$$

est diagonalisable (car normale) sans être hermitienne.

Un exemple moins trivial est fourni par la matrice de permutation $P \in \mathbb{R}^{n\times n}$ de rang n

$$P = \begin{bmatrix} 0 & \dots & \dots & 0 & 1 \\ 1 & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}$$

qui est normale, donc diagonalisable sans même être symétrique :

$$Q^* P Q = \Lambda.$$

Les matrices Q et Λ sont définies en posant $z = e^{\imath\pi/n}$ et $\overline{z} = e^{-\imath\pi/n}$:

$$Q = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \overline{z} & \overline{z}^2 & \vdots & \overline{z}^{(n-1)} \\ 1 & \overline{z}^2 & \overline{z}^4 & \vdots & \overline{z}^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \overline{z}^{n-1} & \overline{z}^{2(n-1)} & \dots & \overline{z}^{(n-1)(n-1)} \end{bmatrix} \quad \text{et} \quad \Lambda = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & z & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & z^{(n-2)} & 0 \\ 0 & \dots & \dots & 0 & z^{(n-1)} \end{bmatrix}.$$

Enfin, toute matrice A diagonalisable n'est pas nécessairement normale ! On a vu qu'il faut que la base de vecteurs propres de A soit *orthonormale* pour avoir cette propriété. L'exercice qui suit prouve qu'il existe des matrices diagonalisables qui ne sont pas normales.

Exercice A.4.1 *Soit la matrice*

$$A = \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix}.$$

Montrer que A est diagonalisable, puis calculer AA^* et A^*A. Conclure.

Complétons ce paragraphe par l'énoncé de quelques résultats utiles sur les matrices hermitiennes.

Théorème A.4.1 (Courant–Fisher) *Soit $A \in \mathbb{C}^{n\times n}$ une matrice hermitienne dont les valeurs propres (réelles) sont rangées suivant*

$$\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$$

alors

$$(i) \quad \lambda_k = \max_{\substack{V \\ dim\, V = k}} \min_{x \in V - \{0\}} \frac{(x, Ax)}{(x, x)},$$

$$(ii) \quad \lambda_k = \min_{\substack{W \\ dim\, W = n-k+1}} \max_{x \in W - \{0\}} \frac{(x, Ax)}{(x, x)}.$$

Preuve : Le rapport $\rho_A(x) = \dfrac{(x, Ax)}{(x, x)}$ est le **quotient de Rayleigh** de la matrice A. L'ensemble

$$F(A) = \{\rho_A(x), x \in \mathbb{C}^n, x \neq 0\}$$

est appelé le **champ des valeurs** de la matrice A. Le champ des valeurs de A contient le spectre de A et aussi les valeurs du quotient de Rayleigh $\rho_A(v) = \lambda$ pour tout vecteur propre v : $Av = \lambda v$.

Pour démontrer le premier point du théorème, on considère le sous–espace engendré par les vecteurs propres u_i associés aux $n-k+1$ valeurs propres λ_i $(k \leq i \leq n)$. Soit $V \subset \mathbb{C}^n$ un sous–espace quelconque de dimension k : $W_k =< u_n, u_{n-1}, \ldots, u_k >$; puisque $\dim W_k = n-k+1$, $W_k \cap V \neq \{0\}$, il existe donc au moins un vecteur commun non nul $x \in W_k \cap V$, que l'on écrit $x = \sum_{i=k}^{n} \alpha_i u_i$. Alors

$$\rho_A(x) = \frac{(x, Ax)}{(x,x)} = \frac{\sum_{i=k}^{n} \lambda_i \alpha_i^2 (u_i, u_i)}{\sum_{i=k}^{n} \alpha_i^2 (u_i, u_i)} \leq \lambda_k.$$

Par conséquent $m(V) = \min_{x \in V - \{0\}} \rho_A(x) \leq \lambda_k$, et on en déduit que le maximum de $m(V)$ sur tous les sous–espaces V de dimension k est plus petit que λ_k. Si on prend en particulier $V =< u_1, u_2, \cdots, u_k >$, alors $\dim V = k$ et $m(V)$ atteint la valeur maximale λ_k pour $x = u_k \in V$.

On procède de même pour le second point du théorème, en introduisant cette fois $V_k =< u_1, u_2, \ldots, u_k >$ le sous–espace de dimension k. Alors pour tout sous–espace W de dimension $n-k+1$, $W \cap V_k \neq \{0\}$ et par le même raisonnement, on en déduit que $M(V) = \max_{x \in W - \{0\}} \frac{(x, Ax)}{(x,x)} \geq \lambda_k$, puis que

$$\min_{\substack{W \\ \dim W = n-k+1}} \max_{x \in W - \{0\}} \frac{(x, Ax)}{(x,x)} \geq \lambda_k.$$

La valeur minimale λ_k est atteinte en prenant pour sous–espace $W =< u_n, u_{n-1}, \ldots, u_k >$ et pour vecteur $x = u_k$. ∎

Théorème A.4.2 *Soit $B = A + E$ la somme de deux matrices hermitiennes de $\mathbb{C}^{n \times n}$, on range les valeurs propres par ordre croissant :*

$$A \ : \quad \lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$$

$$E \ : \quad \varepsilon_n \leq \varepsilon_{n-1} \leq \ldots \leq \varepsilon_2 \leq \varepsilon_1$$

$$B \ : \quad \mu_n \leq \mu_{n-1} \leq \ldots \leq \mu_2 \leq \mu_1.$$

Alors pour tout $k = 1, 2, \ldots, n$

$$(i) \qquad \lambda_k + \varepsilon_n \leq \mu_k \leq \lambda_k + \varepsilon_1$$

$$(ii) \qquad |\mu_k - \lambda_k| \leq \|E\|.$$

Preuve : Soit $u_1, u_2, \ldots, u_n$ une base orthonormée des vecteurs propres de la matrice A, et μ_k une valeur propre de B : on pose $W_k =< u_n, u_{n-1}, \ldots, u_k >$, d'après le théorème de Courant–Fisher

$$\mu_k \leq \max_{x \in W_k - \{0\}} \rho_B(x) \leq \max_{x \in W_k - \{0\}} \rho_A(x) + \max_{x \in W_k - \{0\}} \rho_E(x)$$

soit encore $\mu_k \leq \lambda_k + \varepsilon_1$.

Pour la minoration, on écrit $A = B - E = B + E'$ et le résultat précédent appliqué à la matrice $B + E'$ devient $\lambda_k \leq \mu_k - \varepsilon_n$.

Finalement, pour tout $k = 1, 2, \ldots, n$, $\lambda_k + \varepsilon_n \leq \mu_k \leq \lambda_k + \varepsilon_1$, soit encore $\varepsilon_n \leq \mu_k - \lambda_k \leq \varepsilon_1$. On en déduit (ii) puisque pour toute matrice E et toute norme matricielle $\|E\|$, $|\varepsilon_k| \leq \|E\| \quad \forall k = 1, 2, \ldots, n$. ∎

Ce résultat semble prometteur sur le plan numérique, car il montre que la recherche des valeurs propres d'une matrice A hermitienne est théoriquement stable. Les valeurs propres $\lambda(A)$ dépendent continûment des coefficients de A de la manière suivante : si on pose $A_{\mathcal{E}} = A + \mathcal{E}$ avec $\mathcal{E} \in \mathbb{C}^{n\times n}$ matrice de perturbation **hermitienne**, alors

$$\max_{\lambda} |\lambda(A_{\mathcal{E}}) - \lambda(A)| = \max_{\lambda} |\lambda(\mathcal{E})| \|\mathcal{E}\|_2 \leq \|\mathcal{E}\|_F.$$

Cette majoration donne une borne maximale de variation des valeurs propres de A en fonction des coefficients de $\mathcal{E}$. Malheureusement dans la pratique, les erreurs commises sur les coefficients de la matrice A sont dûes soit à la représentation machine des nombres (erreur de troncature ou d'arrondi), soit aux erreurs de calcul qui en découlent. En conséquence, bien qu'il soit souvent possible d'estimer $\|\mathcal{E}\|_F$, le Théorème A.4.2 n'est pas utilisable pour le calcul numérique car la matrice $\mathcal{E}$ n'est jamais hermitienne !

Pour finir ce paragraphe, nous introduisons la décomposition spectrale d'une matrice diagonalisable. Pour cela, on se souvient que, par définition, lorsqu'une matrice A est diagonalisable, il existe U inversible et Λ diagonale, telles que $A = U\Lambda U^{-1}$. On a vu à la Proposition A.2.3 que les vecteurs colonnes de U sont des vecteurs propres à droite de A. De la même façon, les vecteurs ligne de U^{-1}, appartenant à $\mathbb{C}^{1\times n}$ et notés $(v_i^*)_{1\leq i\leq n}$,

$$U^{-1} = \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix},$$

sont des vecteurs propres à gauche de A. En effet, la relation $U^{-1}A = \Lambda U^{-1}$ peut s'écrire, ligne par ligne, sous la forme $v_i^* A = \lambda_i v_i^*$, pour i variant de 1 à n. En particulier, les ordres de multiplicité de λ_i à gauche et à droite sont *identiques*.

Au final, on peut résumer la relation $A = U\Lambda U^{-1}$ dans les formules

$$A = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \dots \\ v_n^* \end{bmatrix}, \text{ ou bien } A = \sum_{i=1}^{n} \lambda_i \, u_i \cdot v_i^*,$$

en se souvenant que $u_i \cdot v_i^*$ est une matrice de $\mathbb{C}^{n\times n}$. Si on regroupe les $u_i \cdot v_i^*$ correspondant à une même valeur propre λ_k, on peut écrire A sous la forme

$$A = \sum_{k=1}^{d} \lambda_k P_k, \text{ avec } P_k = \sum_{i,\, \lambda_i = \lambda_k} u_i \cdot v_i^*. \tag{A.1}$$

Cette relation est appelée **décomposition spectrale** de la matrice A. Qui plus est, les $(P_k)_{1\leq k\leq d}$ vérifient les relations suivantes.

Proposition A.4.6 *Pour k variant de 1 à n, P_k est la matrice d'un projecteur sur le sous-espace propre* $\mathrm{Ker}\,(A - \lambda_k I_n)$. *De plus, la somme de ces projecteurs est égale à l'identité. En d'autres termes :*

(i) $P_k u_i = u_i$ si $\lambda_i = \lambda_k$, $P_k u_i = 0$ sinon ;

(ii) $\sum_{1\leq k\leq d} P_k = I_n$;

(iii) $P_k^2 = P_k$, $1 \leq k \leq d$, et $P_k P_l = 0$ pour $k \neq l$.

Preuve : La démonstration est relativement aisée, sous réserve que l'on se souvienne de la définition des vecteurs $(u_i)_i$ et $(v_i^*)_i$. En effet, de la relation $U^{-1}U = I_n$, on tire immédiatement

$$v_i^* u_j = \delta_{ij}, \text{ pour } 1 \leq i, j \leq n.$$

Pour prouver (i), on écrit simplement

$$P_k u_i = \sum_{j,\, \lambda_j = \lambda_k} (u_j \cdot v_j^*) u_i = \sum_{j,\, \lambda_j = \lambda_k} u_j (v_j^* u_i) = \sum_{j,\, \lambda_j = \lambda_k} \delta_{ij} u_j = \begin{cases} u_i & \text{si } i \in \{j,\, \lambda_j = \lambda_k\} \\ 0 & \text{sinon} \end{cases} .$$

Les points (ii) et (iii) sont des conséquences simples de (i) car, $(u_i)_i$ étant une base de $\mathbb{C}^n$, les P_k sont complètement déterminées par leur action sur ceux-ci. ■

Pour conclure le cas des matrices diagonalisables, on déduit de (A.1) et de la Proposition ci-dessus que

$$AP_k = P_k A = \lambda_k P_k, \text{ pour } 1 \leq k \leq d.$$

A.5 Matrices défectives et forme de Jordan

Par définition, les matrices défectives ne sont pas diagonalisables. Il faut donc construire d'autres vecteurs associés aux vecteurs propres, appelés **vecteurs principaux**. Cette construction conduit à la **forme de Jordan** dans laquelle la matrice est écrite sous une forme presque diagonale (voir Chatelin [3]).

Théorème A.5.1 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice admettant d valeurs propres distinctes λ_i de multiplicité algébrique m_i et de multiplicité géométrique g_i ($g_i \leq m_i$). Il existe une matrice $X \in \mathbb{C}^{n\times n}$ telle que*

$$A = X \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_d \end{bmatrix} X^{-1}.$$

Dans cette écriture $J_i \in \mathbb{C}^{m_i \times m_i}$ est la boîte de Jordan *associée à la valeur propre λ_i ; chaque boîte de Jordan se décompose elle-même en une matrice diagonale par blocs $g_i \times g_i$, dont les blocs diagonaux $J_{i,j}$ sont appelés* blocs de Jordan *:*

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \dots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{i,g_i} \end{bmatrix} \quad \textit{avec } J_{i,j} = [\lambda_i] \quad \textit{ou} \quad \begin{bmatrix} \lambda_i & 1 & \dots & \dots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{bmatrix}.$$

La démonstration de ce Théorème est effectuée par étapes et requiert plusieurs résultats intermédiaires qui sont proposés comme exercices.

Remarque A.5.1 *Pour toute valeur propre* semi-simple *λ_i, comme on a $m_i = g_i$ blocs diagonaux $J_{i,j}$ sur la diagonale de J_i, une matrice de $\mathbb{C}^{m_i \times m_i}$, on a nécessairement $J_{i,j} = [\lambda_i]$ pour $1 \leq j \leq g_i$, et J_i est une matrice* diagonale *égale à $J_i = \lambda_i I_{m_i}$. On retrouve donc la définition* non défective équivaut à diagonalisable*, dans le sous-espace* $\mathrm{Ker}\,(A - \lambda_i I_n)$.

Exercice A.5.1 *Soit $R \in \mathbb{C}^{n\times n}$ une matrice triangulaire supérieure admettant d valeurs propres distinctes λ_i, montrer qu'il existe une matrice $Z \in \mathbb{C}^{n\times n}$ telle que*

$$R = Z^{-1} \begin{bmatrix} R_1 & 0 & \dots & 0 \\ 0 & R_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & R_d \end{bmatrix} Z,$$

où $R_i = \lambda_i I + U_i$ et U_i est une matrice triangulaire supérieure stricte ($U_{i,j} = 0$ si $j \leq i$).

Exercice A.5.2 *Soient $A \in \mathbb{C}^{p\times p}$, $B \in \mathbb{C}^{q\times q}$ et $C \in \mathbb{C}^{p\times q}$ trois matrices, montrer que l'équation de* **Sylvester**

$$AZ - ZB = C$$

admet une solution unique $Z \in \mathbb{C}^{p\times q}$ si et seulement si les matrices A et B n'ont pas de valeurs propres communes.

Exercice A.5.3 *Pour $k \geq 2$, soit la matrice $E_k \in \mathbb{C}^{k\times k}$ définie par*

$$E_k = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}.$$

Montrer que

$$(i) \quad E_k^k = [0]$$

$$(ii) \quad E_k^* E_k = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & I_{k-1} \end{bmatrix}$$

$$(iii) \quad I_k - E_k^* E_k^k = e_1 \cdot e_1^*$$

$$(iv) \quad E_k e_{j+1} = e_j, \quad j = 1, 2, \dots, k-1$$

où e_j est le $j^{ième}$ vecteur de base de $\mathbb{C}^k$.

Exercice A.5.4 *Soit $U \in \mathbb{C}^{n\times n}$ une matrice strictement triangulaire supérieure. Montrer qu'il existe une matrice inversible Y et g matrices $E_j \in \mathbb{C}^{k_j \times k_j}$ telles que*

$$Y^{-1}UY = \begin{bmatrix} E_1 & 0 & \dots & 0 \\ 0 & E_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & E_g \end{bmatrix} \quad \text{avec } E_j = [0] \text{ ou } \begin{bmatrix} 0 & I_{k_j-1} \\ 0 & 0 \end{bmatrix},$$

avec $k_1 \geq k_2 \geq \dots \geq k_g \geq 1$.

Démonstration du Théorème A.5.1 :
La forme de Jordan d'une matrice $A \in \mathbb{C}^{n\times n}$ est alors obtenue de la façon suivante :
– On commence par mettre A sous forme triangulaire supérieure (forme de Schur de la Proposition A.4.4) $Q^*AQ = R$.

– On applique le résultat de l'Exercice A.5.1 à la matrice R, et on obtient la matrice

$$\tilde{R} = \begin{bmatrix} \tilde{R}_1 & 0 & \dots & 0 \\ 0 & \tilde{R}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \tilde{R}_d \end{bmatrix}.$$

Les d blocs diagonaux $\tilde{R}_i$ correspondent aux d valeurs propres distinctes de R qui est semblable à A par construction.

– On applique ensuite, pour $j = 1, 2, \dots, g_i$, le résultat de l'Exercice A.5.4 à chaque bloc $U_i = \lambda_i I - \tilde{R}_i$:

$$Y_i^{-1}\left(\lambda_i I + U_i\right) Y_i = \lambda_i I + \begin{bmatrix} E_{i,1} & 0 & \dots & 0 \\ 0 & E_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & E_{i,g_i} \end{bmatrix}, \; E_{i,j} = [0] \text{ ou } \begin{bmatrix} 0 & 1 & \dots & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}$$

et les blocs $E_{i,j}$ sont rangés traditionnellement par ordre de rang croissant.

– On pose maintenant

$$\tilde{Y} = \begin{bmatrix} Y_1 & 0 & \dots & 0 \\ 0 & Y_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & Y_d \end{bmatrix},$$

et on obtient

$$\left(\tilde{Y}^{-1} Z^{-1} Q^*\right) A \left(QZ\tilde{Y}\right) = J = \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_d \end{bmatrix}.$$

Dans cette écriture, J_i est la **boîte de Jordan** associée à λ_i :

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \dots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{i,g_i} \end{bmatrix} \text{ avec } J_{i,j} = [\lambda_i] \text{ ou } \begin{bmatrix} \lambda_i & 1 & \dots & \dots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{bmatrix}.$$

Les matrices A et J sont semblables et les λ_i sont les d valeurs propres distinctes de A. De plus par construction, le rang du bloc J_i est égal à la multiplicité algébrique m_i de λ_i dans J, donc dans A. Ainsi si on note M_i le sous–espace de $\mathbb{C}^n$ associé au bloc J_i, comme $\dim M_i = m_i$ et $\sum_{i=1}^{d} m_i = n$, on a

$$\bigoplus_{i=1,d} M_i = \mathbb{C}^n.$$

A chacun des g_i sous–blocs $J_{i,j}$ de rang $m_{i,j} \geq 1$, est associé un sous–espace $M_{i,j}$ de M_i. Cherchons un vecteur propre u dans ce sous–espace. Il doit vérifier les $m_{i,j}$ relations :

$$\lambda_i u_1 + u_2 = \lambda_i u_1, \; \lambda_i u_2 + u_3 = \lambda_i u_2, \; \cdots, \; \lambda_i u_{m_{i,j}-1} + u_{m_{i,j}} = \lambda_i u_{m_{i,j}-1}, \; \lambda_i u_{m_{i,j}} = \lambda_i u_{m_{i,j}}.$$

On en déduit que nécessairement $u_2 = u_3 = \ldots = u_{m_{i,j}} = 0$! Le seul vecteur propre possible dans $M_{i,j}$ s'écrit donc $u = (1, 0, \ldots, 0)^T$. Or, il ne peut y avoir que g_i vecteurs propres linéairement indépendants dans M_i (autant que de sous–espaces $M_{i,j}$) et g_i correspond donc bien à la *multiplicité géométrique* de λ_i.

Soit maintenant $z_0 \in \mathrm{Ker}\,(A - \lambda_i I)$ un vecteur propre de A, existe-t-il un vecteur $z_1 \neq 0$ tel que $(A - \lambda_i I)z_1 = z_0$? Un tel vecteur satisfait nécessairement la relation

$$(A - \lambda_i I)^2 z_1 = (A - \lambda_i I)z_0 = 0 \quad \text{soit} \quad z_1 \in \mathrm{Ker}\,(A - \lambda_i I)^2.$$

On voit donc pour que z_1 existe, il faut et il suffit que $\mathrm{Ker}\,(A - \lambda_i I)^2 \neq \{0\}$. On peut poursuivre en définissant une suite de vecteurs z_k par

$$(A - \lambda_i I)z_k = z_{k-1},$$

et l'on doit chercher z_k dans $\mathrm{Ker}\,(A - \lambda_i I)^{k+1}$. Mais puisque

$$\mathrm{Ker}\,(A - \lambda_i I) \subset \mathrm{Ker}\,(A - \lambda_i I)^2 \subset \ldots \subset \mathrm{Ker}\,(A - \lambda_i I)^k \subset \ldots \subset \mathbb{C}^n,$$

il existe nécessairement un entier $l_i \leq n$ tel que

$$\mathrm{Ker}\,(A - \lambda_i I)^{l_i} = \mathrm{Ker}\,(A - \lambda_i I)^l \quad \forall l \geq l_i.$$

Cet entier est tel que $\mathrm{Ker}\,(A - \lambda_i I)^{l_i} = M_i$: on l'appelle **indice** de la valeur propre λ_i. Les vecteurs z_k sont appelés **vecteurs principaux** associés à z_0 dans M_i. Ils vérifient les relations

$$Az_0 = \lambda_i z_0,\ Az_1 = \lambda_i z_1 + z_0,\ Az_2 = \lambda_i z_2 + z_1,\ \cdots,\ Az_{l_i} = \lambda_i z_{l_i} + z_{l_{i-1}}.$$

Soit encore

$$A\,[\,z_0 \quad z_1 \quad \ldots \quad z_{l_i}\,] = [\,z_0 \quad z_1 \quad \ldots \quad z_{l_i}\,] \begin{bmatrix} \lambda_i & 1 & \ldots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & \ldots & \lambda_i \end{bmatrix}.$$

On reconnaît à droite le bloc de Jordan $J_{i,j}$ associé au vecteur z_0, et ceci montre que le rang de $J_{i,j}$ est inférieur ou égal à l'indice l_i.

On en déduit que la représentation de Jordan peut ne pas être unique, car la décomposition de la boîte J_i en blocs $J_{i,j}$ dépend du choix du vecteur z_0 dans chaque $M_{i,j}$. Par exemple, pour une valeur propre λ_i de multiplicité algébrique $m_i = 7$, de multiplicité géométrique $g_i = 3$ et d'indice $l_i = 3$, on obtient deux formes de Jordan différentes :

$$\left[\begin{array}{c|ccc|ccc} \lambda_i & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \lambda_i & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right] \quad \text{ou} \quad \left[\begin{array}{cc|cc|ccc} \lambda_i & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_i & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right].$$

Soit encore comment écrire 7 ($= m_i$) comme somme de 3 ($= g_i$) entiers naturels non nuls, chaque entier étant inférieur ou égal à 3 ($= l_i$) :

$$7 = 1 + 3 + 3 = 2 + 2 + 3.$$

Dans le cas général, noter que cette écriture contient le cas A diagonalisable pour lequel $l_i = 1$ et $m_i = g_i$ pour tout i.

Exercice A.5.5 *Soit $a \in \mathbb{C}$, $a \neq 0$, on considère la matrice triangulaire supérieure $C(a) \in \mathbb{C}^{n\times n}$*

$$C(a) = \begin{bmatrix} a & 1 & 0 & \dots & 0 \\ 0 & a & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & a \end{bmatrix}.$$

Vérifier que $(v_i, w_i) = 0$ pour tout v_i vecteur propre à droite de $C(a)$ et tout w_i vecteur propre à gauche de $C(a)$.

A.6 Décomposition spectrale d'une matrice quelconque

En généralisant la notion de décomposition spectrale (A.1) introduite pour des matrices diagonalisables, on remarque que l'on peut écrire toute matrice $A \in \mathbb{C}^{n\times n}$ ayant d valeurs propres distinctes suivant $A = XJX^{-1}$, avec

$$X = [\, X_1 \quad X_2 \quad \dots \quad X_d \,].$$

Chaque bloc $X_i \in \mathbb{C}^{n\times m_i}$ correspond aux m_i colonnes de X associées au sous–espace M_i, et on introduit de manière cohérente

$$X^{-1} = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \dots \\ Y_d^* \end{bmatrix}.$$

Les vecteurs colonnes de X_i forment une base de M_i, et les vecteurs lignes $Y_i^* \in \mathbb{C}^{m_i\times n}$ forment une base adjointe. De la relation $X^{-1}X = I_n$, on déduit comme précédemment

$$Y_i^* \cdot X_i = I_{m_i} \text{ et } Y_i^* \cdot X_j = [0] \text{ si } i \neq j.$$

La matrice $P_i = X_i \cdot Y_i^* \in \mathbb{C}^{n\times n}$ est la matrice représentant dans $\mathbb{C}^n$ la projection sur M_i le long de l'ensemble $\{z \in \mathbb{C}^n,\ X_i^* z = 0\} = \bigoplus_{j\neq i} M_j$. On l'appelle **projection spectrale** associée à la valeur propre λ_i. En particulier on vérifie que

$$Y_i^* \cdot X_i = I_{m_i} \Longrightarrow P_i^2 = P_i, \quad Y_i^* \cdot X_j = [0] \Longrightarrow P_i P_j = 0,\ i \neq j,$$

à comparer à la Proposition A.4.6. Finalement on résume ces résultats dans la formule

$$A = \sum_{i=1}^{d} (\lambda_i P_i + D_i), \text{ avec } J_i = \lambda_i I_{m_i} + N_i,\ D_i = X_i N_i Y_i^* \text{ et } \sum_{i=1}^{d} P_i = I_n,$$

qui est la **décomposition spectrale** d'une matrice quelconque. Une forme des N_i est par exemple, pour $m_i = 7$, $g_i = 3$ et $l_i = 3$,

$$N_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

On rappelle que

– le rang de J_i est m_i ;
– le nombre de blocs présents est g_i ;
– le rang maximum d'un bloc est l_i.

On en déduit que si on itère l_i fois, ${N_i}^{l_i} = [0]$ soit ${D_i}^{l_i} = [0]$.

Dans le cas particulier d'une valeur propre λ_i *semi-simple*, on a $m_i = g_i$ et $l_i = 1$. La matrice N_i est nulle, et donc $D_i = [0]$.

Enfin, de la formule générale on déduit que

$$AP_j = \sum_{i=1}^{d} (\lambda_i P_i + D_i) P_j = \lambda_j P_j^2 + D_j P_j = P_j (\lambda_j P_j + D_j)$$

(P_j et D_j commutent par définition), et plus généralement

$$A^k P_j = P_j (\lambda_j P_j + D_j)^k .$$

Proposition A.6.1 *Soit $v \in \mathbb{C}^n$ un vecteur quelconque, pour toute valeur propre λ_i semi–simple, $P_i v$ est vecteur propre associé à λ_i si, et seulement si, $P_i v \neq 0$.*

Preuve : La démonstration découle de l'étude précédente puisque pour tout vecteur $v \in \mathbb{C}^n$, le vecteur $P_i v$ appartient au sous–espace $M_i = \mathrm{Ker}\,(A - \lambda_i)^{l_i}$. Donc

$$AP_i v = \lambda_i P_i^2 v + D_i P_i v = \lambda_i P_i v + D_i P_i v.$$

Mais dire que λ_i est semi–simple est équivalent à $D_i = [0]$, soit

$$AP_i v = \lambda_i P_i v.$$

■

Cette forme générale de la décomposition spectrale d'une matrice va être utilisée au chapitre 5 pour étudier la convergence d'algorithmes de calcul de valeurs propres.

Le lecteur intéressé par cet aspect de l'algèbre linéaire peut se reporter aux livres de F. Chatelin [3], B. N. Parlett [11] et Y. Saad [12].

Annexe B

Normes vectorielles et matricielles

B.1 Introduction

On rappelle dans ce chapitre quelques notions indispensables à l'étude des propriétés des matrices en vue de la résolution de systèmes linéaires par des méthodes itératives, mais aussi pour le calcul des valeurs propres et vecteurs propres. L'outil principal introduit dans ce chapitre est la norme (de vecteur ou de matrice) qui permet de définir la notion de convergence de suites (de vecteurs ou de matrices).

On établit également un lien entre les valeurs propres des matrices et certaines de ces normes ; enfin on introduit la notion de conditionnement d'une matrice, très importante pour les applications numériques.

On raisonne en général dans des espaces vectoriels sur $\mathbb{R}$. Les résultats se transposent sans difficulté au cas d'espaces vectoriels définis sur $\mathbb{C}$. Voir par exemple [5].

B.2 Normes de vecteurs

Définition B.2.1 *soit $\mathbb{E}$ un espace vectoriel sur $\mathbb{R}$, on appelle* **norme** *une application de $\mathbb{E}$ dans $\mathbb{R}^+$, notée $\|\cdot\|$, qui vérifie les trois propriétés suivantes :*

- $\forall x \in \mathbb{E}$, $\forall \lambda \in \mathbb{R}$, $\|\lambda x\| = |\lambda| \ \|x\|$
- $\forall x \in \mathbb{E}$, $\forall y \in \mathbb{E}$, $\|x + y\| \leq \|x\| + \|y\|$
- $\|x\| = 0 \quad \Longleftrightarrow \quad x = 0$.

La seconde propriété est appelée inégalité triangulaire ; un espace vectoriel muni d'une norme est dit espace vectoriel **normé**. Il est immédiat de vérifier que l'application "valeur absolue" est une norme sur l'espace vectoriel $\mathbb{R}$.

D'une façon plus générale, si $\mathbb{E}$ est un espace vectoriel sur $\mathbb{R}$ de dimension finie n, et si $B = \{b_1, b_2, \ldots, b_n\}$ est une base de $\mathbb{E}$, tout vecteur de $\mathbb{E}$ s'écrit de manière unique

$$x = \alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_n b_n.$$

A l'aide des coordonnées $\alpha_i \in \mathbb{R}$, on définit l'application $\|.\|$ de $\mathbb{E}$ dans $\mathbb{R}^+$ par

$$\|x\| = (|\alpha_1|^2 + |\alpha_2|^2 + \ldots + |\alpha_n|^2)^{1/2}.$$

Cette application est une norme, appelée **norme euclidienne** ou encore **norme canonique** associée à la base B.

Si $\mathbb{E}$ est un espace vectoriel sur $\mathbb{R}$ de dimension finie n, muni d'un produit scalaire $(\cdot, \cdot)$, on peut définir une norme par la relation

$$\|x\| = (x, x)^{1/2}.$$

Remarque B.2.1 *La réciproque n'est pas vraie, il existe des espaces vectoriels normés qui ne sont pas euclidiens, car la norme doit posséder des propriétés supplémentaires pour permettre de définir un produit scalaire.*

Proposition B.2.1 *[Inégalité de Schwarz] Pour tout couple de vecteurs x, y d'un espace vectoriel euclidien $\mathbb{E}$ sur $\mathbb{R}$*

$$|(x,y)| \leq \|x\| \times \|y\|.$$

Preuve :

$$\begin{aligned} \forall x, y \in \mathbb{E}, \forall \lambda \in \mathbb{R} \qquad \|\lambda x + y\|^2 &= (\lambda x + y, \lambda x + y) \\ &= \lambda^2 (x,x) + 2\lambda (x,y) + (y,y) \\ &= \lambda^2 \|x\|^2 + 2\lambda (x,y) + \|y\|^2 \end{aligned}$$

en prenant $x \neq 0$, le trinome du second degré en λ garde un signe constant quel que soit la valeur de $\lambda \in \mathbb{R}$, ce qui implique que le discriminant est négatif, soit

$$(x,y)^2 - \|x\|^2 \, \|y\|^2 \leq 0.$$

■

D'une manière générale, à l'aide des coordonnées α_i d'un vecteur x dans la base B, on peut lui associer, pour tout entier $p > 0$ fini, les normes suivantes appelées **normes de Hölder**

$$\|x\|_p = (|\alpha_1|^p + |\alpha_2|^p + \ldots + |\alpha_n|^p)^{1/p}.$$

Cette définition comprend les cas particuliers

$$\|x\|_1 = \sum_{i=1}^{n} |\alpha_i| \quad \text{et} \quad \|x\|_2 = (\sum_{i=1}^{n} |\alpha_i|^2)^{1/2}$$

et s'étend au cas $p = \infty$ avec la norme $\|x\|_\infty = \max_i |\alpha_i|$.

On établit alors la majoration suivante, appelée **inégalité de Hölder**

$$\forall x \in \mathbb{E} \qquad |(x,y)| \leq \|x\|_p \, \|y\|_q$$

pour tout couple d'entiers $p > 0$ et $q > 0$ liés par la relation

$$\frac{1}{p} + \frac{1}{q} = 1,$$

dont l'inégalité de Schwarz est un cas particulier, avec $p = q = 2$.

Définition B.2.2 *on dit que deux normes $\|\cdot\|$ et $|||\cdot|||$ définies sur un ensemble $\mathbb{E}$ sont* **équivalentes** *s'il existe deux constantes positives C_m et C_M telles que :*

$$\forall x \in \mathbb{E} \qquad C_m \, \|x\| \leq |||x||| \leq C_M \, \|x\|.$$

Théorème B.2.1 *Dans un espace vectoriel $\mathbb{E}$ sur $\mathbb{R}$ de dimension finie toutes les normes sont équivalentes.*

On admet ce résultat, qui prend les formes particulières suivantes, dans un espace vectoriel $\mathbb{E}$ de dimension finie n :

$$\begin{aligned} \|x\|_2 \leq \|x\|_1 &\leq \sqrt{n} \, \|x\|_2 \\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n} \, \|x\|_\infty \\ \|x\|_\infty \leq \|x\|_1 &\leq n \, \|x\|_\infty. \end{aligned}$$

Proposition B.2.2 *Soit $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels euclidiens sur $\mathbb{R}$ de dimension finie n et m, et soit f une application linéaire orthogonale de $\mathbb{E}$ dans $\mathbb{F}$, alors*

$$\forall x, y \in \mathbb{E} \qquad \|f(x)\|_m = \|x\|_n.$$

Preuve : En effet par définition de l'orthogonalité, on a

$$\forall x \in \mathbb{E} \qquad \|f(x)\|_m^2 = (f(x), f(x))_m = (x, x)_n = \|x\|_n^2.$$

On dit encore que f conserve la norme, et c'est donc une isométrie. ∎

B.3 Normes de matrices

D'après l'étude des applications linéaires l'ensemble $\mathbb{R}^{m\times n}$ des matrices à m lignes et n colonnes est un espace vectoriel sur $\mathbb{R}$ de dimension $m \times n$; on peut donc considérer toute matrice A de $\mathbb{R}^{m\times n}$ comme un vecteur à $m \times n$ composantes et ainsi utiliser une des normes vectorielles précédentes pour définir $\|A\|_{m,n}$. Il est cependant nécessaire d'introduire une condition supplémentaire pour obtenir un outil de démonstration de la convergence de suites et de séries de matrices.

Définition B.3.1 *On dit que la norme vectorielle $\|\cdot\|$ définie sur $\mathbb{R}^{n\times n}$ est une norme matricielle, si et seulement si elle vérifie pour tout couple (A, B) de matrices de $\mathbb{R}^{n\times n}$*

$$\|AB\| \leq \|A\| \, \|B\|.$$

Il est alors possible de définir une norme matricielle à partir des coefficients de la matrice. C'est la cas de la norme de **Schur-Frobenius** (voir la Proposition B.3.3)

$$\|A\|_F^2 = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} |A_{i,j}|^2.$$

Mais si $\mathbb{E}$ et $\mathbb{F}$ sont deux espaces vectoriels sur $\mathbb{R}$ de dimension finie, il existe une bijection entre l'ensemble $\mathcal{L}(\mathbb{E}, \mathbb{F})$ des applications linéaires de $\mathbb{E}$ dans $\mathbb{F}$ et l'ensemble $\mathbb{R}^{m\times n}$, une fois que l'on a choisi une base de $\mathbb{E}$ et une base de $\mathbb{F}$; on peut alors définir une norme à partir de l'application linéaire associée : si A est la matrice associée à l'application f

$$\|A\| = \max_{x\neq 0} \frac{\|f(x)\|_{\mathbb{F}}}{\|x\|_{\mathbb{E}}}$$

Cette application satisfait aux axiomes de la Définition B.2.1 et aussi à la Définition B.3.1 (voir la Proposition B.3.3) ; on dit que cette norme est **associée** à la norme vectorielle $\|\cdot\|$, ou **induite** par la norme vectorielle, ou encore **subordonnée** à la norme vectorielle.

Dans le cas particulier où $\mathbb{E} = \mathbb{R}^n$ et $\mathbb{F} = \mathbb{R}^m$, la matrice A est rectangulaire $A \in \mathbb{R}^{m\times n}$ et on note

$$\|A\|_{m,n} = \max_{x\neq 0} \frac{\|Ax\|_m}{\|x\|_n}.$$

Enfin dans le cas $m = n$, il est d'usage de noter de la même façon la norme vectorielle et la norme matricielle associée :

$$\|A\|_n = \max_{x\neq 0} \frac{\|Ax\|_n}{\|x\|_n}.$$

Pour éviter toute confusion les vecteurs sont représentés par des lettres minuscules et les matrices par des majuscules.

Proposition B.3.1 *Pour toute matrice carrée $A \in \mathbb{R}^{n\times n}$ les normes matricielles de Hölder vérifient*

(i) $\|A\|_1 = \max\limits_{x\neq 0} \dfrac{\|Ax\|_1}{\|x\|_1} = \max\limits_j \sum\limits_i |A_{i,j}|.$

(ii) $\|A\|_\infty = \max\limits_{x\neq 0} \dfrac{\|Ax\|_\infty}{\|x\|_\infty} = \max\limits_i \sum\limits_j |A_{i,j}|.$

Preuve : Pour tout vecteur $v \in \mathbb{R}^n$

$$\|Av\|_1 = \sum_i |\sum_j A_{i,j}v_j| \leq \sum_i \sum_j |A_{i,j}|\ |v_j| \leq (\max_j \sum_i |A_{i,j}|)\|v\|_1;$$

pour obtenir l'égalité, on construit un vecteur v particulier : soit j_0 un indice pour lequel

$$\sum_i |A_{i,j_0}| = \max_j \sum_i |A_{i,j}|\ ;$$

le vecteur e_{j_0} dont toutes les composantes sont nulles à l'exception de $e_{j_0} = 1$ répond à la question.

De même

$$\|Av\|_\infty = \max_i |\sum_j A_{i,j}v_j| \leq (\max_i \sum_j |A_{i,j}|)\|v\|_\infty;$$

soit i_0 un indice tel que

$$\sum_j |A_{i_0,j}| = \max_i \sum_j |A_{i,j}|\ ;$$

le vecteur v dont les composantes sont

$$v_j = 1 \quad \text{si } A_{i_0,j} = 0 \qquad \text{et } v_j = \frac{A_{i_0,j}}{|A_{i_0,j}|} \quad \text{si } A_{i_0,j} \neq 0$$

permet d'atteindre l'égalité. ■

Proposition B.3.2 *La norme de Frobenius $\|\cdot\|_F$ n'est pas une norme matricielle induite.*

Preuve : On pose $p = \min(m,n)$ et si I_p est la matrice identité d'ordre p, on considère la matrice $I_{m,n} \in \mathbb{R}^{m\times n}$ définie par

$$I_{m,n} = \begin{pmatrix} I_p \\ 0 \end{pmatrix} \quad \text{si } m > n, \quad I_{m,n} = \begin{pmatrix} I_p & 0 \end{pmatrix} \quad \text{si } m < n, \text{ et} \quad I_{m,n} = I_p \quad \text{si } n = m = p.$$

Pour toute norme induite, on a l'égalité

$$\|I_{m,n}\|_{m,n} = \max_{x\neq 0} \frac{\|I_{m,n}x\|_m}{\|x\|_n} = 1,$$

or on trouve $\|I_{n,m}\|_F = \sqrt{p}$. ■

On notera encore que pour toute matrice $A \in \mathbb{R}^{n\times n}$

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2.$$

Proposition B.3.3 *Les normes $\|\cdot\|_{m,n}$ et $\|\cdot\|_F$ vérifient*

$$\forall A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p} \qquad \|AB\| \leq \|A\|\ \|B\|.$$

Preuve : Traitons d'abord le cas de la norme $\|\cdot\|_{m,n}$; pour cela on considère trois entiers m, n et p et les espaces vectoriels $\mathbb{R}^{m\times n}$ et $\mathbb{R}^{n\times p}$ munis de leur norme induite : pour toutes les matrices $A \in \mathbb{R}^{m\times n}$ et $B \in \mathbb{R}^{n\times p}$, le produit AB est une matrice de $\mathbb{R}^{m\times p}$ qui vérifie

$$\|AB\|_{m,p} = \max_{x\neq 0} \frac{\|ABx\|_m}{\|x\|_p} = \max_{x\neq 0, Bx\neq 0} \frac{\|ABx\|_m}{\|Bx\|_n}\frac{\|Bx\|_n}{\|x\|_p} \leq \max_{y\neq 0}\frac{\|Ay\|_m}{\|y\|_n}\max_{x\neq 0}\frac{\|Bx\|_n}{\|x\|_p}.$$

Pour la norme de Frobenius,

$$\|AB\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{p}(AB)_{i,j}^2 = \sum_{i=1}^{m}\sum_{j=1}^{p}(\sum_{k=1}^{n}A_{i,k}B_{k,j})^2 \leq (\sum_{i=1}^{m}\sum_{k=1}^{n}A_{i,k}^2)(\sum_{j=1}^{p}\sum_{k=1}^{n}B_{k,j}^2).$$

■

Proposition B.3.4 *Les normes matricielles $\|\cdot\|_{m,n}$ et $\|\cdot\|_F$ vérifient*

$$\forall A \in \mathbb{R}^{m\times n} \qquad \max_{i,j}|A_{i,j}| \leq \|A\|_{m,n} \leq \|A\|_F \leq \sqrt{mn}\ \|A\|_{m,n}.$$

Preuve : D'après l'inégalité de Cauchy-Schwarz, pour tout vecteur $x \in \mathbb{R}^m$ et tout vecteur $y \in \mathbb{R}^n$, tels que $x \neq 0$ et $Ay \neq 0$, on a

$$|(x, Ay)| \leq \|x\|_m\ \|Ay\|_m$$

soit

$$\frac{|(x,Ay)|}{\|x\|_m\ \|y\|_n} \leq \frac{\|Ay\|_m}{\|y\|_n} \leq \|A\|_{m,n}.$$

Pour tout entier i donné, $1 \leq i \leq m$, on prend $x \in \mathbb{R}^m$ tel que toutes les composantes sont nulles sauf x_i qui vaut 1, et pour tout entier j donné, $1 \leq j \leq n$, on prend $y \in \mathbb{R}^n$ tel que toutes les composantes sont nulles sauf y_j qui vaut 1. Alors $\|x\|_m = 1$, $\|y\|_n = 1$ et $(x, Ay) = A_{i,j}$. Ainsi pour tout couple (i, j) vérifiant $1 \leq i \leq m$, et $1 \leq j \leq n$, $|A_{i,j}| \leq \|A\|_{m,n}$.
D'où la première majoration en passant au maximum.

En utilisant la majoration

$$\|Ax\|_m^2 = \sum_{i=1}^{m}(\sum_{j=1}^{n}A_{i,j}x_j)^2 \leq \sum_{i=1}^{m}[(\sum_{j=1}^{n}A_{i,j}^2)(\sum_{j=1}^{n}x_j^2)] = (\sum_{i=1}^{m}\sum_{j=1}^{n}A_{i,j}^2)(\sum_{j=1}^{n}x_j^2) = \|A\|_F^2\ \|x\|_n^2,$$

on arrive à $\|A\|_{m,n} \leq \|A\|_F$.

Enfin en majorant chaque coefficient de la matrice A, puis en utilisant la première inégalité, on obtient pour finir

$$\|A\|_F \leq \sqrt{mn}\max_{i,j}|A_{i,j}| \leq \sqrt{mn}\|A\|_{m,n}.$$

■

Remarque B.3.1 *Cette proposition illustre encore un cas particulier d'équivalence des normes :*

$$\|A\|_{m,n} \leq \|A\|_F \leq \sqrt{mn}\|A\|_{m,n}$$

et

$$\frac{1}{\sqrt{mn}}\|A\|_F \leq \|A\|_{m,n} \leq \|A\|_F.$$

Définition B.3.2 *Une matrice Q de $\mathbb{R}^{n\times n}$ est dite* **orthogonale** *lorsque $Q\,Q^T = Q^T\,Q = I_n$.*

Proposition B.3.5 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ et toute matrice orthogonale $Q \in \mathbb{R}^{n\times n}$*

$$\|QA\|_2 = \|AQ\|_2 = \|A\|_2.$$

Preuve : Ce résultat découle directement de la Proposition B.2.2 en écrivant

$$\|QA\|_2 = \max_{x\neq 0} \frac{\|QAx\|_2}{\|x\|_2} = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_2;$$

de même

$$\|AQ\|_2 = \max_{x\neq 0} \frac{\|AQx\|_2}{\|x\|_2} = \max_{x\neq 0} \frac{\|AQx\|_2}{\|Qx\|_2} = \|A\|_2.$$

■

Remarque B.3.2 *Cette propriété se généralise dans le cas complexe aux matrices U unitaires :*

$$\|UA\|_2 = \|AU\|_2 = \|A\|_2.$$

B.4 Normes des matrices et valeurs propres

On rappelle que les valeurs propres d'une matrice A de $\mathbb{R}^{n\times n}$, notées λ_i ou $\lambda_i(A)$, appartiennent *a priori* à $\mathbb{C}$.

Définition B.4.1 *le* **rayon spectral** *de la matrice $A \in \mathbb{R}^{n\times n}$ est le réel positif*

$$\rho(A) = \max_i |\lambda_i(A)|.$$

Proposition B.4.1 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ et toute norme matricielle $\|\cdot\|$*

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

Preuve : On fait la démonstration pour les normes induites, la démonstration complète se trouvant dans [7]. Soit λ une valeur propre de A et v un vecteur propre associé :

$$Av = \lambda v \Longrightarrow |\lambda|\|v\| = \|Av\| \leq \|A\|\|v\| \ ;$$

soit, pour toute valeur propre λ

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

■

Proposition B.4.2 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ et tout $\varepsilon > 0$, il existe une norme matricielle $\|\cdot\|$ telle que*

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

Preuve : Pour obtenir ce résultat, on utilise une propriété importante des matrices carrées (voir la Proposition A.4.4) : toute matrice $A \in \mathbb{R}^{n\times n}$ peut s'écrire sous la forme $A = QTQ^*$, où Q est une matrice unitaire ($Q^* = Q^{-1}$) et T une matrice triangulaire supérieure dont la diagonale est formée des valeurs propres de la matrice A (ces valeurs propres peuvent être complexes ou réelles, nulles ou non, distinctes ou non et ne sont pas rangées suivant leur module) :

$$T = \begin{pmatrix} \lambda_1 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \ddots & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Soit δ un nombre réel strictement positif, on construit maintenant la matrice diagonale $D \in \mathbb{R}^{n\times n}$

$$D = \begin{pmatrix} \delta & 0 & 0 & 0 & 0 & 0 \\ 0 & \delta^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta^n \end{pmatrix};$$

alors

$$(QD)^{-1}A(QD) = D^{-1}TD = \begin{pmatrix} \lambda_1 & \delta T_{1,2} & \delta^2 T_{1,3} & \dots & \dots & \delta^{n-1}T_{1,n} \\ 0 & \lambda_2 & \delta T_{2,3} & \ddots & \ddots & \delta^{n-2}T_{2,n} \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \ddots & \delta^2 T_{n-2,n} \\ 0 & 0 & 0 & 0 & \ddots & \delta T_{n-1,n} \\ 0 & 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Pour une matrice $A \in \mathbb{R}^{n\times n}$ et un réel ε donnés, on peut trouver $\delta \in \mathbb{R}$ tel que

$$\max_{i<j} |\delta^{j-i} T_{i,j}| < \varepsilon/n;$$

alors la norme matricielle (dépendant de A et ε) définie pour toute matrice $B \in \mathbb{R}^{n\times n}$ par

$$\|B\|_{A,\varepsilon} = \|(QD)^{-1}B(QD)\|_\infty$$

vérifie l'inégalité

$$\|A\|_{A,\varepsilon} \leq \max_i |\lambda_i| + \varepsilon$$

Cette norme est la norme matricielle subordonnée à la norme vectorielle

$$v \mapsto \|(QD)^{-1}v\|_\infty$$

car

$$\|(QD)^{-1}B(QD)\|_\infty = \max_{y\neq 0} \frac{\|(QD)^{-1}B(QD)y\|_\infty}{\|y\|_\infty} = \max_{x\neq 0} \frac{\|(QD)^{-1}Bx\|_\infty}{\|(QD)^{-1}x\|_\infty}.$$

■

Proposition B.4.3 *Pour toute matrice* $A \in \mathbb{R}^{n\times n}$

$$\rho(A) \leq \|A\|_2 = \sqrt{\rho(A^T A)}.$$

Pour toute matrice $A \in \mathbb{R}^{n\times n}$ symétrique

$$\rho(A) = \|A\|_2.$$

Preuve : D'après la Proposition B.4.1 on a toujours $\rho(A) \leq \|A\|_2$; de plus par définition

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\neq 0} (\frac{(Ax,Ax)}{(x,x)})^{1/2} = \max_{x\neq 0} (\frac{(A^T Ax,x)}{(x,x)})^{1/2} = \rho^{1/2}(A^T A).$$

En effet $A^T A \in \mathbb{R}^{n\times n}$ est une matrice symétrique, qui admet une base de vecteurs propres orthogonaux $\{v_1, v_2, \dots, v_n\}$ (voir la Proposition A.4.3). Ses valeurs propres $\lambda_i(A^T A)$ étant positives, on les range par valeurs décroissantes :

$$\lambda_n(A^T A) \leq \lambda_{n-1}(A^T A) \leq \dots \leq \lambda_2(A^T A) \leq \lambda_1(A^T A), \text{ et } \rho(A^T A) = \lambda_1(A^T A).$$

On obtient pour tout vecteur $u \in \mathbb{R}^n$

$$u = \sum_{i=1}^{n} \alpha_i v_i, \quad \text{et } Au = \sum_{i=1}^{n} \alpha_i \lambda_i(A^T A) v_i.$$

Ainsi

$$\frac{(Au, Au)}{(u, u)} = \frac{\sum_{i=1}^{n} |\alpha_i|^2 \lambda_i(A^T A)^2 (v_i, v_i)}{\sum_{i=1}^{n} |\alpha_i|^2 (v_i, v_i)} \leq \lambda_1(A^T A)^2 = \rho(A^T A).$$

Dans le cas d'une matrice A symétrique, on écrit de même

$$\frac{(Au, Au)}{(u, u)} = \frac{\sum_{i=1}^{n} |\alpha_i|^2 |\lambda_i(A)|^2 (v_i, v_i)}{\sum_{i=1}^{n} |\alpha_i|^2 (v_i, v_i)} \leq \max_{i=1,n} |\lambda_i(A)|^2 = \rho^2(A).$$

■

Définition B.4.2 *Soit A une matrice de $\mathbb{R}^{n\times n}$. A toute valeur propre de $A^T A$, $\lambda(A^T A) \geq 0$, on associe la* **valeur singulière** *$\sigma(A)$, par la relation*

$$\sigma(A) = \sqrt{\lambda(A^T A)}.$$

De manière classique, on range les valeurs singulières de la matrice A par valeur décroissante :

$$\sigma_n(A) \leq \sigma_{n-1}(A) \leq \ldots \leq \sigma_2(A) \leq \sigma_1(A).$$

La Proposition B.4.3 se résume à

$$\|A\|_2 = \sigma_1(A).$$

Enfin, on vérfie facilement le résultat suivant

Proposition B.4.4 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ la norme de Schur-Frobenius vérifie*

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |A_{i,j}|^2 = \sum_{i=1,n} \lambda_i(A^T A) = \sum_{i=1,n} \sigma_i^2(A).$$

Preuve : Par définition de la norme de Frobenius

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |A_{i,j}|^2 = tr(A^T A) = \sum_{i=1,n} \lambda_i(A^T A) = \sum_{i=1,n} \sigma_i^2(A)$$

où $tr(B)$ désigne la trace de la matrice B qui est aussi la somme de ses valeurs propres. ■

Remarque B.4.1 *Ces résultats s'étendent au cas d'une matrice A complexe, en remplaçant dans les formules A^T par A^*.*

B.5 Suites de vecteurs. Suites de matrices

Dans ce paragraphe, on se place explicitement dans $\mathbb{C}^n$. Bien évidemment, les résultats restent valables pour des matrices à coefficients réels !
L'analyse théorique des algorithmes de calcul numérique utilise la notion de suite de vecteurs et étudie leur convergence éventuelle. On notera $\{x_k\}_{k\in\mathbb{N}}$ une suite d'éléments d'un espace vectoriel $\mathbb{E}$ sur $\mathbb{C}$, muni de la norme $\|\cdot\|$, et on dira que la suite $\{x_k\}_{k\in\mathbb{N}}$ converge vers l'élément x de $\mathbb{E}$ si

$$\lim_{k\to\infty} \|x_k - x\| = 0,$$

que l'on écrit de manière classique $\lim_{k\to\infty} x_k = x$. Dans le cas particulier d'un espace vectoriel $\mathbb{E}$ de dimension finie n, cette définition est indépendante de la norme choisie, et la convergence de la suite $\{x_k\}_{k\in\mathbb{N}}$ vers x est équivalente à la convergence de chacune des suites de composantes $\{(x_k)_i\}_{k\in\mathbb{N}}$ vers x_i pour $i = 1, 2, \ldots, n$, par rapport à une base quelconque.

On définit de manière analogue une suite de matrices $\{A_k\}_{k\in\mathbb{N}} \in \mathbb{C}^{m\times n}$ et on peut utiliser la définition de la convergence d'une suite de vecteurs dans l'espace vectoriel $\mathbb{C}^{m\times n}$ de dimension finie $n \times m$. Cependant dans de nombreux algorithmes les éléments de la suite de matrices considérée sont de la forme $A_k = A^k$, puissances successives d'une matrice donnée $A \in \mathbb{C}^{n\times n}$. Dans ce cas particulier, on relie la convergence de cette suite au rayon spectral de la matrice A.

Théorème B.5.1 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ les conditions suivantes sont équivalentes :*
(i) $\lim_{k\to\infty} A^k = 0$ *;*
(ii) $\forall x \in \mathbb{R}^n$, $\lim_{k\to\infty} A^k x = 0$ *;*
(iii) $\rho(A) < 1$ *;*
(iv) il existe une norme induite telle que $\|A\| < 1$.

Preuve : On montre l'équivalence par implication circulaire.

(i) $\Longrightarrow$ (ii) : pour toute norme vectorielle $\|\cdot\|$ et sa norme matricielle induite, on a la majoration

$$\forall x \in \mathbb{R}^n \qquad \|A^k x\| \leq \|A^k\| \times \|x\|.$$

d'où le résultat.

(ii) $\Longrightarrow$ (iii) : soit λ une valeur propre de A telle que $\rho(A) = |\lambda|$ et soit $u \neq 0$ un vecteur propre associé, alors

$$\|A^k u\| = \|\lambda^k u\| = |\lambda|^k \|u\| = \rho(A)^k \|u\|.$$

Si on suppose $\rho(A) \geq 1$ alors $\|A^k u\| \geq \|u\|$, ce qui contredit (ii).

(iii) $\Longrightarrow$ (iv) : est une conséquence de la Proposition B.4.2 (prendre $\varepsilon = (1 - \rho(A))/2$).

(iv) $\Longrightarrow$ (i) : puisque $\|A^k\| \leq \|A\|^k$, on a bien $\lim_{k\to\infty} \|A^k\| = 0$ pour la norme matricielle telle que $\|A\| < 1$. ∎

Remarque B.5.1 *Il faut souligner que l'utilisation d'une norme matricielle arbitraire pour évaluer la convergence d'une suite peut amener à une mauvaise conclusion, comme le montre l'exercice suivant. Par contre, la valeur du rayon spectral de la matrice est toujours pertinente.*

Exercice B.5.1 *Calculer les normes $\|\cdot\|_1$, $\|\cdot\|_\infty$, $\|\cdot\|_F$ ainsi que le rayon spectral des matrices*

$$A = \begin{pmatrix} 0.9 & 0.0 \\ 0.4 & 0.8 \end{pmatrix} \quad \textit{et} \quad B = \begin{pmatrix} 5. & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \begin{pmatrix} 0.9 & 0.0 \\ 0.4 & 0.8 \end{pmatrix} \begin{pmatrix} 0.2 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}.$$

B.6 Critères numériques associés à la convergence

Tout d'abord, il faut être conscient, lorsque l'on effectue un calcul *numérique*, que la précision est **finie**, à la différence du calcul *formel*, par exemple. La première question est donc, pourquoi utilise-t-on une méthode numérique, *a priori* moins précise ? La réponse est pragmatique : on ne sait pas résoudre formellement un système linéaire, dès lors que le dimension de la matrice A est trop grande ; ou de façon encore plus pragmatique, le temps de résolution est de toute façon beaucoup trop important !

Que signifie alors l'association de termes **convergence numérique** ? Avant de répondre à cette question, nous allons détailler quelques problèmes inhérents au calcul numérique, par opposition

au calcul formel.
La finitude de la précision vient de la représentation en machine des nombres réels, sous la forme générique[1]

$$\pm a_0, a_1 \cdots a_p \ 10^d, \text{ avec } (a_0, \cdots, a_p) \in \{0, \cdots, 9\}^{p+1}, \ a_0 \neq 0, \ d \in \{-d_{max}, \cdots, d_{max}\},$$

où p et d_{max} dépendent du microprocesseur qui effectue les calculs. On dit aussi que $p+1$ est le nombre maximal de chiffres significatifs de la représentation en machine, et que $10^{-d_{max}}$ est la précision machine. Cette représentation génère deux difficultés :

Tout nombre dont la valeur absolue est plus grande que $10^{d_{max}+1}$ est considéré comme infini, et symétriquement, tout nombre dont la valeur absolue est strictement plus petite que $10^{-d_{max}}$ est considéré comme étant nul ;

Les opérations sur ces nombres (addition, multiplication, etc. ; extraction de racine, exponentation, etc.) sont effectuées en précision finie. Prenons l'exemple de la multiplication... Si les deux nombres ont respectivement q et q' chiffres significatifs ($q, q' \in \{1, \cdots, p+1\}$), leur produit possède $q+q'-1$ ou $q+q'$ chiffres significatifs. Dès lors que $q+q'-1 > p+1$, une *troncature* est effectuée lors de la mise en mémoire du résultat (même si le calcul était exact), puisque la représentation de tout nombre comporte au plus $p+1$ chiffres significatifs.

C'est la raison pour laquelle les calculs numériques produisent en général des erreurs d'arrondi... Par voie de conséquence, et pour revenir à notre problème, il devient difficile d'obtenir un résultat du type[2] $Au - b = 0$. Par ailleurs, on se contente en général d'une valeur *approchée*, c'est-à-dire à ε près, pour éviter un coût calcul trop élevé (compromis coût calcul-précision). Nous venons d'introduire la notion de *calcul exact à ε près*, qui est très courante chez l'ingénieur. Petit à petit, nous glissons du monde des mathématiques, en passant par celui du calcul scientifique, vers celui de l'art de l'ingénieur. Ces mondes, bien qu'ils ne répondent pas aux mêmes critères, n'en restent pas moins complémentaires, et indissociables.
Revenons aux mathématiques, après cette brêve incursion. Quand on parle de calcul exact à ε près, quel est le sens mathématique sous-jacent ? Typiquement, si on note $\|\cdot\|$ une norme quelconque, pour $\varepsilon \in \mathbb{R}_\star^+$, on cherche v_ε tel que

$$\|Av_\varepsilon - b\| \leq \varepsilon. \tag{B.1}$$

Il est clair que l'ensemble des v_ε qui satisfont à (B.1) n'est pas réduit à un singleton ! Quoiqu'il en soit, à ε près, l'obtention d'un tel v_ε est suffisante... On parle de *convergence numérique*.

Exercice B.6.1 *Quel est l'ensemble défini par (B.1) ?*

Comme nous le verrons dans le cours, les résultats de convergence peuvent être obtenus pour des normes quelconques, ou pour des normes spécifiques, telle que la norme associée à A, et définie par

$$\|v\|_A = (Av, v)^{1/2}.$$

Exercice B.6.2 *Vérifier que, lorsque A est symétrique définie positive, $\|\cdot\|_A$ est bien une norme dans $\mathbb{R}^n$.*

A la notion de calcul à ε près correspond, par dualité, celle de la précision requise, ce qui permet de déterminer un **critère d'arrêt** pour notre méthode. En effet, pour $\varepsilon \in \mathbb{R}_\star^+$ et u_0 donnés, on va effectuer des itérations,

$$\text{Pour } k = 0, 1, \cdots, \textit{tant que } \|Au_k - b\| > \varepsilon \quad \text{itérer } u_k \to u_{k+1}.$$

[1] Plus précisément, la représentation est du type indiqué ci-dessous, mais en base 2.

[2] Et même si l'ordinateur affirme que $Au - b = 0$, ceci signifie uniquement que la différence est plus petite que la précision machine, d'après l'exposé précédent.

(Les itérations sont interrompues pour la plus petite valeur de k telle que $\|Au_k - b\| \leq \varepsilon$.)
A partir de là, la voie est libre pour évaluer le **coût calcul** d'une méthode itérative.
La première quantité est le **nombre d'itérations** nécessaire à la validation du critère d'arrêt. Naturellement, on aura tendance à privilégier une méthode nécessitant peu d'itérations. C'est effectivement un critère, mais ça n'est pas le seul. Baser une analyse de la qualité d'une méthode itérative sur le nombre d'itérations uniquement est *incorrect*. Un second critère, complémentaire du premier, est le **coût d'une itération**. Typiquement, il s'agit du nombre d'opérations nécessaires à la réalisation d'une itération, c'est-à-dire au calcul de u_{k+1}, connaissant u_k. A partir de ces deux critères, on obtient une idée du **coût calcul** en multipliant le nombre d'itérations par le coût d'une itération.
Donnons deux exemples élémentaires d'estimation du nombre d'opérations dans $\mathbb{R}^n$.
1. Le *produit scalaire* de deux vecteurs, qui s'écrit

$$(x, y) = \sum_{i=1}^{n} x_i y_i,$$

est effectué en n multiplications et $(n-1)$ additions. Usuellement, on ne conserve que le terme principal, ce qui signifie que l'on considère que le produit scalaire requiert n additions et n multiplications.
2. La *multiplication matrice vecteur*, qui s'écrit composante par composante,

$$(Ax)_i = \sum_{j=1}^{n} A_{i,j} x_j, \quad 1 \leq i \leq n,$$

requiert n^2 additions et n^2 multiplications, ce qui laisse à penser qu'un produit matrice-vecteur est équivalent à n produits scalaires... Ceci étant, que se passe-t-il si l'on sait que la matrice A est creuse, c'est-à-dire avec K éléments non nuls par ligne, en moyenne, pour K très petit devant n. On ne va stocker que les positions, i. e. les paires d'indices (i, j), et les valeurs $A_{i,j}$ non nulles ! Lorsque l'on multiplie A par x, on n'effectue que les multiplications pour lesquelles $A_{i,j} \neq 0$ (et les additions de termes non nuls). En moyenne, on aura donc effectué Kn additions, et autant de multiplications...
Pourquoi un tel exemple ? Lorsque l'on résout un problème par une méthode de différences finies ou d'élément finis, la matrice obtenue comporte très peu d'éléments non nuls par ligne, de l'ordre d'une dizaine[3]. Si la dimension de l'espace est $n = 10^4$ (ce qui est très courant !), on voit que les deux évaluations du coût calcul donnent

$$2n^2 = 2\,10^8, \text{ et } 2Kn = 14\,10^4,$$

ou l'équivalent de 10.000 produits scalaires, contre 14.

Une autre façon d'estimer le coût du calcul est de mesurer le **temps de calcul**, par l'intermédiaire d'une horloge.
A priori, ces deux méthodes semblent tout à fait similaires. De fait, ceci dépend de la machine sur laquelle on effectue le calcul numérique. La première objection concerne les *opérations*. Une addition, une multiplication, une division ont-elles le même coût ? Une réponse possible consiste à compter précisément le nombre de chaque type d'opérations[4]... Un problème beaucoup plus épineux est que la machine peut (pour simplifier, il existe d'autres modes de fonctionnement), soit travailler *séquentiellement*, soit *en parallèle*. Dans le premier cas, les opérations sont exécutées l'une après l'autre. Dans le second cas, la machine est constituée de plusieurs processeurs, qui peuvent alors exécuter simultanément des opérations, et échanger des données entre eux[5]. Dans ce cas, supposons que l'on teste plusieurs fois le même problème, sur une machine disposant de

[3] On a vu que K est inférieur à 3 (resp. 5, 7) pour la discrétisation d'un Laplacien par différences finies en 1D (resp. 2D, 3D). On peut également prouver que $K \leq 3$ (resp. 7) pour un calcul par éléments finis modélisant le même problème en 1D (resp. 2D).

[4] Ceci étant, on raisonne usuellement en opérations flottantes par seconde, ou **FLOPs** = **FL**oating **OP**erations per **s**econd, pour un processeur donné, sans distinguer les opérations entre elles.

[5] On suppose l'algorithme de calcul le permet. Le fait qu'un algorithme est effectivement exécutable en parallèle, ou *parallélisable*, sort du cadre de ce cours...

plus en plus de processeurs : le temps horloge diminue, alors que le nombre total d'opérations restera constant ! Les deux estimateurs de coût calcul ne sont donc pas si similaires que ça...
Enfin, il peut également être utile de quantifier le **stockage mémoire** requis pour l'exécution de la méthode. Par exemple, lorsque l'on utilise une méthode itérative, on constate que le stockage est beaucoup plus faible que pour une méthode directe. Ceci ne préjuge cependant pas de la supériorité d'une méthode sur une autre...
Cette discussion est volontairement restée très générale, et elle peut être vue comme une introduction à l'algorithmique numérique. Ce qu'il faut retenir, c'est qu'il convient d'être prudent lorsque l'on évalue la qualité d'une méthode numérique, car celle-ci résulte habituellement de compromis entre les divers critères et contraintes que nous avons évoqués ci-dessus. Pour ce type de problèmes, il est fort utile d'acquérir de l'expérience, notamment en réalisant des comparaisons entre plusieurs méthodes.

Bibliographie

[1] **L. M. Adams, H. F. Jordan**, Is SOR color-blind ?, *SIAM Journal on Scientific and Statistical Computing*, **7** (1986).

[2] **A.-S. Bonnet Ben Dhia, C. Hazard, E. Luneville**, *Résolution numérique des équations aux dérivées partielles*, Cours MA 201, ENSTA.

[3] **F. Chatelin**, *Valeurs propres de matrices*, Masson, Paris (1988).

[4] **P. Ciarlet, P. Joly**, *Optimisation quadratique*, Cours AO 101, ENSTA.

[5] **P. G. Ciarlet**, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris (1982).

[6] **P. G. Ciarlet, B. Miara, J.-M. Thomas**, *Exercices d'analyse numérique matricielle et d'optimisation*, Masson, Paris (1982).

[7] **A. S. Householder**, *The theory of matrices in numerical analysis*, Blaisdell Publishing Company (1970).

[8] **R. Krikorian**, *Linéarisation et stabilité des équations différentielles*, Cours AO 102, ENSTA.

[9] **P. Lascaux, R. Théodor**, *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Masson, Paris (1987).

[10] **J. Ortega, R. J. Plemmons**, Extension of the Ostrowski-Reich theorem for SOR iterations, *Linear Algebra and its Applications*, **28** (1987).

[11] **B. N. Parlett**, *The symmetric eigenvalue problem*, Prentice Hall, Englewood Cliffs (1980).

[12] **Y. Saad**, *Numerical methods for large eigenvalue problems*, Manchester University Press, Manchester (1992).

[13] **W. Shakespeare**, *Much ado about nothing* (ca. 1598).

[14] **D. M. Young**, *Iterative solution of large linear systems*, Academic Press, New York (1971).

Index