AO101 2002-2003

Optimisation et algèbre linéaire

Patrick Ciarlet et Pascal Joly

Table des matières

Avant-Propos

Cet ouvrage constitue le support de cours d'un enseignement de première année de l'ENSTA. A ce titre, il s'agit d'une introduction aux matières abordées, l'optimisation et l'algèbre linéaire. Notons tout de suite que, par rapport aux enseignements dispensés en Classes Préparatoires, ce cours présente l'originalité de mêler étroitement les deux disciplines que sont l'algèbre et l'analyse.

Comme ces deux matières sont traitées au sein d'un même cours, en commençant par l'optimisation, et en poursuivant par l'algèbre linéaire, nous avons essayé, autant que faire se peut, de les relier. Notamment, dans la première partie, un certain nombre de problèmes de minimisation sont étudiés, qui peuvent ensuite être résolus à l'aide de méthodes d'algèbre linéaire. Par exemple, nous considérons les moindres carrés *linéaires*, ainsi que des algorithmes de minimisation de fonctionnelles *quadratiques*. On montre que ces problèmes particuliers peuvent être ramenés à la résolution de systèmes linéaires. Dans la deuxième partie, on revient sur ces liens, lorsque l'on aborde les méthodes *itératives* de résolution de systèmes linéaires.

Plus généralement, ces thèmes sont repris, développés, voire plus simplement utilisés comme outils, dans un certain nombre d'autres enseignements dispensés à l'ENSTA, en première, deuxième ou troisième années. Remarquons, et c'est fondamental, que les enseignements évoqués ne se limitent pas aux Mathématiques Appliquées !

Enfin, les auteurs tiennent à remercier chaleureusement les personnes suivantes pour leur relecture attentive, ainsi que pour les conseils qu'ils ont prodigués lors de l'élaboration de ce manuscrit : P. Carpentier, F. Jean, O. Kaber, E. Lunéville et M. Postel.

Patrick Ciarlet et Pascal Joly

Première partie :

Optimisation

Avant de commencer...

Un ouvrage mathématique ne contient pas toute la théorie du sujet qu'il aborde, et à plus forte raison les démonstrations des résultats fondamentaux sur lesquels il repose. Aussi, nous nous contentons d'évoquer les prérequis en algèbre linéaire et en analyse, ainsi que quelques éléments de démonstration, nécessaires à une bonne compréhension des notions abordées dans la Partie 1.

Pour ce qui est de l'algèbre linéaire, nous nous plaçons dans un espace vectoriel $\mathbb{E}$ de dimension finie, et soit $\mathbb{F}$ un sous-espace vectoriel de $\mathbb{E}$. Dans la suite, nous étudierons essentiellement des espaces vectoriels définis sur $\mathbb{R}$.

Rappelons quelques résultats bien connus concernant les sous-espaces vectoriels *supplémentaires*.

Il existe $\mathbb{G}$ un sous-espace vectoriel de $\mathbb{E}$ tel que $\mathbb{E} = \mathbb{F} + \mathbb{G}$, et $\mathbb{F} \cap \mathbb{G} = \{0\}$; de plus, $dim(\mathbb{E}) = dim(\mathbb{F}) + dim(\mathbb{G})$. La preuve de ce résultat repose sur le théorème de la base incomplète.

En particulier, on en déduit que $dim(\frac{\mathbb{E}}{\mathbb{F}}) = dim(\mathbb{E}) - dim(\mathbb{F})$, car $\frac{\mathbb{E}}{\mathbb{F}}$ est isomorphe à $\mathbb{G}$.

De là, on prouve que, pour toute application linéaire u de $\mathbb{E}$ dans un autre espace vectoriel,

$$dim(\mathbb{E}) = dim(Ker\, u) + rg(u), \text{ avec le rang } rg(u) = dim(Im\, u).$$

(On utilise le fait que $\frac{\mathbb{E}}{Ker\, u}$ est isomorphe à $Im\, u$.)

A toute application linéaire d'un espace vectoriel $\mathbb{E}$ dans un espace vectoriel $\mathbb{E}'$ (tous deux de dimension finie), on peut associer une matrice, qui dépend des bases de $\mathbb{E}$ et $\mathbb{E}'$ choisies. Si $\mathbb{E}$ est de dimension m et $\mathbb{E}'$ de dimension n, les matrices appartiennent à $\mathbb{R}^{n\times m}$, c'est-à-dire qu'elles possèdent n lignes et m colonnes.

Ceci permet de définir entre autres la *transposée*, le *rang*, les *vecteurs propres* et *valeurs propres* d'une matrice, par référence aux notions correspondantes pour les applications linéaires. En particulier, si note A une matrice, et $(A_{i,j})_{1\leq i\leq n, 1\leq j\leq m}$ ses éléments, on a par définition $A^{\mathsf{T}}_{i,j} = A_{j,i}$.
Rappelons également que l'on peut démontrer que le rang d'une matrice est égal à la taille maximale d'une famille libre de ses vecteurs colonnes. On peut aisément en déduire que pour toute matrice A,

$$rg(A) = rg(A^{\mathsf{T}}).$$

A partir de la transposition, on définit la classe des matrices *symétriques*, c'est-à-dire le sous-ensemble des matrices carrées, telles que $A = A^{\mathsf{T}}$.

Enfin, dans un espace vectoriel (défini sur $\mathbb{R}$, rappelons-le !), nous considérons que les notions de *produit scalaire* et d'espace vectoriel *euclidien* sont connues (ainsi que celle de base *orthonormale*, etc.). Le résultat fondamental est le suivant :
Toute matrice symétrique est *diagonalisable* dans une base orthonormale de vecteurs propres : si A est symétrique, il existe une base orthonormale, notée $(p_i)_{1\leq i\leq n}$, telle que

$$Ap_i = \lambda_i p_i,\ 1 \leq i \leq n.$$

Ou, de façon équivalente: il existe O une matrice orthogonale et D une matrice diagonale, telles que

$$D = O^{-1}AO.$$

Nous supposons pour finir que le lecteur est familiarisé avec la notion de dérivation d'une fonction de la variable réelle, à valeurs dans $\mathbb{R}$, $\mathbb{R}^n$ ou dans un espace vectoriel normé sur $\mathbb{R}$. Si γ est une telle fonction, on dénotera par $\gamma'(t)$:

la dérivée en t (si γ est à valeurs dans $\mathbb{R}$), ou

le vecteur dérivé en t (dans les autres cas).

Chapitre 1

Introduction

L'optimisation est un concept qui fait partie intégrante de la vie courante. Citons quelques exemples tout à fait banals, mais représentatifs :

Quel est le meilleur itinéraire pour aller d'un point A à un point B en voiture ?

Au tennis, comment maximiser l'effet, la vitesse d'une balle de service ?

Peut-on gagner contre la banque à la roulette au casino ?

A la bourse, comment maximiser les profits tout en minimisant les risques ?

Pourquoi tel composant chimique réagit-il avec tel autre ?

etc.

Une stratégie raisonnable est d'essayer de modéliser chacun de ces problèmes, c'est-à-dire de les reformuler sous une forme mathématique, puis de résoudre/optimiser les modèles mathématiques ainsi obtenus, et enfin de tester les résultats sur les situations pratiques... La modélisation, la mise en équations, ne sera que très marginalement étudiée dans ce cours (voir le chapitre 3). De fait, cette activité est du ressort du physicien, du chimiste, de l'économiste, du joueur (!)... L'ingénieur, à qui revient la charge de résoudre ces modèles, se doit de les bien connaître, notamment en ce qui concerne les hypothèses sous lesquelles le modèle est valide, avant d'envisager leur résolution. Dans cette optique, le thème de ce cours est la construction, et la justification mathématique, de méthodes de résolution de ces modèles. Nous considérerons principalement des modèles simplifiés, que nous nous attacherons à analyser (mathématiquement) en détail. Nous proposerons également des méthodes de résolution approchées, c'est-à-dire leur résolution numérique sur ordinateur. En particulier, nous ferons appel à des outils d'analyse (topologie, calcul différentiel, convexité), mais aussi à de nombreuses branches d'algèbre linéaire. En ce sens, la distinction algèbre/analyse, classique en classes préparatoires, s'estompera.

1.1 Cadre du problème

En pratique, lorsque l'on résout un problème d'optimisation, on utilise des algorithmes permettant d'approcher numériquement la solution d'un problème du type

$$\text{Trouver } u \in K, \text{ tel que } J(u) = \inf_{v \in K} J(v), \text{ ou bien } J(u) = \sup_{v \in K} J(v),$$

où J est une fonctionnelle définie sur un ensemble K non vide, à valeurs dans $\mathbb{R}$.
Avant d'envisager l'utilisation d'un algorithme, il est naturel[1] de répondre aux questions ci-dessous (dans cet ordre !) :

(i) **Existe**-t-il une solution u ? Est-elle **unique** ?

(ii) Comment la **caractériser** ?

(iii) Quel(s) **algorithme(s)** permet(tent) de calculer la solution ?
S'il y a un choix à faire, quel est l'algorithme le plus **efficace** ?

Pour commencer, nous allons démontrer quelques résultats élémentaires concernant le point (i), puis nous étudierons un exemple simple de détermination d'un minimum.

1.2 Existence d'un minimum : résultats généraux

Plaçons nous dans la situation *abstraite* suivante : soit $\mathbb{E}$ un espace vectoriel normé, K un sous-ensemble non vide de $\mathbb{E}$ et J une fonctionnelle définie sur K, à valeurs dans $\mathbb{R}$.

Définition 1.2.1 *u est un* **point de minimum local** *de J sur K si, et seulement si*

$$\exists \eta > 0, \quad \forall v \in K, \quad \|v - u\| < \eta \Longrightarrow J(u) \leq J(v).$$

u est un **point de minimum global** *de J sur K si, et seulement si*

$$\forall v \in K, \quad J(u) \leq J(v).$$

Définition 1.2.2 *On dit qu'une suite $(u_k)_{k\in\mathbb{N}}$ d'éléments de K est une* **suite minimisante** *si, et seulement si,*

$$\lim_{k\to+\infty} J(u_k) = \inf_{v\in K} J(v).$$

Remarque 1.2.1 *Par définition de la notion d'infimum, il existe toujours des suites minimisantes !*

Commençons par rappeler le résultat suivant, bien connu.

Théorème 1.2.1 *Si K est compact et J est continue sur K, elle atteint ses extréma :*

$$\exists (u_{min}, u_{max}) \in K \times K, \text{ tels que } J(u_{min}) = \inf_{v\in K} J(v), \ J(u_{max}) = \sup_{v\in K} J(v).$$

On peut montrer une variante du théorème 1.2.1, valable lorsque $\mathbb{E}$ est de dimension *finie*. Ce résultat est fort utile si K est non compact.

Définition 1.2.3 *On dit qu'une fonctionnelle J est infinie à l'infinie si, et seulement si,*

$$\lim_{v,\|v\|\to+\infty} J(v) = +\infty.$$

Proposition 1.2.1 *Dans $\mathbb{R}^n$, si K est un fermé, et si J est continue et infinie à l'infini, alors elle admet un minimum global sur K. De plus, de toute suite minimisante, on peut extraire une sous-suite qui converge vers un point de minimum.*

1. Même si cette procédure n'est pas toujours respectée en pratique...

Preuve : Soit $(u_k)_k$ une suite minimisante.

- $(u_k)_k$ est bornée : en effet, supposons qu'il existe une sous-suite extraite, $(u_{k'})_{k'}$, telle que $\|u_{k'}\| \to +\infty$; comme J est infinie à l'infini, on infère que $J(u_{k'}) \to +\infty$, ce qui contredit le fait que (u_k) est une suite minimisante (en particulier, $\lim_k J(u_k) < +\infty$.)

- Comme $(u_k)_k$ est bornée, on peut en extraire une sous-suite, toujours notée $(u_{k'})_{k'}$, qui converge vers un point u. C'est une suite d'éléments de K qui est fermé, donc $u \in K$.

- Par ailleurs, comme J est continue, $\lim_{k'} J(u_{k'}) = J(u)$. Enfin, la sous-suite $(u_{k'})_{k'}$ reste minimisante : $J(u) = \inf_{v\in K} J(v)$, et u est un minimum global de J sur K.

■

♠ Lorsque la dimension de $\mathbb{E}$ est *infinie*, la proposition précédente est *fausse*[2] ! On peut en effet construire des contrexemples, lorsque la dimension est infinie.
♠ Il est également indispensable que l'ensemble K soit *fermé*. Si on considère par exemple la fonction $x \mapsto x^2$ sur $K = \mathbb{R}_*^+$, on a bien une fonction continue, et infinie à l'infini, définie sur K non vide, mais pas fermé... Elle n'admet pas de point de minimum sur K.

Remarque 1.2.2 *Dans l'énoncé de la proposition ci-dessus, on se convainc facilement que l'on peut remplacer l'assertion J infinie à l'infini par J infinie à l'infini sur K, c'est-à-dire*

$$\lim_{v\in K, \|v\|\to+\infty} J(v) = +\infty.$$

1.3 Exemple de résolution d'un problème de minimisation

Dans cette section, nous allons résoudre complètement un problème de minimisation "classique". Qui plus est, nous en tirerons des enseignements généraux, enseignements qui nous permettront de développer la théorie liée à l'optimisation de fonctionnelles *différentiables* et/ou *convexes*.
On considère le problème de la minimisation d'une fonctionnelle quadratique qui dépend de n variables ; soit donc un polynôme de degré 2 en $x_1, \cdots, x_n$, c'est-à-dire :

$$P(x_1, \cdots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{i} \alpha_{ij}\, x_i\, x_j - \sum_{k=1}^{n} \beta_j\, x_j + \gamma. \tag{1.1}$$

On va étudier l'existence de *minima locaux ou globaux*, lorsque $(x_1, \cdots, x_n)$ parcourt $\mathbb{R}^n$.
Tout d'abord, essayons de nous servir des résultats généraux de la section précédente. Le théorème 1.2.1 ne s'applique pas, puisque $\mathbb{R}^n$ n'est pas compact. Quant à la proposition 1.2.1, il est difficile de savoir si on peut l'utiliser, car vérifier que P est bien infini à l'infini n'est pas aisé ! Bref, nous allons développer de nouveaux outils, (mieux) adaptés au problème de la minimisation de (1.1).

2. Nous allons expliquer pourquoi la démonstration de la proposition ne s'applique pas dans un espace de dimension infinie. Pour cela, rappelons un théorème dû à Riesz.

Théorème 1.2.2 *Soit $\mathbb{E}$ un espace vectoriel normé et $B(0,1) = \{v \in \mathbb{E} \,:\, \|v\| \leq 1\}$ sa boule unité fermée. Alors, $\mathbb{E}$ est de dimension finie si, et seulement si, $B(0,1)$ est compacte.*

A partir de ce résultat, on voit qu'il ne sert à rien de se ramener à une suite bornée, si l'on reprend la démonstration dans le cas de la dimension infinie. En effet, les éléments de la suite appartiennent bien à une boule fermée et bornée, mais celle-ci n'est plus compacte. On ne peut alors plus considérer une sous-suite qui converge...

1.3.1 Dans $\mathbb{R}$

Considérons brièvement le cas d'un polynôme d'une seule variable. Bien évidemment, si $n = 1$, $P(x) = \alpha\, x^2 - \beta\, x + \gamma$.
Si x_0 est un minimum local de P, il existe $\eta > 0$ tel que, pour tout h vérifiant $|h| < \eta$, on ait $P(x_0 + h) \geq P(x_0)$. Par différence, on obtient $h(2\alpha\, x_0 + \alpha\, h - \beta) \geq 0$.
Si on choisit h dans $]0, \eta[$, on a alors $2\alpha\, x_0 + \alpha\, h - \beta \geq 0$; on fait tendre h vers 0, pour arriver à $2\alpha\, x_0 - \beta \geq 0$.
En prenant h négatif, on obtient cette fois $2\alpha\, x_0 - \beta \leq 0$.
Ainsi, une condition *nécessaire* d'existence de minimum est que

$$2\alpha\, x_0 = \beta. \tag{1.2}$$

Réciproquement, si x_0 est tel que $2\alpha\, x_0 = \beta$, on trouve $P(x_0 + h) = P(x_0) + \alpha\, h^2$. Pour garantir l'existence d'un minimum (qui sera d'ailleurs global), α doit être positif ou nul. Notons enfin que pour que (1.2) possède une solution, il faut soit que $\alpha \neq 0$, soit que $\alpha = \beta = 0$. Dans le premier cas, il existe une solution et une seule, et dans le second cas, x_0 est quelconque.
En conclusion, nous sommes arrivés au résultat suivant :

(I) $\alpha < 0$: la condition *nécessaire* d'existence de minimum (1.2) n'est jamais *suffisante*. Il n'existe pas de minimum.

(II) $\alpha \geq 0$: la condition d'existence de minimum (1.2) est *nécessaire et suffisante*. Qui plus est, la résolution de (1.2) permet de *caractériser* les minima, qui sont automatiquement globaux.

Si $\alpha > 0$, il existe un minimum x_0 unique, égal à $x_0 = \beta/2\,\alpha$.

Si $\alpha = 0$ et $\beta = 0$, tout élément de $\mathbb{R}$ réalise le minimum.

Si $\alpha = 0$ et $\beta \neq 0$, il n'existe pas de minimum.

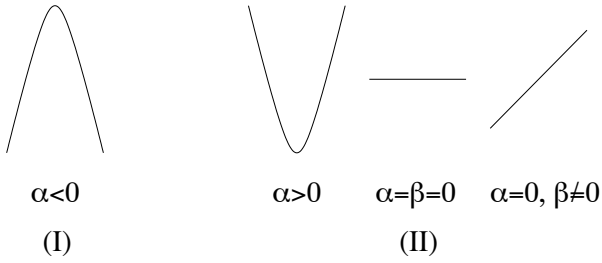


FIG. 1.1 – *Petit récapitulatif 'visuel'.*

1.3.2 Dans $\mathbb{R}^n$

Ci-dessous, nous allons faire usage des prérequis d'algèbre linéaire rappelés dans l'Avant-Propos.
Dans le cas général ($n \geq 2$), on peut se ramener à une forme "condensée", qui permet de *simplifier* l'étude que l'on se propose de réaliser, en la rapprochant du cas à une variable.

En effet, si on note : $v = (x_1, \cdots, x_n)^{\mathsf{T}}$ et $b = (\beta_1, \cdots, \beta_n)^{\mathsf{T}}$, on a $\sum_{k=1}^{n} \beta_j\, x_j = (b, v)$, où $(.,.)$ est le produit scalaire usuel de $\mathbb{R}^n$.

Qu'en est-il pour le terme quadratique de P ? Soit $A = (A_{i,j})_{1\leq i,j\leq n}$ une matrice de $\mathbb{R}^{n\times n}$; comparons $\frac{1}{2}(Av, v)$ au premier terme de P :

$$\frac{1}{2}(Av,v) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}x_j\, x_i = \frac{1}{2}\sum_{i=1}^{n} A_{i,i}x_i\, x_i + \sum_{i=1}^{n}\sum_{j<i}\frac{1}{2}\{A_{i,j}+A_{j,i}\}x_i\, x_j.$$

En identifiant les coefficients terme à terme, on arrive à :

$$\begin{cases} A_{i,i} = 2\alpha_{ii}, \quad 1\leq i\leq n. \\ A_{i,j}+A_{j,i} = 2\alpha_{ij}, \quad 1\leq i\leq n,\ 1\leq j<i. \end{cases}$$

Il y a plus d'inconnues que d'équations. Ceci étant, si l'on suppose que A est *symétrique*, on peut déterminer A, puisqu'on obtient $A_{i,i} = 2\alpha_{ii}$, pour $1\leq i\leq n$, et $A_{i,j} = \alpha_{ij}$, pour $1\leq i\leq n$, $1\leq j<i$. En résumé, on vient de démontrer le résultat élémentaire suivant :

Proposition 1.3.1 *A tout polynôme P de n variables et de degré 2, on peut associer un unique triplet (A, b, c), où A est une matrice symétrique de $\mathbb{R}^{n\times n}$, b un vecteur de $\mathbb{R}^n$, et c un réel, tel que*

$$\forall v = (x_1,\cdots,x_n)^{\mathsf{T}}\in\mathbb{R}^n, \quad P(x_1,\cdots,x_n) = \frac{1}{2}(Av,v) - (b,v) + c.$$

Pour A symétrique, on introduit la fonctionnelle $J_0 : \mathbb{R}^n \to \mathbb{R}$, définie par

$$\forall v\in\mathbb{R}^n, \quad J_0(v) = \frac{1}{2}(Av,v) - (b,v) + c. \tag{1.3}$$

On s'intéresse aux problèmes de minimisation suivants :

$$\text{Problème } \textit{local} \qquad \begin{cases} \text{Trouver } u\in\mathbb{R}^n, \text{ solution de} \\ \exists\eta>0, \quad \forall h\in\mathbb{R}^n, \quad \|h\|<\eta \Longrightarrow J_0(u)\leq J_0(u+h). \end{cases} \tag{1.4}$$

$$\text{Problème } \textit{global} \qquad \begin{cases} \text{Trouver } u\in\mathbb{R}^n, \text{ solution de} \\ \forall h\in\mathbb{R}^n, \quad J_0(u)\leq J_0(u+h). \end{cases} \tag{1.5}$$

Considérons tout d'abord le problème (1.4).

• Si $u\in\mathbb{R}^n$ est un minimum local, il existe $\eta>0$ tel que, pour tout h de norme plus petite que η, $J_0(u+h)\geq J_0(u)$.
Par différence, on obtient :

$$\begin{aligned} J_0(u+h) - J_0(u) &= \frac{1}{2}(A\{u+h\},u+h) - (b,u+h) + c - \frac{1}{2}(Au,u) + (b,u) - c \\ &= \frac{1}{2}(Au,h) + \frac{1}{2}(Ah,u) + \frac{1}{2}(Ah,h) - (b,h) \\ &= (Au-b,h) + \frac{1}{2}(Ah,h), \end{aligned} \tag{1.6}$$

Comme dans le cas monodimensionnel, considérons maintenant des *petites variations*. Pour tout vecteur non nul d de $\mathbb{R}^n$, $\lambda\, d$ est de norme plus petite que η dès lors que $|\lambda| < \eta/\|d\|$. On arrive alors à

$$\forall\lambda \text{ tel que } |\lambda|<\eta/\|d\|, \quad \lambda\{(Au-b,d) + \frac{\lambda}{2}(Ad,d)\}\geq 0.$$

En faisant tendre λ vers zéro par valeurs supérieures ($\lambda>0$), on en déduit qu'une condition *nécessaire* d'existence d'un minimum est

$$\forall d\in\mathbb{R}^n, \quad (Au-b,d)\geq 0, \tag{1.7}$$

ou, de façon équivalente, puisque d parcourt l'ensemble des directions possibles dans $\mathbb{R}^n$,

$$Au = b. \tag{1.8}$$

• Comme dans le cas monodimensionnel, examinons la *réciproque* :

si (1.8) est vérifiée, on a l'égalité $J_0(u+h) = J_0(u) + \frac{1}{2}(Ah, h)$, pour tout h. En conséquence, pour que u soit bien un minimum, A doit être **positive**, c'est-à-dire que

$$\forall h \in \mathbb{R}^n, \quad (Ah, h) \geq 0.$$

Dans ce cas, le minimum est *global*. Si A n'est pas positive, il n'existe pas de minimum.
Bien évidemment, pour que (1.8) possède une solution, il faut et il suffit que b appartienne à l'image de A, notée $Im\, A$. Comme A est symétrique, ceci équivaut à ce que b soit orthogonal à $Ker\, A$. En effet,

Lemme 1.3.1 *Soit A une matrice de $\mathbb{R}^{m\times n}$, alors $Im\, A = (Ker\, A^{\mathsf{T}})^{\perp}$.*

Preuve : Prouvons pour commencer que $Im\, A \subset (Ker\, A^{\mathsf{T}})^{\perp}$. Soit donc x un élément de $Im\, A$; il existe $v \in \mathbb{R}^n$ tel que $x = Av$. Alors, pour tout élément y appartenant à $Ker\, A^{\mathsf{T}}$, on a

$$(x, y)_m = (Av, y)_m = (v, A^{\mathsf{T}} y)_n = 0.$$

Pour prouver l'égalité entre ces deux sous-espaces vectoriels de $\mathbb{R}^m$, vérifions qu'ils ont même dimension. D'une part, puisque $Ker\, A$ et $(Ker\, A^{\mathsf{T}})^{\perp}$ sont supplémentaires,

$$m = dim[Ker\, A^{\mathsf{T}}] + dim[(Ker\, A^{\mathsf{T}})^{\perp}].$$

Et, d'autre part, comme $A^{\mathsf{T}} : \mathbb{R}^m \to \mathbb{R}^n$, d'après le théorème du rang (rg(A)=rg(A^{T})), on trouve

$$m = dim[Ker\, A^{\mathsf{T}}] + dim[Im\, A^{\mathsf{T}}] = dim[Ker\, A^{\mathsf{T}}] + dim[Im\, A].$$

On a bien l'égalité entre les dimensions, $dim[(Ker\, A^{\mathsf{T}})^{\perp}] = dim[Im\, A]$, ce qui permet d'arriver à l'égalité annoncée. ∎

Résumons : lorsque A est positive, on a deux possibilités, selon que A est inversible ou non. Si A est inversible, il existe une unique solution à (1.8). Si A n'est pas inversible, on sait que l'ensemble des solutions de (1.8) est égal à l'espace affine $u_0 + Ker\, A$, où u_0 est une solution particulière de l'équation.
Avant de conclure, relions l'inversibilité de A au fait qu'elle est **définie positive**, i. e.

$$\forall h \in \mathbb{R}^n \setminus \{0\}, \quad (Ah, h) > 0.$$

Proposition 1.3.2 *Soit A une matrice symétrique et positive. Alors A est inversible si et seulement elle est définie positive.*

Preuve : Supposons que A est inversible. Soit w tel que $(Aw, w) = 0$. On va montrer que w est en fait égal à zéro. On va se servir encore une fois de petites variations autour de w, selon une direction d de $\mathbb{R}^n$ ($d \neq 0$) : soit donc enfin $\lambda > 0$. Comme A est positive et symétrique :

$$0 \leq (A\{w + \lambda\, d\}, w + \lambda\, d) = 2\lambda(Aw, d) + \lambda^2(Ad, d).$$

En mettant λ en facteur, puis en faisant tendre λ vers 0 dans le facteur restant, on obtient $(Aw, d) \geq 0$. Comme c'est valable pour toute direction, on en déduit que $Aw = 0$, ce qui conduit enfin à $w = 0$ par hypothèse sur A.
La réciproque est aisée et classique. En effet, si A est définie-positive, $Aw = 0$ entraîne que $(Aw, w) = 0$, et donc que $w = 0$: par conséquent, A est inversible. ∎

Dans $\mathbb{R}^n$, nous avons résolu les problèmes (1.4) et (1.5) :

(III) A n'est pas positive : la condition *nécessaire* d'existence de minimum (1.8) n'est jamais *suffisante*. Il n'existe pas de minimum.

(IV) A est positive : la condition d'existence de minimum (1.8) est *nécessaire et suffisante*. Qui plus est, la résolution de (1.8) permet de *caractériser* les minima, qui sont automatiquement globaux.

Si A est définie-positive, il existe un minimum u unique, égal à $u = A^{-1}b$.

Sinon,

si $b \perp Ker\, A$, l'espace des minima est égal à $u_0 + Ker\, A$.

Si $b \not\perp Ker\, A$, il n'existe pas de minimum.

Exercice 1.3.1 *1. Essayer de résoudre directement le problème de la minimisation du polynôme P défini par (1.1).*
2. Résoudre les problèmes de minimisation avec une fonctionnelle J_0 définie à partir d'une matrice A quelconque (voir (1.3)).
3. Vérifier directement que si A n'est pas positive, il n'existe pas de minimum global.
4. Montrer que A est définie-positive si et seulement si il existe $\nu > 0$ telle que, pour tout vecteur v, $(Av, v) \geq \nu\|v\|^2$.
5. En économie, on maximise le profit : résoudre les problèmes de maximisation quadratiques associés.

La façon dont nous avons résolu le problème posé est riche d'enseignements :

Il est utile de **condenser** les notations, en passant de (1.1) à (1.3). Les sceptiques sont invités à examiner la question 1 de l'exercice 1.3.1 ! Ceci reste vrai lorsque l'on est confronté à un calcul de différentielle.

Or, la condition **nécessaire** d'existence d'un minimum est obtenue à l'aide de calculs de petites variations, autour d'un point de minimum. Ceci nous conduira naturellement, dans la section suivante, à utiliser les notions indispensables de **calcul différentiel** dans des espaces vectoriels normés, et plus particulièrement dans $\mathbb{R}^n$, en vue de résoudre des problèmes d'**optimisation**.

Le fait que la condition d'existence d'un minimum est **suffisante** découle de la **positivité** de A, qui est elle-même équivalente à la **convexité** de la fonctionnelle J_0, comme on le verra section 2.3, qui traite en particulier d'**optimisation convexe**. On prouvera aussi que la condition qui garantit l'existence (et l'unicité) du minimum est l'α**-convexité** de J_0.

Enfin, lorsque A est symétrique définie-positive, ces résultats théoriques, qui lient minimisation (cf. (1.5)) et résolution du système linéaire en A (1.8), ont des conséquences pratiques importantes. En effet, ils sont à la base de nombreux algorithmes de calcul numérique qui seront étudiés dans la suite du cours.

1.4 Par la suite

A partir de l'exemple précédent, nous allons définir les outils mathématiques adaptés à des problèmes plus généraux.
Dans l'Annexe, au chapitre 14, nous rappelons les notions élémentaires, ainsi que les théorèmes fondamentaux, associés à la différentiabilité d'une fonctionnelle définie sur un espace vectoriel normé, à valeurs dans un espace vectoriel normé.
Nous abordons les problèmes de minimisation proprement dits au chapitre 2 : la fonctionnelle est à valeurs réelles. Pour ce qui est de la caractérisation d'un minimum, nous commençons

par les conditions nécessaires d'existence. Nous nous attachons en particulier à l'étude de problèmes posés non pas sur l'espace entier, mais plutôt sur une partie de celui-ci ; on parle alors de problème de minimisation avec contraintes. Dans un second temps, nous déterminons des conditions suffisantes d'existence, qui sont liées à la convexité de la fonctionnelle étudiée.
Dans le chapitre suivant, nous étudions un problème classique de minimisation, appelé moindres carrés linéaires, qui peut être perçu comme une généralisation de la résolution d'un système linéaire tel que (1.8). En effet, pour A une matrice de $\mathbb{R}^{m\times n}$ et b un vecteur de $\mathbb{R}^m$, on considère la minimisation de la fonctionnelle

$$f \ : \ \mathbb{R}^n \to \mathbb{R}, \quad f(v) = \|Av - b\|_m.$$

Le point fondamental est, qu'en général, on ne peut pas déterminer/calculer u tel que $f(u) = 0$, c'est-à-dire qu'il n'existe pas de solution au problème $Au = b$. Malgré tout, nous vérifierons qu'il existe toujours des points de minimum.
Enfin, dans le dernier chapitre, nous construirons des algorithmes permettant de calculer numériquement une approximation du minimum. Deux points-clefs sont à noter dès à présent au sujet de ces algorithmes :

Ils sont basés sur les caractérisations obtenues dans les chapitres théoriques.

Ils sont itératifs : à partir d'une initialisation u^0, on calcule u^1, puis u^2, etc. jusqu'à arriver à une solution numérique correcte.

Dans la suite du cours (cf. la Partie 2), nous revenons en détail sur la résolution de systèmes linéaires comme (1.8), en définissant des algorithmes avancés de résolution. Nous ne nous limitons pas au cas de matrices symétriques ; en effet, nous considérons des matrices quelconques, sous réserve qu'elles soient inversibles.

Bien sûr, la recherche en optimisation est très dynamique, et la théorie en constante évolution. Aussi, les résultats présentés ci-après ne représentent qu'une petite introduction à l'art de l'optimisation. Des généralisations et approfondissements seront proposés lors d'autres cours de l'ENSTA, en deuxième (cf. [10]) et troisième années (Filière optimisation et recherche opérationnelle). Nous renvoyons également le lecteur aux ouvrages [6, 2], qui proposent de nombreuses extensions, tout en restant tout à fait abordable pour le (futur) ingénieur...

Chapitre 2

Minimisation

2.1 Introduction

Dans ce chapitre on considère les fonctionnelles dérivables au sens de Gateaux (cf. le chapitre 14), sauf mention explicite du contraire. Bien évidemment, comme on parle de minimisation, l'espace d'arrivée $\mathbb{F}$ sera égal à $\mathbb{R}$. Dans la première partie, nous allons traiter de **conditions nécessaires** d'existence d'un minimum. Dans la seconde, nous étudierons les **conditions suffisantes**. Soit donc $\mathbb{E} = \mathbb{R}^n$.
Sauf mention explicite du contraire, les résultats, définitions et notations de ce chapitre sont valables lorsque $\mathbb{E}$ est un espace vectoriel normé complet de dimension infinie, muni d'un produit scalaire. On dit alors que $\mathbb{E}$ est un espace de Hilbert (voir par exemple [3, 15]).

Nous commençons par introduire la notion de **chemin**, et rappeler celle des **tangentes**.

Définition 2.1.1 *On appelle chemin réel une fonction dérivable $\gamma : t \mapsto \gamma(t)$ de $\mathbb{R}$ dans $\mathbb{E}$. On appelle tangente au chemin en t_0 la droite passant par $\gamma(t_0)$ et de direction $\gamma'(t_0)$. On appelle chemin une fonction de $[0, \alpha[$ à valeurs dans $\mathbb{E}$, dérivable sur $]0, \alpha[$ et dérivable à droite en 0, avec $\alpha > 0$. Dans ce cas, la tangente en 0 est une demi-droite, passant par $\gamma(0)$, et de direction $\gamma'_d(0)$, c'est-à-dire : $\{w \in \mathbb{E} \ : \ \exists \eta \geq 0, \ w = \gamma(0) + \eta\, \gamma'_d(0)\}$.*

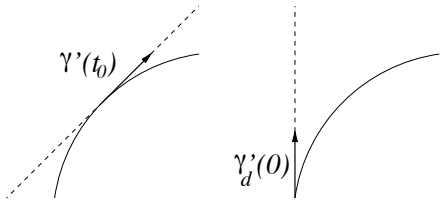


FIG. 2.1 – *Tangentes*

On rappelle que, lorsque la variable est réelle,

$$\gamma'(t_0) = \lim_{\theta \to 0} \frac{\gamma(t_0 + \theta) - \gamma(t_0)}{\theta}, \text{ et } \gamma'_d(t_0) = \lim_{\theta \to 0^+} \frac{\gamma(t_0 + \theta) - \gamma(t_0)}{\theta}.$$

Proposition 2.1.1 *Soit γ un chemin. On a, pour $t_0 \in]0, \alpha[$, avec $\alpha > 0$*

$$d\gamma(t_0) \cdot h = h\, \gamma'(t_0), \qquad \forall h \in \mathbb{R}.$$

Preuve : Il suffit de comparer la définition 14.1.1 à celle de la dérivée usuelle

$$\gamma(t_0 + h) = \gamma(t_0) + h\gamma'(t_0) + o(h), \qquad \forall h \in]-t_0, \alpha - t_0[.$$

On en déduit donc que l'on a l'égalité $d\gamma(t_0)\cdot h = h\,\gamma'(t_0)$, pour h suffisamment petit. Comme $d\gamma(t_0)$ est une application linéaire, l'égalité précédente est vraie pour tout h de $\mathbb{R}$. ∎

Remarque 2.1.1 *La dérivée à droite peut être vue comme une dérivée directionnelle (cf. définition 14.1.3) dans le sens positif. En effet,*

$$\gamma'_d(t_0) = \lim_{\theta\to 0^+} \frac{\gamma(t_0+\theta\,(+1)) - \gamma(t_0)}{\theta} = d\gamma(t_0)\cdot(+1).$$

Soit maintenant f une application Fréchet-différentiable de $\mathbb{E}$ dans $\mathbb{F}$ (cf. le chapitre 14). On construit $\mu = f\circ\gamma$, une fonction de la variable réelle à valeurs dans $\mathbb{F}$. Tous les résultats connus s'appliquent sur une telle fonction (théorème des accroissements finis, formules de Taylor, etc.). μ est dérivable, comme composée d'applications différentiables, et on a

$$\begin{aligned} & d\mu(t_0)\cdot h = df(\gamma(t_0))\cdot(d\gamma(t_0)\cdot h), \qquad \forall h\in\mathbb{R} \\ \Longleftrightarrow\ & h\,\mu'(t_0) = df(\gamma(t_0))\cdot(h\,\gamma'(t_0)), \qquad \forall h\in\mathbb{R} \\ \Longleftrightarrow\ & h\,\mu'(t_0) = h\,df(\gamma(t_0))\cdot\gamma'(t_0), \qquad \forall h\in\mathbb{R}, \end{aligned}$$

puisque $df(\gamma(t_0))$ est linéaire, soit finalement

$$\mu'(t_0) = df(\gamma(t_0))\cdot\gamma'(t_0). \tag{2.1}$$

Si J est une fonctionnelle Fréchet-différentiable de $\mathbb{E}=\mathbb{R}^n$ dans $\mathbb{F}=\mathbb{R}$, si on pose $\mu = J\circ\gamma$, on infère que (cf. (14.5))

$$\mu'(t_0) = (\nabla J(\gamma(t_0)), \gamma'(t_0)), \tag{2.2}$$

et si enfin $\gamma(t) = u + t\,w$, avec $u, w\in\mathbb{R}^n$, on a $\gamma'(t_0) = w$, ce qui donne

$$\mu'(t_0) = (\nabla J(u + t_0\,w), w). \tag{2.3}$$

Proposition 2.1.2 *Si J est de classe $\mathcal{C}^2$, on a*

$$\mu''(t_0) = (\nabla^2 J(u + t_0\,w)w, w). \tag{2.4}$$

Preuve : On écrit

$$\begin{aligned} \mu'(t_0+h) - \mu'(t_0) &= (\nabla J(u + t_0\,w + h\,w) - \nabla J(u + t_0\,w), w) \\ &= (h\nabla^2 J(u + t_0\,w)w + \|h\,w\|\varepsilon(h\,w), w) = h(\nabla^2 J(u + t_0\,w)w, w) + o(h), \end{aligned}$$

puisque ∇J est de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$, cf. (14.6). D'où finalement

$$\mu''(t_0) = \lim_{h\to 0} \frac{\mu'(t_0+h) - \mu'(t_0)}{h} = (\nabla^2 J(u + t_0\,w)w, w).$$

∎

Notons que si J est simplement Gateaux-différentiable, on se limite aux chemins inclus dans des droites, c'est-à-dire de la forme $\gamma(t) = u + t\,w$.

2.2 Conditions nécessaires

On considère K un sous-ensemble non vide de $\mathbb{E}$, et J une fonctionnelle de K dans $\mathbb{R}$. Dans cette section, nous allons déterminer des conditions nécessaires d'existence d'un point de minimum de J sur K. On va commencer par le cas général, poursuivre par celui d'un sous-ensemble convexe, et finir par le cas d'un sous-espace affine.

2.2.1 Cas général

Dans le cas le plus général, il est pratique d'introduire le **cône des directions admissibles**.

Définition 2.2.1 *Soit K un sous-ensemble non vide de $\mathbb{E}$, et v un point de K. On appelle cône des directions admissibles l'ensemble $K(v)$ des tangentes en v aux chemins inclus dans K et commençant en v.*

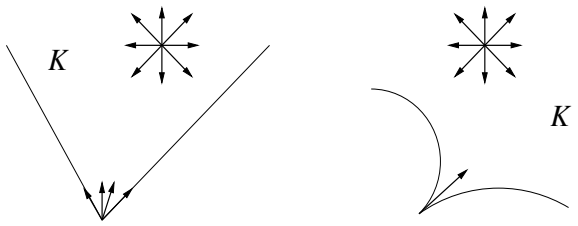


FIG. 2.2 – *Exemples de cônes de directions admissibles*

En d'autres termes, w appartient à $K(v)$, si, et seulement si, il existe un chemin $\gamma : [0, \alpha[\to K$ ($\alpha > 0$) tel que $\gamma(0) = v$ et $w = \gamma'_d(0)$. Dans ce cas, on a

Théorème 2.2.1 *Soient K un sous-ensemble non vide de $\mathbb{E}$, u un point de K et J une fonctionnelle de K à valeurs dans $\mathbb{R}$. On suppose que J est Fréchet-différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$(\nabla J(u), w) \geq 0, \qquad \forall w \in K(u). \tag{2.5}$$

Preuve : Si $w = 0$, l'inégalité est une égalité...
Si w est un élément (non nul) de $K(u)$, il existe un chemin $\gamma : [0, \alpha[\to K$ ($\alpha > 0$) passant par u en $t = 0$ tel que $w = \gamma'_d(0)$. Pour t suffisamment petit, $\gamma(t)$ est proche de u. Ainsi,

$$J \circ \gamma(t) \geq J \circ \gamma(0), \qquad t \in [0, t_0[.$$

Or, $J \circ \gamma$ est dérivable à droite en 0, puisque ceci correspond à la Gateaux-différentiabilité. En effet, on peut écrire γ sous la forme $\gamma(t) = u + t\,w + o(t)$, pour $t \in [0, \alpha[$. On a alors

$$J \circ \gamma(t) = J(u + t\,w + o(t)) = J(u + h), \text{ avec } h = t\,w + o(t).$$

Or, $\|h\|$ tend vers 0 lorsque t tend vers 0^+, puisque $\|h\| \leq 2t\|w\|$, pour t suffisamment petit. Par application de (14.6), on trouve alors

$$J \circ \gamma(t) = J(u) + (\nabla J(u), t\,w + o(t)) + o(t) = J(u) + t(\nabla J(u), w) + o(t).$$

D'où le passage à la limite

$$\frac{J \circ \gamma(t) - J \circ \gamma(0)}{t} = (\nabla J(u), w) + \frac{o(t)}{t} \to (\nabla J(u), w).$$

L'inégalité initiale implique que $(\nabla J(u), w) \geq 0$. ∎

2.2.2 Cas convexe

Définition 2.2.2 *On dit qu'un sous-ensemble K de $\mathbb{E}$ est* **convexe** *si, et seulement si, pour tout couple d'éléments (u, v), le segment $[u, v]$ est inclus dans K : $\forall u, v \in K$, $\forall t \in [0, 1]$, $u + t(v - u) = (1 - t)u + t\,v \in K$.*

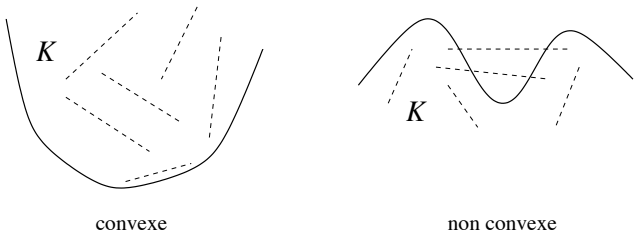


FIG. 2.3 – *Convexe ou non convexe*

Lorsque l'on se place dans un ensemble convexe, on obtient la première condition nécessaire d'existence d'un minimum local, dite **inéquation d'Euler**.

Théorème 2.2.2 *Soient K un sous-ensemble convexe non vide de $\mathbb{E}$, u un point de K et J une fonctionnelle de K à valeurs dans $\mathbb{R}$. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$(\nabla J(u), v-u) \geq 0, \qquad \forall v \in K. \tag{2.6}$$

Preuve : Soit $v \in K$.
Par définition de la convexité, on sait que $u + t(v-u)$ appartient à K, pour $t \in [0,1]$.
Puisque u est un minimum local, il existe $t_0 > 0$ tel que $J(u+t(v-u)) \geq J(u)$, pour tout t dans $]0, t_0[$.
Si l'on remplace $J(u+t(v-u)) - J(u)$ par la valeur donnée par la formule (14.4), on en déduit

$$(\nabla J(u), t(v-u)) + o(t) \geq 0, \qquad \forall t \in]0, t_0[.$$

Suivant la méthodologie des petites variations, on peut mettre t en facteur, ce qui laisse

$$(\nabla J(u), v-u) + \frac{o(t)}{t} \geq 0, \qquad \forall t \in]0, t_0[.$$

En faisant tendre t vers 0, on infère le résultat annoncé. ∎

Remarque 2.2.1 *Si J est* Fréchet-différentiable, *on note que la définition de la convexité permet d'affirmer que $\gamma : [0,1[\to K$ défini par $\gamma(t) = u + t(v-u)$ est un chemin inclus dans K, tel que $\gamma(0) = u$ et $\gamma'_d(0) = v - u$. Le théorème 2.2.1 permet de conclure directement.*

A partir de ce résultat, très simple à démontrer, on arrive à l'**équation d'Euler**.

Corollaire 2.2.1 *Si $K = \mathbb{E}$ (ou si u est intérieur à K), l'inéquation (2.6) devient*

$$\nabla J(u) = 0. \tag{2.7}$$

Preuve : (i) $K = \mathbb{E}$: il suffit de remarquer que (2.6) est valable pour tout v.
On a tout d'abord $(\nabla J(u), w) \geq 0$, $\forall w \in \mathbb{E}$.
En prenant $-w$, on arrive à $(\nabla J(u), w) = 0$, $\forall w \in \mathbb{E}$, et la conclusion suit.
(ii) $u \in \overset{\circ}{K}$: par définition (cf. [3]), il existe une boule ouverte $B(u,\theta)$, $\theta > 0$, telle que $B(u,\theta)$ est incluse dans K. Or, $B(u,\theta)$ est *convexe* et contient *toutes* les directions w issues de u. En effet, pour $w \neq 0$, $u + t\,w$ appartient à $B(u,\theta)$ dès que $\|u + t\,w\| = |t|\,\|w\| < \theta$. En particulier, $v = u + \dfrac{\theta}{2\|w\|} w$ appartient toujours à $B(u,\theta) \subset K$, et on peut appliquer (2.6) avec cet élément pour retrouver $(\nabla J(u), w) \geq 0$. On conclut comme au (i). ∎

Exercice 2.2.1 *Dans le cas d'une fonction dérivable d'un sous-ensemble de $\mathbb{R}$ dans $\mathbb{R}$, que signifie la distinction entre le théorème et son corollaire ?*

Remarque 2.2.2 *Si $\mathbb{R}^n$ est muni d'une base orthonormale, la condition nécessaire d'existence d'un minimum, dans les conditions du corollaire, peut être écrite sous la forme*

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) = \ldots = \frac{\partial J}{\partial x_n}(u) = 0.$$

Corollaire 2.2.2 *Supposons que les hypothèses du corollaire ci-dessus sont vérifiées, et que J est deux fois*[1] *différentiable en u. Alors, si u est un minimum local de J, on a*

$$(\nabla^2 J(u)h, h) \geq 0, \qquad \forall h \in \mathbb{E}. \tag{2.8}$$

Preuve : Comme J est deux fois différentiable en u, on peut écrire le développement limité de J en u à l'ordre 2. Pour h un élément de $\mathbb{E}$ et $\lambda > 0$ petit, on a, cf. (14.10) :

$$J(u + \lambda h) = J(u) + \lambda (\nabla J(u), h) + \frac{\lambda^2}{2}(\nabla^2 J(u)h, h) + r_2(\lambda h).$$

D'après le corollaire précédent, $\nabla J(u) = 0$. Bien sûr, u est un minimum local de J :

$$\frac{\lambda^2}{2}(\nabla^2 J(u)h, h) + r_2(\lambda h) \geq 0, \text{ soit } (\nabla^2 J(u)h, h) + \frac{2}{\lambda^2} r_2(\lambda h) \geq 0.$$

Faisons tendre λ vers 0^+. Le premier terme est indépendant de λ, et $\lim_{\lambda \to 0^+} \frac{2}{\lambda^2} r_2(\lambda h) = 0$, d'après le théorème de Taylor-Young 14.3.5. Ainsi, on trouve bien $(\nabla^2 J(u)h, h) \geq 0$. ∎

2.2.3 Contraintes d'égalité affines

Supposons, ce qui est un cas relativement courant en pratique, que la fonctionnelle J soit définie sur l'espace *affine*

$$K^\star = u_0 + K,$$

où u_0 et K sont respectivement un vecteur et un sous-espace vectoriel de $\mathbb{E}$. En d'autres termes, $v^\star \in K^\star$ si, et seulement si, $\exists v \in K$ tel que $v^\star = u_0 + v$. Alors, d'après le théorème 2.2.2, puisque $K^\star$ *est un convexe non vide*, si $u^\star$ est un point de minimum local de J sur $K^\star$, et si J est différentiable en $u^\star$, on a nécessairement

$$(\nabla J(u^\star), v^\star - u^\star) \geq 0, \quad \forall v^\star \in K^\star.$$

Or, on a $u^\star = u_0 + u$ et $v^\star = u_0 + v$, où u et v sont deux éléments de K : $v^\star - u^\star = v - u \in K$. Qui plus est, l'ensemble $\{w \in \mathbb{E} \ : \ \exists v \in K, \ w = v - u\}$ est égal à K (K est invariant par translation de vecteur un de ses éléments.) Enfin, puisque $-w$ est un élément de K dès lors que w en est un, on en déduit qu'une condition équivalente à l'inégalité ci-dessus est

$$\nabla J(u^\star) \in K^\perp.$$

Soit $(a_1, \cdots, a_p)$ une base[2] de $K^\perp$; ceci revient à affirmer que

$$\exists (\lambda_1, \cdots, \lambda_p) \in \mathbb{R}^p, \quad \nabla J(u^\star) + \sum_{k=1}^{p} \lambda_k a_k = 0.$$

Par ailleurs, la différence $u^\star - u_0$ est orthogonale aux vecteurs $(a_1, \cdots, a_p)$, ce que l'on exprime sous la forme

$$(u^\star - u_0, a_k) = 0, \ 1 \leq k \leq p.$$

1. on considère donc une fonctionnelle Fréchet-différentiable.
2. Si $\mathbb{E}$ est un espace de Hilbert, on suppose ici que K est un sous-espace vectoriel de codimension finie.

On a donc démontré le résultat suivant, dit de Karush, Kuhn et Tucker (K.K.T.).

Théorème 2.2.3 *Soit $(a_1, \cdots, a_p)$ une famille libre de $\mathbb{E}$, K le sous-espace vectoriel tel que $K^\perp = Vect(a_1, \cdots, a_p)$, et u_0 un vecteur de $\mathbb{E}$. Soient $K^\star$ l'espace affine $u_0 + K$, $u^\star$ un point de $K^\star$ et J une fonctionnelle de $K^\star$ à valeurs dans $\mathbb{R}$. On suppose que J est différentiable en $u^\star$. Si $u^\star$ est un point de minimum local de J sur $K^\star$, on a nécessairement*

$$\begin{cases} \exists(\lambda_1, \cdots, \lambda_p) \in \mathbb{R}^p, \quad \nabla J(u^\star) + \displaystyle\sum_{k=1}^{p} \lambda_k a_k = 0 \\ (u^\star - u_0, a_k) = 0, \ 1 \leq k \leq p \end{cases} . \tag{2.9}$$

Une reformulation pour conclure. On a parlé de contraintes affines dans le titre de cette sous-section. Supposons explicitement que $\mathbb{E} = \mathbb{R}^n$; dans ce cas, la seconde suite d'équations de (2.9) est *affine* en les composantes de $u^\star$. En effet, si l'on note C la matrice de $\mathbb{R}^{p\times n}$ telle que $C_{i,j} = (a_i)_j$ et f le vecteur de $\mathbb{R}^p$ de composantes $f_i = (u_0, a_i)$, on peut la réécrire de façon équivalente

$$C\, u^\star - f = 0.$$

De plus,

$$(\sum_{k=1}^{p} \lambda_k a_k)_i = \sum_{k=1}^{p} \lambda_k (a_k)_i = \sum_{k=1}^{p} \lambda_k C_{k,i} = \sum_{k=1}^{p} (C^{\mathsf{T}})_{i,k} \lambda_k = (C^{\mathsf{T}}\lambda)_i,$$

où λ est le vecteur de $\mathbb{R}^p$ de composantes λ_k. Ainsi, lorsque l'on cherche un point de minimum $u^\star$ du problème

$$\min_{v,\ Cv-f=0} J(v),$$

la condition nécessaire (2.9) peut être reformulée sous la forme

$$\begin{cases} \exists\lambda \in \mathbb{R}^p, \quad \nabla J(u^\star) + C^{\mathsf{T}}\lambda = 0 \\ C\, u^\star - f = 0 \end{cases} . \tag{2.10}$$

Remarque 2.2.3 *Pour obtenir un résultat de caractérisation similaire lorsque les contraintes d'égalité ne sont plus affines, mais quelconques, il faut disposer du théorème des fonctions implicites. Le lecteur intéressé est renvoyé à [10].*

2.2.4 Le Lagrangien

Reprenons le problème de la minimisation de J sur $K^\star$, tels que définis dans la sous-section précédente. Nous allons voir qu'il est particulièrement intéressant d'introduire la fonctionnelle

$$\mathcal{L}(v, \mu) = J(v) + \sum_{k=1}^{p} \mu_k \, (a_k, v - u_0), \ \forall (v, \mu) \in \mathbb{E} \times \mathbb{R}^p. \tag{2.11}$$

On l'appelle le **Lagrangien**, associé au problème de la minimisation **de J sur $K^\star$**. Ici, *v parcourt $\mathbb{E}$ entier, et non plus $K^\star$ uniquement.* On dit que l'on a *dualisé les contraintes*, et les éléments μ de $\mathbb{R}^p$ sont appelés **multiplicateurs de Lagrange**.
Pourquoi est-ce utile? Pour le comprendre, étudions la différentielle partielle par rapport à v de $\mathcal{L}$ (c'est-à-dire que l'on raisonne à μ fixé) :

$$\mathcal{L}(v + \theta\, h, \mu) - \mathcal{L}(v, \mu) = J(v + \theta\, h) - J(v) + \sum_{k=1}^{p} \mu_k \, (a_k, \theta\, h).$$

Dès lors que J est différentiable en v, on en déduit que $\mathcal{L}$ est différentiable par rapport à v en (v,μ), puisque

$$\mathcal{L}(v+\theta\,h,\mu)-\mathcal{L}(v,\mu)=\theta(\nabla J(v,h)+\sum_{k=1}^{p}\mu_k a_k,h)+o(\theta).$$

Ainsi, la différentielle et le gradient partiels de $\mathcal{L}$ par rapport à v sont égaux à

$$d_v\mathcal{L}(v,\mu)\cdot h=(\nabla_v\mathcal{L}(v,\mu),h),\quad \nabla_v\mathcal{L}(v,\mu)=\nabla J(v,h)+\sum_{k=1}^{p}\mu_k a_k. \tag{2.12}$$

Qu'en est-il de la différentielle partielle par rapport à μ?

$$\mathcal{L}(v,\mu+\theta\,\eta)-\mathcal{L}(v,\mu)=\theta\sum_{k=1}^{p}\eta_k(a_k,v-u_0).$$

Cette fois, la différentielle et le gradient partiels de $\mathcal{L}$ par rapport à μ valent

$$d_\mu\mathcal{L}(v,\mu)\cdot\eta=(\nabla_\mu\mathcal{L}(v,\mu),\eta),\quad (\nabla_\mu\mathcal{L}(v,\mu))_k=(a_k,v-u_0),\ 1\le k\le p. \tag{2.13}$$

Le point de minimum local $u^\star$ est donc tel que

Corollaire 2.2.3 *On reprend les hypothèses du théorème 2.2.3. Si $u^\star$ est un point de minimum local de J sur $K^\star$, on a nécessairement*

$$\exists\lambda\in\mathbb{R}^p\ \textit{tel que}\ \left\{\begin{array}{l}\nabla_v\mathcal{L}(u^\star,\lambda)=0\\ \nabla_\mu\mathcal{L}(u^\star,\lambda)=0\end{array}\right.. \tag{2.14}$$

Preuve : on reprend (2.9), et les expressions (2.12) et (2.13). ∎

Lorsque l'on considère le Lagrangien associé au problème de minimisation, outre les expressions (2.14), il est primordial de noter que dans la définition de $\mathcal{L}$, v appartient à $\mathbb{E}$ **entier**. Ainsi, on a troqué l'appartenance à $K^\star$ pour un terme additionnel, dans la fonctionnelle à étudier.

Remarque 2.2.4 *Il est loisible de définir le Lagrangien de beaucoup d'autres problèmes de minimisation, que ce soit pour des problèmes avec contraintes d'égalité quelconques, ou pour des problèmes avec contraintes d'inégalité [10]. Encore une fois, nous invitons le lecteur intéressé à patienter pendant quelques mois...*

Exercice 2.2.2 *Vérifier que si J est différentiable en v, alors $\mathcal{L}$ l'est en (v,μ), pour tout μ de $\mathbb{R}^p$, et étudier la réciproque.*

2.2.5 Fonctionnelle quadratique et contraintes d'égalité affines

La fonctionnelle est, dans cette sous-section, quadratique en v ; c'est J_0, définie pour des éléments de $\mathbb{R}^n$, par (1.3), c'est-à-dire

$$J_0=\frac{1}{2}(Av,v)-(b,v)+c,$$

avec A une matrice *symétrique* de $\mathbb{R}^{n\times n}$, b un vecteur de $\mathbb{R}^n$ et c un réel. On considère ses variations sur l'espace affine

$$K^\star=\{v\in\mathbb{R}^n\ :\ C\,v=f\},$$

avec C appartenant à $\mathbb{R}^{p\times n}$ de rang p, et f un élément de $\mathbb{R}^p$.
D'après l'expression (2.9), si $u^\star$ est un point de minimum de J_0 sur $K^\star$,

$$\exists \lambda \in \mathbb{R}^p \text{ tel que } \left\{ \begin{array}{l} A\,u^\star + C^{\mathsf{T}}\lambda = b \\ C\,u^\star = f \end{array} \right. . \tag{2.15}$$

En d'autres termes,

Corollaire 2.2.4 *Si $u^\star$ est un point de minimum de J_0 sur $K^\star$, alors il existe λ un élément de $\mathbb{R}^p$ tel que le couple $(u^\star, \lambda)$ de $\mathbb{R}^n \times \mathbb{R}^p$ soit solution du système linéaire*

$$\left(\begin{array}{cc} A & C^{\mathsf{T}} \\ C & 0 \end{array} \right) \left(\begin{array}{c} u^\star \\ \lambda \end{array} \right) = \left(\begin{array}{c} b \\ f \end{array} \right). \tag{2.16}$$

2.3 Convexité et conditions suffisantes

Une catégorie très importante parmi les fonctionnelles est celle des fonctionnelles convexes, pour lesquelles on peut obtenir la caractérisation des minima. En effet, les conditions nécessaires de la première section deviennent **nécessaires et suffisantes**, lorsque la fonctionnelle est convexe (et différentiable). Qui plus est, le minimum, qui *a priori* peut-être *local*, devient *global*.

2.3.1 Définition, propriétés

Définition 2.3.1 *Soit J une fonctionnelle définie sur un sous-ensemble convexe non vide K de $\mathbb{E}$, à valeurs dans $\mathbb{R}$. On dit que J est* **convexe** *si et seulement si*

$$\forall u, v \in K, \quad u \neq v, \quad \forall \theta \in]0,1[\qquad J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v).$$

Dans le cas d'une inégalité stricte, on dit que la fonctionnelle J est **strictement convexe**.
Enfin, pour être complet, nous dirons que, s'il existe $\alpha > 0$ tel que

$$\forall u, v \in K, \quad u \neq v, \quad \forall \theta \in]0,1[\qquad J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha}{2}\theta(1-\theta)\|u-v\|^2,$$

J est **α-convexe**.

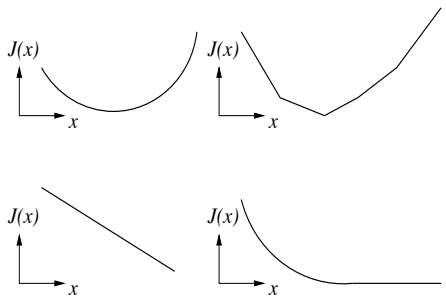


FIG. 2.4 – *Exemples de fonctions convexes*

Remarque 2.3.1 *(***géométrique***) La convexité de J signifie que le graphe de J est au-dessous de toutes ses cordes.*

Exercice 2.3.1 *Montrer que si J est continue et α-convexe, alors elle est strictement convexe.*

Exercice 2.3.2 *Montrer que si J est α-convexe et différentiable en un point, alors elle est infinie à l'infini (NB. On ne fait aucune hypothèse de continuité sur J.)*

Proposition 2.3.1 *Soit J une fonctionnelle convexe définie sur un convexe non vide K.*
Si u et v sont deux points de minimum locaux, alors $J(u) = J(v)$.
Si de plus J est strictement convexe, alors $u = v$.

Preuve : Soient u et v deux minima.
Comme J est convexe,

$$J(u + \theta(v - u)) = J((1 - \theta)u + \theta\, v) \leq (1 - \theta)J(u) + \theta\, J(v), \quad \forall \theta \in]0, 1[.$$

Si l'on suppose que $J(v) < J(u)$, on infère que

$$J(u + \theta(v - u)) < (1 - \theta)J(u) + \theta\, J(u) = J(u), \quad \forall \theta \in]0, 1[,$$

ce qui contredit le fait que u est un minimum local (θ proche de 0).

De la même façon, si l'on suppose que $J(u) < J(v)$, on en déduit cette fois que

$$J(u + \theta(v - u)) < (1 - \theta)J(v) + \theta\, J(v) = J(v), \quad \forall \theta \in]0, 1[,$$

ce qui contredit cette fois le fait que v est un minimum local (θ proche de 1).
On a donc bien $J(u) = J(v)$.
Supposons que J est strictement convexe. Si u et v sont deux points de minimum, on a vu que $J(u) = J(v)$. Si u et v sont distincts,

$$J(u + \theta(v - u)) < (1 - \theta)J(u) + \theta\, J(v) = J(u), \quad \forall \theta \in]0, 1[,$$

ce qui contredit le fait que u est un minimum local.
On a donc bien $u = v$. ■

Remarque 2.3.2 *Ceci signifie en particulier que*

> *tout point de minimum local d'une fonctionnelle convexe est en fait un point de minimum* global. *En effet, si on reprend le raisonnement ci-dessus avec u minimum local et v tel que $J(v) < J(u)$, on trouve $J(u + \theta(v - u)) < J(u)$, pour tout θ dans $]0, 1[$: ceci contredit l'hypothèse lorsque θ est proche de 0 !*
>
> *le point de minimum d'une fonctionnelle strictement convexe, s'il existe, est* unique.

Avant de démontrer les résultats portant sur les conditions suffisantes, lorsque la fonctionnelle est convexe, nous allons commencer par relier cette notion dans $\mathbb{E}$ à celle, plus connue, de la convexité dans $\mathbb{R}$. Pour cela, pour tout couple $(u, v) \in K \times K$, nous introduisons la fonction

$$\mu_{u,v} \ : \ \theta \mapsto J((1 - \theta)u + \theta v) = J(u + \theta(v - u)), \text{de } [0, 1] \text{ dans } \mathbb{R}$$

(la dépendance de $\mu_{u,v}$ par rapport à u et v est sous-entendue dans la suite.)

Théorème 2.3.1 *Soit J définie sur un convexe non vide K. Alors J est (strictement) convexe si, et seulement si, μ est (strictement) convexe pour tout couple (u, v) de $K \times K$.*

Preuve : Supposons pour commencer que J est strictement convexe. Soient u et v deux éléments de K, μ la fonction associée, $0 \leq x < y \leq 1$, et $\beta \in]0,1[$.

$$\begin{aligned}\mu(\beta\, x + (1-\beta)y) &= J((1-\beta\, x - (1-\beta)y)u + (\beta\, x + (1-\beta)y)v) \\ &= J(u + \beta(-x\, u + x\, v) + (1-\beta)(-y\, u + y\, v)) \\ &= J((\beta + (1-\beta))u + \beta(-x\, u + x\, v) + (1-\beta)(-y\, u + y\, v)) \\ &= J(\beta((1-x)u + x\, v) + (1-\beta)((1-y)u + y\, v)) \\ &< \beta\, J((1-x)u + x\, v) + (1-\beta)J((1-y)u + y\, v),\end{aligned}$$

soit finalement $\mu(\beta x + (1-\beta)y) < \beta\, \mu(x) + (1-\beta)\mu(y)$.
Réciproquement, si μ est strictement convexe pour tout couple (u, v), on écrit, $\theta \in]0,1[$,

$$J(\theta\, u + (1-\theta)v) = \mu(1-\theta) < \theta\, \mu(0) + (1-\theta)\mu(1) = \theta\, J(u) + (1-\theta)J(v).$$

Bien évidemment, on peut reprendre le raisonnement ci-dessus pour la convexité simple, en remplaçant les inégalités strictes par des inégalités larges. ■

Rappelons maintenant quelques résultats concernant la (stricte) convexité des fonctions :

Proposition 2.3.2 *Soit μ une application de $[0,1]$ dans $\mathbb{R}$.*
μ est **convexe** *si et seulement si*

$$\forall x_0, x_1, x_2 \in [0,1],\ x_0 < x_1 < x_2,\ \frac{\mu(x_1) - \mu(x_0)}{x_1 - x_0} \leq \frac{\mu(x_2) - \mu(x_0)}{x_2 - x_0} \leq \frac{\mu(x_2) - \mu(x_1)}{x_2 - x_1}.$$

Si de plus μ est dérivable alors μ est convexe si, et seulement si, μ' est croissante.
Si enfin μ est deux fois dérivable alors μ est convexe si, et seulement si, μ'' est positive.
μ est **strictement convexe** *si et seulement si*

$$\forall x_0, x_1, x_2 \in [0,1],\ x_0 < x_1 < x_2,\ \frac{\mu(x_1) - \mu(x_0)}{x_1 - x_0} < \frac{\mu(x_2) - \mu(x_0)}{x_2 - x_0} < \frac{\mu(x_2) - \mu(x_1)}{x_2 - x_1}.$$

Si de plus μ est dérivable alors μ est strictement convexe si, et seulement si, μ' est strictement croissante.

Preuve : Elle est laissée en exercice... ■

A partir de ces rappels, nous sommes en mesure d'énoncer le théorème principal de caractérisation des fonctionnelles différentiables convexes.

Théorème 2.3.2 *Soit J une fonctionnelle différentiable sur un sous-ensemble K convexe.*
Les assertions suivantes sont équivalentes.

(i) J *est convexe sur* K.

(ii) $\forall u, v \in K,\ u \neq v,\ J(v) \geq J(u) + (\nabla J(u), v - u)$.

(iii) $\forall u, v \in K,\ u \neq v,\ (\nabla J(u) - \nabla J(v), u - v) \geq 0$.

De même, les assertions suivantes sont équivalentes.

(iv) J *est strictement convexe sur* K.

(v) $\forall u, v \in K,\ u \neq v,\ J(v) > J(u) + (\nabla J(u), v - u)$.

(vi) $\forall u, v \in K,\ u \neq v,\ (\nabla J(u) - \nabla J(v), u - v) > 0$.

Preuve : Comme pour le théorème 2.3.1, nous considérons uniquement le cas de la stricte convexité.
Montrons tout d'abord que (iv) ⇒ (v).
Soient donc u et v deux éléments distincts de K, et μ la fonction associée ; d'après le théorème 2.3.1, μ est strictement convexe. D'après la proposition ci-dessus, ceci est équivalent au fait que μ' est strictement croissante. Si l'on applique le théorème de Rolle entre 0 et 1, on trouve

$$\exists c \in]0,1[\text{ tel que } \mu(1) - \mu(0) = \mu'(c).$$

Comme μ' est strictement croissante, ceci implique

$$\mu(1) - \mu(0) > \mu'(0).$$

Il ne reste plus qu'à revenir à la fonctionnelle J. Or, par définition (cf. (2.3)), on a

$$\mu'(\theta) = (\nabla J(u + \theta(v-u)), v-u).$$

On en déduit finalement que

$$J(v) - J(u) > (\nabla J(u), v-u),$$

c'est-à-dire (v).
Montrons maintenant que (v) ⇒ (vi).
Pour cela, il suffit d'appliquer (v) au couple (u,v), puis au couple (v,u), et de faire la somme.
Finalement, il reste à vérifier que (vi) ⇒ (iv).
Soient donc u et v deux éléments distincts de K, et μ la fonction associée ; nous allons montrer que μ' est strictement croissante. La proposition ci-dessus et le théorème 2.3.1 permettront alors de conclure. Soient donc $0 \leq x < y \leq 1$:

$$\begin{aligned}
\mu'(y) - \mu'(x) &= (\nabla J(u + y(v-u)) - \nabla J(u + x(v-u)), v-u) \\
&= (\nabla J(u + y(v-u)) - \nabla J(u + x(v-u)), \frac{[u + y(v-u)] - [u + x(v-u)]}{y-x}) \\
&> 0, \text{ d'après (vi), puisque } y > x.
\end{aligned}$$

■

Remarque 2.3.3 *(**géométrique**) La convexité de J (ii) signifie que le graphe de J est au-dessus du graphe de l'application affine tangente à J en u, c'est-à-dire $v \mapsto J(u) + (\nabla J(u), v-u)$, en tout point u de K.*

Remarque 2.3.4 *En ce qui concerne l'α-convexité, on peut prouver les équivalences ci-dessous. Soit J une fonctionnelle différentiable sur un sous-ensemble K.*
Les assertions suivantes sont équivalentes.

(vii) J *est α-convexe sur K.*

(viii) $\forall u, v \in K,\ u \neq v,\ J(v) \geq J(u) + (\nabla J(u), v-u) + \dfrac{\alpha}{2}\|u-v\|^2.$

(ix) $\forall u, v \in K,\ u \neq v,\ (\nabla J(u) - \nabla J(v), u-v) \geq \alpha\|u-v\|^2.$

Théorème 2.3.3 *Soit J une fonctionnelle de $\mathbb{E}$ de classe $\mathcal{C}^2$. Alors J est convexe si, et seulement si,*

$$\forall u, v \in \mathbb{E}, \quad (\nabla^2 J(u)(v-u), v-u) \geq 0$$

Preuve : Si J est convexe, μ l'est également pour tout couple (u, v). Par ailleurs, de (2.4), on tire

$$\mu''(\theta) = (\nabla^2 J(u + \theta(v-u))(v-u), v-u).$$

Comme μ'' est positive par hypothèse, il suffit d'utiliser la formule ci-dessus en $\theta = 0$.

Réciproquement, supposons que $(\nabla^2 J(u)(v-u), v-u) \geq 0$, pour tout couple d'éléments (u, v) de $\mathbb{E} \times \mathbb{E}$.
On peut reformuler cette condition en la condition équivalente $(\nabla^2 J(u)h, h) \geq 0$, pour tout couple d'éléments (u, h) de $\mathbb{E} \times \mathbb{E}$, et remplacer pour finir u par $u + \theta\, h$.
Ceci étant noté, nous appliquons la formule de Taylor-Mac Laurin (cf. théorème 14.3.3), en u et en v.

$$\begin{aligned} J(u+h) &= J(u) + (\nabla J(u), h) + \frac{1}{2}(\nabla^2 J(u + \theta\, h)h, h), \quad \theta \in]0, 1[, \\ &\geq J(u) + (\nabla J(u), h). \end{aligned}$$

On choisit $h = v - u$, pour trouver $J(v) \geq J(u) + (\nabla J(u), v-u)$, ce qui correspond bien à la condition de convexité (ii) de J. ∎

Pour finir, notons que lorsque l'on sait qu'une fonctionnelle est convexe, on peut démontrer des résultats généraux concernant sa *régularité* (nous renvoyons le lecteur à [10]).

2.3.2 Conditions suffisantes

Pourquoi avoir passé quelques pages à établir ces caractérisations de la convexité d'une fonctionnelle différentiable J ? Comme le suggère le titre du paragraphe, la raison principale est que, lorsque J est convexe, les conditions qu'elle doit vérifier en un point u, qui sont simplement nécessaires dans le cas général, deviennent nécessaires et suffisantes.
Reprenons donc pour finir les principaux résultats portant sur les conditions de réalisation d'un point de minimum.

Théorème 2.3.4 *Soient K un sous-ensemble convexe de $\mathbb{E}$, u un point de K et J une fonctionnelle convexe et différentiable de K. Alors u est un point de minimum global de J sur K si, et seulement si,*

$$(\nabla J(u), v-u) \geq 0, \qquad \forall v \in K. \tag{2.17}$$

Preuve : Si u est un point de minimum de J, nous savons déjà que l'inéquation d'Euler (2.17) est vérifiée.
Réciproquement, du fait de la convexité, la condition (ii) nous dit que u est un point de minimum global. ∎

On se place maintenant dans le cas d'un convexe $K^\star$ égal à un espace affine (voir la sous-section 2.2.3), défini par

$$K^\star = \{v \in \mathbb{E} \ : \ (v - u_0, a_k) = 0, \ 1 \leq k \leq p\},$$

où $(a_k)_{1 \leq k \leq p}$ est une famille libre de $\mathbb{E}$, et u_0 un vecteur de $\mathbb{E}$. On infère du théorème ci-dessus un second résultat dû à Karush, Kuhn et Tucker (K.K.T.).

Corollaire 2.3.1 *Soient $u^\star$ un point de $K^\star$ et J une fonctionnelle convexe et différentiable de $K^\star$. Alors $u^\star$ est un point de minimum global de J sur $K^\star$ si, et seulement si,*

$$\exists \lambda \in \mathbb{R}^p, \quad \nabla J(u^\star) + \sum_{k=1}^{p} \lambda_k\, a_k = 0. \tag{2.18}$$

Preuve : Si $u^\star$ est un point de minimum, le théorème 2.2.3 nous permet d'affirmer qu'il existe un élément λ de $\mathbb{R}^p$ tel que (2.18) soit vérifiée.
Réciproquement, s'il existe λ tel que (2.18) soit vraie, on en déduit que le gradient $\nabla J(u^\star)$ est orthogonal à toute différence de deux éléments de $K^\star$, c'est-à-dire que

$$(\nabla J(u^\star), v^\star - u^\star) = 0, \quad \forall v^\star \in K^\star.$$

Le théorème 2.3.4 nous permet de conclure. ■

Enfin, pour en finir ici avec les problèmes avec contraintes, revenons à la minimisation quadratique sous contraintes d'égalité. Considérons la fonctionnelle J_0 définie comme d'habitude par (1.3) pour des vecteurs de $\mathbb{R}^n$, où l'on suppose cette fois que la matrice A est *symétrique et positive.* L'espace $K^\star$ est quant à lui défini par

$$K^\star = \{v \in \mathbb{R}^n \ : \ C\,v = f\},$$

où C est une matrice de $\mathbb{R}^{p\times n}$ de rang p, et f est un vecteur de $\mathbb{R}^p$. On a alors le

Théorème 2.3.5 *$u^\star$ est un point de minimum de J_0 sur $K^\star$ si, et seulement si, il existe un élément λ de $\mathbb{R}^p$ tel que le couple $(u^\star, \lambda)$ soit solution de*

$$\begin{pmatrix} A & C^\mathsf{T} \\ C & 0 \end{pmatrix} \begin{pmatrix} u^\star \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}. \tag{2.19}$$

Si de plus A est (symétrique) définie-positive*, le système linéaire (2.19) admet une solution unique. En d'autres termes, il existe un point de minimum global de J_0 sur $K^\star$ et un seul.*

Preuve : Si $u^\star$ est un point de minimum, on applique le corollaire 2.2.4.

Réciproquement, nous allons vérifier que J_0 est convexe. En effet, on sait que $\nabla J_0(v) = Av - b$, puisque A est symétrique. De plus,

$$(\nabla J_0(v) - \nabla J_0(u), v - u) = (A(v-u), v-u) \geq 0,$$

puisque A est positive. D'après (iii), J_0 est convexe. On se retrouve dans la situation du corollaire précédent, ce qui achève la démonstration du premier point.

On suppose ici que A est symétrique définie-positive. La matrice

$$\begin{pmatrix} A & C^\mathsf{T} \\ C & 0 \end{pmatrix} \text{ appartient à } \mathbb{R}^{(n+p)\times(n+p)}.$$

Pour prouver que le système linéaire (2.19) admet une solution unique, il suffit de vérifier que le noyau de l'application linéaire associée est réduit à $\{0\}$ (pour une application linéaire d'un espace vectoriel de dimension finie dans lui-même, bijectivité $\Longleftrightarrow$ surjectivité $\Longleftrightarrow$ injectivité.) Soit donc un couple (u, λ) de $\mathbb{R}^n \times \mathbb{R}^p$ tel que

$$\begin{pmatrix} A & C^\mathsf{T} \\ C & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Comme A est inversible, on infère, par implications successives que
$u = -A^{-1}C^\mathsf{T}\lambda$ (première ligne) ; $CA^{-1}C^\mathsf{T}\lambda = 0$ (seconde ligne) ;
$(CA^{-1}C^\mathsf{T}\lambda, \lambda)_p = 0$ (produit scalaire par λ) ; $(A^{-1}C^\mathsf{T}\lambda, C^\mathsf{T}\lambda) = 0$ (transposition) ;
Comme A est symétrique définie-positive, A^{-1} l'est également, et ainsi

$$C^\mathsf{T}\lambda = 0.$$

Pour conclure, on note que l'application linéaire associée à C^T va de $\mathbb{R}^p$ dans $\mathbb{R}^n$, et qu'elle est de rang p. Par conséquent, $\dim(Ker(C^\mathsf{T})) = 0$, dont on déduit que $\lambda = 0$, puis finalement que $u = 0$ par retour à la première ligne du système linéaire. La conclusion suit. ■

En ce qui concerne les problèmes sans contraintes, on a immédiatement le

Théorème 2.3.6 *Soient K un sous-ensemble convexe de $\mathbb{E}$, u un point de K et J une fonctionnelle convexe et différentiable de K. Si $K = \mathbb{E}$ (ou si u est intérieur à K), alors u est un point de minimum (global) de J si, et seulement si,*

$$\nabla J(u) = 0. \tag{2.20}$$

La démonstration, immédiate, est laissée au lecteur.

Exercice 2.3.3 *On reprend $J_0(u) = \frac{1}{2}(Av, v) - (b, v) + c$, définie sur $\mathbb{R}^n$.*
1. A quelle(s) condition(s) J_0 est-elle convexe, strictement convexe, α-convexe ?
2. Retrouver les conclusions concernant l'étude des problèmes (1.4) et (1.5).

Exercice 2.3.4 *Soit J une fonctionnelle α-convexe et différentiable sur $\mathbb{R}^n$. Montrer que J admet un minimum global, et le caractériser.*

Chapitre 3

Moindres carrés linéaires

Nous considérons dans ce chapitre un problème de minimisation, relativement courant en pratique, appelé problème de moindres carrés. Nous nous contentons de considérer le cas particulier des moindres carrés linéaires, car on verra que ces problèmes ont de très fortes ramifications avec la Partie 2, qui traite d'algèbre linéaire.

3.1 Problèmatique

De prime abord, il est rassurant (!?) de résoudre exactement un problème. En pratique, cependant, on se rend compte que, dans de nombreux cas, il n'existe pas de solution "exacte" (voir la note de bas de page). C'est souvent le cas lorque l'on désire réaliser l'opération suivante :

> A partir d'un nombre fini (parfois très grand) de mesures, inférer un comportement valable dans tous les cas, passés, présents ou à venir.

Typiquement, d'une part on dispose d'un modèle abstrait, et d'autre part de données, et l'on souhaite fusionner l'un et l'autre, pour disposer d'une modélisation concrète du phénomène étudié, et/ou d'outils de prédiction. Prenons l'exemple suivant.
Carl Friedrichs Gauss (1777-1855) désirait déterminer la trajectoire de planètes, et notamment celle d'Uranus, découverte à la fin du 18ème siècle. D'après les lois de Képler, si l'on néglige la présence des autres planètes autour du Soleil, Uranus décrit une ellipse. Si l'on suppose *connus* le plan de la trajectoire (écliptique) ainsi que la direction du grand axe, sa trajectoire est une ellipse E dans le plan de l'écliptique, dont l'équation est

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1. \tag{3.1}$$

L'ellipse E est donc caractérisée par quatre paramètres, (x_0, y_0, a, b). Dès que l'on dispose de quatre positions (ou plus) d'Uranus dans le ciel, il est possible de caractériser sa trajectoire elliptique[1]... Pour cela, Gauss a inventé le principe dit des **moindres carrés** (en 1801). Disposant de K mesures de la position d'Uranus $M_k(x_k, y_k)_{1\leq k\leq K}$, on choisit (x_0, y_0, a, b), ce qui définit une unique ellipse $E = E_{x_0,y_0,a,b}$. A partir de là, on introduit les points $(M'_k)_{1\leq k\leq K}$: pour chaque valeur de k, M'_k est le point d'intersection de l'ellipse avec la droite passant par M_k et le centre de l'ellipse, le plus proche de M_k, de coordonnées

$$x'_k = p_E(x_k, y_k),\ y'_k = q_E(x_k, y_k), \quad 1 \leq k \leq K. \tag{3.2}$$

1. Caractérisation de la trajectoire... Pour trois mesures ou moins, il existe une infinité de possibilités. Quatre mesures sont idéales, puisque qu'il leur correspond une unique ellipse. A partir de cinq mesures ou plus, il faut *espérer* que tous les points de la trajectoire, à partir du $5^{ème}$, se trouvent sur l'ellipse définie par les quatre premiers ! Cette prise de conscience (existence d'une **surdétermination**) est fondamentale, lorsque l'on résout ce type de problèmes.

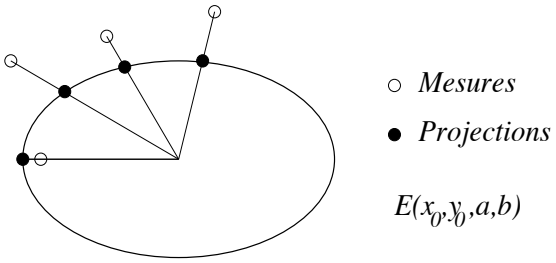


FIG. 3.1 – *Projections sur l'ellipse*

Pour mesurer l'erreur commise entre les positions mesurées et leurs projections sur l'ellipse E, on forme la quantité

$$\nu = \sum_{k=1}^{K} M_k M_k'^2. \tag{3.3}$$

Si tous les points de la trajectoire mesurée se trouvent sur l'ellipse E, on obtient $\nu = 0$; dans le cas contraire, $\nu > 0$. Précisons, avant de continuer, que les *données* sont $(x_k, y_k)_{1\le m\le K}$, et que les *inconnues* sont (x_0, y_0, a, b). Comme les nombres $(x'_k, y'_k)_{1\le k\le K}$ sont caractérisés par les relations (3.2), on peut donc introduire la fonctionnelle

$$\nu(x_0, y_0, a, b) = \sum_{k=1}^{K} \{\|x_k - p_E(x_k, y_k)\|^2 + \|y_k - q_E(x_k, y_k)\|^2\}. \tag{3.4}$$

L'idée est de partir d'une première ellipse, puis de la modifier, de façon à diminuer la valeur de ν correspondante, et ainsi de suite... Le but est de *minimiser* la valeur de $\nu(x_0, y_0, a, b)$, le quadruplet (x_0, y_0, a, b) décrivant $\mathbb{R}^4$:

$$\begin{aligned} &\text{Trouver } (x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) \in \mathbb{R}^4, \\ &\qquad \text{tel que } \nu(x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) = \inf_{(x_0,y_0,a,b)\in\mathbb{R}^4} \nu(x_0, y_0, a, b). \end{aligned} \tag{3.5}$$

Idéalement, comme nous l'avons remarqué plus haut, *si* les mesures sont exactes, *et si* la trajectoire est effectivement elliptique dans le plan de l'ecliptique, on détermine une solution telle que

$$\nu(x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) = 0.$$

Malheureusement, on sait que toute mesure est approchée, ce qui interdit de trouver un tel résultat. *Heureusement*, ceci n'est pas incompatible avec la résolution du problème (3.5).

Dans la suite, nous nous limiterons à l'étude de modèles-type, pour lesquels la dépendance par rapport aux inconnues est *linéaire*. A des fins illustratives, dans le formalisme adopté ci-dessus, on aurait

$$\begin{aligned} &\left\{ \begin{array}{l} x'_k = \alpha_k x_0 + \beta_k y_0 + \gamma_k a + \delta_k b + f(x_1, y_1, \cdots, x_K, y_K), \\ y'_k = \alpha'_k x_0 + \beta'_k y_0 + \gamma'_k a + \delta'_k b + f'(x_1, y_1, \cdots, x_K, y_K), \end{array} \right. \quad 1 \le k \le K, \\ &\text{soit } \nu(v) = \|Av - b\|^2,\ v \in \mathbb{R}^4,\ A \in \mathbb{R}^{2K\times 4},\ b \in \mathbb{R}^{2K}. \end{aligned} \tag{3.6}$$

On parle alors de **moindres carrés linéaires**.

Remarque 3.1.1 *Pour refermer la parenthèse historique (voir [20] pour plus de détails), mentionnons que Gauss a mené à bien ses calculs (sans ordinateur !). A la suite de quoi, on s'est aperçu qu'au cours du temps la trajectoire elliptique optimale variait... Après avoir éliminé les*

incertitudes liées aux erreurs de mesure, on en a déduit que la trajectoire n'était pas une ellipse, mais plutôt une perturbation de trajectoire elliptique. L'influence des autres planètes a été prise en compte, mais cela ne résolvait toujours pas la difficulté. URBAIN LE VERRIER *(1811-1877) a donc eu l'idée de chercher une nouvelle planète, introduisant une nouvelle perturbation, qui validerait le modèle : il a découvert Neptune en 1846.*

3.2 Le formalisme abstrait et son étude : pourquoi des carrés ?

Dans la suite, pour A une matrice *non nulle* de $\mathbb{R}^{m\times n}$ et b un vecteur de $\mathbb{R}^m$, on considère la résolution du problème :

$$\min_{v\in\mathbb{R}^n} f(v), \text{ avec } f(v) = \|Av-b\|_m.$$

m et n sont deux éléments quelconques de $\mathbb{N}^*$, *a priori* distincts. Pour cette raison, on indicera les normes et produits scalaires par $_m$ ou $_n$ si nécessaire, pour éviter les confusions.
On remarque, avant de commencer l'étude proprement dite du problème de minimisation, que f est *convexe*. En effet, on vérifie que pour v et w deux éléments de $\mathbb{R}^n$, et θ dans $]0,1[$, on a l'inégalité

$$f(\theta v + (1-\theta)w) \leq \theta f(v) + (1-\theta) f(w) \ ;$$

comme f est à valeurs positives, il est équivalent de prouver que les carrés sont dans cet ordre. On pose $x = Av - b$ et $y = Aw - b$:

$$\begin{aligned} f(\theta v + (1-\theta)w)^2 &= \|A(\theta v + (1-\theta)w) - b\|^2 \\ &= \|\theta x + (1-\theta)y\|^2 \\ &= \theta^2\|x\|^2 + 2\theta(1-\theta)(x,y) + (1-\theta)^2\|y\|^2 \\ &\leq \theta^2\|x\|^2 + 2\theta(1-\theta)\|x\|\,\|y\| + (1-\theta)^2\|y\|^2 \\ &= [\theta\|x\| + (1-\theta)\|y\|]^2 \\ &= [\theta f(v) + (1-\theta) f(w)]^2. \end{aligned}$$

En conséquence, d'après les résultats du chapitre 2, les conditions d'existence de minimum seront *nécessaires et suffisantes*. Comment caractériser le minimum ? C'est l'objet des deux sous-sections ci-dessous...

3.2.1 L'approche directe

En vue d'appliquer les résultats du chapitre 2, calculons le gradient de f, sans toutefois oublier de *vérifier* que f est différentiable.
Allons-y... Soient donc v et h deux éléments de $\mathbb{R}^n$, et θ un réel destiné à tendre vers 0 par valeurs positives.

$$f(v+\theta h) - f(v) = \|x - \theta Ah\| - \|x\|, \text{ avec } x = Av - b.$$

On se place pour commencer dans le cas général $x \neq 0$.

$$\begin{aligned} \|x-\theta Ah\| - \|x\| &= \frac{1}{\|x-\theta Ah\| + \|x\|}[\|x-\theta Ah\|^2 - \|x\|^2] \\ &= \frac{1}{\|x-\theta Ah\| + \|x\|}[2\theta(x, Ah)_m + \theta^2\|Ah\|^2] \\ &= \frac{1}{\|x-\theta Ah\| + \|x\|}[2\theta(A^{\mathsf{T}}x, h)_n + O(\theta^2)]. \end{aligned}$$

Par ailleurs, $\|x\| - \|\theta Ah\| \leq \|x - \theta Ah\| \leq \|x\| + \|\theta Ah\|$: on a donc $\|x - \theta Ah\| = \|x\| + O(\theta)$. Ainsi, puisque x est fixé (avant-dernière égalité),

$$\frac{1}{\|x - \theta Ah\| + \|x\|} = \frac{1}{2\|x\| + O(\theta)} = \frac{1}{2\|x\|(1 + O(\theta))} = \frac{1}{2\|x\|}(1 + O(\theta)).$$

D'où

$$\|x - \theta Ah\| - \|x\| = \theta \frac{(A^{\mathsf{T}}x, h)_n}{\|x\|} + O(\theta^2) = \theta \frac{(A^{\mathsf{T}}Av - A^{\mathsf{T}}b, h)_n}{\|Av - b\|} + o(\theta).$$

On a donc trouvé

$$\nabla f(v) = \frac{A^{\mathsf{T}}Av - A^{\mathsf{T}}b}{\|Av - b\|}. \tag{3.7}$$

NB. On vérifie que f est Fréchet-différentiable selon une procédure similaire.
Que se passe-t-il dans le cas particulier $x = 0$? Supposons que f soit différentiable, de différentielle $h \mapsto (g, h)$ $(g \in \mathbb{R}^n)$. Par définition de la Gateaux-differentiabilité :

$$f(v + \theta h) - f(v) = \theta \|Ah\| = \theta(g, h) + o(\theta), \quad \forall h \in \mathbb{R}^n.$$

Prenons, pour les deux directions h et $-h$, la même valeur de θ, soit

$$\theta(g, h) + o(\theta) = \theta \|Ah\| = \theta \|A(-h)\| = \theta(g, -h) + o(\theta),$$

et divisons par θ, que l'on fait tendre vers 0. Il reste $2(g, h) = 0$, pour toute direction h de $\mathbb{R}^n$. On infère la nullité de g, ce qui implique finalement

$$\theta \|Ah\| = o(\theta),$$

soit $Ah = 0$ pour tout h, ou encore $A = 0$. Or, on a supposé que A est une matrice non nulle. En conclusion, *f n'est pas différentiable en 0* !
Outre le fait que le calcul n'est pas immédiat, nous sommes confrontés à un problème majeur. f n'est pas différentiable en v_0 si $Av_0 = b$. Mais, si $Av_0 = b$, $f(v_0) = 0$ et v_0 est un point de minimum de f, puisque f est à valeurs positives ! Les résultats du chapitre 2 ne sont donc pas applicables, puisqu'ils requièrent la différentiabilité au point de minimum. Comment remédier à cette difficulté ? C'est l'objet de la sous-section suivante.

3.2.2 Une astuce de calcul

Comme f est à valeurs positives, les minima et points de minimum de f sont identiques à ceux de son carré, f^2 ! On peut donc considérer le problème de minimisation

$$\min_{v \in \mathbb{R}^n} J(v), \text{ avec } J(v) = \|Av - b\|_m^2.$$

On vérifie sans peine que

$$J(v + \theta h) - J(v) = 2\theta(Av - b, Ah)_m + \theta^2 \|Ah\|_m^2 = 2\theta(A^{\mathsf{T}}Av - A^{\mathsf{T}}b, h)_n + o(\theta).$$

Ainsi, J est différentiable en tous points (la Fréchet-différentiabilité est obtenue de même), et l'on a déterminé l'expression suivante du gradient

$$\nabla J(v) = 2A^{\mathsf{T}}Av - 2A^{\mathsf{T}}b. \tag{3.8}$$

Cette fois, on peut appliquer les résultats du chapitre 2. Pour commencer, J est convexe, d'après le point (iii) du théorème 2.3.2, puisque

$$(\nabla J(v) - \nabla J(u), v - u) = 2(A^{\mathsf{T}}A(v - u), v - u)_n = 2\|A(v - u)\|_m^2.$$

Qui plus est, on a le résultat ci-dessous :

Théorème 3.2.1 *u est un point de minimum global de J si, et seulement si, u est solution de*

$$A^{\mathsf{T}}Au = A^{\mathsf{T}}b. \tag{3.9}$$

Preuve : Ceci est une simple application du théorème 2.3.6. ■

Définition 3.2.1 *L'équation $A^{\mathsf{T}}Au = A^{\mathsf{T}}b$ est appelée* **équation normale**.

♠ Il faut faire très attention. Si bien sûr $Au = b$ entraîne (3.9), la réciproque est **fausse** en général...

3.2.3 Existence du point de minimum

On a le

Théorème 3.2.2 *Il existe au moins un point de minimum global.*

Preuve : Ceci revient à montrer que le système linéaire (3.9) admet toujours au moins une solution. Pour cela, nous allons utiliser la relation $Im\, A = (Ker\, A^{\mathsf{T}})^{\perp}$, énoncée et démontrée au lemme 1.3.1.
Pour tout élément b de $\mathbb{R}^m$, on peut écrire $b = b_0 + b_\perp$, avec $b_0 \in Ker\, A^{\mathsf{T}}$ et $b_\perp \in (Ker\, A^{\mathsf{T}})^{\perp}$. Alors, $A^{\mathsf{T}}b = A^{\mathsf{T}}b_\perp$, et d'après la relation ci-dessus, il existe un élément u de $\mathbb{R}^n$ tel que $b_\perp = Au$. On en déduit finalement, pour ce vecteur u :

$$A^{\mathsf{T}}b = A^{\mathsf{T}}b_\perp = A^{\mathsf{T}}Au.$$

■

On peut se servir d'outils différents pour retrouver ce résultat. Nous allons détailler la démarche, car elle est fort instructive, et utile pour la suite du chapitre... La matrice $A^{\mathsf{T}}A$, qui apparaît dans le terme quadratique de J, est une matrice symétrique et positive ; en effet :

$$\begin{aligned}&(A^{\mathsf{T}}A)^{\mathsf{T}} = A^{\mathsf{T}}A, \text{ et}\\ &(A^{\mathsf{T}}Ax, x)_n = (Ax, Ax)_m = \|Ax\|_m^2 \geq 0, \quad x \in \mathbb{R}^n.\end{aligned}$$

Par voie de conséquence, il existe $(v_i)_{1\leq i\leq n}$ une base orthonormale de $\mathbb{R}^n$ de vecteurs propres, de valeurs propres associées $(\lambda_i)_{1\leq i\leq n}$, appartenant à $\mathbb{R}_+$: $A^{\mathsf{T}}Av_i = \lambda_i v_i$, pour $1 \leq i \leq n$. Dans la suite, on les classe par ordre *décroissant*, et l'on définit q, le cardinal de l'ensemble $\{\lambda_i \ : \ \lambda_i > 0\}$, c'est-à-dire que $q = \mathrm{rg}(A^{\mathsf{T}}A)$. Notons que, puisque A n'est pas la matrice nulle, on a $1 \leq q \leq n$. Dans l'expression de J, on a également un terme linéaire, de la forme $-2(b, Av)_m$. Soient donc les vecteurs de $\mathbb{R}^m$ définis par

$$w_i = \frac{1}{\sqrt{\lambda_i}}Av_i, \quad 1 \leq i \leq q.$$

Pourquoi avoir introduit le facteur $1/\sqrt{\lambda_i}$? Parce que, pour $1 \leq i, j \leq q$, on a la relation

$$(w_i, w_j)_m = (\frac{1}{\sqrt{\lambda_i}}Av_i, \frac{1}{\sqrt{\lambda_j}}Av_j)_m = \frac{1}{\sqrt{\lambda_i\lambda_j}}(A^{\mathsf{T}}Av_i, v_j)_n = \sqrt{\frac{\lambda_i}{\lambda_j}}(v_i, v_j)_n = \delta_{ij}.$$

En d'autres termes, $(w_i)_{1\leq i\leq q}$ est une famille orthonormale de $\mathbb{R}^m$.
NB. Au passage, on vient de prouver que
$dim[Im\, A] = dim[Vect(Av_1, \cdots, Av_n)] = dim[Vect(w_1, \cdots, w_q)] = q$.

Ceci signifie en particulier que rg(A)=rg($A^{\mathsf{T}}A$) et $q \leq m$.
On la complète, le cas échéant, en une base orthonormale de $\mathbb{R}^m$. On peut alors décomposer le vecteur courant v ainsi que b sur les bases *ad hoc*, soit $v = \sum_{i=1}^{n} x_i v_i$ et $b = \sum_{i=1}^{m} b_i w_i$, pour obtenir

$$\begin{aligned} J(v) &= \|Av - b\|_m^2 \\ &= \|\sum_{i=1}^{q} \sqrt{\lambda_i} x_i w_i - \sum_{i=1}^{m} b_i w_i\|_m^2 \\ &= \sum_{i=1}^{q} \left(\sqrt{\lambda_i} x_i - b_i\right)^2 + \sum_{i=q+1}^{m} {b_i}^2. \end{aligned} \tag{3.10}$$

NB. Dans (3.10), la seconde somme peut être vide (si $q = m$).
Qu'en déduit-on ?

Proposition 3.2.1 *u est un point de minimum de J si, et seulement si,*

$$u = \sum_{i=1}^{n} x_i^0 v_i, \text{ avec } x_i^0 = \frac{1}{\sqrt{\lambda_i}} b_i,\ 1 \leq i \leq q, \quad x_i^0 \text{ quelconques }, q+1 \leq i \leq n. \tag{3.11}$$

De façon équivalente, si on note $u^0 = \sum_{i=1}^{q} x_i^0 v_i$, u est un point de minimum si, et seulement si,

$$u \in u^0 + Vect(v_{q+1}, \cdots, v_n). \tag{3.12}$$

Par construction (encore une fois !), l'ensemble des points de minimum est non vide...

Exercice 3.2.1 *Vérifier que (3.11) ou (3.12) est équivalent à (3.9).*

Ceci est un bon exemple de la propriété générale suivante. Supposons que, pour un problème posé à l'aide d'une matrice, on puisse prouver que celle-ci est *diagonalisable*. Alors, sous réserve que l'on connaisse ses éléments propres, résoudre le problème initial revient à résoudre un ensemble de problèmes [2] dans $\mathbb{R}$. Bien évidemment, le *défaut majeur* est qu'en général, il est beaucoup trop couteux de calculer l'ensemble des éléments propres d'une matrice ! Dans le cas des moindres carrés linéaires, on choisit plutôt de construire des algorithmes numériques directs ou itératifs permettant d'*"inverser"* l'équation normale (c'est-à-dire de calculer un vecteur u solution de (3.9)).

3.2.4 Moindres carrés contraints

Evoquons brièvement ici, ce qui se passe lorsque le problème est contraint, avec une contrainte du type

$$v \in K^\star = \{v \in \mathbb{R}^n \ :\ C\,v = f\},$$

avec C appartenant à $\mathbb{R}^{p\times n}$ de rang p, et f un élément de $\mathbb{R}^p$.

Théorème 3.2.3 *$u^\star$ est un point de minimum de J sur $K^\star$ si, et seulement si, il existe un élément λ de $\mathbb{R}^p$ tel que le couple $(u^\star, \lambda)$ soit solution de*

$$\begin{pmatrix} A^{\mathsf{T}}A & C^{\mathsf{T}} \\ C & 0 \end{pmatrix} \begin{pmatrix} u^\star \\ \lambda \end{pmatrix} = \begin{pmatrix} A^{\mathsf{T}}b \\ f \end{pmatrix}. \tag{3.13}$$

Si de plus $A^{\mathsf{T}}A$ est inversible*, le système linéaire (3.13) admet une solution unique.*

Preuve : Ceci est une application du théorème 2.3.5. ∎

2. Par exemple (classique), soit à calculer l'action d'un polynome R sur une matrice A de $\mathbb{R}^{n\times n}$, pour laquelle on suppose qu'il existe P inversible et D diagonale de $\mathbb{R}^{n\times n}$ telles que $D = P^{-1}AP$. Alors,

$$A^2 = PDP^{-1}PDP^{-1} = PD^2P^{-1}, \quad A^k = PD^kP^{-1},\ \forall k, \quad \text{ et } R(A) = PR(D)P^{-1}.$$

3.3 Décomposition en valeurs singulières

Dans cette partie, nous allons considérer un aspect algébrique lié aux problèmes de moindres carrés linéaires, celui de la factorisation de la matrice A de $\mathbb{R}^{m\times n}$ sous la forme

$$A = W\Sigma V^{\mathsf{T}} \tag{3.14}$$

où W et V sont deux matrices *orthogonales* (appartenant respectivement à $\mathbb{R}^{m\times m}$ et $\mathbb{R}^{n\times n}$), et Σ une matrice *dont les seuls éléments non nuls sont situés sur la diagonale*, de $\mathbb{R}^{m\times n}$.

Remarque 3.3.1 *Il est tout à fait possible de reprendre le raisonnement qui suit et de l'appliquer à une matrice de $\mathbb{C}^{m\times n}$. Dans (3.14), W et V sont alors des matrices unitaires.*

Pourquoi la décomposition de A (3.14), dite *en valeurs singulières*, est-elle liée aux problèmes de moindres carrés étudiés ci-dessus? Tout simplement parce que V est reliée à la base orthonormale $(v_i)_{1\leq i\leq n}$ de vecteurs propres de $A^{\mathsf{T}}A$, Σ aux valeurs propres $(\lambda_i)_{1\leq i\leq n}$, et W à la base orthonormale $(w_i)_{1\leq i\leq m}$. Ceci est résumé dans le

Théorème 3.3.1 *Soit A une matrice de $\mathbb{R}^{m\times n}$. Il existe W et V deux matrices orthogonales de $\mathbb{R}^{m\times m}$ et $\mathbb{R}^{n\times n}$ respectivement, et Σ une matrice dont les seuls éléments non nuls sont situés sur la diagonale, de $\mathbb{R}^{m\times n}$, telles que (3.14) soit satisfaite.*

Preuve : Par définition des deux bases orthonormales, on a les relations

$$Av_k = \sigma_k w_k, \quad 1 \leq k \leq n,$$

avec $\sigma_k = \sqrt{\lambda_k}$, pour $1 \leq k \leq n$, ce que l'on peut réécrire sous la forme

$$A \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}.$$

Soit

$$AV = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}, \text{ où l'on a posé } V = \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Par construction,V est orthogonale, puisque

$$(V^{\mathsf{T}}V)_{i,j} = \sum_{k=1}^{n} (V^{\mathsf{T}})_{i,k} V_{k,j} = \sum_{k=1}^{n} V_{k,i} V_{k,j} = \sum_{k=1}^{n} (v_i)_k (v_j)_k = (v_i, v_j)_n = \delta_{ij}.$$

Si maintenant, on pose

$$W = \begin{pmatrix} \vdots & \vdots & & \vdots \\ w_1 & w_2 & \cdots & w_m \\ \vdots & \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{m\times m} \text{ et}$$

$$\Sigma \in \mathbb{R}^{m\times n} \text{ telle que } \Sigma_{i,j} = \begin{cases} \sigma_i, \ 1 \leq i,j \leq q, \ i = j \\ 0 \text{ sinon} \end{cases},$$

vérifions que l'on a l'identité

$$W\Sigma = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}.$$

En effet,

pour $1 \le i \le m$, $1 \le j \le q$: $(W\Sigma)_{i,j} = \sum_{k=1}^{m} W_{i,k}\Sigma_{k,j} = \sum_{k=1}^{m} (w_k)_i \sigma_j \delta_{kj} = \sigma_j (w_j)_i$;

pour $1 \le i \le m$, $q+1 \le j \le n$: $(W\Sigma)_{i,j} = \sum_{k=1}^{m} W_{i,k}\Sigma_{k,j} = 0$ (la $j^{ème}$ colonne de Σ est composée de zéros).

Par construction, W est elle aussi orthogonale, et l'on trouve finalement

$$AV = W\Sigma, \text{ soit } A = W\Sigma V^{\mathsf{T}}.$$

■

Définition 3.3.1 *On appelle $(\sigma_k)_k$ les* **valeurs singulières** *de A.*

Remarque 3.3.2 *Quelle est l'apparence de Σ ? Si on appelle $r = \min(n, m)$ on a*

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & \sigma_{r-1} & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_r \end{pmatrix} \quad si\ r = n = m\ ;$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & 0 & \cdots & 0 \\ \vdots & & 0 & \sigma_{r-1} & 0 & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \sigma_r & 0 & \cdots & 0 \end{pmatrix} \quad si\ r = m < n\ ;$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & \sigma_{r-1} & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_r \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} \quad si\ r = n < m\ ;$$

Bien sûr, on a toujours $q \le r$...

Pour aller encore un peu de l'avant, démontrons à présent les identités *matricielles* de la proposition ci-dessous : v de $\mathbb{R}^l$ appartient aussi à $\mathbb{R}^{l\times 1}$, et v^{T} à $\mathbb{R}^{1\times l}$, et le symbole $\cdot$ représente la multiplication matricielle.

Proposition 3.3.1

$$A = \sum_{k=1}^{q} \sigma_k w_k \cdot v_k^{\mathsf{T}}, \quad A^{\mathsf{T}}A = \sum_{k=1}^{q} {\sigma_k}^2 v_k \cdot v_k^{\mathsf{T}}.$$

Preuve : Plutôt que la simple vérification des résultats, construisons les identités, en commençant par la première.
De (3.14), on tire, pour $1 \leq i \leq m$, $1 \leq j \leq n$,

$$\begin{aligned}
A_{i,j} &= \sum_{k=1}^{n} (W\Sigma)_{i,k} V^{\mathsf{T}}_{k,j} \\
&= \sum_{k=1}^{q} (W\Sigma)_{i,k} V_{j,k} \text{ (pour } k > q \text{, la } k^{ème} \text{ colonne de } W\Sigma \text{ est composée de zéros)} \\
&= \sum_{k=1}^{q} \sigma_k W_{i,k} V_{j,k} = \sum_{k=1}^{q} \sigma_k (w_k)_i (v_k)_j \\
&= \sum_{k=1}^{q} \sigma_k (w_k)_{i,1} (v_k^{\mathsf{T}})_{1,j} \text{ (on passe des vecteurs aux matrices)} \\
&= \sum_{k=1}^{q} \sigma_k (w_k \cdot v_k^{\mathsf{T}})_{i,j} = (\sum_{k=1}^{q} \sigma_k w_k \cdot v_k^{\mathsf{T}})_{i,j}.
\end{aligned}$$

Pour la seconde identité, on procède de la même façon. Tout d'abord, on remarque que

$$A^{\mathsf{T}} A = V \Sigma^{\mathsf{T}} W^{\mathsf{T}} W \Sigma V^{\mathsf{T}} = V D V^{\mathsf{T}}, \text{ avec } D = \Sigma^{\mathsf{T}} \Sigma = diag({\sigma_i}^2) \in \mathbb{R}^{n \times n}.$$

(Ce qui exprime aussi le fait que $(v_i)_{1 \leq i \leq n}$ est une base orthonormale de vecteurs propres de $A^{\mathsf{T}} A$, de valeurs propres associées $({\sigma_i}^2)_{1 \leq i \leq n}$.)
A partir de là, on obtient, pour $1 \leq i \leq n$, $1 \leq j \leq n$,

$$\begin{aligned}
(A^{\mathsf{T}} A)_{i,j} &= \sum_{k=1}^{n} (VD)_{i,k} V^{\mathsf{T}}_{k,j} \\
&= \sum_{k=1}^{q} (VD)_{i,k} V_{j,k} \text{ (pour } k > q \text{, la } k^{ème} \text{ colonne de } VD \text{ est composée de zéros)} \\
&= \sum_{k=1}^{q} {\sigma_k}^2 V_{i,k} V_{j,k} = \cdots = (\sum_{k=1}^{q} {\sigma_k}^2 v_k \cdot v_k^{\mathsf{T}})_{i,j}.
\end{aligned}$$

■

Avant de vérifier l'utilité pratique des deux identités ci-dessus, introduisons le **pseudo-inverse** de Σ : soit $\Sigma^{\dagger}$ la matrice de $\mathbb{R}^{n \times m}$ définie par

$$(\Sigma^{\dagger})_{i,j} = \begin{cases} \dfrac{1}{\sigma_i}, \ 1 \leq i, j \leq q, \ i = j \\ 0 \text{ sinon} \end{cases}.$$

On vérifie immédiatement que l'on a

$$\Sigma^{\dagger} \Sigma = \begin{cases} I_q & \text{si } q = n \\ \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} & \text{si } q < n \end{cases}.$$

A l'aide de la décomposition en valeurs singulières, nous pouvons maintenant définir le pseudo-inverse de A.

Définition 3.3.2 *On appelle* **pseudo-inverse** *de la matrice A de $\mathbb{R}^{m \times n}$ la matrice $A^{\dagger}$ de $\mathbb{R}^{n \times m}$ définie par*

$$A^{\dagger} = V \Sigma^{\dagger} W^{\mathsf{T}}.$$

A partir de là, on établit aisément les identités ci-dessous

Lemme 3.3.1

$$A^\dagger = \sum_{k=1}^{q} \frac{1}{\sigma_k} v_k \cdot w_k^\mathsf{T}, \quad AA^\dagger = \sum_{k=1}^{q} w_k \cdot w_k^\mathsf{T}, \quad A^\dagger A = \sum_{k=1}^{q} v_k \cdot v_k^\mathsf{T}.$$

Preuve : La démonstration de la première égalité est semblable à celle de la première identité enoncée pour A.

En ce concerne la deuxième égalité, on a

$$\begin{aligned} AA^\dagger &= \left(\sum_{k=1}^{q} \sigma_k w_k \cdot v_k^\mathsf{T} \right) \cdot \left(\sum_{l=1}^{q} \frac{1}{\sigma_l} v_l \cdot w_l^\mathsf{T} \right) = \sum_{k,l=1}^{q} \frac{\sigma_k}{\sigma_l} w_k \cdot v_k^\mathsf{T} \cdot v_l \cdot w_l^\mathsf{T} \\ &= \sum_{k,l=1}^{q} \frac{\sigma_k}{\sigma_l} w_k \cdot (v_k, v_l)_n \cdot w_l^\mathsf{T} = \sum_{k,l=1}^{q} \delta_{kl} \frac{\sigma_k}{\sigma_l} w_k \cdot w_l^\mathsf{T} = \sum_{k=1}^{q} w_k \cdot w_k^\mathsf{T}. \end{aligned}$$

La troisième et dernière égalité se démontre à l'identique. ■

On peut alors démontrer le résultat élégant ci-dessous.

Théorème 3.3.2 *Un point de minimum du problème de moindres carrés linéaires étudié précédemment est $A^\dagger b$.*

Preuve : On écrit simplement

$$\begin{aligned} A^\dagger b &= \left(\sum_{k=1}^{q} \frac{1}{\sigma_k} v_k \cdot w_k^\mathsf{T} \right) \cdot \left(\sum_{i=1}^{m} b_i w_i \right) = \sum_{k=1}^{q} \frac{1}{\sigma_k} v_k \cdot \left(\sum_{i=1}^{m} b_i w_k^\mathsf{T} \cdot w_i \right) \\ &= \sum_{k=1}^{q} \frac{b_k}{\sigma_k} v_k = \sum_{k=1}^{q} x_k^0 v_k = x^0. \end{aligned}$$

Or, x^0 appartient à l'ensemble des points de minimum, d'après (3.12). ■

Examinons, pour conclure ce chapitre, l'expression du pseudo-inverse dans certains cas particuliers.

Proposition 3.3.2 *Si $rg(A) = n$, on a la relation $A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$.*

Si $rg(A) = n = m$, on a la relation $A^\dagger = A^{-1}$.

Preuve : Supposons que $\mathrm{rg}(A) = n$. On a vu que le rang de A et celui de $A^\mathsf{T}A$ sont identiques (et égaux à q). Dès lors que $\mathrm{rg}(A^\mathsf{T}A) = n$, on peut inverser cette dernière. Cette constatation étant faite, on a les relations :

$$(A^\mathsf{T}A)A^\dagger = (A^\mathsf{T}A) \sum_{k=1}^{n} \frac{1}{\sigma_k} v_k \cdot w_k^\mathsf{T} = \sum_{k=1}^{n} \sigma_k v_k \cdot w_k^\mathsf{T} = A^\mathsf{T}.$$

On a utilisé le fait que les $(v_k)_{1 \le k \le n}$ sont les vecteurs propres de $A^\mathsf{T}A$, ainsi que la transposition de la première égalité de la proposition 3.3.1. Comme $A^\mathsf{T}A$ est inversible, la première égalité suit.

Supposons que $\mathrm{rg}(A) = n = m$. On se trouve ici dans le cas où A est une matrice inversible de $\mathbb{R}^{n \times n}$. D'après ce que l'on vient de prouver, on déduit

$$A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T} = A^{-1}(A^\mathsf{T})^{-1}A^\mathsf{T} = A^{-1}.$$

■

Chapitre 4

Algorithmes numériques de minimisation : fonctionnelles quadratiques

Dans ce chapitre, nous allons étudier des algorithmes qui permettent de calculer **numériquement** la solution du problème de minimisation,

$$\text{Trouver } u \in \mathbb{R}^n \text{ tel que } J_0(u) = \min_{v \in K^\star} J_0(v).$$

Ici, J_0 est la fonctionnelle (cf. (1.3)) qui à v associe $J_0(v) = \frac{1}{2}(Av, v) - (b, v)$. Nous supposons dans la suite que A est une matrice *symétrique définie positive* de $\mathbb{R}^{n\times n}$ et b un vecteur quelconque de $\mathbb{R}^n$. L'ensemble $K^\star$, quant à lui, est soit égal à $\mathbb{R}^n$ entier (problème sans contraintes), soit à un sous-espace affine de $\mathbb{R}^n$ (problème contraint), défini par

$$K^\star = \{v \in \mathbb{R}^n \ : \ Cv = f\},$$

où C appartient à $\mathbb{R}^{p\times n}$ est de rang p, et f est un élément de $\mathbb{R}^p$.
Nous avons vu, aux chapitres 1 et 2, que la solution d'un tel problème existe et est unique (sous l'hypothèse que A est symétrique définie positive), et que ceci revient à résoudre les systèmes linéaires

$$\text{(Problème sans contraintes)} \qquad Au = b, \text{ ou} \tag{4.1}$$

$$\text{(Problème contraint)} \qquad \begin{pmatrix} A & C^{\mathsf{T}} \\ C & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}. \tag{4.2}$$

Pour construire des algorithmes, nous choisissons un **vecteur initial** u_0, puis nous construisons une *suite* de vecteurs, notés $(u_k)_{k\geq 1}$. L'idée de cette méthode **itérative** de résolution est de s'assurer que la suite $(u_k)_k$ converge vers la solution u recherchée. Pour que de telles méthodes soient efficaces, il faut qu'elles possèdent les deux propriétés suivantes :

La convergence de la suite (u_k) est assurée, quel que soit le vecteur initial.

La convergence doit être "suffisamment rapide".

Le premier critère admet une interprétation claire, d'un point de vue mathématique. Le sens du second critère est plus flou, et nous essayerons de le préciser à la section ci-dessous.

Puis, pour les problèmes sans contraintes, nous explorerons l'idée suivante : pour itérer, c'est-à-dire calculer le vecteur u_{k+1} à partir du vecteur u_k, est-il possible de résoudre un problème

de minimisation "simple" ? Dans la suite, nous entendrons, par simple, la minimisation d'une fonction réelle d'une seule variable réelle.
Pour les problèmes avec contraintes, nous étudierons des méthodes qui permettent de se ramener à des problèmes sans contraintes, et donc d'utiliser les résultats obtenus pour celles-ci.

En particulier, les algorithmes que nous construisons ci-dessous permettent de résoudre les problèmes linéaires (4.1) et (4.2) en u, sous les conditions énoncées pour A.

Notons que nous nous concentrons à nouveau sur des problèmes fortement liés à l'algèbre linéaire. Ceci étant, nous indiquerons quelques extensions possibles au cours de l'exposé, lorsque la fonctionnelle n'est pas quadratique.

4.1 Critères 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,

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.

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

$$\|Av_\varepsilon - b\| \leq \varepsilon. \tag{4.3}$$

Il est clair que l'ensemble des v_ε qui satisfont à (4.3) 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 4.1.1 *Quel est l'ensemble défini par (4.3) ?*

Comme nous le verrons dans la Partie 2, 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 4.1.2 *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,

$$\text{Pour } k = 0, 1, \cdots, \textit{tant que } \|Au_k - b\| > \varepsilon \quad \text{itérer } u_k \to u_{k+1}.$$

(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 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, telles que celles qui seront étudiées dans la Partie 2. 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.

4.2 Algorithmes pour problèmes sans contraintes

4.2.1 Principe des méthodes étudiées

Comme mentionné dans la partie introductive, on considère ici des algorithmes basés sur les minimisations successives de fonctions réelles de la variable réelle. Comment ? Supposons l'itéré u_k connu : on choisit une direction, dite **de descente** , d_k, et l'on minimise

$$J_k \quad : \quad \rho \mapsto J_0(u_k + \rho d_k). \tag{4.4}$$

En d'autres termes, on minimise J_0 sur la droite passant par u_k, de direction d_k. Dans le cas qui nous intéresse, plutôt que d'avoir recours à une formule de composition des dérivées [6], telle

3. Typiquement, on peut prouver que $K \leq 7$ pour un calcul par éléments finis modélisant certains problèmes dans le plan.

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...

6. C'est parce que nous étudions un problème quadratique qu'il est possible de faire le raisonnement qui suit. Dans un cas plus général, il est nécessaire de calculer (formellement ou numériquement) le gradient de J_0 pour déterminer le minimum de la fonction J_k. Les problèmes inhérents à ce type de calcul ne sont pas étudiés ici ; ils ont donné naissance à une riche littérature, et sont entre autres abordés dans [10].

que (2.3), on remarque que

$$J_k(\rho) = \frac{\rho^2}{2}(Ad_k, d_k) + \rho(Au_k - b, d_k) + J_0(u_k).$$

C'est un polynôme de degré 2, avec un coefficient strictement positif pour le terme d'ordre 2. Il existe donc un point de minimum unique, ρ_k, caractérisé par $J_k'(\rho_k) = 0$, soit

$$\rho_k = \frac{(b - Au_k, d_k)}{(Ad_k, d_k)}. \tag{4.5}$$

On infère alors immédiatement la caractérisation de u_{k+1}, i.e.

$$u_{k+1} = u_k + \frac{(b - Au_k, d_k)}{(Ad_k, d_k)} d_k. \tag{4.6}$$

Remarquons, avant de poursuivre que, si u_k est égal à u, la solution cherchée, on a $b - Au_k = 0$, ce qui entraîne en particulier que $u_{k+1} = u$. Bref, la suite est stationnaire.
A partir de là, nous sommes en mesure de décrire quelques méthodes numériques de minimisation.

4.2.2 Relaxation

Pour définir la méthode de **relaxation**, une base orthonormale $(e_i)_{1 \le i \le n}$ de $\mathbb{R}^n$ étant donnée, on choisit la suite de directions de descente $d_0 = e_1,\ d_1 = e_2, \cdots$; si l'algorithme n'a pas convergé après n itérations (supposition raisonnable !), on prend $d_n = e_1,\ d_{n+1} = e_2$ et ainsi de suite... Dans ce cas, (4.5) et (4.6) deviennent

$$\begin{aligned} &\text{pour} l \ge 0,\ i \in \{1, \cdots, n\} \quad (k = ln + i - 1) \\ &\qquad \rho_{ln+i-1} = \frac{(b - Au_{ln+i-1}, e_i)}{(Ae_i, e_i)}, \quad u_{ln+i} = u_{ln+i-1} + \rho_{ln+i-1} e_i. \end{aligned} \tag{4.7}$$

Entre les deux itérés successifs u_{ln+i+1} et u_{ln+i}, on en déduit que seule la $i^{ème}$ composante diffère. Comme seule une composante (sur n) évolue, il est raisonnable d'introduire la suite $(\tilde{u}_l)_{l \ge 0}$, avec

$$\begin{cases} \tilde{u}_0 = u_0, & \\ \tilde{u}_1 = u_n, & \text{le résultat des } n \text{ premières itérations,} \\ \tilde{u}_2 = u_{2n}, & \text{le résultat des } n \text{ suivantes, etc.} \end{cases}$$

Ainsi, toutes les composantes de $\tilde{u}_{l+1}$ sont *a priori* distinctes de celles de $\tilde{u}_l$. De plus, par construction, chaque composante est mise à jour une fois et une seule. Plus précisément, on a vu que la $i^{ème}$ composante est modifiée lorsque l'on considère la direction de descente e_i, ce qui donne, d'après (4.7) :

$$(\tilde{u}_{l+1} - \tilde{u}_l, e_i) = \rho_{ln+i-1},\ \text{et}\ \|\tilde{u}_{l+1} - \tilde{u}_l\|^2 = \sum_{i=1}^{n} {\rho_{ln+i-1}}^2 = \sum_{i=1}^{n} \|u_{ln+i} - u_{ln+i-1}\|^2. \tag{4.8}$$

Ces expressions seront fort utiles pour démontrer la proposition 4.2.1 ci-dessous. Avant de l'aborder, établissons le

Lemme 4.2.1 *Soit A une matrice symétrique, et λ_{min} et λ_{max} ses plus petite et plus grande valeurs propres. Alors*

$$\forall v \in \mathbb{R}^n, \lambda_{min}\|v\|^2 \le (Av, v) \le \lambda_{max}\|v\|^2\ ; \tag{4.9}$$

$$\textit{si de plus } A \textit{ est positive}, \lambda_{min}\|v\| \le \|Av\| \le \lambda_{max}\|v\|. \tag{4.10}$$

Preuve : On sait qu'il existe une base orthonormale de vecteurs propres de A ; notons-la $(p_i)_{1\leq i\leq n}$. On pose $v=\sum_{i=1}^n v_i p_i$, et l'on effectue

$$(Av,v)=(\sum_{i=1}^n v_i Ap_i,\sum_{j=1}^n v_j p_j)=(\sum_{i=1}^n \lambda_i v_i p_i,\sum_{j=1}^n v_j p_j)=\sum_{i=1}^n \lambda_i {v_i}^2.$$

On en déduit alors (4.9). En effet :

$$\lambda_{min}\|v\|^2=\lambda_{min}\sum_{i=1}^n {v_i}^2\leq (Av,v)\leq \lambda_{max}\sum_{i=1}^n {v_i}^2=\lambda_{max}\|v\|^2.$$

Comme

$$\|Av\|^2=\sum_{i=1}^n {\lambda_i}^2{v_i}^2,$$

et on déduit de même (4.10), car

$${\lambda_{min}}^2\|v\|^2={\lambda_{min}}^2\sum_{i=1}^n {v_i}^2\leq \|Av\|^2\leq {\lambda_{max}}^2\sum_{i=1}^n {v_i}^2={\lambda_{max}}^2\|v\|^2.$$

■

Proposition 4.2.1 *La méthode de relaxation est convergente.*

Preuve : Etape 1. Commençons par borner $\|\tilde{u}_{l+1}-\tilde{u}_l\|$. Pour cela, on remarque que[7]

$$J(u_k)-J(u_{k+1})=J_k(0)-J_k(\rho_k)=\frac{(Ae_k,e_k)}{2}{\rho_k}^2\geq\frac{\lambda_{min}}{2}{\rho_k}^2=\frac{\lambda_{min}}{2}\|u_k-u_{k+1}\|^2.$$

En conséquence, pour la suite $(\tilde{u}_l)_l$, on arrive à la minoration :

$$\begin{aligned}J(\tilde{u}_l)-J(\tilde{u}_{l+1})&=J(u_{ln})-J(u_{l(n+1)})=\sum_{i=1}^n J(u_{ln+i-1})-J(u_{ln+i})\\&\geq\frac{\lambda_{min}}{2}\sum_{i=1}^n\|u_{ln+i-1}-u_{ln+i}\|^2=\frac{\lambda_{min}}{2}\|\tilde{u}_{l+1}-\tilde{u}_l\|^2.\end{aligned}$$

Par construction, la suite $(J(\tilde{u}_l))_l$ est décroissante et minorée. En conséquence, la différence de deux termes successifs $|J(\tilde{u}_l)-J(\tilde{u}_{l+1})|$ tend vers 0 lorsque l tend vers l'infini. D'après la majoration ci-dessus, on obtient $\lim_{l\to+\infty}\|\tilde{u}_{l+1}-\tilde{u}_l\|=0$. De l'imbrication des suites $(u_k)_k$ et $(\tilde{u}_l)_l$, on en déduit également

$$\lim_{l\to+\infty}\|u_{ln+i}-\tilde{u}_l\|=0,\ \text{pour chaque } i\in\{1,\cdots,n\}. \tag{4.11}$$

Etape 2. Convergence de $(\tilde{u}_l)_l$. Reprenons maintenant (4.9), avec comme vecteur-test $v=\tilde{u}_l-u$, et utilisons l'inégalité de Cauchy-Schwartz

$$\lambda_{min}\|\tilde{u}_l-u\|^2\leq(A(\tilde{u}_l-u),\tilde{u}_l-u)=(A\tilde{u}_l-b,\tilde{u}_l-u)\leq\|A(\tilde{u}_l-u)\|\,\|\tilde{u}_l-u\|.$$

Ainsi, en simplifiant par $\|\tilde{u}_l-u\|$, on trouve

$$\lambda_{min}\|\tilde{u}_l-u\|\leq\|A(\tilde{u}_l-u)\|. \tag{4.12}$$

7. $J(\rho)=\alpha\rho^2+\beta\rho+\gamma$, $\alpha>0$. On a $\rho_{min}=-\frac{\beta}{2\alpha}$, d'où $J(0)-J(\rho_{min})=\frac{\beta^2}{4\alpha}=\alpha{\rho_{min}}^2$.

Que vaut le terme de droite?

$$\|A(\tilde{u}_l - u)\|^2 = \sum_{i=1}^{n} (A\tilde{u}_l - b)_i{}^2 = \sum_{i=1}^{n} (A\tilde{u}_l - b, e_i)^2.$$

Comment faire usage ce qui précède? Revenons aux définitions (4.5)-(4.6). D'après (2.3) (qui est finalement utile ici...)

$$0 = J_k'(\rho_k) = (\nabla J_0(u_k + \rho_k d_k), d_k) = (\nabla J_0(u_{k+1}), d_k) = (Au_{k+1} - b, d_k).$$

Pour $k = ln + i - 1$, on trouve $0 = (Au_{ln+i} - b, e_i)$, soit $(b, e_i) = (Au_{ln+i}, e_i)$. Nous pouvons donc transformer l'expression du terme de droite de (4.12) en

$$\left\{ \sum_{i=1}^{n} (A(\tilde{u}_l - u_{ln+i}), e_i)^2 \right\}^{1/2}.$$

Il nous reste maintenant à utiliser (4.11), ainsi que (4.10). Commençons par

$$\sum_{i=1}^{n} (A(\tilde{u}_l - u_{ln+i}), e_i)^2 \leq \sum_{i=1}^{n} \|A(\tilde{u}_l - u_{ln+i})\|^2 \leq \lambda_{max}{}^2 \sum_{i=1}^{n} \|\tilde{u}_l - u_{ln+i}\|^2, \text{ d'où}$$

$$\|\tilde{u}_l - u\| \leq \frac{\lambda_{max}}{\lambda_{min}} \left(\sum_{i=1}^{n} \|\tilde{u}_l - u_{ln+i}\|^2 \right)^{1/2}. \tag{4.13}$$

Lorsque l tend vers l'infini, chaque terme de la somme tend vers 0. Par ailleurs, le nombre de termes est borné indépendamment de l. On arrive donc finalement à

$$\lim_{l \to +\infty} \|\tilde{u}_l - u\| = 0. \tag{4.14}$$

Etape 3. Convergence de $(u_k)_k$. Nous venons donc de prouver la convergence de $(\tilde{u}_l)_l$ vers u. Bien évidemment, $(u_k)_k$ converge également vers u. En effet,

$$\|u_k - u\| \leq \|u_k - \tilde{u}_l\| + \|\tilde{u}_l - u\|, \text{ avec } l = E(k/n),$$

et (4.11) et (4.14) permettent de conclure! ∎

Exercice 4.2.1 *Le but de cet exercice est de montrer que l'algorithme de relaxation correspond à la méthode itérative de* **Gauss-Seidel**, *de résolution d'un système linéaire. Cette méthode est considérée en détail dans la Partie 2. On note $(u_k{}^j)_{1 \leq j \leq n}$ les composantes du vecteur u_k.*
1. Prouver que l'on peut écrire (4.7) sous la forme

$$A_{i,i} u_{k+1}{}^i = b_i - \sum_{j \neq i} A_{i,j} u_k{}^j, \text{ pour } i \text{ tel que } k = ln + i - 1.$$

2. On découpe A en trois parties : $A = D - E - F$, avec

la partie diagonale : $D_{i,i} = A_{i,i}$, $1 \leq i \leq n$, $D_{i,j} = 0$ sinon ;

la partie triangulaire inférieure : $E_{i,j} = -A_{i,j}$, $1 \leq j < i \leq n$, $E_{i,j} = 0$ sinon ;

la partie triangulaire supérieure : $F_{i,j} = -A_{i,j}$, $1 \leq i < j \leq n$, $F_{i,j} = 0$ sinon.

On revient aux itérés $\tilde{u}_l$, c'est-à-dire ceux dont chaque composante est mise à jour une fois et une seule par itération. Montrer que

$$(D - E)\tilde{u}_{l+1} = b + F\tilde{u}_l.$$

4.2.3 Gradient à pas fixe, à pas optimal

Cette catégorie de méthodes a été conçue à partir de la réponse à la question suivante : dans quelle direction diminue-t-on le plus la valeur d'une fonctionnelle ? Ou, en termes plus mathématiques, si on pose

$$w_\varepsilon = u + \varepsilon d, \text{ avec } d \in \mathbb{R}^n, \ \|d\| = 1, \ \varepsilon > 0,$$

comment maximiser la différence $J(u) - J(w_\varepsilon)$? Pour cela, cf. (14.6), on écrit

$$J(u) - J(w_\varepsilon) = -\varepsilon(\nabla J(u), d) + o(\varepsilon).$$

Lorsque ε est petit, la différence se comporte comme $-\varepsilon(\nabla J(u), d)$ (si $\nabla J(u) \neq 0$), c'est-à-dire qu'elle est maximale pour

$$d = -\frac{\nabla J(u)}{\|\nabla J(u)\|}.$$

L'opposé de la direction du gradient est une direction *privilégiée*. Dans un premier temps, nous allons donc considérer (4.5)-(4.6) avec

$$d_k = -\nabla J_0(u_k) = b - Au_k, \quad \rho_k = \frac{\|d_k\|^2}{(Ad_k, d_k)}. \tag{4.15}$$

Cette méthode est appelée **méthode du gradient à pas optimal**. Notons dès maintenant que, d'après (2.3), on a la propriété

$$0 = J_k'(\rho_k) = (\nabla J_0(u_{k+1}), d_k) = -(d_{k+1}, d_k).$$

En clair, deux directions *consécutives* de descente sont orthogonales. Qui plus est, une itération permet *a priori* la modification de toutes les composantes de l'itéré (d_k n'a pas de raison d'être parallèle à l'un des vecteurs de base, ou d'être combinaison d'une partie d'entre eux uniquement).

Proposition 4.2.2 *La méthode de gradient à pas optimal est convergente.*

Preuve : Elle est notablement plus simple que celle prouvant la convergence de la méthode de relaxation.

Etape 1. Majorons pour commencer la norme $\|u_k - u\|$:

$$\lambda_{min}\|u_k - u\|^2 \leq (A(u_k - u), u_k - u) = (Au_k - b, u_k - u) = -(d_k, u_k - u) \leq \|d_k\| \, \|u_k - u\|.$$

On infère

$$\|u_k - u\| \leq \frac{1}{\lambda_{min}}\|d_k\|.$$

Etape 2. On utilise maintenant *essentiellement* l'orthogonalité entre deux directions consécutives de descente. En effet, cf. (4.10),

$$\|d_k\|^2 = (d_k - d_{k+1}, d_k) = (A(u_{k+1} - u_k), d_k) \leq \|A(u_{k+1} - u_k)\| \, \|d_k\| \leq \lambda_{max}\|u_{k+1} - u_k\| \, \|d_k\|.$$

Ainsi

$$\|d_k\| \leq \lambda_{max}\|u_{k+1} - u_k\|.$$

On arrive alors à la majoration

$$\|u_k - u\| \leq \frac{\lambda_{max}}{\lambda_{min}}\|u_{k+1} - u_k\|.$$

La convergence de $(u_k)_k$ vers u découle donc de la propriété $\lim_{k\to+\infty} \|u_{k+1} - u_k\| = 0$.

Etape 3. $J(u_k) - J(u_{k+1}) = \frac{(Ad_k,d_k)}{2}{\rho_k}^2$. Or, l'égalité $u_{k+1} - u_k = \rho_k d_k$ implique (si $d_k \neq 0$), $|\rho_k| = \frac{\|u_{k+1}-u_k\|}{\|d_k\|}$. Ainsi

$$J(u_k) - J(u_{k+1}) = \frac{(Ad_k, d_k)}{2\|d_k\|^2}\|u_{k+1} - u_k\|^2 \geq \frac{\lambda_{min}}{2}\|u_k - u_{k+1}\|^2.$$

Comme $(J(u_k)_k)$ est décroissante et minorée, la différence de deux termes consécutifs tend vers 0, ce qui implique la convergence de $(u_k)_k$ vers u, puisque

$$\|u_k - u\| \leq \sqrt{2}\frac{\lambda_{max}}{\lambda_{min}^{3/2}}(J(u_{k+1}) - J(u_k))^{1/2}. \tag{4.16}$$

■

NB. La majoration (4.16) fournit une borne explicite de la norme de l'erreur.

La **méthode du gradient à pas fixe** consiste à s'affranchir du calcul du minimum ρ_k. De fait, on fixe, pour tout k, la valeur de ce paramètre à $\rho > 0$, c'est-à-dire que (4.5)-(4.6) deviennent

$$\rho_k = \rho, \quad u_{k+1} = u_k + \rho(b - Au_k). \tag{4.17}$$

NB. Dans le cas de la fonctionnelle J_0, on a vu que le calcul de la valeur optimale ρ_k ne présente aucune difficulté. Il peut en être tout autrement dans le cas général (voir la note de bas de page [6]) !

Proposition 4.2.3 *La méthode de gradient à pas fixe est convergente, sous réserve qu'il existe a et b vérifiant*

$$0 < a \leq \rho \leq b < \frac{2}{\lambda_{max}}.$$

Preuve : Que vaut l'erreur ?

$$u_{k+1} - u = u_k + \rho(b - Au_k) - u = (I_n - \rho A)u_k + \rho Au - u = (I_n - \rho A)(u_k - u).$$

Nous allons majorer la norme de l'erreur à l'itération $k+1$ en fonction de celle de l'itération k, grâce à la relation ci-dessus, en reprenant la démonstration de (4.10) :

$$\begin{aligned}
(I_n - \rho A)v &= \sum_{i=1}^{n}(I_n - \rho A)v_i p_i = \sum_{i=1}^{n}(1 - \rho\lambda_i)v_i p_i \ ; \\
\|(I_n - \rho A)v\|^2 &= (\sum_{i=1}^{n}(1 - \rho\lambda_i)v_i p_i, \sum_{j=1}^{n}(1 - \rho\lambda_j)v_j p_j) = \sum_{i=1}^{n}(1 - \rho\lambda_i)^2 {v_i}^2 \\
&\leq \max_i(1 - \rho\lambda_i)^2 \sum_{j=1}^{n} {v_j}^2 = \left\{\max_i |1 - \rho\lambda_i|\right\}^2 \|v\|^2.
\end{aligned}$$

En regroupant les deux résultats, on trouve

$$\|u_{k+1} - u\| \leq \max_i |1 - \rho\lambda_i| \, \|u_k - u\|.$$

Si on note $\gamma_\rho = \max_i |1 - \rho\lambda_i|$, on a obtenu $\|u_{k+1} - u\| \leq \gamma_\rho \|u_k - u\|$. Par récurrence, on en déduit

$$\|u_k - u\| \leq {\gamma_\rho}^k \|u_0 - u\|. \tag{4.18}$$

Si γ_ρ est strictement plus petit que 1, on aura démontré la convergence... C'est ce que nous allons vérifier maintenant.

$$\begin{aligned} & \lambda_{min} \leq \lambda_i \leq \lambda_{max},\ 1 \leq i \leq n \\ \Longrightarrow\ & 1 - \rho\lambda_{min} \geq 1 - \rho\lambda_i \geq 1 - \rho\lambda_{max},\ 1 \leq i \leq n \\ \Longrightarrow\ & |1 - \rho\lambda_i| \leq \max(|1 - \rho\lambda_{min}|, |1 - \rho\lambda_{max}|),\ 1 \leq i \leq n. \end{aligned}$$

Puisque les bornes sur les valeurs propres λ_{min} et λ_{max} sont atteintes,

$$\gamma_\rho = \max(|1 - \rho\lambda_{min}|, |1 - \rho\lambda_{max}|). \tag{4.19}$$

Pour conclure, nous majorons γ_ρ, à l'aide des hypothèses sur A (positivité), ρ, a et b :

$$1 > 1 - a\lambda_{min} \geq \left\{ \begin{array}{c} 1 - \rho\lambda_{min} \\ 1 - \rho\lambda_{max} \end{array} \right\} \geq 1 - b\lambda_{max} > -1.$$

Si enfin on pose $\gamma_{a,b} = \max(|1 - a\lambda_{min}|, |1 - b\lambda_{max}|)$, on vient de prouver que

$$\gamma_\rho \leq \gamma_{a,b},\ \text{et}\ \gamma_{a,b} < 1.$$

■

A partir de ce résultat, on constate que, pour appliquer la méthode du gradient à pas fixe, il faut connaître la valeur propre λ_{max} ou, au moins, une estimation de cette dernière.

4.2.4 Gradient conjugué

Nous avons remarqué que pour la méthode du gradient à pas optimal, deux directions successives de descente sont orthogonales. Par contre, il n'y a aucune raison pour que trois (ou plus) directions de descente soient orthogonales entre elles. Ainsi, on peut "revenir" dans des directions déjà explorées... Le principe de la **méthode du/des gradient(s) conjugué(s)** est de construire une suite de *directions de recherche* que l'on garde en mémoire, pour essayer d'éviter les retours. Pour cela, si $u_1, \ldots, u_k$ ont déjà été calculés, et si tous les gradients $g_l = \nabla J_0(u_l)$, $0 \leq l \leq k$ sont non nuls, on cherche u_{k+1} tel que

$$J_0(u_{k+1}) = \min_{v \in u_k + G_k} J_0(v),\ \text{avec}\ G_k = Vect(g_0, g_1, \cdots, g_k).$$

NB. Pour respecter l'esprit des méthodes de gradient, on conserve les directions "optimales", c'est-à-dire les gradients de J_0 aux itérés successifs.

Le principe de la méthode est très attrayant, puisqu'on espère éviter les redondances dans le choix de la direction de descente, i. e. on espère que

$$dim(G_{k+1}) = dim(G_k) + 1,\ k = 0, 1, \cdots.$$

Bien sûr, ceci n'est nullement garanti (il faut et il suffit que $g_{k+1} \notin G_k$)... Par ailleurs, la construction de la suite des espaces vectoriels $(G_k)_{k \geq 0}$, et la résolution des problèmes posés sur $u_k + G_k$, semblent très couteuses, puisqu'on doit gérer des espaces vectoriels dont la dimension peut fort bien devenir comparable à n. Heureusement, et c'est la "magie" de la méthode du gradient conjugué, nous allons vérifier qu'aucun de ces deux problèmes ennuyeux n'en est un !

Tout d'abord, d'après (2.9), g_{k+1} est orthogonale à G_k ; ceci signifie en particulier que

$$(g_{k+1}, g_l) = 0,\ 0 \leq l \leq k.$$

Par récurrence, on infère facilement que

$$(g_i, g_j) = 0,\ 0 \leq i < j \leq k+1. \tag{4.20}$$

Les gradients sont tous orthogonaux entre eux. Ce qui, encore une fois, est beaucoup plus intéressant que la propriété d'orthogonalité de deux gradients successifs de la méthode à pas optimal. On obtient alors la

Proposition 4.2.4 *La méthode du gradient conjugué converge en n itérations au plus.*

Preuve : Au bout de $n-1$ itérations, si aucun des $\nabla J_0(u_k) = g_k$, $0 \leq k \leq n-1$, ne s'annule, on a construit une famille libre de $\mathbb{R}^n$ à n éléments ; ou, en d'autres termes, une base de $\mathbb{R}^n$! Comme $g_n = \nabla J_0(u_n)$ est orthogonal à ces n vecteurs, il est nécessairement nul, ce qui signifie que $u_n = u$. ■

Etudions maintenant les aspects *pratiques* de l'algorithme, et notamment la gestion des espaces $(G_k)_k$. Commençons par la propriété suivante, qui porte sur les directions de descentes $(\delta_k)_k$, $\delta_k = u_{k+1} - u_k$.

Lemme 4.2.2 *Les directions $(\delta_k)_k$ sont telles que*

$$(A\delta_i, \delta_j) = 0,\ \textit{pour } i \neq j. \tag{4.21}$$

Preuve : D'après l'expression du gradient de J_0: $g_{k+1} = Au_{k+1} - b = Au_k - b + A\delta_k = g_k + A\delta_k$. D'après (4.20), pour l compris entre 0 et k,

$$0 = (g_{k+1}, g_l) = (g_k, g_l) + (A\delta_k, g_l).$$

Soit, puisque g_k est lui-même orthogonal à g_l dès lors que l est différent de k,

$$0 = (A\delta_k, g_l),\ 0 \leq l \leq k-1.$$

Or, par définition de la méthode du gradient conjugué, $\delta_l = u_{l+1} - u_l$ appartient à $G_l \subset G_{k-1}$, pour l variant de 0 à $k-1$. En d'autres termes, δ_l est une combinaison linéaire de $(g_m)_{0 \leq m \leq k-1}$, ce qui, combiné à l'égalité ci-dessus, permet d'obtenir

$$0 = (A\delta_k, \delta_l),\ 0 \leq l \leq k-1.$$

■

Définition 4.2.1 *On dit que des directions (non nulles) $(\delta_k)_k$ vérifiant (4.21) sont* **conjuguées** *par rapport à la matrice A.*

Bien sûr, si les vecteurs $(\delta_k)_k$ sont tous non nuls, la relation (4.21) implique que la famille $(\delta_k)_k$ est libre, puisque la forme bilinéaire $(\cdot,\cdot)_A : (x,y) \mapsto (Ax, y)$ est un produit scalaire, que l'on peut réécrire sous la forme

$$(x,y)_A = \sum_{i=1}^{n} \lambda_i x_i y_i,\ x = \sum_{i=1}^{n} x_i p_i,\ y = \sum_{i=1}^{n} y_i p_i.$$

Par construction, $Vect(\delta_0, \cdots, \delta_k) = G_k$: ainsi, connaître G_k équivaut à connaître $(\delta_l)_{0 \leq l \leq k}$! Par ailleurs, on a $\delta_k \in G_k$, et $\delta_k \notin G_{k-1}$. On peut donc écrire la décomposition

$$\delta_k = \sum_{l=0}^{k} \beta_l^k g_l,\ \text{avec } \beta_k^k \neq 0.$$

En particulier, si $k = 0$, on a la relation

$$\delta_0 = \beta_0^0 g_0. \tag{4.22}$$

Soit maintenant $k \geq 1$; d'après (4.21) et (4.20), pour m compris entre 0 et $k-1$:

$$\begin{aligned} 0 &= (A\delta_k, \delta_m) = (\delta_k, A\delta_m) = (\delta_k, A(u_{m+1} - u_m)) = (\delta_k, g_{m+1} - g_m) \\ &= \sum_{l=0}^{k} \beta_l^k (g_l, g_{m+1} - g_m) = \beta_{m+1}^k \|g_{m+1}\|^2 - \beta_m^k \|g_m\|^2. \\ \Longrightarrow \beta_m^k &= \beta_{m+1}^k \frac{\|g_{m+1}\|^2}{\|g_m\|^2},\ m = k-1, \cdots, 0, \\ \Longrightarrow \beta_m^k &= \beta_k^k \frac{\|g_k\|^2}{\|g_m\|^2},\ m = k, \cdots, 0. \end{aligned}$$

On a donc

$$\delta_k = \beta_k^k \left\{ \sum_{l=0}^{k} \frac{\|g_k\|^2}{\|g_l\|^2} g_l \right\}.$$

Intégrons, dans cette expression, celle de δ_{k-1}.

$$\delta_k = \beta_k^k \left\{ \sum_{l=0}^{k-1} \frac{\|g_k\|^2}{\|g_l\|^2} g_l + g_k \right\} = \beta_k^k \left\{ \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \sum_{l=0}^{k-1} \frac{\|g_{k-1}\|^2}{\|g_l\|^2} g_l + g_k \right\},$$

$$\text{soit } \delta_k = \beta_k^k \left\{ \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \frac{1}{\beta_{k-1}^{k-1}} \delta_{k-1} + g_k \right\},\ \text{pour } k \geq 1. \tag{4.23}$$

Le calcul des directions $(\delta_k)_k$ est donc particulièrement simple, à l'aide de la récurrence (4.22)-(4.23).

Pour revenir aux algorithmes avec directions de descente, notons que l'on peut définir $d_k = -\dfrac{1}{\beta_k^k}\delta_k$, pour $k \geq 0$. Les relations (4.22)-(4.23) deviennent

$$d_0 = -g_0, \quad d_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} d_{k-1} - g_k,\ \text{pour } k \geq 1. \tag{4.24}$$

L'expression est encore un peu plus simple. Qui plus est, et c'est là une des propriétés remarquables de la méthode du gradient conjugué, regardons ce qu'il advient si l'on minimise la fonctionnelle J_0 sur la droite passant par u_k de direction d_k :

$\min_{\rho \in \mathbb{R}} J_0(u_k + \rho d_k) \geq \min_{v \in u_k + G_k} J_0(v) = J_0(u_{k+1})$, puisque $d_k \in G_k$.

u_{k+1} se trouve sur la droite passant par u_k de direction d_k ($u_{k+1} = u_k - \beta_k^k d_k$).

En d'autres termes, le minimum est bien atteint en $u_{k+1} = u_k + \rho_k d_k$, avec, d'après (4.5),

$$\rho_k = \frac{(b - Au_k, d_k)}{(Ad_k, d_k)} = -\beta_k^k.$$

La boucle est bouclée :

La méthode du gradient conjugué s'apparente à une méthode de descente.

Son nom vient des directions conjuguées par rapport à A car, d'après (4.21), les directions de descente $(d_k)_k$ sont *conjuguées* par rapport à A. Récapitulons. Une fois ε et u_0 définis, l'algorithme du gradient conjugué consiste en

$$
\left\|\begin{array}{l}
\textbf{Initialisation :}\ k=0 \\
\quad \text{Choisir } u_0. \\
\quad \text{Calculer } g_0 = Au_0 - b\ ;\ d_0 = -g_0. \\
\textbf{Tant que } \|g_k\| > \varepsilon,\ \textbf{itérer } k = 1, 2, \cdots \\
\text{(a) approximation de la solution} \\
\quad \text{Calculer } \beta_{k-1} = \dfrac{(g_{k-1}, d_{k-1})}{(Ad_{k-1}, d_{k-1})}\ ;\ u_k = u_{k-1} - \beta_{k-1} d_{k-1}. \\
\text{(b) détermination de la nouvelle direction} \\
\quad \text{Calculer } g_k = Au_k - b\ ;\ \alpha_k = \dfrac{\|g_k\|^2}{\|g_{k-1}\|^2}\ ;\ d_k = \alpha_k d_{k-1} - g_k.
\end{array}\right. \tag{4.25}
$$

NB. On verra, dans la Partie 2, que l'on peut établir des résultats portant sur la vitesse de convergence de la méthode, en norme $\|\cdot\|_A$.

Les propriétés fondamentales de l'algorithme ci-dessus sont au nombre de deux :

+ La minimisation est effective sur un sous-espace vectoriel dont la dimension croît à chaque itération. En conséquence, la solution u est calculée en n itérations au plus (aux erreurs de calcul près !).

+ Les directions $(d_k)_k$ sont faciles à calculer, et il suffit d'en conserver deux en mémoire à tout instant de l'algorithme.

Pour accélérer la vitesse de convergence, on peut *préconditionner* le système linéaire, ce qui conduit en pratique à une réduction notable du nombre d'itérations.
♠ Cependant, comme indiqué à la section 4.1, il faut veiller à ne pas trop augmenter le coût de calcul par itération... Ce type de considération a donné naissance à une littérature considérable (citons notamment [11, 17]).

4.2.5 Extensions

Il est prouvé dans [6] que les méthodes de relaxation et de gradient à pas optimal, ou à pas fixe, sont applicables dans un espace de Hilbert, sous réserve que la fonctionnelle J vérifie certaines propriétés. En clair, la fonctionnelle J_0 n'est qu'un cas très particulier, mais fort utile, puisqu'elle permet de construire des méthodes de résolution de systèmes linéaires. Précisément, si la fonctionnelle J est $\mathcal{C}^1$ et α-convexe, avec une différentielle lipschitzienne, les méthodes convergent. Ceci signifie :

J α-convexe : cf. remarque 2.3.4, points (viii) et (ix).

dJ lipschitzienne : $\exists M > 0,\ \forall u, v \in \mathbb{E},\ \|dJ(u) - dJ(v)\| \leq M\|u - v\|$.

Exercice 4.2.2 *Retrouver les résultats des propositions 4.2.1 et 4.2.2 sous réserve que la fonctionnelle vérifie les hypothèses ci-dessus.* [8]

8. La démonstration de la généralisation de la proposition 4.2.3 est beaucoup plus ardue, notamment parce qu'elle est complètement différente de celle utilisée dans le cas de la fonctionnelle J_0.

Pour ce qui est de la méthode du gradient conjugué, certaines adaptations sont également possibles, dans le cas d'une fonctionnelle plus générale, sous réserve toutefois de modifications de l'algorithme (cf. [21], et [10] pour une discussion détaillée.)
Si maintenant on considère la résolution d'un système linéaire, dont la matrice n'est pas symétrique, il est *impossible* de conserver à la fois les deux propriétés remarquables de l'algorithme du gradient conjugué, à savoir la convergence en n itérations au plus, associée à l'utilisation de récurrences de taille constante (voir [9]) ! Il faut, au choix, soit conserver toutes les directions précédentes de descente, ce qui accroît notablement le coût calcul, soit ne garder que les p (pour p fixé, petit devant n) dernières directions, et raisonner dans

$$u_k + Vect(g_k, \cdots, g_{k-p+1}).$$

Nous renvoyons le lecteur intéressé à [24, 17], article dans lequel la méthode GMRES[9] a été introduite pour résoudre des systèmes linéaires, de matrice non symétrique.

4.3 Algorithmes pour problèmes contraints

De façon similaire, les méthodes de résolution des problèmes contraints sont très nombreuses. Nous allons en présenter trois, qui conduisent à des algorithmes numériques utilisables en pratique. Nous débutons par deux méthodes purement algébriques, avant de revenir à une technique de minimisation.

4.3.1 Elimination des contraintes

Soit donc le système linéaire (4.2) à résoudre, de solution (u, λ) appartenant à $\mathbb{R}^n \times \mathbb{R}^p$:

$$\begin{cases} Au + C^{\mathsf{T}}\lambda = b \\ Cu = f \end{cases}.$$

L'inconnue qui nous intéresse est u. Eliminons λ : $u = A^{-1}(b - C^{\mathsf{T}}\lambda)$, ce qui implique la relation $CA^{-1}b - CA^{-1}C^{\mathsf{T}}\lambda = f$. En d'autres termes, λ est la solution de

$$\text{Trouver } \lambda \in \mathbb{R}^p \text{ tel que } CA^{-1}C^{\mathsf{T}}\lambda = CA^{-1}b - f. \tag{4.26}$$

Si p est très petit devant n, la difficulté est la formation de la matrice $CA^{-1}C^{\mathsf{T}}$ de $\mathbb{R}^{p\times p}$, et du second membre $CA^{-1}b$ appartenant à $\mathbb{R}^p$. En effet, une fois ceux-ci connus, il est raisonnable de supposer que la résolution de (4.26) sera aisée (qui plus est, $CA^{-1}C^{\mathsf{T}}$ est symétrique définie positive). A partir de là, u est solution de

$$\text{Trouver } u \in \mathbb{R}^n \text{ tel que } Au = b - C^{\mathsf{T}}\lambda, \tag{4.27}$$

et l'on en revient aux méthodes de la section précédente. Pour ce qui est de la formation de $CA^{-1}C^{\mathsf{T}}$, notons que l'on peut écrire

$$CA^{-1}C^{\mathsf{T}} = CC', \text{ avec } C' = A^{-1}C^{\mathsf{T}} \in \mathbb{R}^{n\times p}.$$

C' est caractérisée par

$$\text{Trouver } C' \in \mathbb{R}^{n\times p} \text{ telle que } AC' = C^{\mathsf{T}}. \tag{4.28}$$

Ce système linéaire peut être reformulé *colonne par colonne*. En effet, si on note $(c'_i)_{1\le i\le p}$ les colonnes de C' et $(\underline{c}_i)_{1\le i\le p}$ celles de C^{T}, (4.28) est équivalent à

$$\text{Pour } i = 1, \cdots, p, \text{ trouver } c'_i \in \mathbb{R}^n \text{ tel que } Ac'_i = \underline{c}_i. \tag{4.29}$$

9. Algorithme GMRES : **G**eneralized **M**inimum **RES**idual algorithm.

L'obtention de $CA^{-1}C^{\mathsf{T}}$ est alors immédiate, par simple multiplication. Pour ce qui est du calcul de $A^{-1}b$, on procède de façon similaire, en résolvant

$$\text{Trouver } c_{p+1} \in \mathbb{R}^n \text{ tel que } Ac_{p+1} = b, \tag{4.30}$$

puis en construisant $CA^{-1}b$, résultat de la multiplication de C par c_{p+1}.

De cette façon, on a démontré la

Proposition 4.3.1 *On peut ramener le calcul de (u, λ), solution de (4.2), à la résolution de $p+2$ problèmes sans contraintes, de type (4.1).*

Preuve : Il suffit de résoudre (4.29)-(4.30), soit $p+1$ problèmes, puis (4.26), dont le coût est supposé "faible", et enfin (4.27). ∎

Cette méthode présente l'avantage d'être complètement compatible avec les algorithmes proposés à la section 4.2, puisque l'on a uniquement des problèmes sans contraintes à résoudre. En outre, elle est particulièrement indiquée si p est petit... Si p est grand, la même technique n'en reste pas moins valable, sachant que l'étape (4.26) peut devenir prépondérante, et qu'il faut la traiter avec attention.

Nous présentons une seconde technique d'élimination de la contrainte. Rappelons le contexte sous un angle un peu différent, c'est-à-dire sans multiplicateur de Lagrange. Le but est de minimiser J_0 sur $K^\star$, qui est défini par $\{v \in \mathbb{R}^n \ : \ Cv = f\}$. C est une matrice de $\mathbb{R}^{p\times n}$ de rang p. On suppose que l'on peut l'écrire par blocs sous la forme

$$C = \begin{pmatrix} C_{11} & C_{12} \end{pmatrix},\ C_{11} \in \mathbb{R}^{p\times p},\ rg(C_{11}) = p.$$

(Eventuellement après un réarrangement des colonnes.)

$$Cv = f \Longleftrightarrow C_{11}v_1 + C_{12}v_2 = f, \text{avec } v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},\ v_1 \in \mathbb{R}^p,\ v_2 \in \mathbb{R}^{n-p}.$$
$$\text{D'où } v_1 = {C_{11}}^{-1}(f - C_{12}v_2) = g - \underline{C}v_2, \text{ avec } g = {C_{11}}^{-1}f,\ \underline{C} = {C_{11}}^{-1}C_{12}.$$

Nous allons maintenant réécrire $J_0(v)$ sous la forme $\tilde{J}(v_2)$, *pour tout* $v \in K^\star$.

Le terme *linéaire* :

$$\begin{aligned} -(b, v) &= -(b_1, v_1)_1 - (b_2, v_2)_2 = -(b_1, g)_1 + (b_1, \underline{C}v_2)_1 - (b_2, v_2)_2. \\ &= \alpha_{lin} + (\underline{C}^{\mathsf{T}}b_1 - b_2, v_2)_2, \text{ où } \alpha_{lin} = -(b_1, g)_1 \text{ est une constante.} \end{aligned} \tag{4.31}$$

Le terme *quadratique* : (notons que A_{11} et A_{22}, les deux blocs diagonaux, sont symétriques)

$$Av = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^{\mathsf{T}} & A_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} A_{11}v_1 + A_{12}v_2 \\ A_{12}^{\mathsf{T}}v_1 + A_{22}v_2 \end{pmatrix}.$$

On en déduit que :

$$\begin{aligned} \frac{1}{2}(Av, v) &= \frac{1}{2}(A_{11}v_1 + A_{12}v_2, v_1)_1 + \frac{1}{2}(A_{12}^{\mathsf{T}}v_1 + A_{22}v_2, v_2)_2 \\ &= \frac{1}{2}(A_{11}v_1, v_1)_1 + (A_{12}^{\mathsf{T}}v_1, v_2)_2 + \frac{1}{2}(A_{22}v_2, v_2)_2. \end{aligned}$$

Examinons le premier terme :

$$\frac{1}{2}(A_{11}v_1, v_1)_1 = \frac{1}{2}(A_{11}g - A_{11}\underline{C}v_2, g - \underline{C}v_2)_1$$
$$= \alpha_{quad} - (\underline{C}^{\mathsf{T}}A_{11}g, v_2)_2 + \frac{1}{2}(\underline{C}^{\mathsf{T}}A_{11}\underline{C}v_2, v_2)_2,\ \alpha_{quad} = \frac{1}{2}(A_{11}g, g)_1.$$

Le second terme :

$$(A_{12}^{\mathsf{T}}v_1, v_2)_2 = (A_{12}^{\mathsf{T}}g - A_{12}^{\mathsf{T}}\underline{C}v_2, v_2)_2 = (A_{12}^{\mathsf{T}}g, v_2)_2 - (A_{12}^{\mathsf{T}}\underline{C}v_2, v_2)_2.$$

En regroupant le tout, on trouve, avec $\alpha = \alpha_{lin} + \alpha_{quad}$,

$$J_0(v) = \frac{1}{2}(\{A_{22} + \underline{C}^{\mathsf{T}}A_{11}\underline{C} - 2A_{12}^{\mathsf{T}}\underline{C}\}v_2, v_2)_2 - (b_2 + \underline{C}^{\mathsf{T}}A_{11}g - \underline{C}^{\mathsf{T}}b_1 - A_{12}^{\mathsf{T}}g, v_2)_2 + \alpha.$$

Comme on l'a remarqué au chapitre 1, on peut symétriser le terme quadratique,

$$\left.\begin{array}{l} J_0(v) = \tilde{J}(v_2),\ \text{avec}\ \tilde{J}(v_2) = \frac{1}{2}(\tilde{A}_{22}v_2, v_2)_2 - (\tilde{b}_2, v_2)_2 + \alpha \\ \tilde{A}_{22} = A_{22} + \underline{C}^{\mathsf{T}}A_{11}\underline{C} - A_{12}^{\mathsf{T}}\underline{C} - \underline{C}^{\mathsf{T}}A_{12}, \quad \tilde{b}_2 = b_2 + \underline{C}^{\mathsf{T}}A_{11}g - \underline{C}^{\mathsf{T}}b_1 - A_{12}^{\mathsf{T}}g. \end{array}\right\} \tag{4.32}$$

On peut donc remplacer $J_0(v)$ par $\tilde{J}(v_2)$, *pour tout* $v \in K^\star$. Réciproquement, à chaque $v_2 \in \mathbb{R}^{n-p}$, on peut associer un unique $v^\star \in K^\star$, égal à

$$v^\star = \begin{pmatrix} g - \underline{C}v_2 \\ v_2 \end{pmatrix} \in K^\star, \text{et l'on a } \tilde{J}(v_2) = J_0(v^\star).$$

Proposition 4.3.2 *Résoudre le problème avec contraintes est équivalent à*

$$\text{Trouver } u_2 \in \mathbb{R}^{n-p} \text{ tel que } \tilde{J}(u_2) = \min_{v_2 \in \mathbb{R}^{n-p}} \tilde{J}(v_2). \tag{4.33}$$

De plus, la matrice $\tilde{A}_{22}$ intervenant dans la fonctionnelle $\tilde{J}$ est symétrique définie positive, ce qui permet d'utiliser les techniques énoncées auparavant.

Preuve : Il reste à vérifier que $\tilde{A}_{22}$ est bien symétrique définie positive. Bien sûr, $\tilde{A}_{22}$ est symétrique par construction. Par ailleurs,

$$(\tilde{A}_{22}v_2, v_2)_2 = (A_{22}v_2, v_2)_2 + (A_{11}\underline{C}v_2, \underline{C}v_2)_1 - (A_{12}^{\mathsf{T}}\underline{C}v_2, v_2)_2 - (A_{12}v_2, \underline{C}v_2)_1$$
$$= (-A_{11}\underline{C}v_2 + A_{12}v_2, -\underline{C}v_2)_1 + (-A_{12}^{\mathsf{T}}\underline{C}v_2 + A_{22}v_2, v_2)_2$$
$$= (\begin{pmatrix} A_{11} & A_{12} \\ A_{12}^{\mathsf{T}} & A_{22} \end{pmatrix}\begin{pmatrix} -\underline{C}v_2 \\ v_2 \end{pmatrix}, \begin{pmatrix} -\underline{C}v_2 \\ v_2 \end{pmatrix})$$
$$= (A\begin{pmatrix} -\underline{C}v_2 \\ v_2 \end{pmatrix}, \begin{pmatrix} -\underline{C}v_2 \\ v_2 \end{pmatrix}).$$

Le produit scalaire est strictement positif, sauf si $\begin{pmatrix} -\underline{C}v_2 \\ v_2 \end{pmatrix} = 0$, i. e. $v_2 = 0$. $\tilde{A}_{22}$ est bien symétrique définie positive. ∎

Par rapport à l'autre technique, notons que cette méthode est très attractive, puisqu'elle ne requiert pas $p + 2$ résolutions de problèmes sans contraintes de matrice A... Par contre, deux inconvénients potentiels sont à prendre en considération

- Il faut extraire le bloc C_{11} de rang p de la matrice C.
- La structure interne de $\tilde{A}_{22}$ est complètement différente de celle de A.

En particulier, même si A est une matrice creuse, $\tilde{A}_{22}$ peut être une matrice pleine. Le coût d'un produit matrice vecteur (voir l'exemple de la section 4.1) est alors bien plus important lorsque l'on résout (4.33). Ce type de considération doit absolument être examiné, pour évaluer les mérites de la mise en œuvre numérique.

4.3.2 Techniques de pénalisation

Nous allons encore une fois éliminer la contrainte $v \in K^\star$; pour cela, nous introduisons un paramètre $\varepsilon > 0$, la fonctionnelle

$$J_\varepsilon(v) = J_0(v) + \frac{1}{\varepsilon}\|Cv - f\|^2,$$

ainsi que le **problème pénalisé**

$$\text{Trouver } u_\varepsilon \in \mathbb{R}^n \text{ tel que } \tilde{J}(u_\varepsilon) = \min_{v\in\mathbb{R}^n} \tilde{J}_\varepsilon(v). \tag{4.34}$$

Dans la suite, on appelle ψ la fonctionnelle qui à v associe $\|Cv - f\|^2$.
On remarque que ψ est à valeurs dans $\mathbb{R}^+$, convexe, continue et telle que, pour tout élément v de $K^\star$, $\psi(v) = 0$. En particulier, $\psi(u) = 0$, ce qui signifie que, pour tout ε, $J_\varepsilon(u) = J_0(u)$.

Proposition 4.3.3 *Le problème (4.34) admet une solution unique, pour tout $\varepsilon > 0$.*

Preuve : Existence d'une solution. J_ε est continue. Montrons qu'elle est de plus infinie à l'infini. On écrit

$$\begin{aligned} J_\varepsilon(v) = J_0(v) + \frac{1}{\varepsilon}\psi(v) &\geq J_0(v) = \frac{1}{2}(Av, v) - (b, v) \\ &\geq \frac{\lambda_{min}}{2}\|v\|^2 - \|b\|\,\|v\|, \\ &\quad \text{quantité qui tend vers l'infini lorsque } \|v\| \to +\infty. \end{aligned}$$

La proposition 1.2.1 (du début du texte, comme quoi tout peut servir à tout moment !) permet de conclure qu'il existe un point de minimum.

Unicité de la solution. J_0 étant strictement convexe (cf. exercice 2.3.3), et $\frac{1}{\varepsilon}\psi$ étant convexe, leur somme J_ε est strictement convexe. En conséquence, le point de minimum est unique, d'après la proposition 2.3.1. ∎

Nous allons maintenant prouver que la suite $(u_\varepsilon)_\varepsilon$ possède une propriété très intéressante... Dont la preuve est très similaire à celle de la "fameuse" proposition 1.2.1 !

Proposition 4.3.4 *La suite $(u_\varepsilon)_\varepsilon$ converge vers u, lorsque ε tend vers 0^+.*

Preuve : Etape 1. Par définition de ψ, $J_0(u_\varepsilon) \leq J_0(u_\varepsilon) + \frac{1}{\varepsilon}\psi(u_\varepsilon) = J_\varepsilon(u_\varepsilon)$; or, u_ε réalise le minimum de J_ε sur $\mathbb{R}^n$, donc $J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(u)$. Enfin, d'après ce que l'on a remarqué plus haut, $J_\varepsilon(u) = J_0(u)$. Ainsi

$$\forall \varepsilon > 0, \quad J_0(u_\varepsilon) \leq J_0(u). \tag{4.35}$$

La fonctionnelle J_0 étant infinie à l'infini, nous en déduisons que $(u_\varepsilon)_\varepsilon$ est bornée.

Etape 2. Comme nous nous trouvons dans $\mathbb{R}^n$, il existe une sous-suite extraite $(u_{\varepsilon'})_{\varepsilon'}$ qui converge. Appelons u' sa limite. D'après la continuité de J_0 et la relation (4.35), qui s'applique notamment pour tous les termes de la sous-suite :

$$J_0(u') = \lim_{\varepsilon'\to 0+} J_0(u_{\varepsilon'}) \leq J_0(u).$$

Par ailleurs,

$$0 \leq \psi(u_{\varepsilon'}) = \varepsilon'\{J_{\varepsilon'}(u_{\varepsilon'}) - J_0(u_{\varepsilon'})\} \leq \varepsilon'\{J_{\varepsilon'}(u) - J_0(u_{\varepsilon'})\} = \varepsilon'\{J_0(u) - J_0(u_{\varepsilon'})\}.$$

On vient de voir que $(J_0(u_{\varepsilon'}))_{\varepsilon'}$ admet une limite (égale à $J_0(u')$), ce qui entraîne que

$$\lim_{\varepsilon' \to 0+} \left(\varepsilon' \{ J_0(u) - J_0(u_{\varepsilon'}) \} \right) = 0, \text{ et donc } \lim_{\varepsilon' \to 0+} \psi(u_{\varepsilon'}) = 0.$$

Comme ψ est continue : $\psi(u') = 0$, i. e. $u' \in K^\star$. Bien sûr, u réalise le minimum de J_0 sur $K^\star$, ce qui induit $J_0(u) \leq J_0(u')$. On en arrive finalement à l'égalité $J_0(u) = J_0(u')$, et comme J_0 est strictement convexe, $u = u'$.

Etape 3. Pour finir, supposons que $(u_\varepsilon)_\varepsilon$ ne converge pas vers u. Ceci signifie qu'il existe une sous-suite extraite, toujours notée $(u_{\varepsilon'})_{\varepsilon'}$, et $\eta > 0$ tels que $\|u_{\varepsilon'} - u\| \geq \eta$, pour tout ε'.
On reprend le raisonnement de l'étape 2 : $(u_{\varepsilon'})_{\varepsilon'}$ étant bornée, on peut en extraire une sous-suite, $(u_{\varepsilon''})_{\varepsilon''}$, qui converge. En poursuivant le même raisonnement (n'oublions pas que, par construction, $(u_{\varepsilon''})_{\varepsilon''}$ est également une sous-suite extraite de $(u_\varepsilon)_\varepsilon$!), on prouve que $(u_{\varepsilon''})_{\varepsilon''}$ converge nécessairement vers u. Ceci contredit le fait que $\|u_{\varepsilon''} - u\| \geq \eta$, pour tout ε''.

En conclusion, la suite $(u_\varepsilon)_\varepsilon$ converge bien vers u. ∎

Pour cette méthode, le problème central est celui du choix d'une *suite de valeurs de* ε, qui permette d'obtenir rapidement une bonne approximation de u. Par rapidement, on entend sans avoir à résoudre "beaucoup" de problèmes sans contraintes du type (4.34). Notons que J_ε peut être développée sous la forme :

$$\begin{aligned} J_\varepsilon(v) &= \frac{1}{2}(Av, v) - (b, v) + \frac{1}{\varepsilon}\{(C^{\mathsf{T}}Cv, v) - 2(C^{\mathsf{T}}f, v) + \|f\|^2\} \\ &= \frac{1}{2}([A + \frac{2}{\varepsilon}C^{\mathsf{T}}C]v, v) - (b + \frac{2}{\varepsilon}C^{\mathsf{T}}f, v) + \frac{1}{\varepsilon}\|f\|^2. \end{aligned}$$

$A + \frac{2}{\varepsilon}C^{\mathsf{T}}C$ est symétrique définie positive. Cependant, comme pour la seconde méthode d'élimination du paragraphe 4.3.1, sa structure interne peut être très différente de celle de A.

4.3.3 Extensions

La première technique d'élimination peut être appliquée au problème plus général

$$\text{Trouver } (u_1, u_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \text{ tel que } \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

sous réserve que ce problème admette une unique solution. L'équation (4.26) est remplacée par

$$\text{Trouver } u_2 \in \mathbb{R}^{n_2} \text{ tel que } (A_{22} - A_{21}{A_{11}}^{-1}A_{12})u_2 = b_2 - A_{21}{A_{11}}^{-1}b_1. \tag{4.36}$$

$S = (A_{22} - A_{21}{A_{11}}^{-1}A_{12})$ est une matrice de $\mathbb{R}^{n_2 \times n_2}$, appelée **complément de Schur**. Ce type de méthode est très utilisé en conjonction avec une mise en œuvre sur machine parallèle (constituée de plusieurs processeurs), l'extraction de la composante u_2 permettant de construire des problèmes de variable u_1 qui sont parallélisables (cf. [26], [22]).

Il est également possible d'appliquer la méthode de pénalisation dans un cadre beaucoup plus général... On considère une fonctionnelle J de $\mathbb{R}^n$ dans $\mathbb{R}$, α-convexe et différentiable, et K une partie convexe, fermée et non vide de $\mathbb{R}^n$, qui est l'ensemble des choix possibles. Comme dans le cas de la fonctionnelle quadratique, supposons que l'on dispose d'une fonctionnelle ψ de $\mathbb{R}^n$ dans $\mathbb{R}^+$, telle que

ψ est différentiable et convexe ; $\psi(v) = 0$, pour tout élément v de K.

Sous ces hypothèses, on peut approcher la solution u du problème

$$\text{Trouver } u \in K, \text{ tel que } J(u) = \min_{v \in K} J(v) \tag{4.37}$$

en résolvant une suite de problèmes pénalisés : pour $\varepsilon > 0$, on définit

$$\begin{aligned} &J_\varepsilon(v) = J(v) + \frac{1}{\varepsilon}\psi(v), \text{ et l'on résout} \\ &\text{Trouver } u_\varepsilon \in \mathbb{R}^n, \text{ tel que } J_\varepsilon(u_\varepsilon) = \min_{v \in \mathbb{R}^n} J_\varepsilon(v). \end{aligned} \tag{4.38}$$

Exercice 4.3.1 *1. Rappeler pourquoi le problème (4.37) admet une solution et une seule.*
2. Montrer que le problème (4.38), pour ε fixé, admet une solution unique.
3. Vérifier que $\lim_{\varepsilon \to 0+} u_\varepsilon = u$.

Notons pour finir que l'on peut affaiblir l'hypothèse de régularité sur J et ψ, en les supposant simplement continues (cf. [6]).

Deuxième partie :

Algèbre linéaire

Chapitre 5

Un problème modèle

5.1 Introduction

Afin de motiver cette introduction aux méthodes et algorithmes de l'algèbre linéaire utilisés par les ingénieurs, on établit dans ce chapitre un lien entre les valeurs propres d'une matrice et les fréquences de résonance d'une structure déformable.

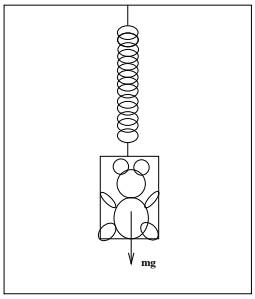
Dans le premier paragraphe on rappelle quelques éléments de dynamique dans le cas simple d'un problème monodimensionnel.

Dans le paragraphe suivant, cette approche est généralisée au cas d'une structure déformable, pour caractériser les fréquences de résonance et les modes propres de cette structure comme valeurs propres et vecteurs propres d'une matrice.

En conclusion, on rappelle l'importance de cette famille de problèmes dans le travail quotidien de l'ingénieur.

5.2 La masse oscillante

On s'intéresse aux oscillations d'une masse suspendue à un ressort vertical, selon le dispositif représenté par la figure suivante :



Il s'agit de déterminer l'amplitude et la fréquence des oscillations de la masse m autour d'une position d'équilibre y_0, quand elle est soumise aux forces suivantes :

- la force de gravité de module mg
- la force de rappel du ressort, proportionnelle à l'élongation $-k(y + y_0)$
- une force d'amortissement liée au frottement de la masse sur l'air $-c\dfrac{dy}{dt}$

– une force d'entrainement éventuelle $F(t)$

L'application de la loi de Newton à ce dispositif se résume en l'équation

$$m\frac{d^2y}{dt^2} = mg - k(y + y_0) - c\frac{dy}{dt} + F(t); \tag{5.1}$$

au repos on peut écrire : $0 = mg - ky_0$, soit finalement :

$$m\frac{d^2y}{dt^2} + c\frac{dy}{dt} + ky = F(t). \tag{5.2}$$

La solution générale d'une telle équation différentielle s'écrit comme la somme d'une solution particulière et de la solution générale de l'équation homogène

$$m\frac{d^2y}{dt^2} + c\frac{dy}{dt} + ky = 0, \tag{5.3}$$

que l'on obtient à partir des racines de l'équation caractéristique

$$mX^2 + cX + k = 0,$$

dont le discrimant est $\delta = c^2 - 4km$.

— si $\delta = c^2 - 4km > 0$ alors l'équation caractéristique admet deux racines réelles négatives r_1 et r_2 ($r_1 + r2 = -c/m < 0$ et $r_1r_2 = k/m > 0$).

La solution générale de (5.3) s'écrit donc

$$y_h(t) = ae^{r_1t} + be^{r_2t}$$

et vérifie $\lim_{t\to\infty} y_h(t) = 0$

— si $\delta = c^2 - 4km = 0$ l'équation caractéristique admet une racine double réelle et négative $r_0 = -c/2m = -\sqrt{k/m}$. La solution générale s'écrit alors $y_h(t) = (a + b\ t)e^{-ct/2m}$ et s'annule encore à l'infini : $\lim_{t\to\infty} y_h(t) = 0$

— enfin si $\delta = c^2 - 4km < 0$ l'équation caractéristique admet deux racines complexes conjuguées, et la solution générale de (5.3) s'écrit

$$y_h(t) = [a\cos(\mu t) + b\sin(\mu t)]e^{-ct/2m}$$

avec $\mu = \sqrt{4km - c^2}/2m$, et on a encore $\lim_{t\to\infty} y_h(t) = 0$ avec des oscillations.

Pour le traitement de l'équation complète, on se limite au cas des oscillations forcées du type $F(t) = F_0\ \cos(\omega\ t)$, et on cherche une solution particulière de la forme $y(t) = y_0\ \cos(\omega\ t - \varphi)$:

$$\begin{aligned} y(t) &= y_0\ \cos(\omega\ t - \varphi), \\ \frac{dy}{dt}(t) &= -\omega\ y_0\ \sin(\omega\ t - \varphi), \\ \frac{d^2y}{dt^2}(t) &= -\omega^2 y_0\ \cos(\omega\ t - \varphi). \end{aligned}$$

En reportant dans l'équation (5.2), on obtient

$$(k - m\omega^2)y_0\ \cos(\omega\ t - \varphi) - c\ \omega\ y_0\ \sin(\omega\ t - \varphi) = F_0\ \cos(\omega\ t),$$

soit encore

$$\begin{aligned}\left[(k-m\omega^2)\cos(\varphi)+c\ \omega\ \sin(\varphi)\right]\ y_0\ \cos(\omega\ t)\\ +\left[(k-m\omega^2)\sin(\varphi)-c\ \omega\ \cos(\varphi)\right]\ y_0\ \sin(\omega\ t)\\ =F_0\ \cos(\omega\ t)\end{aligned}$$

d'où on déduit

$$\begin{aligned}(k-m\omega^2)\cos(\varphi)+c\ \omega\ \sin(\varphi)&=F_0/y_0,\\ (k-m\omega^2)\sin(\varphi)-c\ \omega\ \cos(\varphi)&=0.\end{aligned}$$

En introduisant $\omega_0=\sqrt{k/m}$, ce système d'équations s'écrit :

$$\begin{aligned}m(\omega_0^2-\omega^2)\cos(\varphi)+c\ \omega\ \sin(\varphi)&=F_0/y_0,\\ m(\omega_0^2-\omega^2)\sin(\varphi)-c\ \omega\ \cos(\varphi)&=0.\end{aligned}\tag{5.4}$$

En supposant dans un premier temps $\omega\neq\omega_0$, on obtient

$$\tan(\varphi)=\frac{c\ \omega}{m(\omega_0^2-\omega^2)}.\tag{5.5}$$

En introduisant $z_0=\sqrt{m^2(\omega_0^2-\omega^2)^2+c^2\omega^2}$, on peut encore écrire

$$\begin{aligned}\cos(\varphi)&=m(\omega_0^2-\omega^2)/z_0,\\ \sin(\varphi)&=-c\ \ \omega\ /z_0,\\ y_0&=F_0/z_0.\end{aligned}$$

La solution générale de l'équation (5.2) non homogène s'écrit donc

$$y(t)=y_h(t)+F_0/z_0\ \cos(\omega\ t-\varphi).\tag{5.6}$$

La solution $y(t)$ se comporte à l'infini comme $F_0/z_0\ \cos(\omega\ t-\varphi)$, et les oscillations restent bornées. Dans le cas particulier où il n'y a pas d'amortissement $c=0$, on déduit de (5.5) que $\varphi=0$.

Le phénomène de résonance est obtenu lorsque la fréquence des oscillations forcées prend la valeur ω_0, appelée en conséquence fréquence de résonance. Dans ce cas particulier (5.4) s'écrit

$$\begin{aligned}c\ \omega_0\ \sin(\varphi)&=F_0/y_0,\\ \cos(\varphi)&=0.\end{aligned}$$

soit encore $\varphi=\pi/2\ (2\pi)$ et $y_0=F_0/c\ \omega_0$ quand $c\ \neq 0$.

Ainsi en présence d'un terme d'amortissement, le phénomène de résonance conduit à une solution $y(t)$ du type général (5.6) avec oscillations bornées :

$$y(t)=y_h(t)+F_0/z_0\ \sin(\omega_0 t).$$

Mais que se passe-t-il en l'absence d'amortissement ? Cette fois le système (5.4) n'a pas de solution et il faut chercher une solution particulière de (5.2) sous la forme

$$y(t)=y_0\ t\ \sin(\omega_0 t)$$

alors

$$y(t) = y_0\ t\ \sin(\omega_0 t),$$
$$\frac{dy}{dt}(t) = y_0\ [\omega_0\ t\ \cos(\omega_0 t) + \sin(\omega_0 t)],$$
$$\frac{d^2y}{dt^2}(t) = y_0\ [2\omega_0 \cos(\omega_0 t) - \omega_0^2\ t\ \sin(\omega_0 t)].$$

On en déduit $y_0 = F_0/2m\omega_0$ et la solution générale de (5.2) s'écrit alors

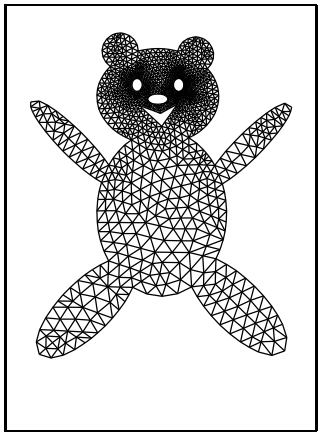
$$y(t) = y_h(t) + (F_0/2m\omega_0)\ t\ \sin(\omega_0 t).$$

L'amplitude des oscillations n'est plus bornée, et le comportement de ce dispositif pour un temps élevé est résumé dans le tableau suivant :

	$c \neq 0$	$c = 0$
$\omega \neq \omega_0$	oscillations bornées	oscillations bornées
$\omega = \omega_0$	oscillations bornées	oscillations **non** bornées

D'un point de vue concret, cela signifie qu'au bout d'un temps fini le ressort casse ! Il est donc très important de déterminer la fréquence de résonance ω_0. Ce phénomène se généralise aux structures plus complexes, comme on va le constater dans le paragraphe suivant.

5.3 Une structure déformable



Une structure déformable

Le lien de cette étude avec les valeurs propres des matrices apparait dans l'exemple d'une structure déformable, encastrée sur une partie de sa surface, et soumise à un ensemble de forces. Le modèle mécanique choisi pour traiter ce problème est la formulation en déplacements du

problème de l'élasticité linéaire, qui s'écrit en dimension n sous la forme

$$\begin{cases} \text{Trouver un déplacement admissible } \overrightarrow{u} \in \mathbb{R}^n \text{ vérifiant} \\ \rho \dfrac{\partial^2 \overrightarrow{u}}{\partial t^2} - div\,(\sigma(\overrightarrow{u})) = \overrightarrow{F} \text{ dans } \Omega \\ \overrightarrow{u} = \overrightarrow{0} \text{ sur } \partial\Omega_D \\ \displaystyle\sum_{j=1}^{n} \sigma_{i,j}(\overrightarrow{u})\nu_j = f_i \text{ sur } \partial\Omega_N \qquad 1 \leq i \leq n \end{cases}$$

le vecteur $\overrightarrow{f}$ représente les forces surfaciques auxquelles est soumise la structure Ω sur $\partial\Omega_N$, une partie de la frontière $\partial\Omega$. La partie $\partial\Omega_D$ complémentaire de la frontière $\partial\Omega$ de $\partial\Omega_N$, représente une partie de la surface de Ω qui est encastrée . Encastrée signifie que les points de cette partie du solide ne bougent pas (le déplacement $\overrightarrow{u}$ y est donc nul). Enfin le vecteur $\overrightarrow{\nu}$ représente la normale extérieure à la surface $\partial\Omega$.

Dans cette formulation la **matrice des contraintes** $\sigma \in \mathbb{R}^{n\times n}$ est est reliée à la **matrice des déformations** $\varepsilon \in \mathbb{R}^{n\times n}$ par la **loi de Hooke** :

$$\sigma(\overrightarrow{u}) = \lambda \text{ trace}[\varepsilon(\overrightarrow{u})] + 2\mu\varepsilon(\overrightarrow{u})$$
$$\varepsilon(\overrightarrow{u}) = \frac{1}{2}\left[\nabla(\overrightarrow{u}) + \nabla^T(\overrightarrow{u})\right]$$

où λ et μ sont appelés coefficients de Lamé.

Cette relation s'écrit encore

$$\sigma_{i,j}(\overrightarrow{u}) = \lambda\delta_{i,j}\sum_{k=1}^{n}\varepsilon_{k,k})(\overrightarrow{u}) + 2\mu\varepsilon_{i,j}(\overrightarrow{u}) \qquad 1 \leq i,j \leq n$$
$$\varepsilon_{i,j}(\overrightarrow{u}) = \frac{1}{2}\left[\frac{\partial \overrightarrow{u}_i}{\partial x_j} + \frac{\partial \overrightarrow{u}_j}{\partial x_i}\right] \qquad 1 \leq i,j \leq n$$

où $\delta_{i,j}$ est le symbole de Kronecker.

Pour étudier l'existence d'une solution au problème précédent, et déterminer ses éventuelles propriétés, on utilise une démarche mathématique appelée **formulation variationnelle** . Il s'agit de considérer un espace fonctionnel X dans lequel on va chercher la solution $\overrightarrow{u}$. Dans cet espace, encore appelé **champ des déplacements admissibles**, on réécrit le problème initial en multipliant la première équation par une fonction $\overrightarrow{v}$ quelconque de X ; on intégre ensuite "par parties" pour obtenir la formulation suivante :

$$\begin{cases} \text{Trouver } \overrightarrow{u} \in X \text{ tel que} \\ \quad \forall t \in [0,T], \forall \overrightarrow{v} \in X \quad \displaystyle\int_\Omega \rho\, \frac{\partial^2 \overrightarrow{u}}{\partial t^2} \cdot \overrightarrow{v}\; dx + \sum_{i,j=1}^{n} \int_\Omega \varepsilon_{i,j}(\overrightarrow{u}) \cdot \varepsilon_{i,j}(\overrightarrow{v})\; dx \\ \qquad\qquad\qquad\qquad = \displaystyle\int_\Omega \overrightarrow{F} \cdot \overrightarrow{v}\; dx + \int_{\partial\Omega_N} \overrightarrow{f} \cdot \overrightarrow{v}\; ds \end{cases} \tag{5.7}$$

Cette écriture présente deux avantages

- le premier est de rentrer dans un cadre mathématique formel pour lequel il existe des résultats théoriques généraux, qui ne dépendent que des données du problème : Ω, $\partial\Omega$, $\overrightarrow{F}$, $\overrightarrow{f}$ etc

- le second concerne le calcul de la solution $\overrightarrow{u}$.

En général il n'est pas possible d'avoir une expression analytique des composantes de $\overrightarrow{u}$, et on doit procéder par approximation. La méthode la plus utilisée par les ingénieurs est la **méthode des éléments finis**, qui consiste à approcher l'espace fonctionnel X (de dimension infinie) par un espace vectoriel X_h de dimension finie p. Cet espace X_h est formé des déplacements admissibles dont les composantes sont**polynomiales par morceaux** ; la formulation variationnelle en dimension finie s'écrit :

$$\left\{\begin{array}{l} \text{Trouver } \overrightarrow{u}_h \in X_h \text{ tel que} \\ \quad \forall t \in [0,T], \forall \overrightarrow{v} \in X_h \quad \displaystyle\int_\Omega \rho\, \frac{\partial^2 \overrightarrow{u}_h}{\partial t^2} \cdot \overrightarrow{v}\; dx + \sum_{i,j=1}^{n} \int_\Omega \varepsilon_{i,j}(\overrightarrow{u}_h) \cdot \varepsilon_{i,j}(\overrightarrow{v})\; dx \\ \qquad\qquad\qquad\qquad = \displaystyle\int_\Omega \overrightarrow{F} \cdot \overrightarrow{v}\; dx + \int_{\partial\Omega_N} \overrightarrow{f} \cdot \overrightarrow{v}\; ds \end{array}\right. \tag{5.8}$$

La différence essentielle entre les problèmes 5.7 et 5.8 est que l'on peut construire une base de l'espace X_h : $\overrightarrow{v}_1, \overrightarrow{v}_2, \ldots, \overrightarrow{v}_p$; on représente alors chaque élément de X_h par un vecteur de $\mathbb{R}^p$. A toute fonction de base $\overrightarrow{v}_k$ est associé un vecteur $\mathcal{V}_k$, et l'inconnue $\overrightarrow{u}_h$ s'écrit comme une combinaison linéaire de ces vecteurs :

$$\overrightarrow{u}_h = \sum_{k=1}^{p} U_h^k \mathcal{V}_k.$$

Dans la formulation (5.8), on prend pour $\overrightarrow{v}$ successivement chacun des vecteurs $\overrightarrow{v}_k$ de la base ; le problème est alors équivalent au système

$$\left\{\begin{array}{l} \text{Trouver } U_h \in \mathbb{R}^p, \text{ tel que} \\ M \dfrac{d^2 U_h}{dt^2} + K U_h = F. \end{array}\right. \tag{5.9}$$

Dans cette écriture,

– U_h est le vecteur des p composantes de $\overrightarrow{u}_h$ dans la base $\{\mathcal{V}\}_{k\in\mathbb{N}}$

– la matrice $M \in \mathbb{R}^{p\times p}$, définie par

$$(M\overrightarrow{v}_k, \overrightarrow{v}_l) = \int_\Omega \rho\, \overrightarrow{v}_k \cdot \overrightarrow{v}_l\; dx \qquad 1 \le k, l \le p$$

est traditionnellement appelée **la matrice de masse**

– la matrice $K \in \mathbb{R}^{p\times p}$, définie par

$$(K\overrightarrow{v}_k, \overrightarrow{v}_l) = \sum_{i,j=1}^{n} \int_\Omega \varepsilon_{i,j}(\overrightarrow{v}_k) \cdot \varepsilon_{i,j}(\overrightarrow{v}_l)\; dx \qquad 1 \le k, l \le p$$

est appelée **la matrice de raideur**

– enfin le vecteur $F \in \mathbb{R}^p$, défini par

$$(F, \overrightarrow{v}_k) = \int_\Omega \overrightarrow{F} \cdot \overrightarrow{v}_k \, dx + \int_{\partial\Omega_N} \overrightarrow{f} \cdot \overrightarrow{v}_k \, ds \qquad 1 \leq k \leq p$$

représente l'action des forces extérieures sur la structure.

Remarque 5.3.1 *le système linéaire (5.9) est une formulation en dimension finie du problème (5.2) dans laquelle l'inconnue est un vecteur $U_h \in \mathbb{R}^p$, et qui ne comporte pas de terme d'amortissement.*

Pour obtenir facilement la solution de (5.9), l'astuce consiste à choisir une base de X_h associée aux matrices M et K de manière à simplifier les calculs. On remarque que ces deux matrices sont toujours symétriques, et que la matrice M est toujours définie positive. Il existe alors une matrice inversible $L \in \mathbb{R}^{p\times p}$ triangulaire inférieure, telle que

$$M = LL^T$$

cette matrice provient de la factorisation de Cholesky de la matrice M (ce résultat est établi par la Proposition 6.15.1), elle permet de définir la matrice

$$A = L^{-1}KL^{-T}.$$

La matrice $A \in \mathbb{R}^{p\times p}$ est symétrique et (semi)définie positive, elle est donc diagonalisable (voir la Proposition 10.3.3). Il existe donc p vecteurs : V_1, V_2,..., V_p de $\mathbb{R}^p$, tels que

$$AV_k = \lambda_k V_k \qquad 1 \leq k \leq p.$$

Cette relation s'écrit encore sous la forme

$$A = [\, V_1 \quad V_2 \quad \dots \quad V_p \,] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{bmatrix} [\, V_1 \quad V_2 \quad \dots \quad V_n \,]^{-1}$$

dans laquelle les $\lambda_k \geq 0$ sont les valeurs propres réelles de la matrice A, et les V_k sont ses vecteurs propres qui vérifient (voir la Proposition 10.3.2)

$$(V_k, V_l) = 0 \qquad \forall k \neq l.$$

On définit maitenant un nouvel ensemble de vecteurs $W_k \in \mathbb{R}^p$ par la relation $L^T W_k = V_k$. Par construction ces vecteurs sont linéairement indépendants et vérifient de plus

$$(L^T W_k, L^T W_l) = 0 \qquad \forall k \neq l$$

soit encore

$$(MW_k, W_l) = 0 \qquad \forall k \neq l.$$

Normalisons ces vecteurs pour que

$$(MW_k, W_k) = 1 \qquad 1 \leq k \leq p,$$

puisque que $W_k = L^{-T}V_k$ avec V_k vecteur propre de la matrice A : la realtion

$$L^{-1}KL^{-T}V_k = \lambda_k V_k \qquad 1 \leq k \leq p$$

s'écrit encore

$$[K - \lambda_k M]\, W_k = 0 \qquad 1 \leq k \leq p \tag{5.10}$$

En conséquence les vecteurs W_k vérifient également

$$(KW_k, W_l) = 0 \qquad \forall k \neq l.$$

Tous ces résultats sont exploités de la manière suivante : on cherche la solution de l'équation homogène associée à (5.9) sous la forme

$$U_h(t) = \sum_{k=1}^{p} \alpha_k \cos(\omega_k\ t + \varphi_k) W_k. \tag{5.11}$$

Les inconnues ω_k et φ_k doivent nécessairement satisfaire la relation

$$M \frac{d^2}{dt^2} \cos(\omega_k\ t + \varphi_k) W_k + K \cos(\omega_k\ t + \varphi_k) W_k = 0,$$

soit encore pour tout $t \in [0, T]$ et $1 \leq k \leq p$

$$\cos(\omega_k\ t + \varphi_k) \left[K - \omega_k^2 M\right] W_k = 0,$$

et finalement

$$\left[K - \omega_k^2 M\right] W_k = 0 \qquad 1 \leq k \leq p. \tag{5.12}$$

En rapprochant cette relation de (5.10), on en déduit que les fréquences propres recherchées sont les racines carrées des valeurs propres de la matrice A :

$$\omega_k^2 = \lambda_k \qquad 1 \leq k \leq p \tag{5.13}$$

Il reste à déterminer les composantes α_k et les phases φ_k ; pour cela on utilise les conditions initiales de l'équation homogène (5.9) : le déplacement initial $U_h(0)$ et la vitesse initiale $U_h'(0)$ sont des données qui définissent le vecteur U_h au temps $t = 0$

$$U_h(0) = U_0 \in \mathbb{R}^p \quad \text{et} \quad \frac{dU_h}{dt}(0) = U_0' \in \mathbb{R}^p,$$

ces vecteurs sont écrits sur la base des W_k

$$U_0 = \sum_{k=1}^{p} \alpha_k \cos(\varphi_k) W_k \quad \text{et} \quad U_0' = \sum_{k=1}^{p} -\omega_k\ \alpha_k \sin(\varphi_k) W_k$$

En utilisant maintenant les relations d'orthogonalité, on obtient

$$(MW_k, U_0) = \alpha_k \cos(\varphi_k) \quad \text{et} \quad (MW_k, U_0') = -\omega_k\ \alpha_k \sin(\varphi_k).$$

Les composantes α_k et les phases φ_k sont donc définies par les relations

$$\begin{cases} \alpha_k^2 = (MW_k, U_0)^2 + \dfrac{1}{\lambda_k}(MW_k, U_0')^2, \\ \tan(\varphi_k) = -\dfrac{1}{\omega_k} \dfrac{(MW_k, U_0')}{(MW_k, U_0)}. \end{cases} \tag{5.14}$$

On a donc obtenu la solution de l'équation homogène (5.9) sous la forme (5.11) ; il reste maintenant à calculer la solution générale. Pour cela il faut calculer une solution particulière de l'équation (5.9) avec second membre. Comme pour l'exemple de la masse oscillante, on se limite au cas d'un terme d'excitation de la forme

$$F(t) = F_0 \ \cos(\omega t),$$

et on écrit le vecteur $F_0 \in \mathbb{R}^p$ sur l'ensemble des vecteurs MW_k, qui forment aussi une base de $\mathbb{R}^p$.

$$F_0 = \sum_{k=1}^{p} \beta_k M W_k, \quad \text{avec} \quad \beta_k = (F_0, W_k).$$

Une solution particulière de l'équation (5.9) est alors donnée par la relation

$$\begin{aligned} U_h(t) &= \sum_{k=1}^{p} \gamma_k \cos(\omega \ t) W_k, \\ \text{avec} \quad \gamma_k &= \frac{\beta_k}{(\omega_k^2 - \omega^2)}. \end{aligned} \tag{5.15}$$

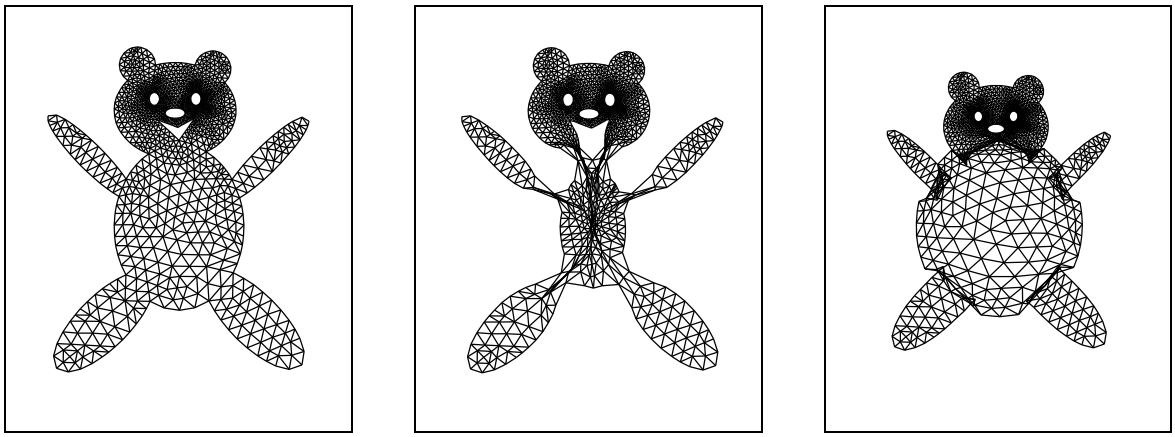
Lorsque la fréquence ω du terme d'excitation est égale à l'une des fréquences propres ω_k, on retrouve un **phénomène de résonance** analogue à celui du paragraphe précédent, c'est-à-dire que **les oscillations** dans la direction associée W_k **ne sont plus bornées** au cours du temps car elles se développent en $t \ \sin(\omega_k \ t) \ W_k$.

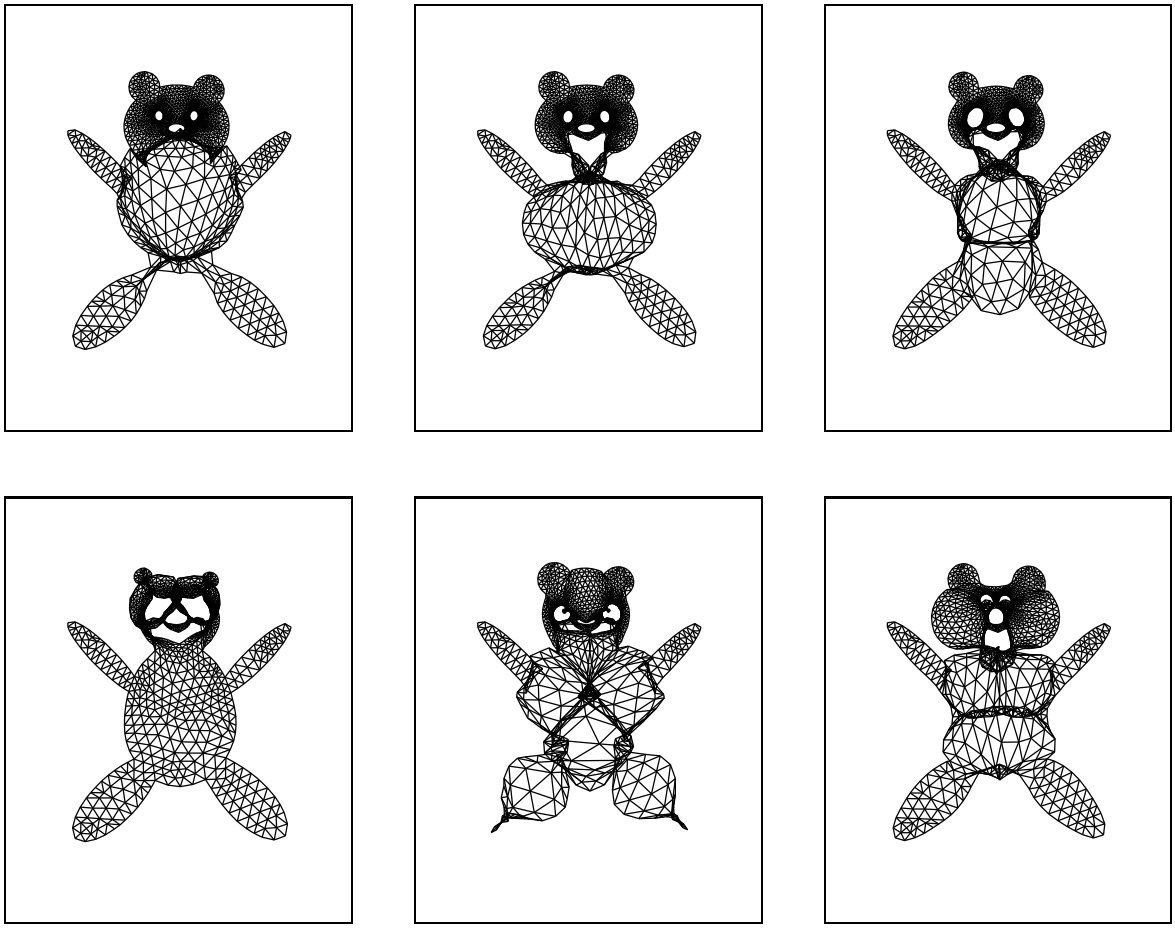
En résumé la recherche des **fréquences de résonance** ω_k d'une structure déformable Ω est équivalent à la résolution d'un problème aux valeurs propres généralisé

$$\left[K - \omega_k^2 M\right] W_k = 0 \qquad 1 \leq k \leq p.$$

Les vecteurs propres W_k associés aux valeurs propres ω_k^2 sont appelés **modes propres** de la structure Ω, ils représentent les directions de l'espace suivant lesquelles les déformations de la structure s'amplifient au cours du temps.

Les figures suivantes représentent les déformations subies par la structure, suivant certains modes propres.





Les déformations de la structure sous l'action des modes propres.

5.4 Conclusion

Le phénomène de résonance est bien connu des ingénieurs, on le rencontre dans de nombreux domaines de la physique : il est particulièrement redouté en mécanique des solides car les vibrations des structures peuvent leur causer de sévères dégats. Les ponts sont spécialement sensibles aux forces d'excitation : le pont de Broughton, près de Manchester, est célèbre pour s'être effrondré en 1831 lors du passage d'un régiment marchant au pas et le pont de Tacoma (Californie) s'est rompu en 1940 sous l'effet de rafales de vent.

D'une manière générale, les situations où la recherche des fréquences de résonance est indispensable sont très nombreuses, citons les vibrations des fusées qui peuvent causer leur destruction, les

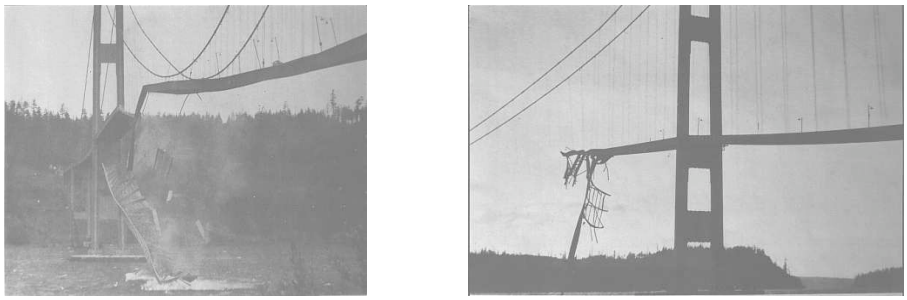


FIG. 5.1 – *Le pont de Tacoma*

vibrations dans les voitures sources de nuisance sonore, les vibrations des machines en rotation rapide comme les turbines et réacteurs, et encore les tremblements de terre... Noter que ce phénomène a parfois des conséquences moins néfastes, et qu'il peut même être recherché, comme dans le cas des circuits RLC des antennes radio.

5.5 Autre motivation

Il existe une autre classe de problèmes aux valeurs propres aussi importante que celle de la recherche des fréquences propres : l'étude de la stabilité des systèmes dynamiques (cf. [15]). Ces systèmes sont décrits par la formulation générale

$$\frac{dy}{dt} = F(y)$$

dans laquelle y est un vecteur de $\mathbb{R}^n$ dont les composantes sont fonctions du temps. F est une application (souvent non linéaire) de $\mathbb{R}^n$ dans $\mathbb{R}^n$.

Parmi les nombreux problèmes qui s'écrivent sous cette forme, citons en particulier les équations de la chimie cinétique, qui décrivent les variations des concentrations des différents constituants d'une réaction chimique. Par exemple la modélisation de la pollution atmosphérique par la circulation automobile conduit à des problémes de cette forme, avec plusieurs dizaines, voire plusieurs centaines d'éléments réactifs suivant le modèle choisi.

Sans entrer plus avant dans les détails, notons que la résolution numérique d'un tel problème nécessite le calcul des valeurs propres de la matrice jacobienne (en général non symétrique) définie par

$$J_{i,j}(y) = \frac{\partial F_i}{\partial y_j}(y) \qquad 1 \leq i, j \leq n.$$

Ce calcul doit être effectué avec une grande précision et nécessite un traitement approprié par des algorithmes efficaces.

Chapitre 6

Les méthodes directes

6.1 Introduction

Dans ce chapitre on montre que le calcul de la solution d'un système linéaire d'ordre 20 par les formules de Cramer est impossible à réaliser sur un ordinateur. Cet exemple simple sert d'introduction à cette partie du cours. Il montre la nécessité d'inventer de nouveaux algorithmes d'algèbre linéaire pour effectuer concrétement des calculs sur ordinateur.

On rappelle ensuite que la forme triangulaire de la matrice d'un système linéaire 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 de ce problème 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 chpaitre. Le dernier paragraphe constitue une introduction à l'utilisation pratique de ces méthodes pour la résolution de systèmes linéaires à matrice creuse.

6.2 Formules de Cramer

Soit à résoudre le système linéaire $Ax = b$, où $A \in \mathbb{R}^{n\times n}$ est une matrice de rang n et $b \in \mathbb{R}^n$ un vecteur. Pour calculer les composantes du vecteur x, on peut utiliser les formules de Cramer, rappelées sans démonstration :

Théorème 6.2.1 *Le système linéaire $Ax = b$ où $A \in \mathbb{R}^{n\times n}$ est une matrice de rang n, a pour solution unique le vecteur $x \in \mathbb{R}^n$ déterminé par les formules de* **Cramer** *: pour $1 \le k \le n$*

$$x_k = \frac{\begin{vmatrix} A_{1,1} & \cdot & A_{1,k-1} & b_1 & A_{1,k+1} & \cdot & A_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{n,1} & \cdot & A_{n,k-1} & b_n & A_{n,k+1} & \cdot & A_{n,n} \end{vmatrix}}{\begin{vmatrix} A_{1,1} & \cdot & A_{1,k-1} & A_{1,k} & A_{1,k+1} & \cdot & A_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{n,1} & \cdot & A_{n,k-1} & A_{n,k} & A_{n,k+1} & \cdot & A_{n,n} \end{vmatrix}}.$$

Coût calcul: pour calculer un déterminant d'ordre n il faut faire la somme de $n!$ produits de n facteurs, un tel calcul nécessite donc $n \times n!$ opérations. Le coût calcul des n composantes du vecteur x, solution du système linéaire $Ax = b$, est donc de $n(n + 1) \times n!$ opérations. Ainsi

pour la résolution d'un système linéaire d'ordre 10, sachant que 10! = 3.628.800, environ 400 millions d'opérations sont nécessaires pour calculer la solution par les formules de Cramer.

Exercice 6.2.1 *Un ordinateur rapide effectue environ* 10^{12} *opérations élémentaires par seconde (soit une puissance de calcul de 1 tera-flops) ; calculer le temps nécessaire au calcul de la solution d'un système linéaire d'ordre* 20 *(on ne tiendra compte ni des pannes de courant, ni des grèves).*

Réponse : environ 32 ans !

6.3 Déterminant d'une matrice triangulaire

Il existe un cas particulier où le calcul par ces formules est beaucoup moins coûteux.

Définition 6.3.1 *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* $1 \le i < j \le n$.

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* $1 \le j < i \le n$.

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

Proposition 6.3.1 *Le déterminant d'une matrice triangulaire (supérieure ou inférieure) est égal au produit des coefficients diagonaux :*

$$dét(T) = \prod_{i=1}^{n} T_{i,i}.$$

Preuve :

$$T = \begin{pmatrix} T_{1,1} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & T_{2,2} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_{n,n} \end{pmatrix}.$$

Il suffit de développer le déterminant de la matrice triangulaire supérieure suivant la première colonne (la première ligne pour une matrice triangulaire inférieure) :

$$\det(T) = \begin{vmatrix} T_{1,1} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & T_{2,2} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_{n,n} \end{vmatrix} = T_{1,1} \times \begin{vmatrix} T_{2,2} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & T_{n,n} \end{vmatrix}$$

soit

$$\text{dét}(T) = T_{1,1}\, T_{2,2} \times \begin{vmatrix} T_{3,3} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & \text{x} & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & \text{x} & \text{x} & \text{x} \\ 0 & 0 & 0 & 0 & \text{x} & \text{x} \\ 0 & 0 & 0 & 0 & 0 & T_{n,n} \end{vmatrix} = T_{1,1}\, T_{2,2}\, \ldots T_{n,n}.$$

■

6.4 Système linéaire à matrice triangulaire

Proposition 6.4.1 *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

Preuve : Il suffit d'écrire pour chaque composante y_k, la ligne correspondante du système linéaire à matrice triangulaire inférieure. Ces relations montrent que l'on peut calculer le vecteur y de proche en proche, en commençant par y_n ; on dit alors que l'on résout le système linéaire **en remontant**. Pour le système linéaire à matrice triangulaire supérieure, on calcule également le vecteur x de proche en proche, en commençant par x_1 ; on dit que l'on résout le système linéaire **en descendant**. Noter que l'hypothèse d'inversibilité entraine que les coefficients diagonaux de deux matrices sont tous différents de zéro, puisque

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

■

Coût calcul : dans le cas particulier d'un système linéaire à matrice triangulaire T, la Proposition 6.3.1 montre que le calcul du déterminant de T nécessite n multiplications ; mais il n'est même pas utile de calculer ce déterminant pour obtenir la solution du système linéaire $Tu = v$.

D'après les formules de la Proposition 6.4.1, pour un système linéaire à matrice triangulaire inférieure $Ly = b$, le calcul de la composante y_k nécessite k multiplications, k additions et une division, soit un total de $2k$ opérations. Le coût calcul des n composantes du vecteur y, est donc de $n(n+1)$ opérations, ce qui devient raisonnable ! On obtient le même coût pour un système linéaire à matrice triangulaire supérieure $Ux = y$.

Cette propriété nous conduit naturellement à une autre approche de la résolution des systèmes linéaires dont la matrice A est inversible mais sans structure particulière ; on essaie de se ramener au cas des matrices triangulaires ; cette technique est développée plus loin.

6.5 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 appelées **blocs**, dont seuls les p blocs diagonaux sont nécessairement carrés. Les définitions précédentes se généralisent naturellement suivant :

Définition 6.5.1 *On appelle matrice* **triangulaire inférieure par blocs** *une matrice $L \in \mathbb{R}^{n\times n}$ telle que $[L]_{i,j} = 0 \in \mathbb{R}^{m\times n}$ pour tout couple (i,j) tel que $1 \le i < j \le p$;*

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 $1 \le j < i \le p$.*

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

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 !

6.6 Exercices sur les matrices triangulaires

On établira les résultats suivants :

Exercice 6.6.1 *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 $T \in \mathbb{R}^{n\times n}$.*

Remarque 6.6.1 Attention *le produit d'une matrice triangulaire inférieure (resp. supérieure) par une matrice triangulaire supérieure (resp. inférieure) est une matrice à structure quelconque !*

Exercice 6.6.2 *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 6.6.3 *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

Exercice 6.6.4 *Montrer que la matrice inverse d'une matrice triangulaire inférieure (resp. supérieure) inversible est une matrice triangulaire inférieure (resp. supérieure) inversible.*

6.7 Déterminant d'une matrice carrée

La Proposition 6.3.1 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

$$\text{dét}(A) = \text{dét}(L)\ \text{dét}(U) = \text{dét}(U)$$

et utiliser la Proposition 6.3.1 pour le calcul de dét(U).

Coût calcul :
On montre dans la suite de ce chapitre (voir la Proposition 6.14.1) 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 développement classique :

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

pour laquelle le coût est en $O(n(n+1)n!)$.

6.8 La méthode d'élimination

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

$$\begin{pmatrix} A_{1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{k,1} & \cdot & \cdot & A_{k,k} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{n,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_k \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_k \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$$

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 :

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

en reportant cette identité dans les lignes 2 à n du système linéaire, on obtient pour tout i, $(2 \leq i \leq n)$

$$\begin{aligned} &[A_{i,2} - A_{i,1} \times A_{1,2}/A_{1,1}]x_2 \\ +&[A_{i,3} - A_{i,1} \times A_{1,3}/A_{1,1}]x_3 \\ &\cdots \\ +&[A_{i,n} - A_{i,1} \times A_{1,n}/A_{1,1}]x_n = [b_i - A_{i,1} \times b_1]/A_{1,1} \end{aligned}$$

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 :

$$\begin{aligned} &\text{pour la première ligne } 1 \leq j \leq n \qquad A^{(1)}_{1,j} = A^{(0)}_{1,j} \quad \text{et} \qquad b^{(1)}_1 = b^{(0)}_1 \\ &\text{et pour les } n-1 \text{ dernières lignes} \\ &\text{pour } 2 \leq i \leq n, \text{ et } 1 \leq j \leq n \qquad A^{(1)}_{i,j} = A^{(0)}_{i,j} - A^{(0)}_{i,1} \times A^{(0)}_{1,j}/A^{(0)}_{1,1} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad b^{(1)}_i = b^{(0)}_i - A^{(0)}_{i,1} \times b^{(0)}_1/A^{(0)}_{1,1} \end{aligned}$$

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 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 $j = 3, 4, \ldots, n$, et ainsi de suite... On génère ainsi 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

$$A_{i,j}^{(k)} = A_{i,j}^{(k-1)} \qquad 1 \le i \le k, \quad 1 \le j \le n$$

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

(2) modification du second membre

$$b_i^{(k)} = b_i^{(k-1)} - A_{i,k}^{(k-1)} \times b_k^{(k-1)} / A_{k,k}^{(k-1)} \qquad k < i \le n$$

fin

Noter que les coefficients $A_{i,j}^{(k)}$ pour $k < i \le n$ et $1 \le j < k$ ne sont pas définis par ces formules, car ils sont nuls par construction. Après $n-1$ itérations de cet algorithme (en supposant que les différents coefficients $A_{k,k}^{(k)}$ sont non nuls) 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 des formules du chapitre précédent et 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.

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

$$L_{i,j}^{(k)} = \begin{cases} 1 & \text{, si } i = j \\ 0 & \text{, si } i \ne j \text{ et } j \ne k \\ 0 & \text{, si } i < k \text{ et } j = k \\ -A_{i,k}^{(k-1)} / A_{k,k}^{(k-1)} & \text{, si } i > k \text{ et } j = k \end{cases}$$

soit encore

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

et on vérifie 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 $\tilde{L} = L^{(n-1)} \ldots L^{(1)}$, on peut écrire

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

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

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 $A^{(k-1)}_{i,k}$ non nul pour $i > k$, et s'il en existe un, on échange alors les lignes i et k de la matrice. Cette modification revient à multiplier à gauche la matrice courante $A^{(k-1)}$ par une matrice de permutation P qui amène le coefficient $A^{(k-1)}_{i,k}$ sur la diagonale

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

soit

$$PA^{(k-1)} = \begin{pmatrix} A^{(k-1)}_{1,1} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & A^{(k-1)}_{2,2} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & A^{(k-1)}_{3,3} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & A^{(k-1)}_{i,k} & \mathrm{x} & \mathrm{x} & A^{(k-1)}_{i,k'} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & A^{(k-1)}_{k,k} & \mathrm{x} & \mathrm{x} & A^{(k-1)}_{k,k'} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

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 $b^{(k-1)}_i$ et $b^{(k-1)}_k$ 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 $A^{(k-1)}_{i,k}$ de la colonne k sont nuls ? Cela veut dire que l'on a obtenu une matrice de la forme

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

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

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

entraîne

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

En effet, par construction des matrices $L^{(k')}$, dét$(L^{(k')}) = 1$. On a donc obtenu le résultat suivant

Proposition 6.8.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible, les matrices $A^{(k)}$ générées par l'algorithme d'élimination de Gauss vérifient (aux permutations de lignes près)*

$$dét A = \prod_{k=1}^{n} A_{k,k}^{(k)} \neq 0.$$

Remarque 6.8.1 *Cette méthode fournit donc aussi la valeur du déterminant de la matrice A.*

Exercice 6.8.1 *Montrer que pour tout k l'inverse $L^{-(k)}$ de la matrice $L^{(k)}$ est une matrice triangulaire inférieure, définie par les relations*

$$L_{i,j}^{-(k)} = \begin{cases} 1 & , \text{ si } i = j \\ 0 & , \text{ si } i \neq j \text{ et } j \neq k \\ 0 & , \text{ si } i < k \text{ et } j = k \\ A_{i,k}^{(k-1)}/A_{k,k}^{(k-1)} & , \text{ si } i > k \text{ et } j = k \end{cases}$$

6.9 La méthode de factorisation

Dans le paragraphe précédent on a obtenu pour une matrice A donnée, les matrices triangulaires inversibles U et $\tilde{L}$. A l'aide du résultat de l'exercice 6.8.1, on écrit pour tout k, $\tilde{L}^{-(k)} = 2I_n - L^{(k)}$. En posant alors

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

on définit une triangulaire inférieure à diagonale unité, qui vérifie la relation

$$A = LU$$

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 $Ux = y$.

6.10 Le complément de Schur

Une autre manière d'introduire cette factorisation consiste à utiliser le partitionnement par blocs de la matrice, en supposant le problème partiellement résolu : on écrit la matrice $A \in \mathbb{R}^{n\times n}$ sous la forme

$$A = \begin{pmatrix} [A]_{1,1} & [A]_{1,2} \\ [A]_{2,1} & [A]_{2,2} \end{pmatrix}$$

avec $[A]_{1,1} \in \mathbb{R}^{n_1\times n_1}$, $[A]_{2,1} \in \mathbb{R}^{n_2\times n_1}$, $[A]_{1,2} \in \mathbb{R}^{n_1\times n_2}$, $[A]_{2,2} \in \mathbb{R}^{n_2\times n_2}$, où n_1 et n_2 sont deux entiers naturels non nuls, tels que $n = n_1 + n_2$. On suppose que dans cette écriture on connait une factorisation du bloc : $[A]_{1,1} = [L]_{1,1}[U]_{1,1}$, avec $[L]_{1,1}$, $[U]_{1,1} \in \mathbb{R}^{n_1\times n_1}$ matrices triangulaires inversibles, $[L]_{1,1}$ à diagonale unité.

Alors on écrit formellement

$$A = \begin{pmatrix} [A]_{1,1} & [A]_{1,2} \\ [A]_{2,1} & [A]_{2,2} \end{pmatrix} = \begin{pmatrix} [L]_{1,1} & 0 \\ [L]_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & \mathcal{S} \end{pmatrix}$$

en identifiant les termes du produit, on obtient

$$[A]_{1,1} = [L]_{1,1}[U]_{1,1} \quad [A]_{1,2} = [L]_{1,1}[U]_{1,2}$$
$$[A]_{2,1} = [L]_{2,1}[U]_{1,1} \quad [A]_{2,2} = [L]_{2,1}[U]_{1,2} + \mathcal{S}.$$

les matrices $[L]_{1,1}$ et $[U]_{1,1}$ étant connues et inversibles, on obtient ainsi

$$[U]_{1,2} = [L]_{1,1}^{-1}[A]_{1,2} \text{ et } [L]_{2,1} = [A]_{2,1}U_{1,1}^{-1}$$

d'où

$$[L]_{2,1}[U]_{1,2} = [A]_{2,1}U_{1,1}^{-1}L_{1,1}^{-1}[A]_{1,2} = [A]_{2,1}A_{1,1}^{-1}[A]_{1,2},$$

et

$$\mathcal{S} = [A]_{2,2} - [A]_{2,1}A_{1,1}^{-1}[A]_{1,2}.$$

La matrice $\mathcal{S} \in \mathbb{R}^{n_2 \times n_2}$ est donc directement définie à partir du découpage par blocs (partition) de A ; on l'appelle **complément de Schur** associé à cette partition. Si on suppose maintenant que l'on sait factoriser la matrice $\mathcal{S}$ sous la forme $\mathcal{S} = [L]_{2,2}[U]_{2,2}$, avec $[L]_{2,2}$, $[U]_{2,2} \in \mathbb{R}^{n_2 \times n_2}$ matrices triangulaires inversibles, $[L]_{2,2}$ à diagonale unité, on vérifie immédiatement que

$$A = \begin{pmatrix} [L]_{1,1} & 0 \\ [L]_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & \mathcal{S} \end{pmatrix} = \begin{pmatrix} [L]_{1,1} & 0 \\ [L]_{2,1} & [L]_{2,2} \end{pmatrix} \times \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & [U]_{2,2} \end{pmatrix}$$

On peut élaborer maintenant une méthode de factorisation de la matrice A de la manière suivante : supposons que le coefficient diagonal $[A]_{1,1}$ soit différent de zéro, alors on peut écrire $[A]_{1,1} = 1 \times [A]_{1,1}$, et on pose $[L]_{1,1} = 1$ et $[U]_{1,1} = [A]_{1,1}$; c'est-à-dire que l'on fait une partition de A avec $n_1 = 1$ et $n_2 = n - 1$. L'étude précédente montre alors que dans ce cas

$$[U]_{1,2} = L_{1,1}^{-1}[A]_{1,2} = [A]_{1,2} \text{ et } [L]_{2,1} = [A]_{2,1}U_{1,1}^{-1} = [A]_{2,1}/[A]_{1,1}$$

$$\mathcal{S} = [A]_{2,2} - [A]_{2,1} \times [A]_{1,2}/[A]_{1,1}$$

cette dernière relation peut encore s'écrire

$$\forall i, j \quad 2 \leq i, j \leq n \qquad \mathcal{S}_{i-1,j-1} = [A]_{i,j} - \frac{[A]_{i,1} \times [A]_{1,j}}{[A]_{1,1}}.$$

Ainsi à partir des coefficients de A, on a obtenu un début de factorisation :
– en reprenant telle quelle la première ligne de la matrice : $[U]_{1,2} = [A]_{1,2}$
– en divisant la première colonne de la matrice : $[L]_{2,1} = [A]_{2,1}/[A]_{1,1}$
– en calculant le complément de Schur : $\mathcal{S} = [A]_{2,2} - [A]_{2,1} \times [A]_{1,2}/[A]_{1,1}$.

A ce stade on voit que l'on peut continuer la factorisation de la matrice A, s'il est possible d'effectuer les mêmes opérations sur la matrice $\mathcal{S}$, et pour cela il faut supposer $\mathcal{S}_{1,1} \neq 0$! Si cela est vrai, alors on applique la méthode précédente à la matrice $\mathcal{S}$ de rang $n-1$ et ainsi de suite... le rang de la matrice à factoriser diminuant d'une unité à chaque étape. En notant $A^{(k)}$ et $\mathcal{S}^{(k+1)}$ les matrices générées successivement par la méthode (on prend $A^{(0)} = A$, et $\mathcal{S}^{(1)} = \mathcal{S}$), on résume les opérations dans l'algorithme

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

si $A_{k,k}^{(k-1)} \neq 0$

$[L]_{k,k} = 1, \quad [U]_{k,k} = A_{k,k}^{(k-1)}$

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

$[L]_{i,k} = A_{i,k}^{(k-1)} / A_{k,k}^{(k-1)}$

$[U]_{k,i} = A_{k,i}^{(k-1)}$

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

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

$\mathcal{S}_{i-k,j-k}^{(k)} = A_{i,j}^{(k)}$

fin

fin

fin

Si en cours de calcul, on rencontre un coefficient diagonal $A_{k,k}^{(k-1)}$ nul, on recherche dans la colonne k de la matrice $A^{(k-1)}$ s'il existe un coefficient non nul, soit $A_{i,k}^{(k-1)}$ pour $i > k$, et on échange alors les lignes i et k de la matrice. Cette modification revient à multiplier à gauche la matrice courante $\mathcal{S}^{(k)}$ par une matrice de permutation $P^{(k)}$:

$$P^{(k)}\mathcal{S}^{(k)} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathcal{S}_{k,k}^{(k)} & \cdot & \cdot & \mathcal{S}_{k,i}^{(k)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathcal{S}_{i,k}^{(k)} & \cdot & \cdot & \mathcal{S}_{i,i}^{(k)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \mathcal{S}_{i,k}^{(k)} & \cdot & \cdot & \mathcal{S}_{i,i}^{(k)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathcal{S}_{k,k}^{(k)} & \cdot & \cdot & \mathcal{S}_{k,i}^{(k)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Replacée dans le contexte de la matrice de rang n, cette échange de lignes revient à multiplier à gauche la matrice $A^{(k-1)}$ par une matrice de permutation $P(i,k) \in \mathbb{R}^{n \times n}$

$$P(i,k) = \begin{pmatrix} I_{n-k} & 0 \\ 0 & P^{(k)} \end{pmatrix}$$

soit encore

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

Dans la factorisation en cours, cette multiplication n'affecte donc pas les lignes d'indice inférieur à k déjà calculées, mais seulement les lignes i et k des matrices L et $A^{(k)}$.

Que se passe-t-il si pour un k donné, on ne trouve aucun coefficient $A_{i,k}^{(k-1)}$ différent de zéro dans la colonne k? Cela veut dire que l'on a obtenu une factorisation de la forme

$$A = \begin{pmatrix} [A]_{1,1} & [A]_{1,2} \\ [A]_{2,1} & [A]_{2,2} \end{pmatrix} = \begin{pmatrix} [L]_{1,1} & 0 \\ [L]_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} [U]_{1,1} & [U]_{1,2} \\ 0 & \mathcal{S}^{(k)} \end{pmatrix}$$

et la première colonne de la matrice $\mathcal{S}^{(k)}$ est nulle ($\det\mathcal{S}^{(k)} = 0$), alors

$$\det A = \det[U]_{1,1}\ \det\mathcal{S}^{(k)} = 0$$

ce qui est contraire à l'hypothèse A inversible !

Au cours des calculs on peut être obligé de faire plusieurs permutations de lignes pour amener des coefficients non nuls sur la diagonale ; à la fin des opérations on a alors factorisé la matrice

$$P\ A = P(i_m, k_m)\ P(i_{m-1}, k_{m-1})\ \ldots P(i_1, k_1)\ A = L\ U$$

La matrice P est une matrice de permutation qui prend en compte l'historique de ces modifications ; on a ainsi démontré le résultat suivant

Théorème 6.10.1 *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.

6.11 Stabilité numérique

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 $A^{(k-1)}_{k,k}$ soit petit, de l'ordre de ε, alors les formules de calcul

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

montrent que dans la matrice $\mathcal{S}^{(k)}$ résultante 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 6.11.1 *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.*

Ce résultat montre que la matrice $\mathcal{S}^{(k)} \in \mathbb{R}^{n_2\times n_2}$ est singulière si $n_2 > 1$! 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 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 $A^{(k-1)}_{i,k}$ 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 $A^{(k-1)}_{i,j}$ pour $i \geq k$ et $j \geq k$. Si on choisit un coefficient $A^{(k-1)}_{i,j}$ en dehors de la colonne k, il faut ajouter à la permutation de lignes $P(i,k)$ une permutation de colonnes qui correspond à une multiplication à droite de la matrice A par une matrice de permutation $Q(j,k)$. En fin de factorisation, on a obtenu

$$P\ A\ Q = L\ U.$$

Remarque 6.11.1 *En reprenant les notations précédentes, on voit qu'une permutation de lignes $P(i,k)$ ou de colonnes $Q(k,j)$ de la matrice $\mathcal{S}^{(k)}$ à l'étape k, ne modifie pas les blocs déjà calculés $[L]_{1,1}$ et $[U]_{1,1}$:*

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

soit encore

$$P(i,k)\ A\ Q(k,j) = \begin{pmatrix} [L]_{1,1} & 0 \\ P[L]_{2,1} & I \end{pmatrix} \begin{pmatrix} [U]_{1,1} & [U]_{1,2}Q \\ 0 & P\mathcal{S}^{(k)}Q \end{pmatrix} = \begin{pmatrix} [L]_{1,1} & 0 \\ P[L]_{2,1} & P[L]_{2,2} \end{pmatrix} \begin{pmatrix} [U]_{1,1} & [U]_{1,2}Q \\ 0 & [U]_{2,2}Q \end{pmatrix}.$$

Proposition 6.11.1 *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 L et L', U et U' qui vérifient

$$P\ A\ Q = L\ U = L'\ U';$$

alors

$$(L')^{-1}L = U'U^{-1} = M \quad \text{soit} \quad L'M = L, \quad \text{et} \quad U' = MU.$$

Si on écrit la relation $L'M = L$ colonne par colonne, on obtient pour la colonne j

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \mathrm{x} & 1 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & 1 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 \end{pmatrix} \begin{pmatrix} M_{1,j} \\ M_{2,j} \\ \vdots \\ \vdots \\ M_{n,j} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \mathrm{x} \end{pmatrix} j$$

on voit donc que nécessairement dans la colonne M_j, les composantes d'indice strictement inférieur à j sont nulles et que $M_{j,j} = 1$; M est donc une matrice triangulaire inférieure à diagonale unité. Le même raisonnement sur la relation $(U')^T = U^T M^T$ montre que M est aussi une matrice triangulaire supérieure à diagonale unité, soit finalement $M = I_n$ et

$$L = L' \quad \text{et} \quad U = U'.$$

■

6.12 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 dans la suite du cours - 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 de l'ordre de $2n(n+1)$ 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 Cholesky $A = LL^T$, avec L matrice triangulaire inférieure. A doit être symétrique définie positive.

– 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 Gauss $A = LU$, avec L matrice triangulaire inférieure à diagonale unité, et U matrice triangulaire supérieure. A doit être inversible.

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.

6.13 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

Remarque 6.13.1 *ces formules sont différentes de celles de la méthode de Schur, mais elles calculent les mêmes matrices car la factorisation est unique.*

Exercice 6.13.1 *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

6.14 Coût calcul

Proposition 6.14.1 *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 $n^3/3$.*

Preuve : Traitons d'abord le coût de la factorisation de Gauss par les formules précédentes : pour déterminer la colonne k de L (respectivement la ligne k de U), il faut calculer $n-k-1$ coefficients (respectivement $n-k$ coefficients), chacun d'eux nécessitant k opérations élémentaires (+,*) soit à l'étape k de l'ordre de $2(n-k)k$ opérations élémentaires. Pour k variant de 1 à n, on a finalement un nombre total d'opérations en $n^3/3$. Maintenant que les matrices L et U sont calculées, le calcul de la composante k du vecteur y vérifiant $Ly = b$ comprend k opérations élémentaires (+,*), de même que celui de la composante k du vecteur x vérifiant $Ux = y$. Chaque résolution de système linéaire triangulaire nécessite donc $n(n+1)/2$ opérations élémentaires (+,*).

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

A noter que la partie la plus coûteuse de l'algorithme est la factorisation $A = LU$, en conséquence lorsque l'on a plusieurs systèmes linéaires à résoudre avec la même matrice A, on ne calcule les matrices L et U qu'une seule fois !

Exercice 6.14.1 *Quel est le temps calcul de la solution d'un système linéaire d'ordre* 20 *en utilisant cette méthode ?*

Réponse : $2,33 \times 10^{-9}$ s = 2,33 nanosecondes !

6.15 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

$$A = LU = LD\tilde{U}$$

encore 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 on a alors

Proposition 6.15.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible et symétrique, à une matrice de permutation P près, la factorisation $PAP^T = LDL^T$ est unique.*

$$A = LD\tilde{U} = A^T = \tilde{U}^T D^T L^T$$

et en utilisant la Proposition 6.11.1 sur l'unicité de la factorisation de Gauss, on a

$$L = \tilde{U}^T \quad \text{et} \quad D\tilde{U} = D^T L^T$$

on obtient ainsi la **factorisation de Crout** :

$$A = LDL^T.$$

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

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

$L_{k,k} = 1 \quad$ et $\quad D_{k,k} = A_{k,k} - \sum_{j<k} L_{k,j}^2$

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

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

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.

Noter que pour transmettre la propriété de symétrie aux différentes matrices qui interviennent dans le calcul précédent, il faut effectuer une permutation simultanée des lignes et colonnes de la matrice $A^{(k)}$; ainsi à la fin des opérations la matrice Q de la Proposition 6.11.1 n'est autre que P^T.

Ceci implique 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.

6.16 Factorisation de Cholesky

Supposons maintenant que $A \in \mathbb{R}^{n\times n}$ soit une matrice symétrique définie positive, il résulte de la Proposition 6.15.1 que l'on peut écrire $P\ A\ P^T = L\ D\ L^T$; de plus on peut écrire

$$\forall x \in \mathbb{R}^n,\ x \neq 0 \qquad (x, Ax) = (x, P^{-1}LDL^T Px) > 0$$

pour tout $y = L^T Px \neq 0$, on a donc $= (y, Dy) > 0$, la matrice diagonale D est donc 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 $D_{i,i}^{1/2} = \sqrt{D_{i,i}}$. En posant alors $\mathcal{L} = LD^{1/2}$, on obtient finalement

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

Proposition 6.16.1 *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 $k = 1, \dots n$ **faire**

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

pour $i = k+1, \dots n$ **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 6.16.1 *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 6.11.1, 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 la Proposition 6.11.1, 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 6.11.1 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 $A_{1,1} = A_{1,1}^T \in \mathbb{R}^{n_1 \times n_1}$, $A_{2,1} \in \mathbb{R}^{n_2 \times n_1}$, $A_{1,2} = A_{2,1}^T \in \mathbb{R}^{n_1 \times n_2}$, $A_{2,2} = A_{2,2}^T \in \mathbb{R}^{n_2 \times n_2}$, $n = n_1 + n_2$ et 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

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

Proposition 6.16.2 *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

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

■

Remarque 6.16.1 *Une conséquence importante de ce résultat est que les permutations de lignes ou de colonnes sont inutiles pour effectuer la factorisation de Cholesky; la matrice P de la Proposition 6.16.1 est donc la matrice identité I_n.*

6.17 Coût calcul

Proposition 6.17.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique (éventuellement 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 Crout (ou Cholesky) est de l'ordre de $n^3/6$.*

Preuve : A partir de la Proposition 6.14.1, il suffit de remarquer les matrices L ou $\mathcal{L}$ ont moitié moins de coefficients que les matrices L **et** U, en conséquence le coût calcul de la factorisation est divisé par deux ! ■

6.18 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**

$[L]_{k,k}\ [U]_{k,k} = [A]_{k,k} - \sum_{j<k} [L]_{k,j}[U]_{j,k}$ *factorisation ponctuelle* du bloc $[A]_{k,k}$

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 :

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

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.

Il existe également la factorisation de Crout par blocs

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

$[L]_{k,k}[D]_{k,k}[L]^T_{k,k} = [A]_{k,k} - \sum_{j<k} [L]_{k,j}[L]^T_{k,j}$ *factorisation ponctuelle* du bloc $[A]_{k,k}$

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 la factorisation de Cholesky par blocs

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

$[L]_{k,k}[L]_{k,k}^T = [A]_{k,k} - \sum_{j<k} [L]_{k,j}[L]_{k,j}^T$ *factorisation ponctuelle* du bloc $[A]_{k,k}$

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

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

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 par exemple 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.

6.19 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 de coefficients non nuls par ligne est petit (de l'ordre de la dizaine) 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 (différentes formes de représentation de telles matrices sont détaillées dans le chapitre 16). 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. Pour chaque ligne k de la matrice A, on peut définir $il(k)$ le plus petit indice de colonne l inférieur ou égal à k tel que $A_{k,l} \neq 0$; on définit de même pour chaque colonne k, $ic(k)$ le plus petit indice de ligne l inférieur ou égal à k, tel que $A_{l,k} \neq 0$. En supposant la matrice A inversible, même si tous les coefficients extra-diagonaux sont nuls, on a $il(k) = k$ (respectivement $ic(k) = k$).

On introduit alors les ensembles

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

et

$$Pc(A) = \{(l,k), 1 \leq k \leq n, ic(k) \leq l \leq k\}$$

Quand la matrice A est symétrique

$$(k,l) \in Pl(A) \Longleftrightarrow (l,k) \in Pc(A)$$

dans le cas d'une matrice non symétrique, on supposera que cette propriété est encore satisfaite (c'est vrai pour les cas qui nous intéressent).

Définition 6.19.1 *on appelle* **profil** *de la matrice A*
– *l'ensemble $Pr(A) = \{(k,l), 1 \leq k \leq n, il(k) \leq l \leq k\} = Pl(A)$ si A est symétrique*
– *l'ensemble $Pr(A) = Pl(A) \cup Pc(A)$ sinon*

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

Remarque 6.19.1 *1) dans le cas A symétrique, la définition $Pr(A) = Pl(A)$ fait intervenir la partie triangulaire inférieure de la matrice A, on peut utiliser de manière équivalente la partie triangulaire supérieure : $Pr(A) = Pc(A)$.*

2) on voit que 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.*

3) 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 6.19.1 *Les factorisations de Gauss, Crout et Cholesky conservent le profil.*

Preuve : Dans la factorisation de Gauss

$$LU = A$$

on examine la colonne k de la matrice U, qui est solution d'un système linéaire triangulaire de la forme

$$\begin{pmatrix} \cdot & 0 & 0 & 0 & 0 \\ \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \mathrm{x} \\ \mathrm{x} \\ \mathrm{x} \\ \mathrm{x} \\ \mathrm{x} \end{pmatrix} = \begin{pmatrix} 0 \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{pmatrix} \; il(k)$$

dans lequel le second membre est la colonne k de la matrice A. Comme pour la Proposition 6.11.1, on en déduit que les $il(k) - 1$ premières composantes de la colonne k de la matrice U sont nulles. Cette propriété est vraie pour toutes les colonnes, $k = 1, \ldots, n$. Les matrices U et A ont donc même profil (pour la partie triangulaire supérieure), en conséquence $Pc(U) = Pc(A)$; on montre de même $Pl(L) = Pl(A)$. Cette propriété est commune aux trois factorisations, la structure des calculs étant identique. ■

Remarque 6.19.2 *ce résultat est valable pour le profil par points comme pour le profil par blocs.*

Remarque 6.19.3 *Pour les matrices symétriques 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 $l = [\sum_{k=1}^{n} l(k)]/n$ (la notion de largeur de bande est détaillée au chapitre 16). Pour les matrices creuses la largeur de bande moyenne l est en général petite devant n ; pour ce type de matrice, on montre que le coût calcul de la factorisation de Cholesky est de l'ordre de nl^2.*

Ce qu'il faut retenir

1. les formules classiques de calcul de la solution d'un système linéaire (type formules de Cramer) ne sont pas adaptées à résolution sur ordinateur des grands systèmes linéaires.
2. des algorithmes spécifiques doivent être utilisés ; leur principe repose sur l'utilisation de méthodes de résolution de systèmes linéaires à matrice triangulaire.
3. toute matrice carrée A peut s'écrire sous la forme d'un produit de deux matrices triangulaires. Quand A est inversible cette propriété permet de définir
   (a) la factorisation de Gauss : $A = LU$, avec L matrice triangulaire inférieure avec des 1 sur la diagonale, et U matrice triangulaire supérieure inversible,
   (b) si A est symétrique, la factorisation de Crout : $A = LDL^T$, avec L matrice triangulaire inférieure avec des 1 sur la diagonale, et D matrice diagonale inversible,
   (c) si A est symétrique définie positive, la factorisation de Cholesky : $A = LL^T$, avec L matrice triangulaire inférieure inversible.
4. les matrices L, D et U sont uniques aux permutations de lignes et colonnes près.
5. les factorisations de Gauss, Crout et Cholesky conservent le profil.

Chapitre 7

Normes vectorielles et matricielles

7.1 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.

7.2 Normes de vecteurs

Définition 7.2.1 *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**.

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

$$\|x\| = (x, x)^{1/2}$$

par exemple le produit scalaire euclidien dans $\mathbb{R}^n$ est défini suivant :

$$\|x\| = (x,x)^{1/2} = (\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2)^{1/2}$$

Remarque 7.2.1 *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 7.2.1 *[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 :

$$\begin{aligned} \forall x, y \in \mathbb{E}, \forall \lambda \in \mathbb{R} \qquad (\lambda x + y, \lambda x + y) &= \|\lambda x + y\|^2 \\ &= \lambda^2 (x,x) + 2\lambda (x,y) + (y,y) \\ &= \lambda^2 \|x\|^2 + 2\lambda (x,y) + \|y\|^2 \end{aligned}$$

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 **norme 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 montre 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

$$\frac{1}{p} + \frac{1}{q} = 1.$$

L'inégalité de Schwarz est donc un cas particulier de l'inégalité de Hölder.

Définition 7.2.2 *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 7.2.1 *Dans un espace vectoriel $\mathbb{E}$ sur $\mathbb{R}$ de dimension finie toutes les normes sont équivalentes.*

On admettra ce résultat qui prend les formes particulières suivantes, dans un espace vectoriel $\mathbb{E}$ de dimension finie n :

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\ \|x\|_2$$
$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\ \|x\|_\infty$$
$$\|x\|_\infty \leq \|x\|_1 \leq n\ \|x\|_\infty.$$

Proposition 7.2.2 *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. ∎

7.3 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 7.3.1 *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 de matrices A, B 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 7.3.3)

$$\|A\|_F^2 = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} |A_{i,j}|^2.$$

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 $\mathbb{R}^{m\times n}$; 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 7.2.1 et aussi à la Définition 7.3.1 (voir la Proposition 7.3.3) ; 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 7.3.1 *Pour toute matrice carrée $A \in \mathbb{R}^{n\times n}$ les normes matricielles de Hölder vérifient*

i) $$\|A\|_1 = \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_j \sum_i |A_{i,j}|.$$

ii) $$\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_i \sum_j |A_{i,j}|.$$

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 7.3.2 *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

$$I_{m,n} = \begin{pmatrix} I_p \\ 0 \end{pmatrix} \quad \text{si } m > n, \quad I_{m,n} = \begin{pmatrix} I_p & 0 \end{pmatrix} \quad \text{si } m < n, \text{ et} \quad I_{m,n} = I_p \quad \text{si } n = m = p;$$

par définition $\|I_{n,m}\|_F = \sqrt{p}$, alors que pour toute norme induite on doit avoir

$$\|I_{m,n}\|_{m,n} = \max_{x\neq 0} \frac{\|I_{m,n}x\|_m}{\|x\|_n} = 1.$$

∎

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 7.3.3 *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 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 7.3.4 *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

$$\|Ax\|_m^2 = \sum_{i=1}^{m}(\sum_{j=1}^{n}A_{i,j}x_j)^2 \leq \sum_{i=1}^{m}[(\sum_{j=1}^{n}A_{i,j}^2)(\sum_{j=1}^{n}x_j^2)] = (\sum_{i=1}^{m}\sum_{j=1}^{n}A_{i,j}^2)(\sum_{j=1}^{n}x_j^2) = \|A\|_F^2\ \|x\|_n^2,$$

donc $\|A\|_{m,n} \leq \|A\|_F$.

Enfin en majorant chaque coefficient de la matrice A, puis en utilisant la première inégalité, on obtient

$$\|A\|_F \leq \sqrt{mn}\max_{i,j}|A_{i,j}| \leq \sqrt{mn}\|A\|_{m,n}$$

■

Remarque 7.3.1 *cette proposition illustre encore un cas particulier d'équivalence des normes :*

$$\|A\|_{m,n} \leq \|A\|_F \leq \sqrt{mn}\|A\|_{m,n}$$

et

$$\frac{1}{\sqrt{mn}}\|A\|_F \leq \|A\|_{m,n} \leq \|A\|_F.$$

Proposition 7.3.5 *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 7.2.2 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 7.3.2 *cette propriété se généralise dans le cas complexe aux matrices U unitaires :*

$$\|UA\|_2 = \|AU\|_2 = \|A\|_2.$$

7.4 Valeurs propres

Il existe de nombreuses relations entre les valeurs propres et les normes des matrices carrées. Avant de les étudier il faut faire quelques rappels (le chapitre 10 est dédié à l'étude des valeurs propres et vecteurs propres des matrices, certaines définitions sont reprises ici pour simplifier la lecture du polycopié).

Définition 7.4.1 *soit $A \in \mathbb{R}^{n\times n}$ une matrice, on appelle* **polynôme caractéristique** *de A le polynome de degré n en la variable λ défini par*

$$p(\lambda) = dét(A - \lambda I_n).$$

On appelle **valeur propre** *λ_i de la matrice A toute racine du polynôme caractéristique. Noter que si A est une matrice à coefficients réels, le polynôme caractéristique admet n racines λ_i complexes, distinctes ou non.*

Si λ_i est valeur propre de A, alors la matrice $A(\lambda_i) = A - \lambda_i I_n$ est singulière et le noyau de l'application linéaire associée n'est pas réduit au vecteur nul. Soit v_i un vecteur **non nul** de Ker $f(\lambda_i)$, il vérifie

$$A(\lambda_i)v_i = Av_i - \lambda_i v_i = 0$$

soit

$$Av_i = \lambda_i v_i.$$

Un tel vecteur est appelé **vecteur propre** associé à la valeur propre λ_i.

Définition 7.4.2 *le* **rayon spectral** *de la matrice $A \in \mathbb{R}^{n\times n}$ est le réel positif*

$$\rho(A) = \max_i |\lambda_i(A)|.$$

7.5 Normes des matrices et valeurs propres

Proposition 7.5.1 *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 est dans [13]. 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 7.5.2 *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 10.3.4) : 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

$$\|(QD)^{-1} B (QD)\|_\infty = \max_{y\neq 0} \frac{\|(QD)^{-1} B (QD) y\|_\infty}{\|y\|_\infty} = \max_{x\neq 0} \frac{\|(QD)^{-1} B x\|_\infty}{\|(QD)^{-1} x\|_\infty}.$$

■

Proposition 7.5.3 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$*

$$\rho(A) \leq \|A\|_2 = \rho^{1/2}(A^T A);$$

pour toute matrice $A \in \mathbb{R}^{n\times n}$ symétrique

$$\rho(A) = \|A\|_2.$$

Preuve : D'après la Proposition 7.5.1 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 $\{v_1, v_2, \ldots, v_n\}$ (voir la Proposition 10.3.3); en rangeant les valeurs propres associées $\lambda_i(A^T A)$ par ordre de module décroissant :

$$\lambda_n(A^T A) \leq \lambda_{n-1}(A^T A) \leq \ldots \leq \lambda_2(A^T A) \leq \lambda_1(A^T A)$$

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

$$\frac{(Au, Au)}{(u,u)} = \frac{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i(A^T A)|^2 (v_i, v_i)}{\sum_{i=1}^n |\alpha_i^2| (v_i, v_i)} \leq \max_{i=1,n} |\lambda_i(A^T A)|^2 = \rho(A^T A).$$

Dans le cas d'une matrice A symétrique, on écrit de même

$$\frac{(Au, Au)}{(u,u)} = \frac{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i^2(A)|^2 (v_i, v_i)}{\sum_{i=1}^n |\alpha_i^2| (v_i, v_i)} \leq \max_{i=1,n} |\lambda_i^2(A)|^2 = \rho^2(A).$$

■

Plus généralement, pour toute matrice $A \in \mathbb{R}^{n\times n}$ la matrice A^TA est symétrique, et ses valeurs propres sont réelles positives. A toute valeur propre $\lambda(A^TA) \geq 0$, on associe la **valeur singulière** $\sigma(A)$ de la matrice A par la relation

$$\sigma(A) = \lambda^{1/2}(A^TA).$$

De manière classique, on range les valeurs propres $\lambda_i(A^TA)$ par ordre de module décroissant :

$$\lambda_n(A^TA) \leq \lambda_{n-1}(A^TA) \leq \ldots \leq \lambda_2(A^TA) \leq \lambda_1(A^TA)$$

et de façon cohérente, les valeurs singulières de la matrice A :

$$\sigma_n(A) \leq \sigma_{n-1}(A) \leq \ldots \leq \sigma_2(A) \leq \sigma_1(A).$$

La Proposition 7.5.3 peut alors se résumer sous la forme $\|A\|_2 = \sigma_1(A)$.

Proposition 7.5.4 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ la norme de Schur-Frobenius vérifie*

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |A_{i,j}|^2 = \sum_{i=1,n} \lambda_i(A^TA) = \sum_{i=1,n} \sigma_i^2(A).$$

Preuve : par définition de la norme de Frobenius

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |A_{i,j}|^2 = tr(A^TA) = \sum_{i=1,n} \lambda_i(A^TA) = \sum_{i=1,n} \sigma_i^2(A)$$

où $tr(B)$ désigne la trace de la matrice B qui est aussi la somme de ses valeurs propres. ■

Remarque 7.5.1 *ces résultats s'étendent au cas d'une matrice A complexe, en remplaçant dans les formules A^T par A^*.*

7.6 Suites de vecteurs. Suites de matrices

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 $\mathbb{R}$, 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

$$\lim_{k\mapsto\infty} \|x^k - x\| = 0$$

que l'on écrit de manière classique $\lim_{k\mapsto\infty} x^k = x$. 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 $\{x_i^k\}_{k\in\mathbb{N}}$ vers x_i pour $i = 1, 2, \ldots, n$.

On définit de manière analogue une suite de matrices $\{A_k\}_{k\in\mathbb{N}} \in \mathbb{R}^{m\times n}$ et on peut utiliser la définition de la convergence d'une suite de vecteurs dans l'espace vectoriel $\mathbb{R}^{m\times n}$ 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 $A \in \mathbb{R}^{n\times n}$. Dans ce cas particulier, on relie la convergence de cette suite au rayon spectral de la matrice A.

Théorème 7.6.1 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ 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 la valeur propre de A telle que $\rho(A) = |\lambda|$ et soit $u \neq 0$ un vecteur propre associé, alors

$$(A^* Au, u) = \overline{\lambda}\lambda(u, u) \qquad \text{soit encore } \|Au\|_2 = \rho(A)\|u\|_2.$$

La suite de vecteurs $x^k = A^k u$ vérifie donc

$$\|x^k\|_2 = \|A^k u\|_2 = \rho(A)^k \|u\|_2$$

si on suppose $\rho(A) \geq 1$ alors $\lim_{k\to\infty} \|A^k u\|_2 = +\infty$, soit pour k assez grand $\|A^k u\|_2 \neq 0$.

iii) $\Longrightarrow$ iv) : est une conséquence de la Proposition 7.5.2 (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 7.6.1 *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 7.6.1 *Calculer les normes $\|\cdot\|_1$, $\|\cdot\|_\infty$, $\|\cdot\|_F$ ainsi que le rayon spectral des matrices*

$$A = \begin{pmatrix} 0.9 & 0.0 \\ 0.4 & 0.8 \end{pmatrix} \quad \textit{et} \quad B = \begin{pmatrix} 5. & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \begin{pmatrix} 0.9 & 0.0 \\ 0.4 & 0.8 \end{pmatrix} \begin{pmatrix} 0.2 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}.$$

7.7 Conditionnement des matrices

Un problème important qui se pose très souvent dans la résolution numérique des systèmes linéaires est la sensibilité de la solution par rapport aux variations des données : matrice ou second membre. Il s'agit d'une notion fondamentale qu'il est indispensable de préciser si l'on ne veut pas être surpris par la mauvaise qualité d'un résultat numérique, ou par le mauvais comportement d'un algorithme.

Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible et $b \in \mathbb{R}^n$ un vecteur. On cherche à résoudre le système linéaire $Ax = b$, et l'on suppose que les données sont connues (ou représentées dans un ordinateur) avec une certaine imprécision : $b + \delta b$, la solution associée est elle -même calculée avec une imprécision $x + \delta x$ qui vérifie

$$A(x + \delta x) = b + \delta b.$$

Comme $Ax = b$, on en déduit que

$$\delta x = A^{-1}\delta b.$$

Ainsi pour toute norme vectorielle $\|\cdot\|$:

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \times \|A^{-1}\| \, \frac{\|\delta b\|}{\|b\|}.$$

De manière analogue, si on considère maintenant une perturbation de la matrice A, on a

$$(A + \delta A)(x + \delta x) = b.$$

et de la relation

$$\delta x = -A^{-1}\delta A(x + \delta x)$$

on tire

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A\| \times \|A^{-1}\| \, \frac{\|\delta A\|}{\|A\|}.$$

Le produit $\|A\| \times \|A^{-1}\|$ est appelé **nombre de conditionnement** de la matrice A :

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

Par sa définition, la valeur précise de ce nombre dépend de la norme choisie, en général on utilise $\mathrm{cond}_2(A) = \|A\|_2 \times \|A^{-1}\|_2$ qu'il est assez facile d'estimer. Le nombre de conditionnement est aussi noté $\kappa(A)$, et joue un rôle important dans la résolution numérique des systèmes linéaires, comme le montre l'exemple suivant : pour tout réel a la matrice triangulaire inversible $A \in \mathbb{R}^{n\times n}$

$$A = \begin{pmatrix} 1 & -a & 0 & \dots & \dots & 0 \\ 0 & 1 & -a & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -a \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

admet 1 comme valeur propre d'ordre n, quel que soit n. Son inverse est la matrice

$$A^{-1} = \begin{pmatrix} 1 & a & a^2 & \dots & \dots & a^{n-1} \\ 0 & 1 & a & a^2 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & a \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Si on veut calculer la solution du système linéaire $Ax = b$, on constate qu'une variation δ de la seule composante b_n du second membre produit une variation de $a^{n-1}\delta$ de x_1, ce qui est catastrophique dès que $|a| > 1$! Ce mauvais comportement est mesuré par le nombre de conditionnement

$$\mathrm{cond}_2(A) = \|A\|_2 \times \|A^{-1}\|_2 \geq |a|^{n-1}.$$

Remarque 7.7.1 *comme le montre cet exemple le nombre de conditionnement n'a aucun lien avec les valeurs propres de la matrice dans le cas général. Le théorème suivant établit un lien dans le cas des matrices normales.*

Théorème 7.7.1 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ inversible, on a les propriétés suivantes :*

i) $\mathrm{cond}(A) \geq 1$

ii) $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$

iii) $\forall \alpha \in \mathbb{R}, \alpha \neq 0 \qquad \mathrm{cond}(\alpha A) = \mathrm{cond}(A)$

iv) $\mathrm{cond}_2(A) = \sigma_1(A)/\sigma_n(A)$

v) *si A est normale* $\mathrm{cond}_2(A) = |\lambda_1(A)|/|\lambda_n(A)|$

vi) *si A est orthogonale ou unitaire* $\mathrm{cond}_2(A) = 1$

vii) *pour toute matrice Q unitaire* $\mathrm{cond}_2(QA) = \mathrm{cond}_2(AQ) = \mathrm{cond}_2(A)$.

Preuve : Où $\sigma_1(A)$ et $\sigma_n(A)$ sont respectivement la plus grande et la plus petite valeur singulière de la matrice A, $\lambda_1(A)$ et $\lambda_n(A)$ sont respectivement les valeurs propres de A de plus grand et plus petit module. La notion de valeur singulière d'une matrice est étudiée en détails dans la Partie 1, on se contentera ici de la définition $\sigma_i(A) = \lambda_i^{1/2}(A^T A)$.

De l'identité $AA^{-1} = I_n$, on tire pour toute norme matricielle,

$$\|I_n\| \leq \|A\| \times \|A^{-1}\|.$$

Ce qui démontre le point i), les propriétés ii) et iii) sont évidentes. Par ailleurs

$$\|A\|_2^2 = \max_{x \neq 0} \frac{(Ax, Ax)}{(x, x)} = \max_i |\lambda_i(A^* A)| = \rho(A^* A) = \sigma_1^2(A)$$

et

$$\|A^{-1}\|_2^2 = \rho([A^{-1}]^*[A^{-1}]) = \max_i |\lambda_i([A^{-1}]^*[A^{-1}])| = \frac{1}{|\lambda_n(A^* A)|} = \frac{1}{\sigma_n^2(A)}.$$

Si la matrice A est normale, alors $\lambda_i(A) = \sigma_i(A)$; cette classe de matrices comprend le cas particulier important des matrices symétriques ou hermitiennes.

Enfin, pour tout matrice Q orthogonale ou unitaire $\|Q\|_2 = 1 = \|Q^{-1}\|_2$ et pour tout matrice $A \in \mathbb{R}^{n \times n}$, $\|QA\|_2 = \|A\|_2 = \|AQ\|_2$. ∎

7.8 Séries de vecteurs. Séries de matrices

De même que la notion de suites de vecteurs et de matrices découle naturellement de la notion de suites de scalaires, on définit les séries de vecteurs et de matrices à partir des séries des composantes et des coefficients, et les résultats classiques de convergence des séries scalaires peuvent être utilisés. Dans le cas des séries de matrices, on peut aussi tirer profit de la connaissance de leurs valeurs propres.

Théorème 7.8.1 *Soit $\sum_{k \in \mathbb{N}} \alpha_k z^k$ une série entière de nombres complexes, de rayon de convergence $r > 0$, alors pour toute matrice $A \in \mathbb{R}^{n \times n}$ de rayon spectral $\rho(A) < r$ la série $\sum_{k \in \mathbb{N}} \alpha_k A^k$ est convergente.*

Preuve : Il suffit de constater que pour toute norme matricielle

$$\| \sum_{k \in \mathbb{N}} \alpha_k A^k \| \leq \sum_{k \in \mathbb{N}} |\alpha_k| \|A^k\| \leq \sum_{k \in \mathbb{N}} |\alpha_k| \|A\|^k.$$

en supposant maintenant $\rho(A) < r$ et en utilisant la Proposition 7.5.2, il existe une norme pour laquelle $\|A\| < r$ et la série majorante converge car une série entière est absolument convergente à l'intérieur de son disque de convergence. ∎

On peut à l'aide de ce Théorème introduire la notion de **fonction de matrice** : soit f une fonction développable en série entière au voisinage de $z_0 = 0$ et de rayon de convergence $r > 0$: pour tout z tel que $|z| < r$ on écrit

$$f(z) = \sum_{k \in \mathbb{N}} \alpha_k z^k,$$

avec $|f(z)| < \infty$; à partir de cette relation, on définit pour toute matrice $A \in \mathbb{R}^{n \times n}$ telle que $\rho(A) < r$, la quantité

$$f(A) = \sum_{k \in \mathbb{N}} \alpha_k A^k.$$

$f(A)$ est une matrice de $\mathbb{R}^{n\times n}$ qui vérifie $\|f(A)\| < \infty$. Avec cette méthode, on peut par exemple introduire la notion d'exponentielle de matrice :

$$e^A = \sum_{k\in\mathbb{N}} \frac{A^k}{k!}$$

qui est toujours convergente puisque la série entière correspondante a un rayon de convergence infini !

Théorème 7.8.2 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ de rayon spectral $\rho(A) < 1$ la série $\sum_{k\in\mathbb{N}} A^k$ est convergente, et vérifie*

$$(I_n - A)^{-1} = \sum_{k\in\mathbb{N}} A^k.$$

Preuve : Par application du Théorème 7.8.1 à la fonction $f(z) = 1/(1-z)$, la série

$$f(z) = \sum_{k\in\mathbb{N}} z^k$$

étant convergente dans le disque unité, la série $\sum_{k\in\mathbb{N}} A^k$ est convergente pour toute matrice $A \in \mathbb{R}^{n\times n}$ telle que $\rho(A) < 1$. Il suffit alors d'écrire l'identité

$$(I_n - A) \times \sum_{k=0}^{p} A^k = I_n - A^{p+1}$$

puisque la série est convergente pour $\rho(A) < 1$, la matrice A^{p+1} tend vers la matrice nulle quand p tend vers l'infini. ■

Ce qu'il faut retenir

1. les normes vectorielles et matricielles sont les outils indispensables pour étudier la convergence et la précision des méthodes de résolution des systèmes linéaires, ainsi que des algorithmes de calcul des valeurs propres.
2. il existe des liens entre ces normes et les valeurs propres des matrices.

Chapitre 8

Les méthodes itératives

8.1 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.

Les méthodes classiques : Jacobi, Gauss-Seidel, relaxation, Richardson et S.S.O.R. sont ensuite présentées, avec les principaux résultats de convergence.

8.2 Décomposition régulière

Définition 8.2.1 *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

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Mx^{k+1} = Nx^k + b \\ \textbf{fin} \end{array}\right. \tag{8.1}$$

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

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

Cette méthode peut aussi s'écrire sous la forme

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + M^{-1} r^k \\ \textbf{fin} \end{array} \right. \tag{8.2}$$

avec $r^k = b - Ax^k$ vecteur résidu.

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 8.2.1 *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 7.6.1. ∎

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 8.2.1 *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 propriété générale $\rho(M^{-1}N) \leq \|M^{-1}N\|$ en choisissant la norme vectorielle induite par la norme $\|v\|_A = \sqrt{(v, Av)}$ qui est bien une norme vectorielle puisque A est symétrique définie positive.

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

Soit $v \in \mathbb{R}^n$ tel que $\|v\|_A = 1$, on introduit $w = M^{-1}Av$; 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

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

■

Remarque 8.2.1 *ce résultat est encore valable si la matrice A n'est plus symétrique, mais reste définie positive [18].*

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.

8.3 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

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

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 :

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

où cette fois les indices k et l varient de 1 à p nombre de blocs de la partition. Dans ce formalisme certains blocs peuvent être rectangulaires, mais les blocs diagonaux sont nécessairement carrés :

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

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 !

8.4 Critère de convergence

Dans la Partie 1, on a introduit la notion de convergence d'un algorithme à la précision ε après k itérations, par la majoration

$$\|r^k\| \leq \varepsilon \|r^0\| \tag{8.3}$$

pour une norme vectorielle $\|\cdot\|$ déterminée. Noter que la majoration

$$\frac{\|x - x^k\|}{\|x\|} \leq \|A\| \times \|A^{-1}\| \frac{\|r^k\|}{\|b\|}$$

entraine en prenant $x^0 = 0$

$$\frac{\|x - x^k\|}{\|x - x^0\|} \leq cond(A) \frac{\|r^k\|}{\|r^0\|}$$

Si la suite satisfait au critère de convergence (8.3), elle satisfait aussi critère *classique*

$$\frac{\|x - x^k\|}{\|x - x^0\|} \leq \varepsilon_A \tag{8.4}$$

avec $\varepsilon_A = \varepsilon \, cond(A)$; noter que le facteur multiplicatif $cond(A)$ (toujours supérieur ou égal à 1 peut être très grand !

8.5 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

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Dx^{k+1} = (E+F)x^k + b \\ \textbf{fin} \end{array} \right. \tag{8.5}$$

la matrice d'itération associée est notée

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

8.6 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

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (D-E)x^{k+1} = Fx^k + b \\ \textbf{fin} \end{array} \right. \tag{8.6}$$

la matrice d'itération associée est notée

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

8.7 Méthode de relaxation

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

$$x_i^{k+1} = (1-\omega)x_i^k + \omega\tilde{x}_i^{k+1}$$

$$\Longrightarrow A_{i,i}x_i^{k+1} = (1-\omega)A_{i,i}x_i^k + \omega\left(b_i - \sum_{j<i} A_{i,j}x_j^{k+1} - \sum_{j>i} A_{i,j}x_j^k\right).$$

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

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (D - \omega E)x^{k+1} = \left(\omega F + (1-\omega)D\right)x^k + \omega b \\ \textbf{fin} \end{array}\right. \tag{8.7}$$

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 : $L_1 = G$.

Pour cette matrice, on peut écrire les relations

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

et

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

car les matrices M et N sont triangulaires ; 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 8.7.1 *[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 8.2.1. ∎

8.8 Matrices tridiagonales par blocs

Les matrices tridiagonales par blocs sont très courantes dans le cadre de l'approximation de solution d'équations différentielles par la méthode des éléments finis; en général une simple renumérotation des inconnues par la méthode de Cuthill-McKee (voir le chapitre 16) permet d'obtenir cette propriété. 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 (noter que cette écriture contient le cas des matrices tridiagonales par points pour $p = n$).

Proposition 8.8.1 *Si la matrice A est tridiagonale par blocs, alors $\rho(G) = \rho(J)^2$.*

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

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

avec $[D']_j \in \mathbb{R}^{q_j \times q_j}$ et $\sum_{j=1}^{p} q_j = n$. 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 dét$(E-D)=$ 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 8.8.1 *Dans ce calcul on a défini la matrice $C(\lambda)$ à partir de la matrice A, de la manière suivante :*

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

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 8.8.1 *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 8.8.1 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 dét$(E - \dfrac{1}{\omega}D) = \omega^{-n}$dét$(-D)$,

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

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

On se contente de reproduire la courbe représentative de ces variations (figure 8.1)

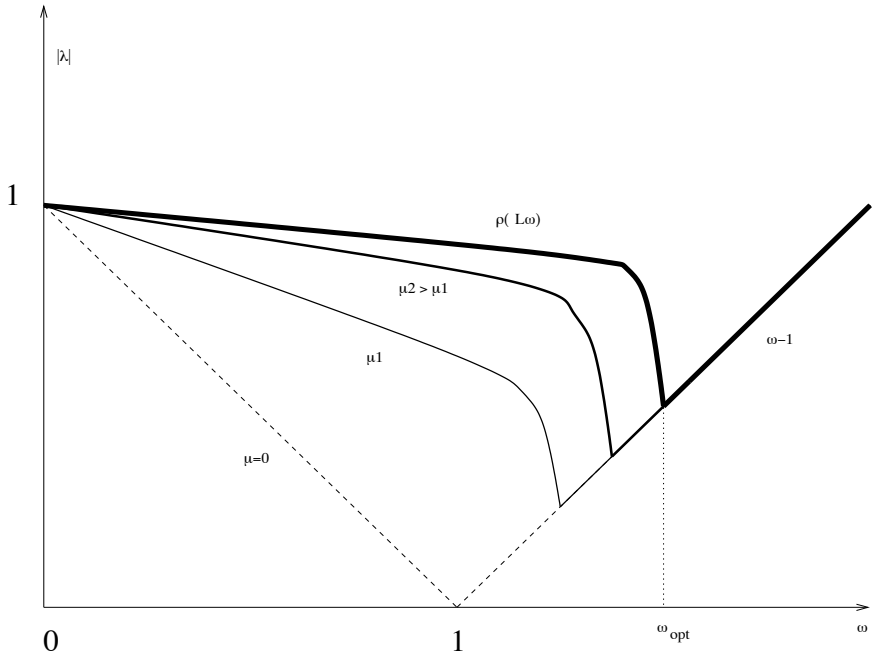


FIG. 8.1 – *Les variations du rayon spectral $\rho(L_\omega)$*

Théorème 8.8.2 *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 8.8.1 il suffit de vérifier que les valeurs propres μ de J sont réelles :

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

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 8.8.1 les méthodes de Jacobi, de Gauss-Seidel et de relaxation pour $0 < \omega < 2$ convergent. ∎

Remarque 8.8.2 *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 $\omega \to \omega_{opt-}$!*

8.9 Méthode de Jacobi relaxée

Comme pour la méthode de Gauss-Seidel, 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

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Dx^{k+1} = (\omega E + \omega F + (1-\omega)D)\, x^k + \omega b \\ \textbf{fin} \end{array} \right. \tag{8.8}$$

Et 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 : $J_1 = J$.

Proposition 8.9.1 *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 8.8.2 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 8.8.1 :

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

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. ■

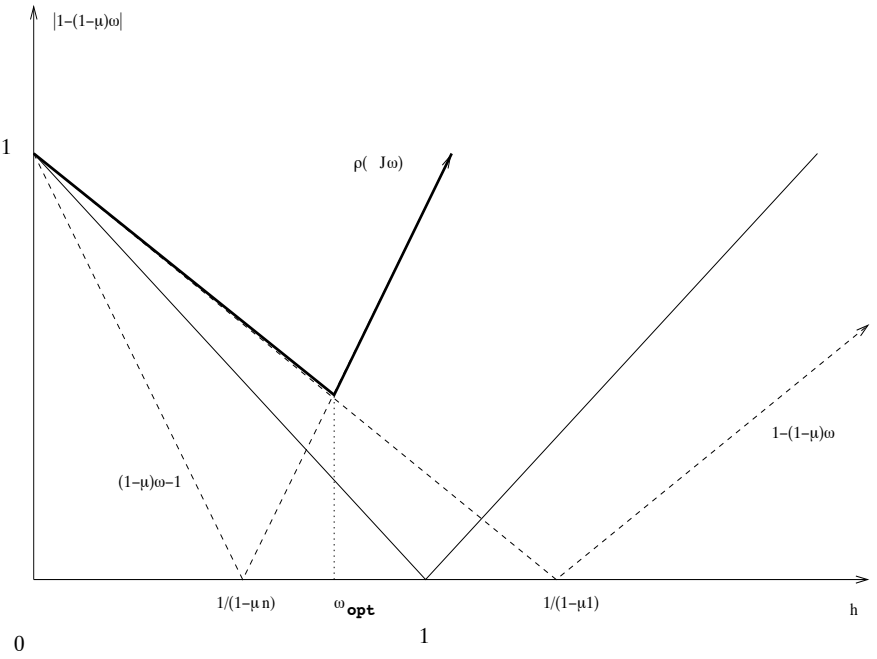


FIG. 8.2 – *Les variations du rayon spectral* $\rho(J)$

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 !

8.10 Méthode de Richardson

Une méthode itérative très utilisée en optimisation (voir la Partie 1) 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. 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

$$\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \alpha r^k \\ \textbf{fin} \end{array} \tag{8.9}$$

Proposition 8.10.1 *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

$$\alpha = \frac{2}{(\lambda_1 + \lambda_n)}.$$

■

8.11 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

$$\left\|\begin{array}{l}\textbf{initialisation}\\ \quad x^0 \in \mathbb{R}^n\\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire}\\ \quad x^{k+1} = x^k + \alpha^k r^k\\ \textbf{fin}\end{array}\right. \tag{8.10}$$

Proposition 8.11.1 *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.

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 8.11.1 *Montrer que pour toute matrice A, symétrique définie positive*

$$\forall v \in \mathbb{R}^n \quad \frac{\|v\|_A^2}{\|v\|_2^2} \times \frac{\|v\|_{A^{-1}}^2}{\|v\|_2^2} \leq \kappa(A)$$

avec $\kappa(A) = \dfrac{\lambda_1}{\lambda_n} = \|A\|_2 \, \|A^{-1}\|_2$ nombre de conditionnement de la matrice A.

En utilisant ce résultat, on obtient pour $v = r^k$ la majoration

$$\begin{aligned}\|r^{k+1}\|_{A^{-1}}^2 &= \|r^k\|_{A^{-1}}^2 \left[1 - \frac{\|r^k\|_2^4}{\|r^k\|_A^2 \, \|r^k\|_{A^{-1}}^2}\right]\\ &\leq \|r^k\|_{A^{-1}}^2 \left[1 - 1/\kappa(A)\right]\end{aligned}$$

Puisque par définition $\kappa(A) \geq 1$, on voit donc que $\|r^{k+1}\|_{A^{-1}}$ tend vers 0 quand k tend vers $+\infty$, et la méthode converge.

Cette méthode peut être considérée comme un premier pas vers la méthode du gradient conjugué (voir la Partie 1).

8.12 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 8.12.1 *une matrice $A \in \mathbb{R}^{n\times n}$ est dite à* **diagonale dominante** *si et seulement si*

$$\forall i, \quad 1 \le i \le n \qquad \sum_{j\neq i} |A_{i,j}| \le |A_{i,i}|.$$

Une matrice $A \in \mathbb{R}^{n\times n}$ est dite à **diagonale strictement dominante** *si et seulement si*

$$\forall i, \quad 1 \le i \le n \qquad \sum_{j\neq i} |A_{i,j}| < |A_{i,i}|.$$

Exercice 8.12.1 *Montrer qu'une matrice à diagonale strictement dominante est inversible.*

Proposition 8.12.1 *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

$$\rho(J) \le \|J\|_\infty < 1.$$

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

$$\lambda A_{i,i} v_i = \lambda \sum_{j<i} A_{i,j} v_j - \sum_{j>i} A_{i,j} v_j$$

$$\Longrightarrow |\lambda||A_{i,i}| \le |\lambda| \sum_{j<i} |A_{i,j}| + \sum_{j>i} |A_{i,j}|.$$

$$\Longrightarrow |\lambda| \left(|A_{i,i}| - \sum_{j<i} |A_{i,j}| \right) \le \sum_{j>i} |A_{i,j}|.$$

Par ailleurs on sait que

$$|A_{i,i}| - \sum_{j<i} |A_{i,j}| > \sum_{j>i} |A_{i,j}| \ge 0,$$

$$\Longrightarrow |\lambda| \le \frac{\displaystyle\sum_{j>i} |A_{i,j}|}{\displaystyle |A_{i,i}| - \sum_{j<i} |A_{i,j}|} < 1;$$

ainsi $\rho(G) < 1$ et la méthode converge. ∎

8.13 Double décomposition régulière de matrices

Continuons le catalogue des méthodes itératives avec une nouvelle définition :

Définition 8.13.1 *On appelle* **double décomposition régulière** *d'une matrice $A \in \mathbb{R}^{n\times n}$, la donnée de quatre matrices $M, N, P, Q \in \mathbb{R}^{n\times n}$ telles que*
(i) $A = M - N = P - Q$
(ii) M et P sont inversibles.

On associe à cette double décomposition régulière la méthode itérative

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad Mx^{k+1/2} = Nx^k + b \\ \quad Px^{k+1} = Qx^{k+1/2} + b \\ \textbf{fin} \end{array}\right. \tag{8.11}$$

Ce que l'on peut encore écrire

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = P^{-1}QM^{-1}Nx^k + P^{-1}QM^{-1}b + P^{-1}b \\ \textbf{fin} \end{array}\right. \tag{8.12}$$

On remarque que si x est solution du système linéaire $Ax = b$ alors

$$\begin{aligned} Mx = Nx + b &\Longrightarrow x = M^{-1}Nx + M^{-1}b \\ Px = Qx + b &\Longrightarrow x = P^{-1}Qx + P^{-1}b, \end{aligned}$$

soit

$$x = P^{-1}QM^{-1}Nx + P^{-1}QM^{-1}b + P^{-1}b.$$

Théorème 8.13.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie positive et soit une double décomposition régulière $A = M - N = P - Q$, si les matrices $M^T + N$ et $P^T + Q$ sont définies positives, alors*

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

Preuve : En posant

$$\left\{\begin{array}{c} e^k = x - x^k \\ e^{k+1/2} = x - x^{k+1/2} \end{array}\right. \quad \text{et} \quad \left\{\begin{array}{c} f^k = M^{-1}Ae^k \\ g^k = P^{-1}Ae^{k+1/2} \end{array}\right. .$$

on obtient

$$\left\{\begin{array}{c} Me^{k+1/2} = Ne^k \\ Pe^{k+1} = Qe^{k+1/2} \end{array}\right. \quad \text{et} \quad \left\{\begin{array}{c} f^k = e^k - e^{k+1/2} \\ g^k = e^{k+1/2} - e^{k+1} \end{array}\right. .$$

En utilisant la norme vectorielle associée à A : $\|v\|_A^2 = (v, Av) = (v, v)_A$, on écrit

$$\|e^{k+1/2}\|_A^2 = \|e^k - f^k\|_A^2 = \|e^k\|_A^2 + \|f^k\|_A^2 - (e^k, Af^k) - (f^k, Ae^k);$$

mais

$$\begin{aligned}\|f^k\|_A^2 - (e^k, Af^k) - (f^k, Ae^k) &= \|f^k\|_A^2 - (f^k, M^T f^k) - (f^k, Mf^k)\\ &= (f^k, (A - M^T - M)f^k) = -(f^k, (M^T + N)f^k).\end{aligned}$$

On en déduit que

$$\|e^k\|_A^2 = \|e^{k+1/2}\|_A^2 + (f^k, (M^T + N)f^k),$$

puis on montre de même que

$$\|e^{k+1/2}\|_A^2 = \|e^{k+1}\|_A^2 + (g^k, (P^T + Q)g^k).$$

Ainsi

$$\|e^k\|_A^2 = \|e^{k+1}\|_A^2 + (f^k, (M^T + N)f^k) + (g^k, (P^T + Q)g^k).$$

La suite $\{\|e^k\|_A\}_{k\in\mathbb{N}}$ est décroissante minorée par 0, elle est donc convergente. Ainsi les suites positives $\{(f^k, (M^T+N)f^k)\}_{k\in\mathbb{N}}$ et $\{(g^k, (P^T+Q)g^k)\}_{k\in\mathbb{N}}$ tendent vers 0 ; les matrices M^T+N et P^T+Q étant définies positives, cela entraine à son tour que les suites $\{f^k\}_{k\in\mathbb{N}}$ et $\{g^k\}_{k\in\mathbb{N}}$ tendent vers 0, ce qui entraine finalement que $\{e^k\}_{k\in\mathbb{N}}$ tend vers 0. ■

8.14 Méthode de relaxation symétrique (S.S.O.R.)

L'ordre des inconnues a-t-il une influence sur la convergence de la méthode ? 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

$$\left\|\begin{array}{l}\textbf{initialisation}\\ \quad x^0 \in \mathbb{R}^n\\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire}\\ \quad (D - \omega E)x^{k+1/2} = (\omega F + (1-\omega)D)\, x^k + \omega b\\ \quad (D - \omega F)x^{k+1} = (\omega E + (1-\omega)D)\, x^{k+1/2} + \omega b\\ \textbf{fin}\end{array}\right. \tag{8.13}$$

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 il est immédiat de vérifier que cette méthode itérative correspond à la double décomposition

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

Théorème 8.14.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie positive, la méthode S.S.O.R. converge si et seulement si $\omega \in]0,2[$.*

Preuve : Par construction

$$|\text{dét } (S_\omega)|^{1/n} = |1-\omega|^2 \leq \rho(S_\omega).$$

La condition est donc nécessaire puisque pour $\omega \notin]0,2[$ la méthode diverge. Pour démontrer que la condition est suffisante, on vérifie que la matrice $B = M^T + N = P^T + Q = \dfrac{2-\omega}{\omega}\, D$ est symétrique définie positive quand A est symétrique définie positive et $\omega \in]0,2[$, et on applique le Théorème 8.13.1. ∎

Remarque 8.14.1 *on peut encore montrer qu'il existe une valeur optimale du paramètre ω (voir Young [27]), 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.

8.15 Etude d'un exemple simple

Comparons maintenant les différentes méthodes présentées dans ce chapitre sur un problème académique, la résolution de l'équation de Laplace sur l'intervalle $]0,1[$, qui s'écrit

$$\begin{cases} \text{Trouver } u \in H_0^1(\Omega) \text{ telle que} \\ \\ -\dfrac{d^2u}{dx^2} = f \quad \text{sur } \Omega =]0,1[\\ u(0) = u(1) = 0 \end{cases}$$

On approche la solution de cette équation par la méthode des différences finies, l'opérateur de Laplace est représenté par le schéma à trois points :

$$-\frac{d^2u}{dx^2}(i*h) \approx \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2}$$

dans lequel $u_i \approx u(i*h)$ est la valeur approchée de u au point d'abscisse $i*h$. Le pas d'espace h est pris uniforme $h = 1/2^p$, il y a donc $n = 2^p - 1$ inconnues, les valeurs au bord étant fixées. Le système linéaire résultant s'écrit

$$A_h x_h = b_h \tag{8.14}$$
$$\tag{8.15}$$

avec

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & \dots & -1 & 2 \end{pmatrix},$$

Exercice 8.15.1 *Montrer que les valeurs propres de la matrice A_h sont les*

$$\lambda_h^k = \frac{2}{h^2}\left[1 - \cos(k\pi h)\right] \quad 1 \le k \le n,$$

associées aux vecteurs propres φ_h^k

$$\varphi_h^k = [\, sin(k\pi h), \quad sin(2k\pi h), \quad \dots, \quad sin((n-1)k\pi h), \quad sin(nk\pi h)\,]^T.$$

On envisage de résoudre le système linéaire (8.14) par les méthodes itératives décrites dans ce chapitre :

– la matrice de la méthode de Jacobi est

$$J = D^{-1}(E + E^T) = I - D^{-1}A_h.$$

Cette matrice a les mêmes vecteurs propres que A_h et les valeurs propres correspondantes sont les

$$\mu_h^k = 1 - \frac{h^2}{2}\lambda_h^k = \cos(k\pi h) = 1 - 2\sin^2(k\pi h/2) \quad 1 \le k \le n.$$

Ainsi pour h petit,

$$\rho(J) = 1 - 2\sin^2(\pi h/2) \approx 1 - \frac{\pi^2 h^2}{2}$$

et la vitesse de convergence de la méthode de Jacobi est donnée par la formule

$$R(J) = -\log \rho[J] \approx \pi^2 h^2/2\ ;$$

elle diminue quand le nombre d'inconnues augmente !

– il en est de même pour la méthode de Gauss-Seidel ; en appliquant la Proposition 8.8.1 à la matrice A_h qui est tridiagonale, on obtient pour la matrice

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

$$\rho(G) = \rho(J)^2 = [1 - 2\sin^2(\pi h/2)]^2 \approx 1 - \pi^2 h^2$$

et la vitesse de convergence de la méthode de Gauss-Seidel est

$$R(G) = -\log \rho[G] \approx \pi^2 h^2.$$

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

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

on utilise le résultat du Théorème 8.8.1 avec la valeur optimale ω_{opt}

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

quand h tend vers 0,

$$1-\rho(J)^2 = 1-\rho(G) \approx 4\sin^2(\pi h/2) \approx \pi^2 h^2$$

$$\rho(L_{\omega_{opt}}) \approx \frac{1-\pi h}{1+\pi h} \approx 1-2\pi h.$$

Finalement la vitesse de convergence de la méthode de relaxation est

$$R(L_{\omega_{opt}}) = -\log \rho[L_{\omega_{opt}}] \approx 2\pi h.$$

– pour la méthode S.S.O.R. avec le paramètre "optimal" ω'_{opt}

$$\rho(S_{\omega'_{opt}}) = \omega'_{opt} - 1 = \frac{2}{1+\sqrt{2\left(1-\rho(J)\right)^2}} - 1 = \frac{1-\sqrt{2\left(1-\rho(J)\right)^2}}{1+\sqrt{2\left(1-\rho(J)\right)^2}};$$

quand h tend vers 0, la vitesse de convergence de la méthode S.S.O.R. vérifie

$$R(S_{\omega'_{opt}}) = -\log \rho[S_{\omega'_{opt}}] \approx 2\sqrt{2}\pi h = \sqrt{2}R(L_{\omega_{opt}}).$$

– 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

$$\alpha_{opt} = \frac{2}{\lambda_1+\lambda_n} = \frac{h^2}{2} \quad \text{et} \quad \rho(R_{\alpha_{opt}}) = 1-\alpha_{opt}\lambda_n = \cos \pi h$$

soit pour la vitesse de convergence de la méthode de Richardson à pas constant

$$R(R_{\alpha_{opt}}) = -\log \cos \pi h \approx \pi^2 h^2/2.$$

Pour compléter cette comparaison, on cherche à estimer le nombre d'itérations nécessaires pour obtenir la convergence à une précision ε donnée. D'un point de vue pratique, on ne dispose que du rapport $\|r^k\|/\|r^0\|$, mais on sait par ailleurs que l'erreur $e^k = x - x^k = [M^{-1}N]^k e^0$ vérifie d'une part

$$\frac{\|e^k\|}{\|e^0\|} \leq \rho^k(M^{-1}N),$$

et d'autre part

$$\frac{\|x-x^k\|}{\|x-x^0\|} \leq cond(A)\frac{\|r^k\|}{\|r^0\|}.$$

Le nombre de conditionnement $cond(A)$ étant en général "grand", on considére d'un point de vue pratique que le nombre d'itérations pour obtenir la convergence de la méthode à une précision ε est estimé suivant

$$\rho^k(M^{-1}N) \leq \frac{\|r^k\|}{\|r^0\|} \leq \varepsilon,$$

soit encore

$$k \approx \frac{\log \varepsilon}{\log \rho(M^{-1}N)} = -\frac{\log \varepsilon}{R(M^{-1}N)}.$$

Exercice 8.15.2 *Montrer que pour la matrice A_h de l'exemple, 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 8.15.1 *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.

Le tableau suivant rassemble toutes ces estimations en ordre de grandeur

Méthode	$R(B)$	Nb. itérations	Nb. opérations
Jacobi	$\pi^2 h^2/2$	$2n^2$	$2n^3$
Gauss-Seidel	$\pi^2 h^2$	n^2	n^3
S.O.R. ω_{opt}	$2\pi h$	$2\pi n$	$2\pi n^2$
S.S.O.R. ω'_{opt}	$2\sqrt{2}\pi h$	$\sqrt{2}\pi n$	$\sqrt{2}\pi n^2$
Richardson avec α_{opt}	$\pi^2 h^2/2$	$2n^2$	$2n^3$

Sur cet exemple, les méthodes les moins coûteuses sont donc la méthode S.S.O.R. et la méthode semi–itérative de Richardson.

Remarque 8.15.2 *on connait explicitement les valeurs propres des matrices utilisées, et il s'agit donc d'un cas particulier. Il n'existe pas de résultat de comparaison dans le cas général pour lequel les valeurs optimales ω_{opt} et ω'_{opt} sont souvent inaccessibles.*

8.16 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. C'est le cas de la matrice

$$A_1 = \begin{bmatrix} 1 & 0 & -1/4 & 1/4 \\ 0 & 1 & -1/4 & 1/4 \\ -1/4 & -1/4 & 1 & 0 \\ 1/4 & 1/4 & 0 & 1 \end{bmatrix}$$

pour laquelle le rayon spectral de la matrice de Gauss-Seidel par points est $\rho(G_p) = 0.25$. Si on définit la décomposition

$$D_1 = \begin{bmatrix} 1 & 0 & -1/4 & \vdots & 0 \\ 0 & 1 & -1/4 & \vdots & 0 \\ -1/4 & -1/4 & 1 & \vdots & 0 \\ \dots & \dots & \dots & \cdot & \dots \\ 0 & 0 & 0 & \vdots & 1 \end{bmatrix} \quad E_1 = \begin{bmatrix} 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ \dots & \dots & \dots & \cdot & \dots \\ -1/4 & -1/4 & 0 & \vdots & 0 \end{bmatrix}.$$

alors le rayon spectral de la matrice de Gauss-Seidel par blocs

$$G_b = (D_1 - E_1)^{-1} E_1^T$$

vaut $\rho(G_b) = 0.1429$; il est bien plus petit que $\rho(G_p)$.

Mais cette propriété n'est pas toujours satisfaite comme le montre le contre–exemple suivant tiré de [27] : la matrice A_2

$$A_2 = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix}$$

est symétrique définie positive et on définit la décomposition

$$D_2 = \begin{bmatrix} 5 & 2 & \vdots & 0 \\ 2 & 5 & \vdots & 0 \\ \dots & \dots & \cdot & \dots \\ 0 & 0 & \vdots & 5 \end{bmatrix} \quad E_2 = \begin{bmatrix} 0 & 0 & \vdots & 0 \\ 0 & 0 & \vdots & 0 \\ \dots & \dots & \dots & \dots \\ -2 & -3 & \vdots & 0 \end{bmatrix};$$

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 $\rho(G_p) = 0.3098$!

Remarque 8.16.1 *en dehors de l'influence du partitionnement des matrices sur la convergence des méthodes itératives, il faut signaler que le traitement "par blocs" des systèmes linéaires a des conséquences pratiques souvent très importantes. En effet sur les ordinateurs vectoriels ou parallèles, la vitesse de réalisation des opérations algébriques peut être considérablement augmentée lorsque les données sont regroupées sous forme de vecteurs ou de matrices (blocs de vecteurs). Sans entrer dans les détails (voir le chapitre 16) il faut savoir que sur certains ordinateurs la version par blocs d'un algorithme peut être beaucoup plus rapide que la version par points.*

Ce qu'il faut retenir

1. pour toute matrice inversible A, qui peut s'écrire $A = M - N$ avec M matrice inversible, on peut définir un algorithme de calcul de la solution du système linéaire $Ax = b$ par les formules

$$\begin{array}{|l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \dots, \textbf{faire} \\ \quad Mx^{k+1} = Nx^k + b \\ \textbf{fin} \end{array} \tag{8.16}$$

2. une condition nécessaire et suffisante pour que cet algorithme converge est que le rayon spectral $\rho(M^{-1}N)$ soit strictement inférieur à 1.

3. les méthodes de Jacobi, Gauss-Seidel, relaxation, SSOR et Richardson sont des algorithmes de la forme (8.16)

4. les méthodes de Jacobi, Gauss-Seidel, relaxation et SSOR (pour $0 < \omega < 2$) convergent lorsque la matrice A est symétrique définie positive.

Chapitre 9

Les méthodes de Krylov

9.1 Introduction

Ce chapitre est une introduction à l'étude des méthodes de Krylov, appellation générale de méthodes de résolution de systèmes linéaires et de méthodes de calcul de valeurs propres de matrices. La plus célèbre d'entre elles est la méthode du gradient conjugué déjà présentée comme méthode de minimisation d'une fonctionnelle quadratique convexe, au chapitre 4.

9.2 Un problème modèle : rappel

Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie positive, et $b \in \mathbb{R}^n$ un vecteur quelconque. On veut résoudre le système linéaire :

$$\left\{ \begin{array}{l} \text{Trouver } x \in \mathbb{R}^n, \text{ tel que} \\ A\,x \;= b \end{array} \right. \tag{9.1}$$

On introduit la fonctionnelle $J : \mathbb{R}^n \to \mathbb{R}$, définie par

$$\forall v \in \mathbb{R}^n \quad J(v) = \frac{1}{2}(Av, v) - (b, v). \tag{9.2}$$

On sait que J est une fonctionnelle strictement convexe sur $\mathbb{R}^n$. Elle admet un minimum unique en $x \in \mathbb{R}^n$, (unique) solution du système linéaire $Ax = b$. Le problème (9.1) est donc équivalent au problème

$$\left\{ \begin{array}{l} \text{Trouver } x \in \mathbb{R}^n, \text{ tel que} \\ \forall v \in \mathbb{R}^n, \qquad J(x) \leq J(v) \end{array} \right. \tag{9.3}$$

Pour construire un algorithme *itératif* efficace de résolution du système linéaire (9.1) on va caractériser le vecteur u. Pour cela on suppose connue une base de $\mathbb{R}^n$, $D = \{d^0, d^1, \ldots, d^{n-1}\}$ et on écrit

$$\forall x, x^0 \in \mathbb{R}^n, \qquad x - x^0 = \sum_{k=0}^{n-1} \alpha_k d^k.$$

On en déduit, pour $j(\alpha) = J(u)$, avec $\alpha = (\alpha_0, \cdots, \alpha_{n-1})$,

$$j(\alpha) = J(x^0 + \sum_{k=0}^{n-1} \alpha_k d^k) = J(x^0) - \sum_{k=0}^{n-1} (b - Ax^0, d^k)\alpha_k + \frac{1}{2}\sum_{k,l=0}^{n-1} (d^k, Ad^l)\alpha_k\alpha_l.$$

Supposons de plus que la base soit orthogonale pour le produit scalaire $(.,.)_A$, c'est-à-dire que les vecteurs de base satisfont la relation

$$(d^k, Ad^l) = 0 \quad \text{ pour } k \neq l.$$

Les vecteurs d^k sont alors appelés **directions conjuguées** (cf. définition 4.2.1) et on obtient

$$j(\alpha) = J(x^0) - \sum_{k=0}^{n-1}(b - Ax^0, d^k)\alpha_k + \frac{1}{2}\sum_{k=0}^{n-1}(d^k, Ad^k){\alpha_k}^2.$$

Pour x^0 fixé, $j(\alpha)$ est une fonctionnelle strictement convexe, puisque $(d^k, Ad^k) > 0$ pour tout k, et

$$\frac{\partial j}{\partial \alpha_k}(\alpha) = -(b - Ax^0, d^k) + \alpha_k(d^k, Ad^k).$$

On en déduit que son gradient $\nabla j(\alpha)$ est nul, si et seulement si

$$\alpha_k = \frac{(b - Ax^0, d^k)}{(d^k, Ad^k)} = \frac{(r^0, d^k)}{(d^k, Ad^k)} \quad \text{pour } k = 1, 2, \ldots, n.$$

Ainsi

$$x = x^0 + \sum_{k=0}^{n-1}\alpha_k d^k = x^0 + \sum_{k=0}^{n-1}\frac{(r^0, d^k)}{(d^k, Ad^k)}d^k. \tag{9.4}$$

On a donc obtenu l'expression du minimum x de J sur la base des directions conjuguées.

9.3 Un algorithme de résolution

La méthode du gradient conjugué est un algorithme itératif de résolution du système linéaire (9.1) qui consiste, à partir d'un vecteur x^0 quelconque de $\mathbb{R}^n$, à construire de manière itérative une base A-orthogonale de $\mathbb{R}^n$ puis à approcher la solution x par la suite de vecteurs

$$x^k = x^0 + \sum_{l=0}^{k-1}\frac{(r^0, d^l)}{(d^l, Ad^l)}d^l, \text{ pour } k = 1, 2, \cdots \tag{9.5}$$

En utilisant la caractérisation du minimum de J, cet algorithme peut s'écrire

Initialisation :

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

Calculer $r^0 = b - Ax^0$; $d^0 = r^0$

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

(a) approximation de la solution (9.6)

Calculer $\alpha_k = (r^0, d^k)/(d^k, Ad^k)$; $x^{k+1} = x^k + \alpha_k d^k$; $r^{k+1} = r^k - \alpha_k Ad^k$

(b) détermination de la nouvelle direction

Calculer $\beta_{k+1} = -(Ad^k, r^{k+1})/(d^k, Ad^k)$; $d^{k+1} = r^{k+1} + \beta_{k+1}d^k$

fin

9.4 Propriétés de l'algorithme

Les vecteurs générés par l'algorithme (9.6) vérifient les relations suivantes

Théorème 9.4.1 (Propriétés de l'algorithme)

$$(Ad^k, d^l) = 0 \qquad \textit{pour } k \neq l \tag{9.7a}$$
$$(d^l, r^k) = 0 \qquad 0 \leq l < k \tag{9.7b}$$
$$(d^k, r^l) = (d^k, r^0) \qquad 0 \leq l \leq k \tag{9.7c}$$
$$(d^k, r^k) = (r^k, r^k) \tag{9.7d}$$
$$(Ad^k, r^k) = (Ad^k, d^k) \tag{9.7e}$$
$$(r^k, r^l) = 0 \qquad \textit{pour } k \neq l \tag{9.7f}$$
$$\mathcal{E}^k = Vect(d^0, d^1, \ldots, d^k) = Vect(r^0, r^1, \ldots, r^k) = Vect(r^0, Ar^0, \ldots, A^k r^0) \tag{9.7g}$$
$$\textbf{si } r^k \neq 0 \textbf{ alors } dim\ \mathcal{E}^k = k+1 \tag{9.7h}$$
$$x^{k+1} \textit{réalise le minimum de } J \textit{ sur } x^0 + \mathcal{E}^k \tag{9.7i}$$
$$\textbf{et } \alpha_k = (r^k, r^k)/(d^k, Ad^k) \tag{9.7j}$$
$$\beta_{k+1} = (r^{k+1}, r^{k+1})/(r^k, r^k) \tag{9.7k}$$
$$\textbf{si } r^k \neq 0 \textbf{ pour tout } k < n \textbf{ alors } r^n = 0 \tag{9.7l}$$

Preuve : Les propriétés sont démontrées par récurrence.

Commençons par les quatre premières :
1) A l'ordre $k = 1$:
(9.7a) $(Ad^0, d^1) = (Ad^0, r^1 + \beta_1 d^0) = 0$ par définition de β_1.
(9.7b) $(d^0, r^1) = (d^0, r^0 - \alpha_0 Ad^0) = 0$ par définition de α_0.
(9.7c) $(d^1, r^0) = (d^1, r^0)$!
(9.7d) $(d^0, r^0) = (r^0, r^0)$ par définition de d^0 et $(d^1, r^1) = (r^1 + \beta_1 d^0, r^1) = (r^1, r^1)$.
2) A l'ordre k : supposons ces propriétés vraies jusqu'à l'ordre $k-1$ inclus, remarquons d'abord que de la relation

$$r^l = r^0 - \sum_{j=0}^{l-1} \alpha_j Ad^j,$$

on peut déduire de l'hypothèse de récurrence que pour $0 \leq l \leq k-1$

$$\alpha_l = \frac{(r^0, d^l)}{(d^l, Ad^l)} = \frac{(r^l, d^l)}{(d^l, Ad^l)} = \frac{(r^l, r^l)}{(d^l, Ad^l)}. \tag{9.8}$$

Il vient ensuite
(9.7a) $(Ad^{k-1}, d^k) = 0$ par définition de β_k ;
(9.7a) pour $0 \leq l < k-1$

$$(Ad^l, d^k) = (Ad^l, r^k + \beta_k d^{k-1}) = (Ad^l, r^k) = (\frac{1}{\alpha_l}(r^l - r^{l+1}), r^k) = 0$$

d'après l'hypothèse de récurrence ;
(9.7b) $(d^{k-1}, r^k) = (d^{k-1}, r^{k-1} - \alpha_{k-1} Ad^{k-1}) = 0$ d'après (9.8) ;
(9.7b) pour $0 \leq l < k-1$

$$(d^l, r^k) = (d^l, r^{k-1} - \alpha_{k-1} Ad^{k-1}) = (d^l, r^{k-1}) - \alpha_{k-1}(d^l, Ad^{k-1}) = 0$$

d'après l'hypothèse de récurrence ;
(9.7c) pour $0 \leq l < k-1$

$$(d^k, r^l) = (d^k, r^0 - \sum_{j=0}^{l-1} \alpha_j Ad^j) = (d^k, r^0)\ ;$$

(9.7d) $(d^k, r^k) = (r^k + \beta_{k-1} d^{k-1}, r^k) = (r^k, r^k)$.

Les propriétés (9.7a) à (9.7d) sont donc vraies pour tout k ; les autres propriétés en sont déduites de la manière suivante :

(9.7e) la propriété est satisfaite à l'ordre 0, puisque $r^0 = d^0$; supposons-la vraie jusqu'à l'ordre $k-1$ inclus, alors $(Ad^k, r^k) = (Ad^k, d^k - \beta_k d^{k-1}) = (Ad^k, d^k)$ d'après (9.7a) ;

(9.7f) par construction $(r^0, r^1) = (d^0, r^1) = 0$ d'après (9.7b). De même d'après (9.7b) $(r^l, r^k) = (d^l - \beta_l d^{l-1}, r^k) = 0$ pour $0 \leq l < k$, Les propriétés (9.7e) et (9.7f) sont donc vraies pour tout k.

Pour établir (9.7g) et (9.7h) on vérifie que les propriétés sont vraies pour $k = 0$, puis en les supposant vraies jusqu'à l'ordre $k-1$ inclus, il vient par construction

$$\left.\begin{array}{c} r^{k-1} \in Vect(r^0, Ar^0, \ldots, A^{k-1}r^0) \\ d^{k-1} \in Vect(r^0, Ar^0, \ldots, A^{k-1}r^0) \\ Ad^{k-1} \in Vect(Ar^0, A^2r^0, \ldots, A^k r^0) \\ r^k = r^{k-1} - \alpha_{k-1} Ad^{k-1} \end{array}\right\} \Longrightarrow r^k \in Vect(r^0, Ar^0, \ldots, A^k r^0)$$

on en déduit que $d^k = r^k + \beta_k d^{k-1}$ appartient lui aussi à $Vect(r^0, Ar^0, \ldots, A^k r^0)$.

En combinant ces résultats, on obtient

$$\begin{array}{c} Vect(d^0, d^1, \ldots, d^k) \subset Vect(r^0, Ar^0, \ldots, A^k r^0) \\ Vect(r^0, r^1, \ldots, r^k) \subset Vect(r^0, Ar^0, \ldots, A^k r^0). \end{array} \tag{9.9}$$

Les directions d^0 à d^k forment une famille libre d'après (9.7a), de même pour les résidus r^0 à r^k, en supposant $r^k \neq 0$. Le sous-espace $Vect(r^0, Ar^0, \ldots, A^k r^0)$, est engendré par $k+1$ vecteurs et il contient deux sous-espaces de dimension $k+1$, on en déduit que

$$Vect(d^0, d^1, \ldots, d^k) = Vect(r^0, r^1, \ldots, r^k) = Vect(r^0, Ar^0, \ldots, A^k r^0).$$

Remarque 9.4.1 *Le cas $r^k = 0$ sera traité dans le théorème 9.5.1*

Pour démontrer les dernières propriétés :

(9.7i) d'après la relation (9.4) le vecteur $x^{k+1} - x^0$ réalise le minimum de la fonctionnelle J sur $\mathcal{E}^k$;

(9.7j) puisque les propriétés (9.7a) à (9.7d) sont vraies pour tout k,

$$\alpha_k = \frac{(r^0, d^k)}{(d^k, Ad^k)} = \frac{(r^k, d^k)}{(d^k, Ad^k)} = \frac{(r^k, r^k)}{(d^k, Ad^k)} \ ;$$

(9.7k) enfin par définition

$$\beta_{k+1} = -\frac{(Ad^k, r^{k+1})}{(d^k, Ad^k)} = (\frac{1}{\alpha_k}(r^{k+1} - r^k), r^{k+1})/(d^k, Ad^k),$$

soit

$$\beta_{k+1} = \frac{(r^{k+1}, r^{k+1})}{(d^k, Ad^k)} \times \frac{(d^k, Ad^k)}{(r^k, r^k)} = \frac{(r^{k+1}, r^{k+1})}{(r^k, r^k)} \ ;$$

(9.7l) enfin les vecteurs résidus vérifient (9.7f) et définissent aussi une base orthogonale de $\mathcal{E}^n = \mathbb{R}^n$. Le vecteur r^n est alors orthogonal dans $\mathbb{R}^n$ aux n vecteurs de base r^0, r^1, … r^{n-1}, il est nécessairement nul. ∎

Remarque 9.4.2 *Une lecture attentive de cette démonstration montre que les relations d'orthogonalité $(r^k, r^l) = 0$ et $(d^k, Ad^l) = 0$ pour $k \neq l$ sont suffisantes pour assurer la convergence de l'algorithme (9.6). Ceci laisse entendre que l'argument de minimisation d'une fonctionnelle convexe n'est pas la seule voie possible pour obtenir la solution du système linéaire (9.1).*

9.5 L'algorithme du gradient conjugué

En utilisant les propriétés (9.7j) et (9.7k), on écrit une seconde version de l'algorithme du **gradient conjugué**, à comparer à celle du chapitre 4,

$$
\begin{array}{l}
\textbf{Initialisation :} \\
\quad \text{Choisir } x^0 \in \mathbb{R}^n \\
\quad \text{Calculer } r^0 = b - Ax^0 \ ; \ d^0 = r^0 \\
\textbf{Itérations : pour } k = 0, 1, \ldots, \textbf{ faire} \\
\text{(a) approximation de la solution} \\
\quad \text{Calculer } \alpha_k = (r^k, r^k)/(d^k, Ad^k) \ ; \ x^{k+1} = x^k + \alpha_k d^k \ ; \ r^{k+1} = r^k - \alpha_k A d^k \\
\text{(b) détermination de la nouvelle direction} \\
\quad \text{Calculer } \beta_{k+1} = (r^{k+1}, r^{k+1})/(r^k, r^k) \ ; \ d^{k+1} = r^{k+1} + \beta_{k+1} d^k \\
\textbf{fin}
\end{array}
\tag{9.10}
$$

Par rapport à la formulation (9.6), cette nouvelle écriture de l'algorithme ne requiert plus à chaque itération que le calcul de deux produits scalaires (d^k, Ad^k) et (r^{k+1}, r^{k+1}) au lieu de trois (d^k, Ad^k), (r^0, d^k) et (r^{k+1}, Ad^k). En effet, le produit (r^k, r^k) provient de l'itération précédente. De plus la formulation (9.6) nécessite le stockage du vecteur r^0, qui n'est plus utilisé dans (9.10).

Théorème 9.5.1 *Si $A \in \mathbb{R}^{n\times n}$ est une matrice symétrique définie positive, l'algorithme du gradient conjugué (9.10) converge en au plus n itérations vers la solution du système linéaire $Ax = b$.*

Preuve : Le cas $r^n = 0$ a déjà été traité en supposant que $r^k \neq 0$ pour $k < n$. Si en cours d'itérations on obtient un résidu r^k nul, alors l'algorithme a convergé en $k < n$ itérations. ∎

9.6 Sous–espace de Krylov

Dans l'étude précédente est apparu le sous-espace $Vect(r^0, Ar^0, \ldots, A^{k-1}r^0)$. La structure de ce sous-espace joue un rôle important dans la méthode du gradient conjugué.

Définition 9.6.1 *On appelle* **sous-espace de Krylov** *d'ordre k, relatif à la matrice $A \in \mathbb{R}^{n\times n}$ et au vecteur $v \in \mathbb{R}^n$, le sous-espace de $\mathbb{R}^n$ engendré par les vecteurs $A^j v$ pour $j = 0, 1, \ldots, k-1$. On le note*

$$K_k(A, v) = Vect(v, Av, A^2, \ldots, A^{k-1}v).$$

Dans la méthode du gradient conjugué la solution x du système linéaire $Ax = b$ est approchée à l'itération k par le vecteur x^k de la forme $x^k = x^0 + v^k$ avec $v^k \in K_k(A, r^0)$. D'après l'étude des propriétés la méthode du gradient conjugué, on a par construction, cf. (9.7f)

$$K_k(A, v) = Vect(d^0, d^1, \ldots, d^{k-1}) = Vect(r^0, r^1, \ldots, r^{k-1}) = Vect(r^0, Ar^0, \ldots, A^{k-1}r^0) \tag{9.11}$$

Si $r^{k-1} \neq 0$ la dimension du sous-espace $K_k(A, r^0)$ est k.
D'une manière générale, à l'itération k si $r^{k-1} \neq 0$, alors $dim[K_k(A, v)] = dim[K_{k-1}(A, v)] + 1$: la dimension du sous-espace de Krylov généré par la méthode du gradient conjugué augmente d'une unité à chaque itération, tant que le résidu n'est pas nul.

La formule (9.4) montre qu'à l'itération k, le vecteur $x^k - x^0$ est la projection dans $K_k(A, r^0)$ du vecteur erreur $x - x^0$.

Cette idée d'approcher la solution d'un système linéaire par sa projection dans un sous-espace de Krylov $K_k(A, v)$ est commune à de nombreuses méthodes de résolution de systèmes linéaires et de calcul des valeurs propres référencées sous l'appellation générique de **méthodes de Krylov**.

Remarquons maintenant qu'un sous-espace de Krylov $K_k(A, v)$, relatif à une matrice A et un vecteur v donnés, n'est pas toujours de dimension k. Par exemple lorsque v est un vecteur propre de la matrice A, on a $dim[K_k(A, v)] = 1$ pour tout $k \geq 1$. Mais cette situation n'est pas pénalisante (voir la proposition 9.6.3), au contraire ! Elle montre qu'il convient d'étudier de plus près la construction de $K_k(A, v)$. Nous savons que

1. Le polynôme caractéristique d'une matrice $A \in \mathbb{R}^{n\times n}$ est un polynôme p de degré n qui vérifie
$$p(A) = \mu_0 I + \mu_1 A + \ldots + \mu_n A^n = 0.$$
Quand $\mu_0 = \det A \neq 0$, on déduit de cette relation que la matrice A^{-1} s'écrit sous la forme
$$A^{-1} = -\sum_{j=0}^{n-1} \frac{\mu_j}{\mu_0} A^j. \tag{9.12}$$

2. Le vecteur x, solution du système linéaire (9.1), s'écrit (en posant $r^0 = b - Ax^0$)
$$x = x^0 + A^{-1}r^0 = x^0 - \sum_{j=0}^{n-1} \frac{\mu_j}{\mu_0} A^j r^0. \tag{9.13}$$
Le vecteur $x - x^0$ appartient donc au sous–espace de Krylov $K_n(A, r^0)$, quel que soit r^0, c'est-à-dire quel que soit le vecteur initial x^0.

3. Dans la méthode du gradient conjugué, à l'itération k, on approche le vecteur $x = A^{-1}b$ par le vecteur x^k défini par la relation (cf. (9.5) et (9.11))
$$x^k = x^0 + \sum_{j=0}^{k-1} \alpha_j d^j = x^0 + \sum_{j=0}^{k-1} \gamma_j A^j r^0. \tag{9.14}$$
Cette approximation est consistante avec la relation (9.13).

4. Soit q un polynôme de degré d ($d \leq n$) tel que $q(A) = 0$. Le vecteur $x = A^{-1}b$ s'écrit encore
$$x = x^0 + A^{-1}r^0 = x^0 + \sum_{j=0}^{d-1} \tilde{\alpha}_j d^j. \tag{9.15}$$
Mais le vecteur x^{d-1}, obtenu après d itérations de la méthode du gradient conjugué en partant du vecteur x^0, s'écrit
$$x^{d-1} = x^0 + \sum_{j=0}^{d-1} \alpha_j d^j. \tag{9.16}$$

Pour montrer que $x^{d-1} = x$, il suffit d'utiliser la propriété de minimisation (9.7i) de la méthode du gradient conjugué : la plus petite valeur de la fonctionnelle J est atteinte quand $\alpha_j = \tilde{\alpha}_j$ pour $j = 0, 1, \ldots, d-1$.

Il ressort de cette étude que pour diminuer le nombre d'itérations de la méthode du gradient conjugué, il faut rechercher un polynôme annulateur de A de degré le plus petit possible. Ceci nous renvoie à la notion de polynôme minimal, qui est un polynôme annulateur de A de plus bas degré.

Proposition 9.6.1 *Soit la matrice $A \in \mathbb{R}^{n\times n}$ et soit d le degré de son polynôme minimal. Alors pour tout vecteur $v \in \mathbb{R}^n$*

$$\max_{k\in\mathbb{N}} dim[K_k(A, v)] \leq d.$$

Preuve : Soit q le polynôme minimal de la matrice A, de degré d. Alors

$$0 = q(A) = \xi_0 I + \xi_1 A + \ldots + \xi_{d-1} A^{d-1} + A^d \quad \text{et} \quad A^d = -\sum_{j=0}^{d-1} \xi_j A^j. \tag{9.17}$$

Tout vecteur $v \in K_{d+1}(A, r^0)$ s'écrit donc sous la forme

$$v = \sum_{j=0}^{d-1} \gamma_j A^j r^0 + \gamma_d A^d r^0 = \sum_{j=0}^{d-1} \zeta_j A^j r^0. \tag{9.18}$$

On a donc l'inclusion $K_{d+1}(A, r^0) \subset K_d(A, r^0)$ et par suite $K_k(A, r^0) \subset K_d(A, r^0)$, pour tout $k \geq d$. ∎

Or, si la matrice A est hermitienne, elle est diagonalisable et le degré de son polynôme minimal est égal au nombre de valeurs propres distinctes (on déduit ce résultat de façon élémentaire de la théorie développée au chapitre 10).

Proposition 9.6.2 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie positive. Le nombre d'itérations nécessaires à la résolution d'un système linéaire $Ax = b$ par la méthode du gradient conjugué est majoré par le nombre de valeurs propres distinctes de la matrice A.*

Preuve : C'est une simple conséquence des résultats précédents. ∎

Il est évident que si la dimension du sous–espace de Krylov $K_k(A, v)$ dépend de la matrice A par l'intermédiaire de son polynôme minimal q, elle dépend aussi du vecteur v. On introduit ainsi la notion de degré d'un vecteur : le plus petit entier k tel qu'il existe un polynôme p_k de degré k, vérifiant $p_k(A)v = 0$, est appelé **degré à droite** du vecteur v pour la matrice A. Le degré à droite d'un vecteur propre à droite de A est 1, car pour v vecteur propre à droite de A, $Av - \lambda v = 0$ et ainsi $dim[K_k(A, v)] = 1$ quel que soit $k \geq 1$! La proposition 9.6.3 montre que la méthode du gradient conjugué appliquée à la résolution du problème (9.1) converge en une seule itération si le résidu initial r^0 est un vecteur propre à droite de la matrice A.

Proposition 9.6.3 *Si r^0 est un vecteur propre de la matrice $A \in \mathbb{R}^{n\times n}$ symétrique définie positive, alors la méthode du gradient conjugué appliquée à la résolution du système linéaire $Ax = b$ converge en une seule itération.*

Preuve : Il suffit d'écrire que $r^0 = b - Ax^0 = A(x - x^0)$ pour constater que l'hypothèse $Ar^0 = \lambda r^0$ revient à supposer que $A(x - x^0) = \lambda(x - x^0)$. On remarque alors que $d^0 = r^0 = A(x - x^0)$, et on écrit

$$x^1 = x^0 + \frac{(r^0, r^0)}{(d^0, Ad^0)} d^0 = x^0 + \frac{(r^0, r^0)}{(r^0, Ar^0)} r^0 = x^0 + \frac{1}{\lambda} A(x - x^0) = x^0 + (x - x^0) = x. \tag{9.19}$$

∎

Chapitre 10

Valeurs propres et vecteurs propres

10.1 Introduction

Dans le chapitre 5 de ce cours, on a montré que la recherche des fréquences de résonance d'une structure est une source très importante de problèmes de calcul de valeurs propres et de vecteurs propres. Dans ce chapitre, on donne quelques résultats théoriques fondamentaux qui seront 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. Ces deux classes de matrices feront l'objet dans les chapitres suivants d'algorithmes de calcul spécifique.

10.2 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 imaginaires. A l'exception de certains cas particuliers, tous les calculs présentés dans ce chapitre sont donc effectués avec des nombres complexes. En particulier on notera

$$(u, v) = u^* v = \sum_i \overline{u}_i \; v_i$$

le produit scalaire complexe de $\mathbb{C}^n$, $\overline{x}$ désignant le complexe conjugué de x.

On associe alors à toute matrice $A \in \mathbb{C}^{n \times m}$, la matrice "transconjuguée", notée A^* (ou encore A^H), 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 10.2.1

- *on appelle* **valeur propre** *d'une matrice A, toute racine λ_i du polynôme caractéristique $p(\lambda) = det(A - \lambda I)$; à 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 $p(\lambda)$*

- *mais 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 $Au = \lambda u$, ce qui peut encore s'exprimer par $u_i \in Ker\,(A - \lambda_i I)$. On introduit donc naturellement la notion de* **multiplicité géométrique** *de λ_i par $g_i = dim\,(Ker\,(A - \lambda_i I))$.*

Supposons que la matrice $A \in \mathbb{C}^{n\times n}$ admette d valeurs propres λ_i distinctes,

$$\text{on a toujours} \qquad \sum_{i=1}^{d} m_i = n, \ \text{et} \qquad \sum_{i=1}^{d} g_i \leq n$$

Pour de nombreuses matrices $\sum_{i=1}^{d} g_i < n$, et cela montre que l'on peut pas construire une base de $\mathbb{C}^n$ avec les vecteurs propres de telles matrices. On introduit donc les notions suivantes :

- *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 $m_i > g_i$).*
- *une matrice A qui admet au moins une valeur propre défective et elle-même appelée* **matrice défective** *; de telles matrices ne sont pas diagonalisables.*
- *l'ensemble des valeurs propres d'une matrice A s'appelle le* **spectre** *de A.*

Remarque 10.2.1 *seule une valeur propre multiple peut être défective, puisque pour une valeur propre simple λ_i, on a $m_i = 1$ et $g_i = dim\ Ker(A - \lambda_i I) \geq 1$!*

Enfin pour en terminer avec les définitions, rappelons encore que

Définition 10.2.2 *on dit que $v \in \mathbb{C}^n$ 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 vérifie $A^*v = \overline{\mu}v$, donc un vecteur propre à gauche de A est vecteur propre à droite de A^*. Rappelons par ailleurs que

$$det(A - \lambda I)^* = det(A^* - \overline{\lambda} I)$$

ainsi les valeurs propres de A^* sont les complexes conjugués des valeurs propres de A, et en conséquence tout vecteur propre à gauche de A, vérifie $v^*A = \lambda v^*$, avec λ valeur propre de A.

Exemple 10.2.1 *Pour mieux assimiler ces notions, considérons les deux matrices suivantes :*

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} \quad et \quad A_2 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix};$$

elles ne différent que par le dernier coefficient diagonal, mais ont des propriétés spectrales distinctes

$$Spe(A_1) = \{1, 2, 3\} \quad et \quad Spe(A_2) = \{1, 2\}.$$

La matrice A_1 ayant ses valeurs propres réelles distinctes, ses vecteurs propres forment une base de $\mathbb{R}^3$, et elle est donc semblable à une matrice diagonale (voir Proposition 10.5.3) :

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}^{-1}.$$

La matrice A_2 n'est pas diagonalisable : 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.

10.3 Cas des matrices diagonalisables

On s'intéresse d'abord au cas des matrices diagonalisables, que l'on va caractériser, puis on étudiera le cas général.

Proposition 10.3.1 *Les valeurs propres d'une matrice hermitienne sont réelles.*

Preuve : En effet on a la suite d'implications :

$$\left.\begin{array}{c} Au = \lambda u \Longrightarrow u^*A^* = \overline{\lambda} u^* \Longrightarrow (u^*A^*)u = \overline{\lambda}\ u^*u \\ (u^*A^*)u = u^*(Au) = u^*(\lambda u) = \lambda\ u^*u \end{array}\right\} \Longrightarrow \lambda = \overline{\lambda}.$$

■

Proposition 10.3.2 *Les vecteurs propres d'une matrice hermitienne correspondant à des valeurs propres distinctes sont orthogonaux.*

Preuve : On peut écrire

$$\left.\begin{array}{l} Au = \lambda u \Longrightarrow u^*A^*v = \overline{\lambda} u^*v = \lambda\ u^*v \\ (u^*A^*)v = u^*(Av) = u^*(\mu v) = \mu\ u^*v \end{array}\right\} \Longrightarrow (\lambda - \mu)\ u^*v = 0 \Longrightarrow u^*v = 0 \text{ si } \lambda \neq \mu.$$

■

Proposition 10.3.3 *Toute matrice hermitienne est diagonalisable dans une base orthonormale.*

Cette propriété découle d'un résultat plus général

Proposition 10.3.4 *[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, semblables, ont 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 10.3.3.

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 $\mathbb{C}^n$, et ce résultat est vrai, que les valeurs propres soient distinctes ou non, ce qui constitue une extension de la Proposition 10.3.2.

Proposition 10.3.5 *Toute matrice $A \in \mathbb{C}^{n\times n}$ normale est diagonalisable dans une base orthonormale.*

Preuve : 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

$$(TT^*)_{2,2} = \sum_{j\geq 2} |T_{2,j}|^2 = (T^*T)_{2,2} = |T_{1,2}|^2 + |T_{2,2}|^2$$

en tenant compte de $T_{1,2} = 0$, cette relation entraine que tous les coefficients extra-diagonaux $T_{2,j}$ sont nuls... En réitérant ce procédé, on montre que la matrice T est diagonale.

La réciproque est évidente : si A est diagonalisable dans une base orthonormale, alors $A = QDQ^*$, donc $A^* = QD^*Q^*$ et $AA^* = QDD^*Q^* = QD^*DQ^* = A^*A$. Toute matrice diagonalisable dans une base orthonormale est normale. ■

Remarque 10.3.1 *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 du paragraphe précédent, car la matrice A définie par

$$A = \begin{bmatrix} i & 0 & \dots & 0 \\ 0 & i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & i \end{bmatrix}$$

est diagonalisable (car normale) sans être hermitienne.

Un exemple moins trivial est fourni par la matrice de permutation P 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 :

$$Q^*PQ = \Lambda$$

Les matrices Q et Λ sont définies en posant $z = e^{i\pi/n}$ et $\overline{z} = e^{-i\pi/n}$:

$$Q = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \overline{z} & \overline{z}^2 & \vdots & \overline{z}^{(n-1)} \\ 1 & \overline{z}^2 & \overline{z}^4 & \vdots & \overline{z}^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \overline{z}^{n-1} & \overline{z}^{2(n-1)} & \dots & \overline{z}^{(n-1)(n-1)} \end{bmatrix} \quad \text{et} \quad \Lambda = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & z & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & z^{(n-2)} & 0 \\ 0 & \dots & \dots & 0 & z^{(n-1)} \end{bmatrix}.$$

Remarque 10.3.2 *Attention : toute matrice A diagonalisable n'est pas nécessairement normale ! Il faut que la base de vecteurs propres de A soit orthonormale pour avoir cette propriété, comme le montre l'exercice suivant.*

Exercice 10.3.1 *Soit la matrice*

$$A = \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix}.$$

Montrer que A est diagonalisable, puis calculer une forme de Schur de A : $A = Q^*TQ$. Que peut-on en conclure ?

Terminons ce paragraphe par l'énoncé de quelques résultats utiles sur les matrices hermitiennes.

Théorème 10.3.1 (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

$$(i) \quad \lambda_k = \max_{\substack{V \\ dimV=k}} \min_{x \in V-\{0\}} \frac{x^*Ax}{x^*x},$$

$$(ii) \quad \lambda_k = \min_{\substack{W \\ dimW=n-k+1}} \max_{x \in W-\{0\}} \frac{x^*Ax}{x^*x}.$$

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 $V =< u_1, u_2, \ldots, u_k >$, 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 posant $V_k =< u_1, u_2, \ldots, u_k >$ 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 $M(V) = \max_{x \in W-\{0\}} \dfrac{x^*Ax}{x^*x} \geq \lambda_k$, puis que

$$\min_{\substack{W \\ dimW=n-k+1}} \max_{x \in W-\{0\}} \frac{x^*Ax}{x^*x} \geq \lambda_k$$

. 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 10.3.2 *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 intéressant 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 10.3.2 n'est pas utilisable pour le calcul numérique car la matrice $\mathcal{E}$ n'est jamais hermitienne !

10.4 Localisation des valeurs propres

Théorème 10.4.1 *[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

$$\sum_j A_{i,j}\ u_j = \lambda u_i \Longleftrightarrow \lambda - A_{i,i} = \sum_{j \neq i} A_{i,j} \frac{u_j}{u_i} \Longleftrightarrow |\lambda - A_{i,i}| \leq \sum_{j \neq i} |A_{i,j}|$$

donc 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 10.1 que les valeurs propres de la matrice S

(représentées par des croix) sont situées à l'intérieur de deux disques

$$S = \begin{pmatrix} 0.5000 & 0.1667 & 0.0417 & 0.0083 \\ 0.1667 & 0.0417 & 0.0083 & 0.0014 \\ 0.0417 & 0.0083 & 0.0014 & 0.0002 \\ 0.0083 & 0.0014 & 0.0002 & 0.0000 \end{pmatrix}$$

$$\begin{aligned} \lambda_1(S) &= 0.5575 \text{ dans le disque de centre } (0.5000, 0.) \quad \text{de rayon } 0.2167 \\ \lambda_2(S) &= -0.0146 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \\ \lambda_3(S) &= 0.0001 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \\ \lambda_4(S) &= 0.0000 \text{ dans le disque de centre } (0.0417, 0.) \quad \text{de rayon } 0.1764 \end{aligned}$$

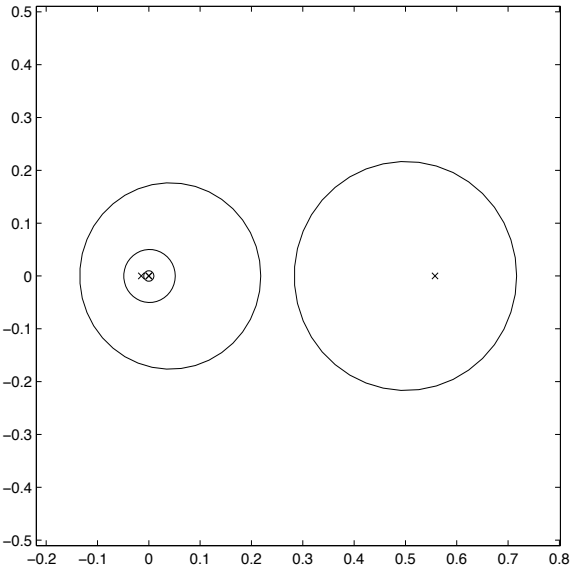


FIG. 10.1 – *Le spectre de la matrice S*

Dans le cas d'une matrice non symétrique N, on peut appliquer le Théorème 10.4.1 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 :

$$N = \begin{pmatrix} 0.5000 & -0.1000 & 0.0417 & 0.0083 \\ 0.1667 & 0.0417 & 0.0083 & 0.0014 \\ 0.0417 & 0.0000 & 0.0014 & 0.0002 \\ 0.0083 & 0.0014 & 0.0000 & 0.0100 \end{pmatrix}$$

Il existe de nombreuses configurations, l'exemple de la matrice circulante C illustre un cas particulier qui fait l'objet de l'exercice 10.4.1

$$C = \begin{pmatrix} 1. & 2. & 3. & 4. \\ 4. & 1. & 2. & 3. \\ 3. & 4. & 1. & 2. \\ 2. & 3. & 4. & 1. \end{pmatrix}$$

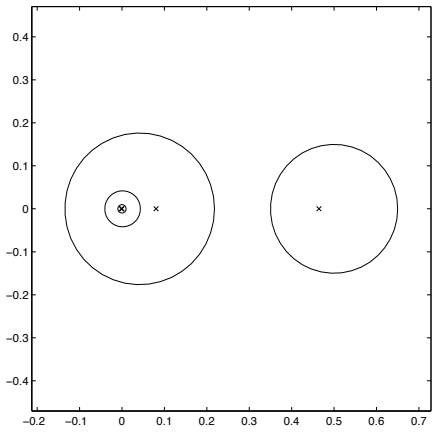


FIG. 10.2 – *Le spectre de N*

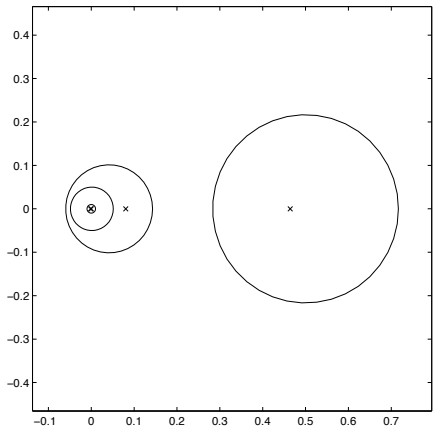


FIG. 10.3 – *Le spectre de* N^T

Pour cette matrice l'ensemble D est contenu tout entier dans un seul disque (voir figure 10.4), avec une valeur propre sur la frontière

$$\begin{aligned}
\lambda_1(C) = &\quad 10. \qquad \text{sur le cercle de centre } (1.,0.) \quad \text{de rayon } 9.\\
\lambda_2(C) = &\quad -2.+2.*i \text{ dans le disque de centre } (1.,0.) \quad \text{de rayon } 9.\\
\lambda_3(C) = &\quad -2.-2.*i \text{ dans le disque de centre } (1.,0.) \quad \text{de rayon } 9.\\
\lambda_4(C) = &\quad -2. \qquad \text{dans le disque de centre } (1.,0.) \quad \text{de rayon } 9.
\end{aligned}$$

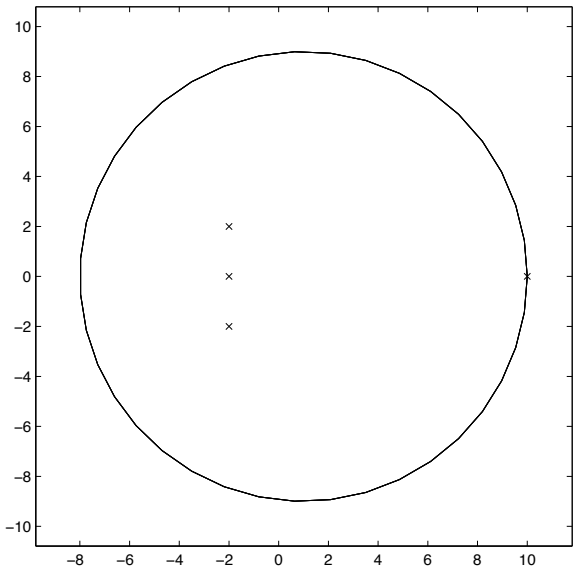


FIG. 10.4 – *Le spectre de C*

Exercice 10.4.1 *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 :

Théorème 10.4.2 *Soit une matrice $A \in \mathbb{C}^{n\times n}$ et soient les n disques D_i du plan complexe définis par*

$$D_i = \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

Si 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 10.4.1, 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 car les valeurs propres de $A(\varepsilon)$ dépendent continûment de ε. ∎

On peut voir sur la figure 10.5 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

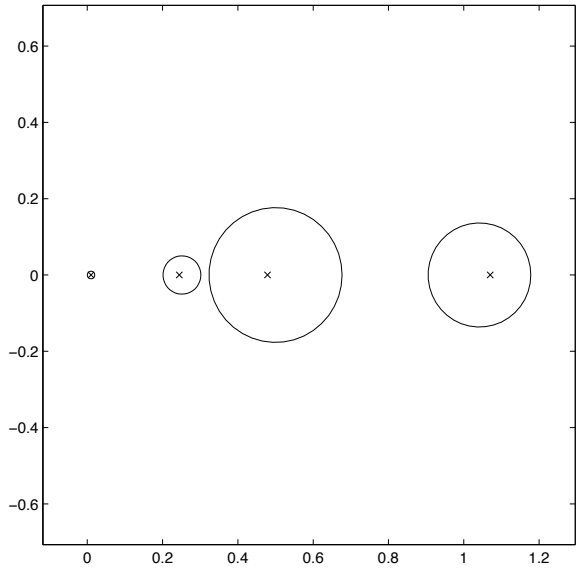


FIG. 10.5 – *Le spectre de A*

10.5 Le cas général

On s'intéresse maintenant au cas d'une matrice $A \in \mathbb{C}^{n\times n}$, sans propriété particulière :

Proposition 10.5.1 *Les vecteurs propres de la matrice $A \in \mathbb{C}^{n\times n}$ associés à des valeurs propres distinctes sont linéairement indépendants.*

Preuve : Soient λ_1,λ_2,...,λ_d les valeurs propres distinctes d'une matrice $A \in \mathbb{C}^{n\times n}$ et soient u_1,u_2,...,u_d des vecteurs propres associés. On démontre la proposition par récurrence :

• pour $d = 2$: $Au_1 = \lambda_1 u_1$ et $Au_2 = \lambda_2 u_2$, si on suppose $u_1 = \alpha u_2$ alors

$$\left.\begin{array}{c} Au_1 = \lambda_1 u_1 = \lambda_1(\alpha u_2) = \alpha\lambda_1 u_2 \\ Au_1 = A(\alpha u_2) = \alpha(Au_2) = \alpha\lambda_2 u_2 \end{array}\right\} \Longrightarrow \alpha(\lambda_1 - \lambda_2)u_2 = 0 \Longrightarrow \alpha = 0.$$

• on suppose la propriété vérifiée jusqu'à l'ordre k inclus : on considère u_{k+1} tel que $Au_{k+1} = \lambda_{k+1}u_{k+1}$, et on suppose que u_{k+1} dépend linéairement des k vecteurs propres précédents, alors

$$Au_{k+1} = A\sum_{i=1}^{k} \alpha_i u_i = \lambda_{k+1}u_{k+1} = \sum_{i=1}^{k} \lambda_{k+1}\alpha_i u_i = \sum_{i=1}^{k} \alpha_i\lambda_i u_i$$

ainsi

$$\sum_{i=1}^{k} \alpha_i(\lambda_i - \lambda_{k+1})u_i = 0.$$

De l'hypothèse de récurrence, on tire $\alpha_i(\lambda_i - \lambda_{k+1}) = 0$ pour $i = 1, 2, \ldots, k$, soit finalement $\alpha_i = 0$ pour $i = 1, 2, \ldots, k$. ■

Proposition 10.5.2 *Une matrice $A \in \mathbb{C}^{n\times n}$ est diagonalisable si et seulement si elle possède n vecteurs propres linéairement indépendants $u_1, u_2, \ldots, u_n$*

$$A = [\,u_1 \quad u_2 \quad \ldots \quad u_n\,] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} [\,u_1 \quad u_2 \quad \ldots \quad u_n\,]^{-1}.$$

Preuve : Si A est diagonalisable, on peut écrire $A = U\Lambda U^{-1}$, en posant

$$U = [\, u_1 \quad u_2 \quad \dots \quad u_n \,] \quad \text{et} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

Les vecteurs colonnes u_i sont donc les vecteurs propres de la matrice A : $AU = U\Lambda$ et ils sont linéairement indépendants puisque la matrice U est inversible.

Réciproquement si les vecteurs propres $u_1, u_2, \dots, u_n$ sont linéairement indépendants alors U est inversible et $AU = U\Lambda \Longrightarrow A = U\Lambda U^{-1}$. ∎

Remarque 10.5.1 *Pour toute matrice A non défective, on a $n = \sum_i^d g_i$; il existe donc n vecteurs propres linéairement indépendants, et la matrice A est donc diagonalisable.*

Proposition 10.5.3 *Si toutes les valeurs propres d'une matrice $A \in \mathbb{C}^{n\times n}$ sont distinctes, alors A est diagonalisable.*

C'est la conséquence des Propositions 10.3.5 et 10.5.4.

Proposition 10.5.4 *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 : 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

$$\left.\begin{array}{r} Au_i = \lambda_i u_i \Longrightarrow v_j^* A u_i = \lambda_i v_j^* u_i \\ A^* v_j = \overline{\lambda}_j v_j \Longrightarrow u_i^* A^* v_j = \overline{\lambda}_j u_i^* v_j \end{array}\right\} \Longrightarrow v_j^* A u_i = \lambda_i v_j^* u_i = (u_i^* A^* v_j)^* = \lambda_j v_j^* u_i$$

soit finalement

$$(\lambda_i - \lambda_j) v_j^* u_i = 0.$$

∎

Exercice 10.5.1 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice admettant n vecteurs propres à droite $v_1, v_2, \dots, v_n$ linéairement indépendants, et n vecteurs propres à gauche $w_1, w_2, \dots, w_n$ linéairement indépendants. Montrer que $(v_i, w_i) \neq 0 \qquad 1 \leq i \leq n$.*

Lorsque la matrice A est diagonalisable, la relation $AU = U\Lambda$, est équivalente à $A^*(U^{-1})^* = (U^{-1})^*\Lambda^*$, les vecteurs colonnes de U sont les vecteurs propres à droite de la matrice A et les vecteurs colonnes de la matrice $(U^{-1})^*$ en sont les vecteurs propres à gauche. Posons

$$U = [\, u_1 \quad u_2 \quad \dots \quad u_n \,] \quad \text{et} \quad (U^{-1}) = \begin{bmatrix} v_1^* \\ v_2^* \\ \dots \\ v_n^* \end{bmatrix};$$

La relation $(U^{-1})U = I$ s'écrit encore $v_j^* u_i = 0$ si $i \neq j$. Ce qui généralise le résultat de la Proposition 10.5.4 au cas des matrices diagonalisables (on n'a plus besoin de l'hypothèse $\lambda_i \neq \lambda_j$). On résume la relation $A = U\Lambda U^{-1}$ dans la formule

$$A = [\, u_1 \quad u_2 \quad \dots \quad u_n \,] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \dots \\ v_n^* \end{bmatrix}$$

que l'on écrit encore

$$A = \sum_{i=1}^{n} \lambda_i \; u_i \cdot v_i^*,$$

cette relation est appelée **décomposition spectrale** de la matrice A. Noter que dans cette relation les termes $u_i \cdot v_j^*$ représentent des matrices de $\mathbb{C}^{n\times n}$.

10.6 Forme de Jordan

On étudie maintenant le cas des matrices défectives ; de telles matrices ne sont pas diagonalisables, car il n'est pas possible de construire une base de $\mathbb{C}^n$ à l'aide de leurs vecteurs propres, comme pour les matrices du le paragraphe précédent. Il faut donc construire d'autres vecteurs associés aux vecteurs propres, ce sont les **vecteurs principaux**. Cette construction conduit à la **forme de Jordan** dans laquelle la matrice est écrite sous une forme presque diagonale (voir Chatelin [5]).

Théorème 10.6.1 *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. 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, elle se décompose en g_i blocs de Jordan $J_{i,j}$:

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \dots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{i,g_i} \end{bmatrix} \quad \text{avec } J_{i,j} = \begin{bmatrix} \lambda_i & 1 & \dots & \dots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{bmatrix}$$

La démonstration de ce Théorème est effectuée par étapes et nécessite plusieurs résultats intermédiaires qui sont proposés comme exercices.

Remarque 10.6.1 *Pour toute valeur propre λ_i telle que $m_i = g_i$ chaque bloc J_i est une matrice diagonale puisque $J_{i,j} = [\lambda_i]$ pour $1 \le j \le g_i$. On retrouve donc la propriété* non défective = diagonalisable, *localisée au sous-espace Ker $(A - \lambda_i I)$.*

Exercice 10.6.1 *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 $j \le i$).

Exercice 10.6.2 *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 10.6.3 *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 $\mathbb{R}^k$.

Exercice 10.6.4 *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*

$$Y^{-1}UY = \begin{bmatrix} E_1 & 0 & \dots & 0 \\ 0 & E_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & E_g \end{bmatrix} \quad \text{avec } E_j = \begin{bmatrix} 0 & I_{k_j-1} \\ 0 & 0 \end{bmatrix}$$

avec $k_1 \geq k_2 \geq \dots \geq k_g$.

Démonstration du Théorème 10.6.1 :
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 10.3.4) $Q^*AQ = R$.
- On applique le résultat de l'Exercice 10.6.1 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 $j = 1, 2, \ldots, g_i$, le résultat de l'Exercice 10.6.4 à chaque bloc $U_i = \lambda_i I - \tilde{R}_i$:

$$Y_i^{-1}\left(\lambda_i I + U_i\right) Y_i = \lambda_i I + \begin{bmatrix} E_1 & 0 & \ldots & 0 \\ 0 & E_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & E_{g_i} \end{bmatrix} \quad \text{avec } E_{i,j} = \begin{bmatrix} 0 & 1 & \ldots & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{bmatrix}$$

et les blocs $E_{i,j}$ sont rangés traditionnellement par ordre de rang croissant.

– On pose maintenant

$$\tilde{Y} = \begin{bmatrix} Y_1 & 0 & \ldots & 0 \\ 0 & Y_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & Y_d \end{bmatrix}$$

et on obtient

$$\left(\tilde{Y}^{-1} Z^{-1} Q^*\right) A \left(Q Z \tilde{Y}\right) = J = \begin{bmatrix} J_1 & 0 & \ldots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & J_d \end{bmatrix}$$

Dans cette écriture

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \ldots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & J_{i,g_i} \end{bmatrix} \text{ avec } J_{i,j} = \begin{bmatrix} \lambda_i & 1 & \ldots & \ldots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \ldots & \ldots & \ldots & \lambda_i \end{bmatrix}$$

est la **boîte de Jordan** associée à λ_i.

Les matrices A et J sont semblables et ainsi 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, alors $dim M_i = m_i$ et

$$\sum_{i=1}^{d} m_i = n \quad \text{soit} \quad \bigoplus_{i=1,d} M_i = \mathbb{C}^n$$

enfin

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \ldots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & J_{i,g_i} \end{bmatrix} \text{ avec } J_{i,j} = \begin{bmatrix} \lambda_i & 1 & \ldots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & \ldots & \lambda_i \end{bmatrix}$$

à chacun des g_i sous–blocs $J_{i,j}$ de rang $m_{i,j}$, est associé un sous–espace $M_{i,j}$ de M_i. Cherchons un vecteur propre u dans ce sous–espace : u doit vérifier les $m_{i,j}$ relations :

$$\begin{aligned}
\lambda_i u_1 + u_2 &= \lambda_i u_1 \\
\lambda_i u_2 + u_3 &= \lambda_i u_2 \\
\dots &= \dots \\
\lambda_i u_{m_{i,j}-1} + u_{m_{i,j}} &= \lambda_i u_{m_{i,j}-1} \\
\lambda_i u_{m_{i,j}} &= \lambda_i u_{m_{i,j}}
\end{aligned}$$

On en déduit que nécessairement $u_2 = u_3 = \dots = u_{m_{i,j}} = 0$! Le seul vecteur propre possible dans $M_{i,j}$ s'écrit donc $u = (1, 0, \dots, 0)^T$; ainsi 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 est donc la multiplicité géométrique de λ_i.

Soit maintenant $z_0 \in \text{Ker } (A - \lambda_i I)$ 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

$$(A - \lambda_i I)^2 z_1 = (A - \lambda_i I)z_0 = 0 \quad \text{soit} \quad z_1 \in \text{Ker } (A - \lambda_i I)^2.$$

On voit donc pour que z_1 existe, il faut et il suffit que Ker $(A - \lambda_i I)^2 \neq \{0\}$.

On peut continuer ainsi 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 Ker $(A - \lambda_i I)^{k+1}$; mais puisque

$$\text{Ker } (A - \lambda_i I) \subset \text{Ker } (A - \lambda_i I)^2 \subset \dots \subset \text{Ker } (A - \lambda_i I)^k \subset \dots \subset \mathbb{C}^n,$$

il existe nécessairement un entier $l_i \leq n$ tel que

$$\text{Ker } (A - \lambda_i I)^{l_i} = \text{Ker } (A - \lambda_i I)^l \quad \forall l \geq l_i.$$

cet entier est tel que 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. Ces vecteurs vérifient les relations

$$\begin{aligned}
Az_0 &= \lambda_i z_0 \\
Az_1 &= \lambda_i z_1 + z_0 \\
Az_2 &= \lambda_i z_2 + z_1 \\
\dots &= \dots \\
Az_{l_i} &= \lambda_i z_{l_i} + z_{l_{i-1}}
\end{aligned}$$

soit encore

$$A \begin{bmatrix} z_0 & z_1 & \dots & z_{l_i} \end{bmatrix} = \begin{bmatrix} z_0 & z_1 & \dots & z_{l_i} \end{bmatrix} \begin{bmatrix} \lambda_i & 1 & \dots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & \dots & \lambda_i \end{bmatrix}$$

On reconnait 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 nécessairement inférieur ou égal à l'indice l_i.

On voit donc que la représentation de Jordan n'est pas unique car la décomposition de la boîte J_i en blocs $J_{i,j}$ dépend du choix du vecteur z_0 dans chaque $M_{i,j}$. Ainsi 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 :

$$\left[\begin{array}{c|ccc|ccc} \lambda_i & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \lambda_i & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right]$$

ou

$$\left[\begin{array}{cc|cc|ccc} \lambda_i & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_i & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right]$$

soit comment écrire 7 (m_i la multiplicité algèbrique de λ_i) comme somme de 3 (g_i le nombre de vecteurs propres linéairement indépendants) entiers naturels, chaque entier étant inférieur ou égal à 3 (l_i la dimension maximale d'un sous–espace propre) :

$$7 = 1 + 3 + 3 = 2 + 2 + 3.$$

Noter que cette écriture contient le cas A diagonalisable pour lequel $l_i = 1$ et $m_i = g_i$ pour tout i.

Exercice 10.6.5 *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 $C(a)$.

10.7 Décomposition spectrale d'une matrice

Ce travail préparatoire va permettre de simplifier la suite de l'étude des matrices défectives en généralisant la notion de décomposition spectrale introduite auparavant : on écrit toute matrice $A \in \mathbb{C}^{n\times n}$ ayant d valeurs propres distinctes suivant

$$A = XJX^{-1}$$

où

$$X = [\, X_1 \quad X_2 \quad \dots \quad X_d \,].$$

Chaque bloc X_i correspond aux m_i colonnes de X associées au sous–espace M_i, et on introduit de manière cohérente

$$X^{-1} = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \cdots \\ Y_d^* \end{bmatrix}.$$

Les vecteurs colonnes de X_i forment une base de M_i et les vecteurs Y_i forment une base adjointe. Comme précédemment,

$$X^{-1}X = I \Longleftrightarrow Y_i^* X_i = I_{m_i} \text{ et } Y_i^* X_j = [0] \text{ si } i \neq j$$

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

$$\begin{aligned} Y_i^* X_i = I_{m_i} &\Longrightarrow P_i^2 = P_i \\ Y_i^* X_j = [0] &\Longrightarrow P_i P_j = 0 \quad i \neq j \end{aligned}$$

Finalement on résume ces résultats dans la formule

$$A = \sum_{i=1}^{d} \left(\lambda_i P_i + D_i\right),$$

qui est la décomposition spectrale d'une matrice quelconque, avec

$$J_i = \lambda_i I_{m_i} + N_i, \quad D_i = X_i N_i Y_i^* \quad \text{et} \quad \sum_{i=1}^{d} P_i = I_n$$

où par exemple

$$N_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

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 $N_i^{l_i} = [0]$ soit $D_i^{l_i} = [0]$.

Dans le cas particulier d'une valeur propre λ_i semi-simple $m_i = g_i$, $l_i = 1$; la matrice N_i est nulle, et donc $D_i = [0]$.

Enfin, de la formule générale on déduit que

$$AP_j = \sum_{i=1}^{d} \left(\lambda_i P_i + D_i\right) P_j = \lambda_j P_j^2 + D_j P_j = P_j \left(\lambda_j P_j + D_j\right)$$

(P_j et D_j commutent par définition), et plus généralement

$$A^k P_j = P_j \left(\lambda_j P_j + D_j\right)^k.$$

Proposition 10.7.1 *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 $M_i = \text{Ker } (A - \lambda_i)^{l_i}$, 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 suivant 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 [5], B. N. Parlett [19] et Y. Saad [23].

Ce qu'il faut retenir

Ce chapitre ne fait que rappeler des notions que vous devez déjà connaitre...

Chapitre 11

Méthode de la puissance itérée

11.1 Introduction

A partir des notions générales introduites dans le chapitre précédent, il s'agit maintenant 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, la puissance itérée inverse et enfin la méthode du sous–espace et la méthode QR.

11.2 Etude d'un exemple

Avant de présenter en détails 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, évoquée à la fin du chapitre 5.

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?

Si on appelle e^k le nombre d'étudiants diplômés pour le mois k, 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 :

$$A = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 3/4 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}^{-1}.$$

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, qui fait intervenir les puissances successives d'une matrice A pour calculer une valeur propre de plus grand module et un vecteur propre associé :

1) **initialisation :**

$v_0 \neq 0 \in \mathbb{C}^n$ quelconque

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

$v_k = Av_{k-1}/\alpha_k$

α_k est une composante de module maximum de Av_{k-1}

fin

Par construction, on a pour tout $k > 1$, $\|v_k\|_\infty = 1$; 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, on vérifie que

$$v_0 = \sum_{i=1}^{d} P_i v_0$$

$$Av_0 = \sum_{i=1}^{d} AP_i v_0 = \sum_{i=1}^{d} (\lambda_i P_i + D_i) P_i v_0$$

$$A^k v_0 = \sum_{i=1}^{d} (\lambda_i P_i + D_i)^k P_i v_0$$

$$\text{donc } v_k = \frac{1}{\alpha_k} Av_{k-1} = \frac{1}{\tilde{\alpha}_k} \sum_{i=1}^{d} (\lambda_i P_i + D_i)^k P_i v_0$$

$$\text{avec } \tilde{\alpha}_k = \prod_{l=1}^{k} \alpha_l$$

Théorème 11.2.1 *Soit la matrice $A \in \mathbb{C}^{n\times n}$ 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. Alors si v_0 a une composante non nulle sur le sous–espace M_1, alors le vecteur v_k tend vers un vecteur propre associé à λ_1 et α_k tend vers $|\lambda_1|$.*

Preuve : Puisque λ_1 est supposée semi–simple, $D_1 = [0]$ et on écrit

$$v_k = \frac{1}{\tilde{\alpha}_k} \left[\lambda_1^k P_1 v_0 + \sum_{i=2}^{d} (\lambda_i P_i + D_i)^k P_i v_0 \right]$$

$$= \frac{\lambda_1^k}{\tilde{\alpha}_k} \left[P_1 v_0 + \sum_{i=2}^{d} \frac{1}{\lambda_1^k} (\lambda_i P_i + D_i)^k P_i v_0 \right]$$

Le rayon spectral de la matrice $\dfrac{1}{\lambda_1^k} P_i(\lambda_i P_i + D_i)$ est plus petit que 1 pour $i \neq 1$ puisque $|\lambda_i| < |\lambda_1|$ pour $i \neq 1$, ainsi $v_k \simeq \dfrac{\lambda_1^k}{\tilde{\alpha}_k} P_1 v_0$ quand $k \to +\infty$. D'après la Proposition 10.7.1 $P_1 v_0$ est un vecteur

propre associé à λ_1, à moins que $P_1 v_0 = 0$. Si on suppose $P_1 v_0 \neq 0$, par construction et pour tout k, v_k est un vecteur normé qui reste colinéaire au vecteur $P_i v_0$. On en déduit successivement (en supposant $\lambda_1 > 0$) que

$$\begin{aligned} v_k &\to v = P_1 v_0 / \|P_1 v_0\|_\infty \text{ quand } k \to +\infty \\ Av_{k-1} &\to Av = \lambda_1 P_1 v_0 / \|P_1 v_0\|_\infty = \lambda_1 v \text{ quand } k \to +\infty \\ \alpha_k v_k &= Av_{k-1} \to \lambda_1 v \text{ quand } k \to +\infty \\ v_k &\to v \text{ quand } k \to +\infty \\ \alpha_k &\to \lambda_1 \text{ quand } k \to +\infty. \end{aligned}$$

Si $\lambda_1 < 0$, le vecteur v_k est de la forme $(-1)^k v$, et on constate le changement de signe des composantes non nulles de ce vecteur à chaque itération. ■

On note que
1) l'algorithme fournit une valeur propre et un vecteur propre associé ;
2) la vitesse de convergence de l'algorithme est lié au rapport $\rho = |\lambda_2|/|\lambda_1|$.

11.3 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} : $Au = \lambda u \iff A^{-1}u = \frac{1}{\lambda}u$. 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 :**
   $v_0 \neq 0 \in \mathbb{C}^n$ quelconque
2) **itérations : pour** $k = 1, 2, \ldots$ **faire**
   $v_k = A^{-1} v_{k-1}/\alpha_k$
   α_k est une composante de module maximum de $A^{-1}v_{k-1}$
   fin

Cet algorithme converge vers la valeur propre de plus grand module de A^{-1} : soit $\frac{1}{\lambda_n}$.

Dans la pratique pour calculer v_k, on effectue une factorisation de la matrice A par la méthode de Cholesky (respectivement par la méthode de Gauss), et on résout le système linéaire $LL^T v_k = v_{k-1}$ (respectivement $LUv_k = v_{k-1}$).

11.4 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 de converger vers une valeur propre $\lambda_j \neq \lambda_1$.

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 $\lambda_2 = 3$!

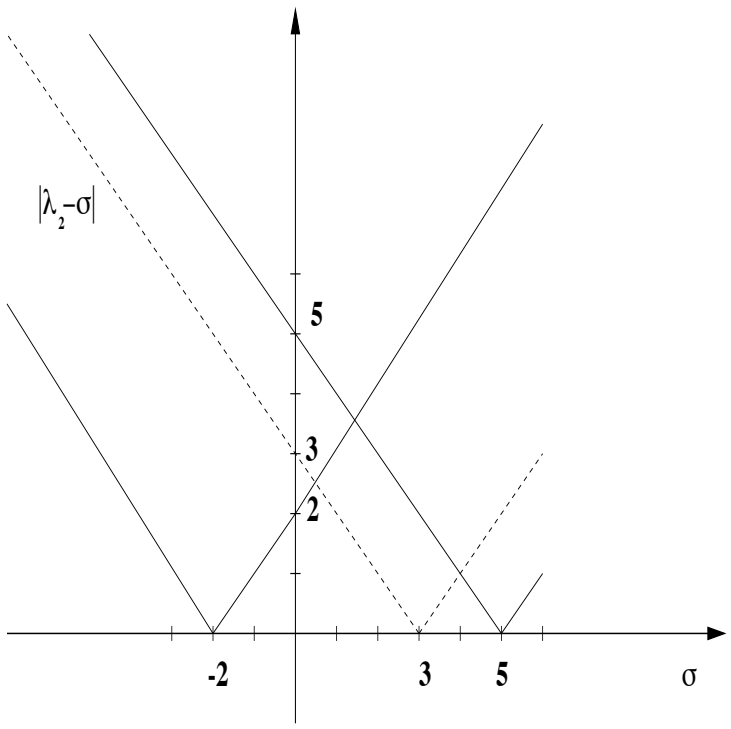


FIG. 11.1 – *Les variations de* $|\lambda - \sigma|$

Sur le graphique 11.1, 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 $|\lambda_2 - \sigma|$ 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 11.2

On peut également appliquer la technique de translation à cet algorithme en factorisant la matrice $\tilde{A} = A - \sigma I$; 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

$$\frac{\dfrac{1}{|\lambda'-\sigma|}}{\dfrac{1}{|\lambda-\sigma|}} = \frac{|\lambda-\sigma|}{|\lambda'-\sigma|}$$

ce rapport peut être très petit si σ est proche de λ (et assez éloignée de λ'). La convergence de la méthode est donc très rapide (quelques itérations) si on dispose d'une bonne estimation de λ.

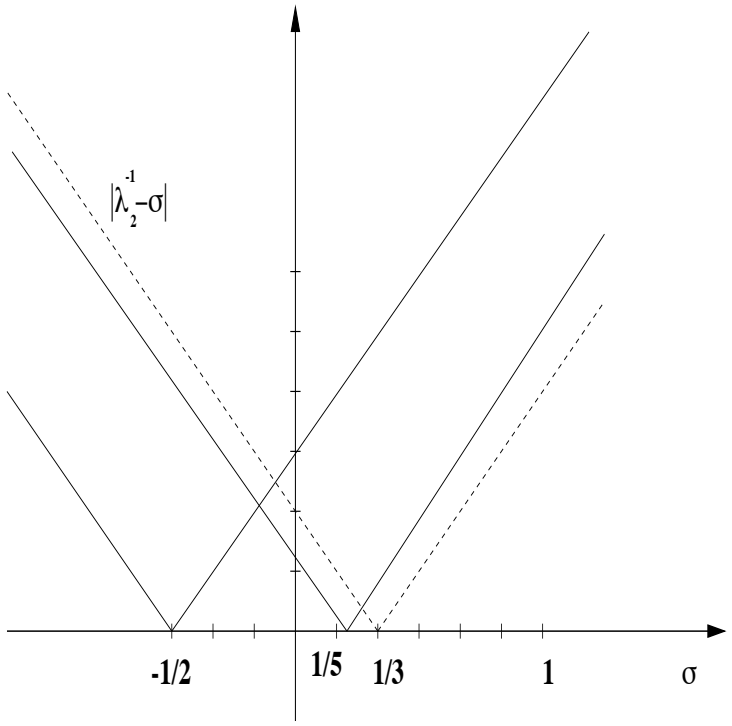


FIG. 11.2 – *Les variations de* $|\lambda^{-1}-\sigma|$

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. C'est le cas des matrices symétriques (ou hermitiennes) pour lesquelles on peut utiliser la méthode de tridiagonalisation associée au calcul des valeurs propres par la méthode de bissection (voir au chapitre 11). 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 11.4.1 *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 [19].*

11.5 Méthode de l'itération inverse de Rayleigh

L'idée d'utiliser un paramètre de translation σ est intéressante du point de vue de la convergence, mais que faire si on n'a pas d'estimation des valeurs propres ? Au cours des itérations, le coefficient de normalisation α_k tend vers la valeur propre dominante en module :

$\alpha_k \to \dfrac{1}{\lambda - \sigma}$. Une estimation de λ à l'itération k est donc $\sigma + \dfrac{1}{\alpha_k}$. On peut prendre en compte cette estimation pour obtenir l'algorithme suivant

1) **initialisation :**

$v_0 \in \mathbb{C}^n, \quad \|v_0\|_2 = 1$

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

$\sigma_k = (v_{k-1}, A^{-1} v_{k-1})$

$v_k = [A - \sigma_k I]^{-1} \, v_{k-1}/\alpha_k$

α_k calculé pour que $\|v_k\|_2 = 1$ (soit $\sigma_k \simeq \lambda$)

fin

Rappelons que par construction, $\sigma_k = \rho_{A^{-1}}(v_{k-1})$ est le quotient de Rayleigh associé au vecteur v_{k-1}, d'où le nom de cet algorithme. Cette méthode est coûteuse car il faut factoriser la matrice $A - \sigma_k I$ à chaque itération ! En pratique on fige donc σ_k pendant quelques itérations avant d'utiliser une nouvelle estimation de λ. On peut aussi utiliser cet algorithme comme méthode d'accélération de convergence quand on a obtenu une estimation σ d'une valeur propre par une autre méthode.

11.6 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 $v \in \mathbb{C}^n$ un vecteur tel que $v^* u_j = 1$.

Théorème 11.6.1 *Si la matrice $A \in \mathbb{C}^{n\times n}$ a pour spectre*

$$Spe(A) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_2, \lambda_1\}$$

et si les vecteurs u_j et v vérifient $Au_j = \lambda_j u_j$ et $v^* u_j = 1$, alors la matrice $\tilde{A} = A - \sigma u_j \cdot v^*$ a pour spectre

$$Spe(\tilde{A}) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_j - \sigma, \ldots, \lambda_2, \lambda_1\}$$

Preuve : D'après la Proposition 10.5.4 tout w_i vecteur propre à gauche de A associé à une valeur propre λ_i distincte de λ_j est orthogonal à u_j, donc

$$\tilde{A}^* w_i = (A^* - \overline{\sigma} v \cdot u_j^*) w_i = A^* w_i = \overline{\lambda}_i w_i$$

ainsi λ_i est valeur propre de $\tilde{A}$ pour $i \neq j$ et w_i vecteur propre à gauche de A est aussi vecteur propre à gauche de $\tilde{A}$. D'autre part

$$\tilde{A} u_j = (A - \sigma u_j \cdot v^*) u_j = (\lambda_j - \sigma) u_j$$

donc $\lambda_j - \sigma$ est valeur propre de $\tilde{A}$ et u_j est un vecteur propre associé. ∎

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 $i \neq j$:

$$\begin{aligned}\tilde{A}\tilde{u}_i &= (A - \sigma u_j \cdot v^*)(u_i - \gamma_i u_j) \\ &= \lambda_i u_i - (\gamma_i(\lambda_j - \sigma) + \sigma v^* u_i)\, u_j.\end{aligned}$$

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 11.6.1 *1) le choix du vecteur v du théorème ne pose pas de difficulté a priori, on peut par exemple prendre $v = w_j$ vecteur propre à gauche associé à λ_j, ce choix conduit à $\gamma_i = 0$, car dans ce cas $v^* u_i = 0$.*

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 $\tilde{A}v_k = Av_k - \sigma u_j(v^* v_k)$.

11.7 Factorisation QR d'une matrice

Avant d'envisager l'étude d'autres algorithmes, on introduit un résultat très utile, qui découle de la construction d'une base orthogonale de $\mathbb{C}^n$ par le procédé de Gram–Schmidt :

Proposition 11.7.1 *Soit $A \in \mathbb{C}^{n\times m}$ $(m \leq n)$ il existe une matrice $Q \in \mathbb{C}^{n\times m}$ et une matrice $R \in \mathbb{C}^{m\times m}$ telles que*

$$A = QR$$

avec $Q^*Q = I_m$ et R triangulaire supérieure.

Preuve : On note $A_k \in \mathbb{C}^n$ pour $k = 1, 2, \ldots, m$ les m colonnes de la matrice A, et on suppose qu'il y a au moins un de ces vecteurs qui est différent de 0 (sinon le résultat est vrai avec $R = A = [0]$) et sans limiter la généralité on suppose que cette colonne est la première : $A_1 \neq 0$.

Alors on commence à construire une base orthogonale de $\mathbb{C}^n$ suivant l'algorithme

1) **initialisation :**

$\tilde{q}_1 = A_1 \neq 0$

$q_1 = \tilde{q}_1 / \|\tilde{q}_1\|_2$

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

$$\tilde{q}_k = A_k - \sum_{l=1}^{k-1} (q_l, A_k) q_l$$

si $\|\tilde{q}_k\|_2 \neq 0 \quad q_k = \tilde{q}_k / \|\tilde{q}_k\|_2$

sinon : Arrêt des calculs

fin

On vérifie que par construction $(q_i, q_j) = \delta_{i,j}$ et on définit la matrice $R \in \mathbb{C}^{m \times m}$:

$$R_{l,k} = \begin{cases} (q_l, A_k) & \text{si } 1 \leq l \leq k-1 \\ \|\tilde{q}_k\| & \text{si } k = l \\ 0 & \text{si } k < l \leq m \end{cases}$$

alors les trois matrices sont liées par la relation

$$A = QR$$

dans laquelle R est une matrice triangulaire supérieure

$$R = Q^* A = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} \end{bmatrix}$$

Les matrices A et R ont même rang $r \leq m$, donc si $r = \mathrm{rang}(A) < m$ les $m - r$ dernières lignes de R sont nulles et les $m - r$ dernières colonnes de Q sont construites à partir de la matrice I_{m-r}. ■

Remarque 11.7.1 *En pratique il est possible d'obtenir un vecteur $\tilde{q}_k$ nul pour $k < r \leq m$; ceci se produit dès que le vecteur q_k est combinaison linéaire des vecteurs $q_1, q_2, \ldots, q_{k-1}$. Pour continuer l'orthogonalisation, au lieu de l'option Arrêt des calculs, il faut effectuer une permutation des vecteurs $q_k, q_{k+1}, \ldots, q_m$ pour remplacer q_k par un vecteur $q_j \notin < q_1, q_2, \ldots, q_{k-1} >$.*

On améliore la stabilité numérique des calculs avec la formulation suivante

1) **initialisation :**

$R_{1,1} = \|A_1\|_2$

$q_1 = A_1/R_{1,1}$

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

$\tilde{q}_k = A_k$

pour $l = 1, 2, \ldots, k-1$ faire

$R_{l,k} = (q_l, \tilde{q}_k)$

$\tilde{q}_k = \tilde{q}_k - R_{l,k} q_l$

fin boucle l

$R_{k,k} = \|\tilde{q}_k\|_2$

si $R_{k,k} = 0$: Stop

sinon $q_k = \tilde{q}_k / R_{k,k}$

fin

Cet algorithme est équivalent au précédent du point de vue algébrique, mais donne de meilleurs résultats numériques car les directions calculées respectent mieux les relations d'orthogonalité. Le résultat précédent est plus souvent utilisé sous la forme

Théorème 11.7.1 *Soit $A \in \mathbb{C}^{n\times m}$ $(m \leq n)$ il existe une matrice unitaire $Q \in \mathbb{C}^{n\times n}$ et une matrice rectangulaire $R \in \mathbb{C}^{n\times m}$ telles que*

$$A = QR$$

Pour passer de la Proposition 11.7.1 au Théorème 11.7.1, il suffit de compléter si nécessaire (cas $\text{rang}(A) < n$) l'ensemble des vecteurs $\{q_k\}_{k=1,m}$ pour obtenir une base orthogonale de $\mathbb{C}^n$. Les n vecteurs ainsi obtenus sont les colonnes d'une matrice $Q \in \mathbb{C}^{n\times n}$ unitaire. De manière cohérente, on ajoute éventuellement $n-m$ lignes de zéros à la matrice R de la Proposition 11.7.1 pour obtenir une matrice rectangulaire $R \in \mathbb{C}^{n\times m}$; dans le cas $m = n$, la matrice R est une matrice triangulaire supérieure.

Remarque 11.7.2 *La construction de la Proposition 11.7.1 définit les matrices Q et R de manière unique à partir des éléments de la matrice A ; mais cette factorisation n'est pas unique, comme le montre l'exercice suivant.*

Exercice 11.7.1 *Soit $A \in \mathbb{C}^{n\times n}$, on suppose qu'il existe deux matrices unitaires $Q \in \mathbb{C}^{n\times n}$, $Q' \in \mathbb{C}^{n\times n}$ et deux matrices triangulaires supérieures $R \in \mathbb{C}^{n\times n}$, $R' \in \mathbb{C}^{n\times n}$ telles que*

$$A = QR = Q'R'$$

Que peut-on dire des matrices Q et Q' ?

11.8 Méthode du sous–espace

La méthode de la puissance itérée est une méthode de calcul efficace, mais elle a quelques limites d'utilisation dans la pratique :

1) la méthode de la puissance itérée et ses variantes ne permettent de calculer les valeurs propres qu'une par une; est-il possible de construire une méthode qui calcule plusieurs valeurs propres en même temps ?

2) une matrice réelle A peut admettre des valeurs propres complexes conjuguées qui ont donc le même module ; dans le cas où la valeur propre dominante λ_1 est complexe, alors la valeur propre $\overline{\lambda}_1$ perturbe la convergence de l'algorithme. En particulier le vecteur $v_k = \dfrac{1}{\alpha_k} A v_{k-1}$ n'a pas de limite. Pourtant on observe numériquement que le sous–espace vectoriel engendré par les vecteurs v_k et v_{k-1} contient les vecteurs propres associés à λ_1 et $\overline{\lambda}_1$; il semble donc intéressant dans ce cas d'itérer sur le sous-espace vectoriel engendré par ces deux vecteurs.

On envisage donc une généralisation de la méthode de la puissance itérée, dans laquelle on itére sur plusieurs vecteurs à la fois. Dans la suite on note $\mathcal{X}_k =< x_1^k, x_2^k, \ldots, x_m^k >$ le sous–espace vectoriel engendré par les m vecteurs x_j^k, et

$$X_k = [\, x_1^k \quad x_2^k \quad \ldots \quad x_m^k \,]$$

la matrice dont les m colonnes sont les vecteurs x_j^k. On considère maintenant l'algorithme

1) **initialisation :**

choix d'un sous–espace $\mathcal{X}_0 \subset \mathbb{C}^n$ de dimension m

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

a) $\tilde{X}_k = A X_{k-1}$

b) $Q_k R_k = \tilde{X}_k$ (factorisation QR de $\tilde{X}_k$)

c) $X_k = Q_k$

fin

L'étape *b*) de l'algorithme fait appel à la factorisation QR de $\tilde{X}_k$, cette opération a pour but de préserver la dimension du sous–espace initial $\mathcal{X}_0$. En effet l'étude de la méthode de la puissance itérée montre que n'importe quel vecteur courant x_j^k tend à s'aligner sur le vecteur propre v_1 associé à la valeur propre dominante quel que soit le vecteur initial x_0^k (non orthogonal à v_1). En conséquence, même en choisissant un sous–espace $\mathcal{X}_0$ engendré par m vecteurs linéairement indépendants, au cours des itérations les vecteurs de $\mathcal{X}_k$ vont s'aligner sur v_1, en conséquence la dimension du sous–espace $\mathcal{X}_k$ deviendra inférieure à m ! Malheureusement cette étape de réorthogonalisation est très coûteuse et dans la pratique on ne l'effectue pas à chaque itération. On utilise plutôt la variante suivante

1) **initialisation :**

choix d'un sous–espace $\mathcal{X}_0 \subset \mathbb{C}^n$ de dimension m

choix du paramètre *iter*

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

a) $\tilde{X}_k = A^{iter} X_{k-1}$

b) $Q_k R_k = \tilde{X}_k$

c) $X_k = Q_k$

d) modification éventuelle de *iter*

fin

Le choix du paramètre *iter* est important : si on le prend trop grand on risque de diminuer la dimension du sous–espace $\mathcal{X}_k$, si on le prend trop petit le coût calcul de l'algorithme est trop élevé ! Seule l'expérimentation numérique peut permettre de trouver une bonne valeur du paramètre en fonction des données A, m et $\mathcal{X}_0$!

Théorème 11.8.1 *Soient λ_1, λ_2, ..., λ_m les m valeurs propres dominantes de la matrice $A \in \mathbb{C}^{n\times n}$, on suppose que*

$$|\lambda_{m+1}| < |\lambda_m| < \ldots < |\lambda_1|.$$

Soient q_1, q_2,..., q_m les vecteurs de Schur associés, on note P_i la projection spectrale associée à λ_i. Soient m vecteurs linéairement indépendants x_1, x_2,..., x_m on introduit les matrices

$$X_0 = [\, x_1 \quad x_2 \quad \ldots \quad x_m \,] \qquad et \qquad Q_m = [\, q_1 \quad q_2 \quad \ldots \quad q_m \,]$$

Si on suppose de plus, que pour tout $i = 1, 2, \ldots, m$ le rang de la matrice $P_i[x_1, x_2, \ldots, x_i]$ est i, alors la suite $\{X_k\}_{k\in\mathbb{N}}$ converge vers la matrice Q_m.

Preuve : Rappelons d'abord que si la matrice A est hermitienne, les vecteurs de Schur définis au chapitre 10 sont les vecteurs propres de A, et le Théorème 11.8.1 dit que la matrice X_k converge alors vers la matrice Q_m des vecteurs propres de A associés à λ_1, λ_2,..., λ_m.

Examinons maintenant l'hypothèse sur les opérateurs de projection spectrale P_i ; chaque P_i représente la projection orthogonale de $\mathbb{C}^n$ sur le sous–espace engendré par i vecteurs de Schur, et dire que le rang de P_iX_0 est i, revient à dire que $Q_m^* x_i^0 \neq 0$ pour tout vecteur initial x_i^0, c'est-à-dire qu'aucun des vecteurs x_i^0 n'est orthogonal au sous–espace engendré par les m vecteurs de Schur qui forment les colonnes de Q_m.

On note $Q = [\, Q_m \quad W \,]$ la matrice associée à la forme de Schur de A : $Q^*AQ = R$, R est une matrice triangulaire supérieure et

$$\begin{bmatrix} Q_m^* \\ W^* \end{bmatrix} A \, [\, Q_m \quad W \,] = \begin{bmatrix} Q_m^*AQ_m & Q_m^*AW \\ W^*AQ_m & W^*AW \end{bmatrix} = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \tag{11.1}$$

avec R_1 et R_2 matrices triangulaires supérieures. En effet, les m valeurs propres dominantes de A étant supposées distinctes, elles sont simples et les vecteurs de Schur associés sont des vecteurs propres ; de plus l'hypothèse $|\lambda_{m+1}| < |\lambda_m|$ entraine que le sous–espace engendré par les m premiers vecteurs q_i est orthogonal au sous–espace engendré par les vecteurs q_j avec $m < j \leq n$. De la relation (11.1) on tire :

$$AQ_m = Q_mR_1 \qquad Q_mQ_m^* = I_m$$

$$AW = WR_2 \qquad Q_m^*W = 0$$

On peut donc écrire $X_0 = Q_mG_1 + WG_2$ décomposition de X_0 dans la base formée par les m vecteurs de Schur $\{q_1, q_2, \ldots, q_m\}$ et les $n - m$ vecteurs colonnes de W : $\{w_1, w_2, \ldots, w_{n-m}\}$. D'après l'hypothèse sur les opérateurs de projection spectrale P_i, la matrice $G_1 \in \mathbb{R}^{m\times m}$ est une matrice triangulaire supérieure inversible car $\text{rang}(X_0) = \text{rang}(Q_m^*X_0) = \text{rang}(G_1)$. On peut alors écrire

$$AX_0 = AQ_mG_1 + AWG_2$$

$$A^kX_0 = A^kQ_mG_1 + A^kWG_2$$

et encore

$$Q_m^*A^kQ_m = R_1^k \Longrightarrow A^kQ_m = Q_mR_1^k$$

$$W^*A^kW = R_2^k \Longrightarrow A^kW = WR_2^k$$

soit finalement

$$\begin{aligned} A^k X_0 &= Q_m R_1^k G_1 + W R_2^k G_2 \\ &= [Q_m + W R_2^k G_2 G_1^{-1} R_1^{-k}] R_1^k G_1 \\ &= [Q_m + E_k] R_1^k G_1 \end{aligned}$$

Par construction, le rayon spectral de la matrice R_1^{-1} est inférieur ou égal à $|\lambda_m|^{-1}$ tandis que celui de la matrice R_2 est égal à $|\lambda_{m+1}| < |\lambda_m|$. La matrice E_k tend donc vers $[0]$ quand k tend vers $+\infty$.

D'autre part, par construction

$$\begin{aligned} X_k = Q_k &= A X_{k-1} R_k^{-1} \\ &= A^k X_0 R_1^{-1} \dots R_k^{-1} \end{aligned}$$

Finalement quand k tend vers $+\infty$

$$X_k \approx Q_m R_1^k G_1 R_1^{-1} \dots R_k^{-1} \approx Q_m \mathcal{R}$$

avec $\mathcal{R}$ matrice triangulaire supérieure. X_k tend donc vers la matrice Q_m des m premiers vecteurs de Schur et les valeurs propres cherchées sont sur la diagonale de la matrice $\mathcal{R}$.

Remarque 11.8.1 *La factorisation QR d'une matrice n'étant pas unique, il est plus exact de dire que la $j^{ème}$ colonne de X_k converge vers la $j^{ème}$ colonne de Q_m multipliée par un nombre complexe de la forme $e^{i\theta_j}$.*

Pour simplifier cette démonstration, seule l'idée principale a été retenue, le raisonnement rigoureux est détaillé par Saad [23]. ■

11.9 Méthode QR

C'est le cas extrême de la méthode du sous–espace dans laquelle on prend $\mathcal{X}_0 = \mathbb{C}^n$! L'algorithme QR s'écrit

$$\left\| \begin{array}{l} 1) \textbf{ initialisation :} \\ \quad X_0 = I \\ 2) \textbf{ itérations : pour } k = 1, 2, \dots \textbf{ faire} \\ \quad \tilde{X}_k = A X_{k-1} \\ \quad Q_k R_k = \tilde{X}_k \\ \quad X_k = Q_k \\ \quad \textbf{fin} \end{array} \right. \tag{11.2}$$

La méthode consiste donc en une succession de factorisations QR, d'où elle tire son nom. C'est bien un algorithme de type sous–espace dans lequel le sous-espace initial est $\mathbb{C}^n$, c'est-à-dire que la méthode QR calcule toutes les valeurs propres et tous les vecteurs propres de la matrice A. Les hypothèses du Théorème 11.8.1 sont automatiquement satisfaites, ce qui entraine la convergence de la méthode.

Classiquement, on définit la méthode QR par les formules

$$
\begin{array}{l}
\text{1) } \textbf{initialisation :} \\
\quad A_1 = A \\
\text{2) } \textbf{itérations : pour } k = 1, 2, \ldots \textbf{ faire} \\
\quad Q_k R_k = A_k \\
\quad A_{k+1} = R_k Q_k \\
\quad \textbf{fin}
\end{array}
\tag{11.3}
$$

Pour lesquelles on montre les propriétés suivantes :

Proposition 11.9.1 *Les matrices générées par l'algorithme QR vérifient les relations*

$$A_{k+1} = (Q_1 Q_2 \ldots Q_k)^* \; A \; (Q_1 Q_2 \ldots Q_k)$$
$$(Q_1 Q_2 \ldots Q_k)(R_k R_{k-1} \ldots R_1) = A^k$$

Preuve : On commence par poser $\tilde{Q}_k = Q_1 Q_2 \ldots Q_k$ et $\tilde{R}_k = R_k R_{k-1} \ldots R_1$, il suffit alors de vérifier que par construction

$$A_{k+1} = R_k Q_k = [Q]_k^* A_k Q_k = \tilde{Q}_k^* A \tilde{Q}_k.$$

Puis par récurrence on montre la seconde identité

- Pour $k = 1$, $A = Q_1 R_1 = A_1$
- Supposons la propriété vraie jusqu'à l'ordre pour $k-1$ inclus : $\tilde{Q}_{k-1}\tilde{R}_{k-1} = A^{k-1}$

$$\begin{aligned}
\tilde{Q}_k(R_k R_{k-1} \ldots R_1) &= \tilde{Q}_{k-1} Q_k R_k (R_{k-1} \ldots R_1) \\
&= \tilde{Q}_{k-1} A_k \tilde{R}_{k-1} \\
&= \tilde{Q}_{k-1} \tilde{Q}_{k-1}^* A \tilde{Q}_{k-1} \tilde{R}_{k-1} \\
&= A \; A^{k-1} = A^k.
\end{aligned}$$

■

Pour montrer l'équivalence des formulations (11.2) et (11.3), il suffit de définir à partir des matrices Q_k et R_k de l'algorithme (11.3) les matrices $Y_k = \tilde{Q}_k$ et $\tilde{Y}_k = \tilde{Q}_k R_k$; en utilisant alors la Proposition 11.9.1 on obtient

$$\begin{aligned}
\tilde{Y}_k &= (Q_1 Q_2 \ldots Q_{k-1})(Q_k R_k) \\
&= \tilde{Q}_{k-1} A_k = A \tilde{Q}_{k-1} = A Y_{k-1}
\end{aligned}$$

La suite de matrices $\{Y_k\}_{k \in \mathbb{N}}$ vérifie donc (en posant $Y_0 = I$)

$$\begin{aligned}
Y_0 &= I \\
\tilde{Y}_1 &= \tilde{Q}_1 R_1 = Q_1 R_1 = A = A Y_0 \\
\tilde{Y}_k &= \tilde{Q}_k R_k = A Y_{k-1} \\
Y_k &= \tilde{Q}_k
\end{aligned}$$

Ainsi $X_k = Y_k$ pour tout k, et la matrice Q_k de l'algorithme (11.2) est égale à la matrice $\tilde{Q}_k$ de l'algorithme (11.3).

Lorsque la méthode a convergé, on pose $Q = \lim_{k\to\infty} Q_k$ et $R = \lim_{k\to\infty} R_k$.

11.10 Méthode QR avec translation

Comme chaque itération de cette méthode est coûteuse, il faut chercher à accélérer sa convergence. Pour cela, il est facile d'adapter la technique de translation à cet algorithme en remarquant que si on effectue une translation sur la matrice A_k, suivie d'une factorisation QR

$$A_k - \sigma_k I = Q_k R_k,$$

alors

$$\begin{aligned} A_{k+1} &= R_k Q_k + \sigma_k I \\ &= Q_k^* Q_k R_k Q_k + \sigma_k I \\ &= Q_k^* [A_k - \sigma_k I] Q_k + \sigma_k I \\ &= Q_k^* A_k Q_k \end{aligned}$$

les matrices A_k et A_{k+1} sont donc semblables pour tout σ_k, ce qui laisse la possibilité de changer σ_k à chaque itération. Pour le choix du paramètre σ_k, on peut utiliser

(i) la translation de Rayleigh : $\sigma_k = [A_k]_{n,n}$ à partir d'un certain rang k, qui permet une convergence cubique dans le cas hermitien.

(ii) la translation de Wilkinson : σ_k est pris comme la valeur propre de la matrice

$$\begin{bmatrix} [A_k]_{n-1,n-1} & [A_k]_{n,n-1} \\ [A_k]_{n-1,n} & [A_k]_{n,n} \end{bmatrix}$$

la plus proche du coefficient $[A_k]_{n,n}$. Cette technique est efficace pour le traitement des valeurs propres doubles (voir Saad [23]).

Ce qu'il faut retenir

1. la méthode de la puissance itérée permet de calculer la valeur propre de plus grand module d'une matrice, et un vecteur propre associé.

2. pour calculer la valeur propre de plus petit module, on utilise la méthode de la puissance itérée inverse

3. pour calculer les valeurs propres intermédiaires, on a recours à différentes techniques comme la translation ou la déflation.

4. la méthode du sous-espace permet de calculer plusieurs valeurs propres et vecteurs propres à la fois.

Chapitre 12

Matrices tridiagonales

12.1 Introduction

Les algorithmes de calcul des vecteurs propres présentés dans le chapitre précédent constituent un premier ensemble de méthodes utilisables pour toutes les matrices. Dans ce chapitre on introduit un algorithme très puissant pour le calcul des valeurs propres des matrices symétriques ; cet algorithme ne s'applique qu'aux matrices symétriques tridiagonales, on verra au chapitre 13 que dans le cas d'une matrice A symétrique, l'algorithme de Lanczos permet de calculer une matrice tridiagonale dont les valeurs propres sont proches de celles de A.

Pour être complet, deux paragraphes en fin de chapitre sont consacrés à la méthode de Householder, qui permet de mettre toute matrice symétrique sous forme tridiagonale.

12.2 Méthode de la bissection (théorie)

Puisque l'on sait contruire une matrice tridiagonale semblable à une matrice symétrique (hermitienne) donnée, on cherche maintenant un algorithme rapide de calcul des valeurs propres d'une telle matrice. Pour cela, on considère la matrice

$$T_n = \begin{bmatrix} a_1 & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 & b_2 & 0 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & b_{n-1} \\ 0 & 0 & 0 & 0 & b_{n-1} & a_n \end{bmatrix}$$

à coefficients réels, et on suppose que $b_k \neq 0$ pour tout k. La matrice T est alors dite **irréductible**, c'est-à-dire qu'il n'existe pas de matrice de permutation $P \in \mathbb{R}^{n\times n}$ telle que

$$T_n = P^T \begin{bmatrix} T' & 0 \\ 0 & T'' \end{bmatrix} P.$$

S'il existait un b_k nul, la matrice T_n serait **réductible**, et le calcul des valeurs propres d'une telle matrice se ramène au calcul des valeurs propres des matrices tridiagonales T' et T''. Pour tout $k \leq n$, on considère la matrice

$$T_k = \begin{bmatrix} a_1 & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 & b_2 & 0 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & b_{k-1} \\ 0 & 0 & 0 & 0 & b_{k-1} & a_k \end{bmatrix}$$

et on introduit $p_k(\lambda) = \text{dét } (T_k - \lambda I_k)$, le polynôme caractéristique de la matrice T_k d'ordre k. Les valeurs propres cherchées étant les racines de p_n, on étudie les propriétés des polynômes p_k, pour $k = 1, 2, \ldots, n$.

Proposition 12.2.1 *Les polynômes p_k vérifient les relations suivantes :*

$$(i) \quad p_0(\lambda) = 1, \quad p_1(\lambda) = a_1 - \lambda$$

$$(ii) \quad p_k(\lambda) = (a_k - \lambda)p_{k-1}(\lambda) - b_{k-1}^2 p_{k-2}(\lambda) \quad \forall k \geq 2$$

Preuve : La démontration est effectuée par récurrence en développant le déterminant de la matrice $T_k - \lambda I_k$ par rapport à la ligne k

$$T_k - \lambda I_k = \begin{bmatrix} a_1 - \lambda & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 - \lambda & b_2 & 0 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & a_{k-2} - \lambda & b_{k-2} & 0 \\ 0 & 0 & 0 & b_{k-2} & a_{k-1} - \lambda & b_{k-1} \\ 0 & 0 & 0 & 0 & b_{k-1} & a_k - \lambda \end{bmatrix}$$

d'où dét $(T_k - \lambda I_k) = (a_k - \lambda)$dét $(T_{k-1} - \lambda I_{k-1}) - b_{k-1}$dét B_{k-1}, avec

$$B_{k-1} = \begin{bmatrix} a_1 - \lambda & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 - \lambda & b_2 & 0 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & a_{k-3} - \lambda & b_{k-3} & 0 \\ 0 & 0 & 0 & b_{k-3} & a_{k-2} - \lambda & 0 \\ 0 & 0 & 0 & 0 & b_{k-2} & b_{k-1} \end{bmatrix}$$

soit dét $B_{k-1} = b_{k-1}$dét $(T_{k-2} - \lambda I_{k-2})$. ■

Proposition 12.2.2 *Les racines du polynôme p_k sont simples et réelles.*

Preuve : Puisque la matrice $T_k - \lambda I_k$ est symétrique, le polynôme p_k admet k racines réelles. Par ailleurs, en reprenant la définition de $T_k - \lambda I_K$:

$$T_k - \lambda I_k = \begin{bmatrix} a_1 - \lambda & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 - \lambda & b_2 & 0 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & a_{k-2} - \lambda & b_{k-2} & 0 \\ 0 & 0 & 0 & b_{k-2} & a_{k-1} - \lambda & b_{k-1} \\ 0 & 0 & 0 & 0 & b_{k-1} & a_k - \lambda \end{bmatrix}$$

la matrice extraite

$$S_{k-1} = \begin{bmatrix} b_1 & a_2 - \lambda & b_2 & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & a_{k-2} - \lambda & b_{k-2} \\ 0 & 0 & 0 & b_{k-2} & a_{k-1} - \lambda \\ 0 & 0 & 0 & 0 & b_{k-1} \end{bmatrix}$$

a pour déterminant dét $S_{k-1} = \prod_{l=1,k-1} b_l \neq 0$ quel que soit λ, donc le rang de $T_k - \lambda I_k$ est au moins $k - 1$, chacune des racines de p_k est simple. On en déduit que p_k admet k racines réelles distinctes. ■

Proposition 12.2.3 *Les polynômes p_k et p_{k-1} n'ont pas de racine commune.*

Preuve : En effet supposons qu'il existe α tel que $p_k(\alpha) = p_{k-1}(\alpha) = 0$, d'après la formule de récurrence,

$$b_{k-1}^2 p_{k-2}(\alpha) = p_k(\alpha) - (a_k - \alpha)p_{k-1}(\alpha) = 0$$

soit $p_{k-2}(\alpha) = 0$ puisque par hypothèse $b_{k-1} \neq 0$. De proche en proche on montre que

$$b_1^2 p_0(\alpha) = p_2(\alpha) - (a_1 - \alpha)p_1(\alpha) = 0$$

soit $b_1 = 0$, ce qui est contraire à l'hypothèse. ■

Proposition 12.2.4 *Les racines du polynôme p_{k-1} séparent strictement les racines du polynôme p_k.*

Preuve : On montre le résultat par récurrence :

• Pour $k = 1$

$$p_1(\lambda) = a_1 - \lambda$$

$$p_2(\lambda) = (a_1 - \lambda)(a_2 - \lambda) - b_1^2 \Longrightarrow p_2(a_1) = -b_1^2 < 0$$

d'autre part $\lim_{\lambda \to \pm\infty} p_2(\lambda) > 0$, donc a_1 est strictement comprise entre les racines de p_2.

• Supposons le résultat vrai jusqu'à l'ordre k inclus. Les racines de p_k : $\alpha_1, \alpha_2, \ldots, \alpha_k$ sont simples et réelles, ainsi que les racines de p_{k-1} : $\beta_1, \beta_2, \ldots, \beta_k$; par hypothèse de récurrence on a

$$\alpha_1 < \beta_1 < \alpha_2 < \beta_2 < \ldots < \alpha_{k-1} < \beta_{k-1} < \alpha_k$$

donc pour toute racine α_i,

$$p_{k+1}(\alpha_i) = (a_{k+1} - \alpha_i)p_k(\alpha_i) - b_k^2 p_{k-1}(\alpha_i) = -b_k^2 p_{k-1}(\alpha_i)$$

ainsi $p_{k+1}(\alpha_i)p_{k-1}(\alpha_i) < 0$, et de même $p_{k+1}(\alpha_{i+1})p_{k-1}(\alpha_{i+1}) < 0$.

Mais par ailleurs $p_{k-1}(\alpha_i)p_{k-1}(\alpha_{i+1}) < 0$ puisque l'intervalle $]\alpha_i, \alpha_{i+1}[$ contient β_i racine simple de p_{k-1}, on en déduit que $p_{k+1}(\alpha_i)p_{k+1}(\alpha_{i+1}) < 0$, soit qu'il existe une racine γ_{i+1} de p_{k+1} strictement comprise entre α_i et α_{i+1}.

Il y a $k-1$ intervalles $]\alpha_i, \alpha_{i+1}[$, on a donc localisé $k-1$ racines de p_{k+1} et on recherche maintenant des racines à l'extérieur de $[\alpha_1, \alpha_k]$.

D'une manière générale, le terme dominant de p_l est $(-\lambda)^l$, la plus petite racine de p_{k-1} étant $\beta_1 > \alpha_1$, le signe de $p_{k-1}(\alpha_1)$ est positif; mais par ailleurs $p_{k+1}(\alpha_1) = -b_k^2 p_{k-1}(\alpha_1)$ donc le signe de $p_{k+1}(\alpha_1)$ est négatif ! Il est donc nécessaire que p_{k+1} s'annule sur $]-\infty, \alpha_1[$ pour que $\lim_{\lambda \to -\infty} p_{k+1}(\lambda) = +\infty$.

De même à l'autre extrêmité de l'intervalle, le signe de $p_{k-1}(\alpha_k)$ est celui de $(-1)^{k-1}$, donc celui de $p_{k+1}(\alpha_k)$ est $(-1)^k$ puisque $p_{k+1}(\alpha_k) = -b_k^2 p_{k-1}(\alpha_k)$, on en déduit que p_{k+1} s'annule nécessairement sur $]\alpha_k, +\infty[$ pour que le signe de $\lim_{\lambda \to -\infty} p_{k+1}(\lambda)$ soit $(-1)^{k+1}$. ■

On définit maintenant la notion de concordance de signes d'une suite $\{s_k\}_{k \in \mathbb{N}}$ par récurrence : pour $k = 0$ on pose $C_0 = 0$, et pour $k > 0$

$$C_{k+1} = C_k + \begin{cases} 0 & \text{si } s_{k+1}s_k < 0 \\ 1 & \text{sinon} \end{cases}$$

On utilise de plus la convention suivante : si $s_k = 0$, on définit son signe comme l'opposé de celui de s_{k-1} ; on démontre alors le

Théorème 12.2.1 *Soit A une matrice réelle symétrique tridiagonale et irréductible, et soit r un nombre réel quelconque : le nombre $C(r)$ de concordance de signes de la suite $p_0(r)$, $p_1(r)$, ..., $p_n(r)$ est égal au nombre $\Lambda_>(r)$ de valeurs propres de A strictement supérieures à r.*

Preuve : Dans la démonstration qui suit $\ominus$ et $\oplus$ signalent l'application de la convention de signe. On montre le résultat par récurrence :

• Pour $k = 1$ $p_0(r) = 1$ et $p_1(r) = a_1 - r$ (donc $\alpha_1 = a_1$) .

r	$-\infty$	α_1	$+\infty$
signe de $p_0(r)$	$+$	$+$	$+$
signe de $p_1(r)$	$+$	$\ominus$	-
$C(r)$	1	0	0
$\Lambda_>(r)$	1	0	0

• Supposons maintenant le résultat vrai jusqu'à l'ordre k inclus, on reprend les notations précédentes :

$$\gamma_1 < \alpha_1 < \beta_1 < \alpha_2 < \beta_2 < \ldots < \alpha_{k-1} < \beta_{k-1} < \alpha_k < \gamma_{k+1}$$

et on refait le tableau en appliquant la règle suivante

$$C_{k+1}(r) = C_k(r) + \begin{cases} 0 & \text{si } p_{k+1}(r)p_k(r) < 0 \\ 1 & \text{sinon} \end{cases}$$

(le tableau complet, qui ne tient pas en largeur, a été découpé en deux parties)

r	$-\infty$	γ_1	α_1	γ_2	α_2	$\ldots$	γ_j	α_j
signe de $p_k(r)$	$+$	$+$	$\ominus$	-	$\oplus$	$\ldots$	$(-1)^{j-1}$	$(\ominus 1)^j$
signe de $p_{k+1}(r)$	$+$	$\ominus$	-	$\oplus$	$+$	$\ldots$	$(\ominus 1)^j$	$(-1)^j$
$C_k(r)$	k	k	$k-1$	$k-1$	$k-2$	$\ldots$	$k-j+1$	$k-j$
$C_{k+1}(r)$	$k+1$	k	k	$k-1$	$k-1$	$\ldots$	$k-j+1$	$k-j+1$
$\Lambda_>(r)$	$k+1$	k	k	$k-1$	$k-1$	$\ldots$	$k-j+1$	$k-j+1$

r	α_j	$\ldots$	γ_k	α_k	γ_{k+1}	$+\infty$
signe de $p_k(r)$	$(\ominus 1)^j$	$\ldots$	$(-1)^{k-1}$	$(\ominus 1)^k$	$(-1)^k$	$(-1)^k$
signe de $p_{k+1}(r)$	$(-1)^j$	$\ldots$	$(\ominus 1)^k$	$(-1)^k$	$(\ominus 1)^{k+1}$	$(-1)^{k+1}$
$C_k(r)$	$k-j$	$\ldots$	1	0	0	0
$C_{k+1}(r)$	$k-j+1$	$\ldots$	1	1	0	0
$\Lambda_>(r)$	$k-j+1$	$\ldots$	1	1	0	0

On vérifie bien que $C_{k+1}(r)$ est égal au nombre de racines $\gamma_j > r$, le résultat est donc vrai à l'ordre $k+1$ et ainsi il est vrai pour tout n. ■

Théorème 12.2.2 *Soit A une matrice réelle symétrique tridiagonale et irréductible, le nombre de valeurs propres de A contenues dans l'intervalle $]a, b]$ est égal à $C(a) - C(b)$.*

Preuve : Puisque $C(a)$ est le nombre de valeurs propres strictement supérieures à a, et $C(b)$ le nombre de valeurs propres strictement supérieures à b, $C(a) - C(b)$ est donc le nombre de valeurs propres strictement supérieures à a et inférieures à b. ■

12.3 Méthode de la bissection (pratique)

Les suites $\{p_k(r)\}_{k\in\mathbb{N}}$ sont appelées **suites de Sturm**, et la technique qui consiste à isoler chaque valeur propre de la matrice dans un intervalle est appelée méthode de **bissection** .

Cette méthode est employée pour calculer de manière efficace les valeurs propres des matrices symétriques (hermitiennes) de la façon suivante :

- on commence par isoler le spectre de A à l'aide du Théorème de Gerschgorin-Hadamard (Théorème 10.4.1) ; il est contenu dans un intervalle $[a, b]$, avec
$$a = \min_i (A_{i,i} - \sum_{j \neq i} |A_{i,j}|) \quad \text{et} \quad b = \max_i (A_{i,i} + \sum_{j \neq i} |A_{i,j}|).$$
- l'intervalle $[a, b]$ est ensuite découpé en sous-intervalles dans lesquels on cherche à isoler les valeurs propres par applications successives du Théorème 12.2.2 : si on sait que $\lambda \in]a_k, b_k]$, on détermine si $\lambda \in]a_k, (a_k + b_k)/2]$ ou si $\lambda \in](a_k + b_k)/2, b_k]$.
- à ce stade on a obtenu d intervalles $]a_k, b_k]$ contenant chacun une seule valeur propre λ_k. Chaque valeur propre λ_k est donc localisée à la précision $|a_k - b_k|/2$. Si $h = \max |a_k - b_k|$ est assez petit, on peut se contenter de cette première approximation. Sinon on peut envisager de continuer la bissection, dont le taux de convergence est $t = 0.5$. Mais à ce stade, il est souvent plus rapide de procéder autrement : on calcule λ_k comme racine de l'équation non linéaire $p_n(\lambda) = 0$ par une méthode itérative de type Newton–Raphson (taux de convergence $t = 1,618$) ou la méthode dite Regula Falsi ($t = 1,642$) ou encore la méthode Pegasus ($1,642 \leq t \leq 1.695$), en initialisant la méthode choisie à la valeur $\lambda_k^0 = (a_k + b_k)/2 \ldots$

Pour calculer les valeurs propres d'une matrice A symétrique on peut aussi associer plusieurs algorithmes :

- on commence par calculer à l'aide de la méthode de Lanczos (voir chapitre 13) une matrice $H_m \in \mathbb{C}^{m \times m}$ tridiagonale
- on calcule ensuite les valeurs propres $\tilde{\lambda}_j$ de la matrice H_m par la méthode de bissection
- à l'aide de la méthode de la puissance itérée inverse avec les paramètres de translation $\sigma = \tilde{\lambda}_j$ (voir chapitre 11) on calcule très rapidement m valeurs propres de la matrice A et les vecteurs propres correspondants.

Remarque 12.3.1 *cette méthode permet aussi de limiter le calcul des valeurs propres d'une matrice A symétrique au sous–ensemble contenu dans un intervalle donné, sans avoir à calculer tout le spectre.*

12.4 Un calcul explicite

On peut déterminer explicitement les valeurs propres d'une matrice tridiagonale symétrique $A_n \in \mathbb{R}^{n \times n}$ dans le cas où les coefficients sont constants :

$$A_n = \begin{bmatrix} \alpha & \beta & 0 & 0 & 0 & 0 \\ \beta & \alpha & \beta & 0 & 0 & 0 \\ 0 & \beta & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & \beta \\ 0 & 0 & 0 & 0 & \beta & \alpha \end{bmatrix}.$$

On suppose que α et β sont différents de zéro, et on introduit le polynôme caractéristique de A_n : $p_n(\lambda) = \text{dét } (A_n - \lambda I_n)$. On écrit ensuite la formule de récurrence

$$p_n(\lambda) = (\alpha - \lambda) p_{n-1}(\lambda) - \beta^2 p_{n-2}(\lambda).$$

qui incite à chercher le polynôme $p_n(\lambda)$ sous la forme $p_n(\lambda) = aX_1^n(\lambda) + bX_2^n(\lambda)$. En introduisant alors l'angle θ par $2\beta\cos(\theta) = \lambda - \alpha$, on obtient

$$X_1(\lambda) = -\beta e^{-i\theta} \quad \text{et} \quad X_2(\lambda) = -\beta e^{i\theta}.$$

Les coefficients a et b sont déterminés par les deux premiers termes de la suite :

$$p_1(\lambda) = \alpha - \lambda \quad \text{et} \quad p_2(\lambda) = (\alpha - \lambda)^2 - \beta^2.$$

On constate que $p_0(\lambda) = 1$ est un choix cohérent avec la formule de récurrence, qui permet d'en déduire plus rapidement les valeurs

$$a = -\frac{e^{-i\theta}}{2\sin(\theta)} \quad \text{et} \quad b = \frac{e^{i\theta}}{2\sin(\theta)}.$$

Le polynôme caractéristique de la matrice A_n s'écrit donc

$$p_n(\lambda) = (-\beta)^n \frac{\sin[(n+1)\theta]}{\sin(\theta)}.$$

et ses racines sont les

$$\lambda_j = \alpha + 2\beta\cos(\frac{j\pi}{n+1}) \quad \text{pour } j = 1, 2, \ldots, n.$$

Noter que la valeur $\theta = 0$, solution de l'équation $\sin[(n+1)\theta] = 0$ ne convient pas car on montre par récurrence que

$$p_n(\alpha + 2\beta) = (n+1)(-\beta)^n \neq 0.$$

On vérifie alors que pour $k = 1, 2, \ldots, n$, le vecteur

$$u_k = (\, e^{i\theta_k}, e^{i2\theta_k}, \ldots, e^{in\theta_k}\,)^T$$

est le vecteur propre associé à la valeur propre

$$\lambda_k = \alpha + 2\beta\cos\theta_k \quad \text{avec} \quad \theta_k = \frac{k\pi}{n+1}.$$

12.5 Méthode de Householder

Le problème posé dans ce paragraphe consiste, à partir d'une matrice $A \in \mathbb{R}^{n\times n}$ donnée, à calculer une matrice T tridiagonale, semblable à la matrice A. Cette opération s'appelle couramment "mettre A sous forme tridiagonale" ou encore "tridiagonalisation de la matrice A". Elle repose sur la définition d'une transformation orthogonale : la transformation de Householder.

Soit $u \in \mathbb{R}^n$, tel que $\|u\| = 1$, on considère la matrice

$$H = I - 2u \cdot u^T$$

$u \cdot u^T \in \mathbb{R}^{n\times n}$ est une matrice symétrique, et H est une matrice symétrique orthogonale :

$$\begin{aligned} H^T H &= (I - 2u\cdot u^T)(I - 2u\cdot u^T) = I - 4u\cdot u^T + 4(u\cdot u^T)(u\cdot u^T) \\ &= I - 4u\cdot u^T + 4u\cdot(u^T u)\cdot u^T = I. \end{aligned}$$

Proposition 12.5.1 *Soient $a, b \in \mathbb{R}^n$ deux vecteurs non nuls, il existe un vecteur $u \in \mathbb{R}^n$ et un scalaire $\alpha \in \mathbb{R}$, tels que*

$$Ha = [I - 2u\cdot u^T]a = \alpha b.$$

Preuve : Soient $a, b \in \mathbb{R}^n$, deux vecteurs non nuls ; on va les supposer non colinéaires, car si $a = \alpha b$, il suffit de prendre $u = 0$ pour obtenir le résultat cherché.

On note ensuite que la matrice H étant orthogonale, $\|Ha\| = \|a\|$; on peut donc se limiter sans perte de généralité au cas où $\|b\| = 1$ et $\alpha = \|a\|$.

Si un tel vecteur u existe, il vérifie nécessairement

$$Ha = (I - 2u \cdot u^T)a = a - 2u(u^T a) = \alpha b$$

soit encore

$$a - \alpha b = 2u(u^T a) = 2\lambda u,$$

avec $\lambda = u^T a$.

On en déduit la relation $a^T a - \alpha a^T b = 2\lambda a^T u$, que l'on écrit encore $\alpha^2 - \alpha a^T b = 2\lambda^2$.

Par hypothèse les vecteurs a et b ne sont pas colinéaires, donc

$$|a^T b| < \|a\|\|b\| = \alpha,$$

et

$$\alpha^2 - \alpha a^T b > 0.$$

En posant alors

$$\lambda = \pm\sqrt{\frac{\alpha^2 - \alpha a^T b}{2}},$$

on obtient le vecteur cherché

$$u = \frac{1}{2\lambda}(a - \alpha b).$$

■

Noter que la matrice H ne dépend pas du signe de λ !

Dans la pratique on procède de la façon suivante :
- on calcule le vecteur $v = a - \|a\|b = a - \alpha b$,
- puis le scalaire $\beta = a^T v = a^T(a - \alpha b)$,
- alors $H = I - \frac{1}{\beta} v \cdot v^T$.

Le produit d'un vecteur quelconque $w \in \mathbb{R}^n$ par la matrice H est ainsi obtenu suivant le calcul de $c = \frac{1}{\beta} v^T w$ soit $2n$ opérations, puis de $Hw = w - cv$ soit encore $2n$ opérations.

Remarque 12.5.1 *1) la matrice H est associée à une transformation géométrique, qui est la symétrie par rapport à l'hyperplan de $\mathbb{R}^n$ orthogonal au vecteur u. En effet pour tout vecteur $a \in \mathbb{R}^n$, $Ha = (I - 2u \cdot u^T)a = a - 2u(u^T a)$, or a s'écrit de manière unique $a = a_H + a_{H^\perp}$ et de $u^T a = u^T a_{H^\perp}$, on déduit $Ha = a_H - a_{H^\perp}$.*

2) il est possible de généraliser cette transformation pour traiter le cas des matrices A non symétriques. On utilise alors la matrice $H = I - 2u \cdot v^T$, avec $u, v \in \mathbb{R}^n$ tels que $(u, v) = 1$, .

3) on peut aussi travailler dans $\mathbb{C}^{n\times n}$ à l'aide de la matrice hermitienne unitaire $H = I - 2u \cdot u^*$ avec $u \in \mathbb{C}^n$ tel que $(u, u) = 1$.

12.6 Tridiagonalisation d'une matrice

Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique, on cherche une matrice de Householder H telle que

$$B = HAH = \begin{bmatrix} \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{bmatrix}.$$

Il suffit de prendre H de la forme

$$H = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H} \end{bmatrix}$$

avec $\tilde{H} = I_{n-1} - 2\tilde{u}\cdot\tilde{u}^T$, $\tilde{u} \in \mathbb{R}^{n-1}$. En effet

$$B = HAH = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H} \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,.} \\ A_{.,1} & \tilde{A} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H} \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,.}\tilde{H} \\ \tilde{H}A_{.,1} & \tilde{H}\tilde{A}\tilde{H} \end{bmatrix}.$$

$A_{.,1} \in \mathbb{R}^{n-1}$, $\tilde{H}$ est une matrice de Householder de rang $n-1$, en utilisant la Proposition 12.5.1, on peut trouver un vecteur $\tilde{u} \in \mathbb{R}^{n-1}$ tel que $\tilde{H}A_{.,1} = \alpha[1,0,\ldots,0]^T \in \mathbb{R}^{n-1}$.

On a ainsi mis à zéro en une seule étape tous les coefficients de la première colonne de A en dessous de la sous-diagonale ! Ce calcul nécessite de l'ordre de $4n^2$ opérations puisqu'il faut modifier toutes les colonnes de A.

Il est possible d'appliquer le même procédé à la matrice $\tilde{H}\tilde{A}\tilde{H} \in \mathbb{R}^{n-1\times n-1}$ pour faire apparaitre des zéros dans la deuxième colonne, sans modifier ceux de la première, et ainsi de suite... La méthode de tridiagonalisation de Householder consiste donc à calculer $n-2$ matrices H^k telles que

$$H_{n-2}\ldots H_1\ A\ H_1\ldots H_{n-2} = \begin{bmatrix} \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \end{bmatrix} = T$$

T est une matrice tridiagonale semblable à A, dont le coût est de l'ordre de n^3 opérations ; il reste maintenant à calculer les valeurs propres de T.

Remarque 12.6.1 *on peut obtenir le même résultat par la méthode des rotations de Givens (voir [16]).*

12.7 Matrice compagnon

Un autre exemple de lien entre les valeurs propres d'une matrice et les racines d'un polynôme est illustré par la **matrice compagnon**

Proposition 12.7.1 *Les racines du polynôme à coefficients complexes*

$$P_n(z) = (-1)^n(z^n - \alpha_{n-1}z^{n-1} - \ldots - \alpha_1 z - \alpha_0)$$

sont les valeurs propres de la matrice compagnon $C_n \in \mathbb{C}^{n\times n}$ définie par

$$C_n = \begin{pmatrix} 0 & 0 & \dots & 0 & \alpha_0 \\ 1 & 0 & \ddots & \ddots & \alpha_1 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & \alpha_{n-1} \end{pmatrix}.$$

Preuve : On démontre par récurrence que $P_n(z) = \det(C_n - zI_n)$:

- Pour $n = 2$:

$$C_2 = \begin{pmatrix} 0 & \alpha_0 \\ 1 & \alpha_1 \end{pmatrix} \quad \text{et} \quad \text{dét}(C_2 - zI_2) = \begin{vmatrix} -z & \alpha_0 \\ 1 & -z+\alpha_1 \end{vmatrix} = z^2 - \alpha_1 z - \alpha_0.$$

- Supposons la propriété satisfaite jusqu'à l'ordre $n-1$ inclus, alors :

$$\text{dét}(C_n - zI_n) = \begin{vmatrix} -z & 0 & \dots & 0 & \alpha_0 \\ 1 & -z & \ddots & \ddots & \alpha_1 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -z+\alpha_{n-1} \end{vmatrix}.$$

on développe ce déterminant par rapport à la première colonne

$$\text{dét}(C_n - zI_n) = -z \begin{vmatrix} -z & 0 & \dots & 0 & \alpha_1 \\ 1 & -z & \ddots & \ddots & \alpha_2 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -z+\alpha_{n-1} \end{vmatrix} - \begin{vmatrix} 0 & 0 & \dots & 0 & \alpha_0 \\ 1 & -z & \ddots & \ddots & \alpha_2 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -z+\alpha_{n-1} \end{vmatrix}$$

soit encore

$$\begin{aligned} \text{dét}(C_n - zI_n) &= -z\text{dét}(\tilde{C}_{n-1} - zI_{n-1}) - (-1)^{n-1}\alpha_0 \\ &= -z\,(-1)^{n-1}(z^{n-1} - \alpha_{n-1}z^{n-2} - \dots - \alpha_1) - (-1)^{n-1}\alpha_0 \\ &= (-1)^n(z^n - \alpha_{n-1}z^{n-1} - \dots - \alpha_1 z - \alpha_0) \end{aligned}$$

■

Cette propriété peut être utilisée pour calculer les racines du polynôme P_n à l'aide d'un algorithme de calcul des valeurs propres de la matrice compagnon C_n. En effet le calcul des racines d'un polynôme de degré n est numériquement très délicat quand n est grand...

Exercice 12.7.1 *On suppose que les racines du polynôme à coefficients complexes*

$$P_n(z) = (-1)^n(z^n - \alpha_{n-1}z^{n-1} - \dots - \alpha_1 z - \alpha_0)$$

sont distinctes. Montrer (sans utiliser la Proposition 12.7.1) que la matrice compagnon C_n est diagonalisable.

Ce qu'il faut retenir

1. il existe un algorithme efficace pour calculer les valeurs propres d'une matrice tridiagonale symétrique; il utilise une suite de polynômes définis par une récurrence à trois termes.
2. toute matrice symétrique étant semblable à une matrice tridiagonale symétrique, on peut calculer ses valeurs propres à l'aide de l'algorithme précédent.

Chapitre 13

Méthodes de projection

13.1 Introduction

Au lieu de calculer tous les vecteurs propres d'une matrice, successivement ou par groupes, on peut aussi rechercher les composantes des vecteurs propres présentes dans un sous-espace vectoriel connu à l'avance. Cette présélection permet de limiter les calculs aux seuls vecteurs propres (et valeurs propres associées) qui ont une importance pour l'application en vue, et ainsi de réduire le coût calcul. Ce type d'algorithme entre dans la catégorie des méthodes de projection, qui sont très utilisées dans la pratique. On présente dans ce chapitre les algorithmes classiques : méthode du sous-espace avec projection, méthodes de Lanczos et Arnoldi. A la fin du chapitre, un lien entre la méthode de Lanczos et l'algorithme du gradient conjugué est établi, qui permet de relier les méthodes de résolution de systèmes linéaires aux méthodes de calcul des valeurs propres d'une matrice.

13.2 Méthode de projection

Le principe général des méthodes de projection est d'essayer de calculer les vecteurs propres de A appartenant à un sous–espace vectoriel donné $H \subset \mathbb{C}^n$, ou plus exactement la projection dans H des vecteurs propres de A. Une formulation "faible" de la relation

$$Au = \lambda u$$

s'écrit en effet

$$\forall \tilde{v} \in H \qquad (Au - \lambda u, \tilde{v}) = 0.$$

On exprime ainsi que la projection de $Au - \lambda u$ sur le sous–espace H est nulle : plus grande est la dimension de H, meilleure est l'approximation car la partie "manquante" du vecteur $Au - \lambda u$ se trouve dans $H^\perp$

$$H^\perp = \{v \in \mathbb{C}^n,\ \forall w \in H,\ v^* w = 0\}.$$

13.3 Méthode de Rayleigh–Ritz

Soient $A \in \mathbb{C}^{n \times n}$ et H un sous–espace vectoriel de $\mathbb{C}^n$ de dimension $m < n$ admettant une base orthonormée $\{q_1, q_2, \ldots, q_m\}$. On pose

$$Q = [\, q_1 \quad q_2 \quad \ldots \quad q_m \,] \in \mathbb{C}^{n \times m}$$

Par définition la matrice Q vérifie $Q^* Q = I_m$, mais elle n'est pas unitaire car elle n'est pas carrée ! La matrice $P = QQ^* \in \mathbb{C}^{n \times n}$ est la matrice de projection orthogonale de $\mathbb{C}^n$ sur H : on vérifie facilement que Im $P = H$ et Ker $P = H^\perp$.

On cherche donc à résoudre le problème suivant :

$$\begin{cases} \text{Trouver } (\tilde{\lambda}, \tilde{u}) \in \mathbb{C} \times H \text{ tel que} \\ P(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0 \end{cases} \tag{13.1}$$

toute solution $(\tilde{\lambda}, \tilde{u})$ de ce problème est appelée **élément de Ritz**.

L'idée de ce type de formulation est d'obtenir une bonne approximation d'un couple (λ, u), valeur propre–vecteur propre, en faisant tendre progressivement H vers $\mathbb{C}^n$, avec l'avantage de pouvoir calculer tous les éléments de Ritz contenus dans H à un moindre coût puisque $m < n$. On peut ainsi espérer approcher par exemple les m valeurs propres dominantes de A en faisant les calculs dans un sous–espace H de dimension $m \ll n$.

Pour commencer, rappelons le résultat classique suivant

Proposition 13.3.1 *Si $P \in \mathbb{C}^{n\times n}$ est la matrice de projection orthogonale de $\mathbb{C}^n$ sur le sous–espace $H \subset \mathbb{C}^n$, alors*

$$\forall x \in \mathbb{C}^n,\ \forall y \in H \qquad y^*(x - Px) = 0$$

Preuve : En effet en notant $x = x_H + x_{H^\perp}$, on a $Px = x_H$ et donc $x - Px = x_{H^\perp}$, d'où le résultat. ∎

En conséquence, le problème (13.1) est équivalent au problème

$$\begin{cases} \text{Trouver } (\tilde{\lambda}, \tilde{u}) \in \mathbb{C} \times H \text{ tel que} \\ \forall \tilde{v} \in H \quad \tilde{v}^*(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0. \end{cases} \tag{13.2}$$

Cette formulation correspond bien à une méthode de projection orthogonale, puisqu'on cherche $\tilde{u}$ dans H avec un résidu $A\tilde{u} - \tilde{\lambda}\tilde{u}$ dans $H^\perp$. Le problème (13.2) peut s'écrire dans une base orthonormée de H : q_1, q_2, …, q_m. On pose alors $\tilde{u} = Qx$ et $\tilde{v} = Qy$ avec $x, y \in \mathbb{C}^m$, et on obtient

$$\begin{cases} \text{Trouver } (\tilde{\lambda}, x) \in \mathbb{C} \times \mathbb{C}^m \text{ tel que} \\ \forall y \in \mathbb{C}^m \quad (Qy)^*(AQx - \tilde{\lambda}Qx) = 0, \end{cases}$$

que l'on écrit encore

$$\text{Trouver } (\tilde{\lambda}, x) \in \mathbb{C} \times \mathbb{C}^m \text{ tel que } Q^*AQx = \tilde{\lambda}x. \tag{13.3}$$

La matrice $Q^*AQ \in \mathbb{C}^{m\times m}$ est appelée **matrice de Rayleigh** et l'on peut résumer tout cette démarche dans le résultat suivant

Proposition 13.3.2 *$(\tilde{\lambda}, x)$ est un élément propre de la matrice de Rayleigh Q^*AQ, si et seulement si $(\tilde{\lambda}, Qx)$ est un élément de Ritz de A.*

Pratiquement la méthode de Rayleigh–Ritz peut donc se décomposer en quatre étapes :

1) on choisit un sous–espace vectoriel $H \subset \mathbb{C}^n$ de dimension m, muni d'une base orthonormée $\{q_1, q_2, \ldots, q_m\}$. Soit Q la matrice

$$Q = [\, q_1 \quad q_2 \quad \ldots \quad q_m \,] \in \mathbb{C}^{n\times m}$$

2) on calcule la matrice de Rayleigh $Q^*AQ \in \mathbb{C}^{m\times m}$

3) on calcule les éléments propres $(\tilde{\lambda}_i, x_i)$ de la matrice de Rayleigh Q^*AQ

4) pour $i = 1, 2, \ldots, m$, les couples $(\tilde{\lambda}_i, Qx_i) \in \mathbb{C} \times \mathbb{C}^m$ sont des approximations des éléments propres (λ_i, u_i) de A.

Il existe un cas particulier intéressant pour l'application de cette méthode :

Proposition 13.3.3 *Si H est un sous–espace vectoriel de $\mathbb{C}^n$ invariant par A, alors tout élément de Ritz $(\tilde{\lambda}, \tilde{u})$ relatif à H vérifie*

$$A\tilde{u} = \tilde{\lambda}\tilde{u}.$$

Preuve : Par construction $P(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0$, avec $\tilde{u} \in H$; si on suppose le sous–espace H invariant par A alors $AH = H$, et en particulier $A\tilde{u} \in H$. Il en résulte que

$$P(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0 = A\tilde{u} - \tilde{\lambda}\tilde{u}.$$

■

Ainsi lorsque H est invariant par A, tout élément de Ritz $(\tilde{\lambda}, \tilde{u})$ définit une valeur propre et un vecteur propre de la matrice A !

Remarque 13.3.1 *puisque par hypothèse, $\tilde{u} \in H$, le problème (13.1)*

$$P(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0$$

s'écrit encore

$$PAP\tilde{u} = \tilde{\lambda}\tilde{u}.$$

$\tilde{u}$ apparait ainsi comme un vecteur propre de la matrice $A_m = PAP \in \mathbb{C}^{n\times n}$ qui représente la restriction de l'opérateur linéaire associé à la matrice A au sous–espace $H \subset \mathbb{C}^n$.

On remarque encore que pour tout vecteur $\tilde{v} \in H^\perp$, $P\tilde{v} = 0$ soit $A_m\tilde{v} = 0 \cdot \tilde{v}$: tout vecteur de $H^\perp$ est vecteur propre de la matrice A_m pour la valeur propre 0.

13.4 Cas particulier : A est hermitienne

Pour tout vecteur $\tilde{v} \in \mathbb{C}^n$, on a la relation

$$\tilde{v}^*(A_m\tilde{v}) = \tilde{v}^*(PAP\tilde{v}) = (P\tilde{v})^*(AP\tilde{v})$$

la matrice P est hermitienne par construction car $P^* = (Q^*Q)^* = Q^*Q = P$, ainsi quand $\tilde{v} \in H$, $P\tilde{v} = \tilde{v}$ et $\tilde{v}^*A_m\tilde{v} = \tilde{v}^*A\tilde{v}$; on en déduit le résultat suivant

Proposition 13.4.1 *Si A est hermitienne*

$$\tilde{\lambda}_i = \max_{\substack{S\subset H \\ dimS=i}} \min_{x\in S-\{0\}} \frac{x^*A\ x}{x^*\ x} \quad et \quad \tilde{\lambda}_i \leq \lambda_i.$$

Preuve : On applique le Théorème 10.3.1 (Courant–Fisher) à la matrice $A_m = PAP$ qui est aussi hermitienne :

$$\tilde{\lambda}_i = \max_{\substack{S\subset H \\ dimS=i}} \min_{x\in S-\{0\}} \frac{x^*A_m\ \mathrm{x}}{x^*\ \mathrm{x}}$$

et on remarque que $x^* A_m x = x^* A x$ puisque $x \in S \subset H$. D'autre part, toujours d'après ce théorème, la "vraie" valeur propre λ_i vérifie

$$\lambda_i = \max_{\substack{S \\ dim S = i}} \min_{x \in S - \{0\}} \frac{x^* A \text{ x}}{x^* \text{ x}}$$

on en déduit que $\tilde{\lambda}_i \leq \lambda_i$, puisqu'on prend le maximum sur un choix plus important de sous–espaces : les méthodes de projection donnent toujours des estimations des valeurs propres par valeurs inférieures. ■

Dans la pratique deux types de méthodes utilisent ces résultats pour calculer simultanément plusieurs éléments propres d'une matrice A :

1) la méthode du sous–espace avec projection, dans laquelle H est choisi a priori.

2) la méthode de Lanczos–Arnoldi dans laquelle H est un espace de Krylov

$$H = K_{m-1}(A, x) =< x, Ax, \ldots, A^{m-1}x >$$

dont la dimension m varie au cours des itérations.

13.5 Méthode du sous–espace avec projection

Pour améliorer la convergence de la méthode du sous–espace (voir le Théorème 11.8.1) on remplace l'étape $X_k = Q_k$ en calculant les vecteurs de Schur d'une matrice de rang m par la méthode de Rayleigh–Ritz. En effet on a vu que la méthode de Rayleigh–Ritz a pour effet d'accélérer la convergence des algorithmes de calcul des vecteurs propres ; cette étape supplémentaire permet d'améliorer la précision du calcul des vecteurs de Schur, et en conséquence d'accélérer la convergence de la méthode du sous–espace.

1) **initialisation :**

choix d'un sous–espace $\mathcal{X}_0 \subset \mathbb{C}^n$ de dimension m

choix du paramètre *iter*

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

a) $\tilde{X}_k = A^{iter} X_{k-1}$

b) $Q_k R_k = \tilde{X}_k$

c) calcul de $B_k = Q_k^* A Q_k$

d) calcul des vecteurs de Schur $Y_k^* B_k Y_k = R'$

e) $X_k = Q_k Y_k$

f) modification éventuelle de *iter*

fin

On notera encore une fois que si A est hermitienne, la matrice de Rayleigh $Q^* A Q$ l'est aussi, et ainsi les vecteurs colonnes de Y_k sont les vecteurs propres de B_k

13.6 Méthode d'Arnoldi

La méthode d'Arnoldi fait partie des méthodes de Krylov, qui constituent une famille de méthodes du type sous–espace pour lesquelles le sous–espace H a une forme particulière :

$$H = K_{m-1}(A, v) =< v, Av, A^2 v, \ldots, A^{m-1} v >$$

$K_{m-1}(A, v)$ est le **sous–espace de Krylov** de dimension m, associé à la matrice A et au vecteur v.

La méthode d'Arnoldi consiste à construire le sous–espace $H = K_{m-1}(A, v)$ de la manière suivante

1) **initialisation :**

soit $v_1 \in \mathbb{C}^n$ tel que $\|v_1\|_2 = 1$

2) **itérations : pour** $j = 1, 2, \ldots m$ **faire**

$H_{i,j} = (Av_j, v_i)$ pour $i = 1, 2, \ldots j$

$$w_j = Av_j - \sum_{i=1}^{j} H_{i,j} v_i$$

$H_{j+1,j} = \|w_j\|_2$

si $w_j \neq 0, \quad v_{j+1} = w_j / H_{j+1,j}$

sinon arrêt des calculs

fin

On utilise en fait dans la pratique une méthode équivalente à la précédente, mais numériquement plus stable :

1) **initialisation :**

$v_1 \in \mathbb{C}^n \ \|v_1\|_2 = 1$

2) **itérations : pour** $j = 1, 2, \ldots m$ **faire**

$\tilde{v}_{j+1} = Av_j$

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

$H_{i,j} = (\tilde{v}_{j+1}, v_i)$

$\tilde{v}_{j+1} = \tilde{v}_{j+1} - H_{i,j} v_i$

$H_{j+1,j} = \|\tilde{v}_{j+1}\|_2$

si $v_{j+1} \neq 0, \quad v_{j+1} = \tilde{v}_{j+1} / H_{j+1,j}$

sinon arrêt des calculs

fin

Le cas $H_{j+1,j} = 0$ sera traité plus loin, examinons d'abord le cas général

Proposition 13.6.1 *Si on suppose* $v_j \neq 0$ *pour* $j = 1, 2, \ldots, m$ *les vecteurs* $v_1, v_2, \ldots, v_m$ *forment une base orthonormée du sous–espace*

$$K_m(A, v_1) = < v_1, Av_1, A^2 v_1, \ldots, A^{m-1} v_1 >$$

Preuve : On commence par vérifier que par construction les v_j sont orthonormés :

- Pour $k = 2$ $\|w_1\|_2 v_2 = Av_1 - (Av_1, v_1)v_1$, soit $(v_2, v_1) = 0$ et $(v_2, v_2) = 1$ par définition.
- Supposons la propriété vraie jusqu'à l'ordre pour $k - 1$ inclus, alors

$$w_{k-1} = Av_{k-1} - \sum_{i=1}^{k-1} (Av_{k-1}, v_i) v_i$$

et

$$v_k = w_{k-1}/\|w_{k-1}\|_2.$$

Donc $(v_k, v_l) = 0$ si $l < k-1$ par construction, et

$$(v_k, v_{k-1})\|w_{k-1}\|_2 = (Av_{k-1}, v_{k-1}) - \sum_{i=1}^{k-1}(Av_{k-1}, v_i)v_i = 0.$$

Montrons maintenant que $< v_1, v_2, \ldots, v_m >= K_m(A, v_1)$:
par construction $< v_1, v_2, \ldots, v_m > \subset K_m(A, v_1)$ et dim $K_m(A, v_1) \leq m$,
mais les vecteurs v_j sont orthonormés donc dim $< v_1, v_2, \ldots, v_m >= m$.
Finalement $< v_1, v_2, \ldots, v_m >= K_m(A, v_1)$ et dim $K_m(A, v_1) = m$. ■

On pose $V_m = [\, v_1 \quad v_2 \quad \ldots \quad v_m\,]$, $V_m \in \mathbb{C}^{n\times m}$ et $H_m = [H_{i,j}]$. $H_m \in \mathbb{C}^{m\times m}$ est une matrice de **Hessenberg** supérieure :

$$H_m = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \end{bmatrix}.$$

En effet les coefficients $H_{i,j}$ pour $i > j+1$ ne sont pas définis dans l'algorithme, mais d'après ce qui précéde

$$H_{i,j} = (Av_j, v_i) = (w_j + \sum_{l=1}^{j} H_{l,i}v_l, v_i) = 0 \text{ pour } i > j+1.$$

Proposition 13.6.2 *Si on suppose $v_j \neq 0$ pour $j = 1, 2, \ldots, m$, alors*

$$AV_m = V_m H_m + w_m \cdot e_m^* \quad \text{et} \quad V_m^* A V_m = H_m.$$

Preuve : Par construction des vecteurs v_j :

$$Av_j = \sum_{i=1}^{j} H_{i,j}v_i + w_j = \sum_{i=1}^{j} H_{i,j}v_i + H_{j+1,j}v_{j+1}.$$

En regroupant toutes les égalités jusqu'à $j \leq m-1$ on trouve bien une relation de la forme $AV_m = V_m H_m$. Pour le dernier terme il suffit de remarquer que $H_{m+1,m}v_{m+1} = w_m$ et que $w_m \cdot e_m^*$ est une matrice à n lignes dont les $n-1$ premières colonnes sont nulles et dont la dernière colonne est w_m.

Enfin puisque les vecteurs v_j sont orthonormés :

$$V_m^* V_m = I_m \quad \text{et} \quad V_m^* w_m = 0$$

Avec cette écriture on remarque que $H_m = V_m^* A V_m$ joue le rôle d'une matrice de Rayleigh associée au sous–espace $K_{m-1}(A, v_1)$. ■

Proposition 13.6.3 *Soit $(\tilde{\lambda}, \tilde{u})$ un élément propre de H_m, alors*

$$(A - \tilde{\lambda} I_n)V_m\tilde{u} = w_m \cdot e_m^*\tilde{u}.$$

De l'égalité

$$AV_m\tilde{u} = V_m H_m\tilde{u} + w_m \cdot e_m^*\tilde{u}$$

on tire

$$(A - \tilde{\lambda}I_n)V_m\tilde{u} = w_m \cdot e_m^*\tilde{u}.$$

On en déduit en particulier que

$$\|(A - \tilde{\lambda}I_n)V_m\tilde{u}\|_2 \leq H_{m+1,m}|e_m^*\tilde{u}|$$

où $e_m^*\tilde{u}$ est la dernière composante du vecteur $\tilde{u}$. On peut ainsi mesurer très exactement la convergence de la méthode lorsque l'on calcule des approximations des éléments propres de A à partir de ceux de H_m. Encore une fois il est hors de question de calculer toutes les valeurs propres de A par cette méthode, dans la pratique on l'utilise de la manière suivante

1) **initialisation :**
   choix de $v_1 \in \mathbb{C}^n$ $\|v_1\|_2 = 1$
   choix de $m > 1$
2) **itérations : pour** $j = 1, 2, \ldots m$ **faire**
   $\tilde{v}_{j+1} = Av_j$
   pour $i = 1, \ldots, j$ **faire**
      $H_{i,j} = (\tilde{v}_{j+1}, v_i)$
      $\tilde{v}_{j+1} = \tilde{v}_{j+1} - H_{i,j}v_i$
   $H_{j+1,j} = \|\tilde{v}_{j+1}\|_2$
   si $v_{j+1} \neq 0, \quad v_{j+1} = \tilde{v}_{j+1}/H_{j+1,j}$
   sinon arrêt des calculs
3) **projection :** $(\tilde{\lambda}, \tilde{u})$
   calcul de $\tilde{\lambda}$ plus grande valeur propre de H_m
4) **si** le critère d'arrêt est satisfait : Stop
   sinon $v_1 = V_m\tilde{u}$ aller en 1
   fin

Proposition 13.6.4 *S'il existe un indice $j < m$ tel que $w_j = 0$ alors tous les éléments propres de H_j sont éléments propres de A.*

Preuve : Par construction $v_{j+1} \in K_{j+1}(A, v_1) =< v_1, Av_1, \ldots, A^j v_1 >$. Si $w_j = 0$ il existe un polynôme de degré j tel que $p_j(A)v_1 = 0$. Dans ces conditions le sous-espace de Krylov $K_j(A, v_1)$ est invariant par A et d'après la Proposition 13.3.3 tous les éléments propres de H_j sont éléments propres exacts de A. ∎

Si on veut calculer d'autres valeurs propres, il faut donc choisir un nouveau vecteur initial v_1 ! Le plus petit degré de polynôme l tel que $p_l(A)v_1 = 0$ est appelé **degré à droite** du vecteur v_1 pour la matrice A. Par analogie, on appelle **degré à gauche** du vecteur v_1 pour la matrice A, le plus petit degré de polynôme l tel que $p_l(A^*)v_1 = 0$. Dans la construction du sous–espace

de Krylov le degré à droite du vecteur v_1 pour la matrice A représente la dimension maximale de $K_m(A, v_1)$.

Par exemple dans le cas particulier où l'on a choisi pour vecteur v_1 un vecteur propre de A, alors $Av_1 = \lambda v_1$: le degré à droite du vecteur propre v_1 est 1, et en conséquence dim $K_m(A, v_1) = 1$ quel que soit m ! Le choix d'un vecteur propre comme vecteur de départ ne semble donc pas idéal du point de vue de la méthode d'Arnoldi, mais en fait comme v_1 est vecteur propre, on a atteint le but fixé en une seule itération !

13.7 Méthode de Lanczos

Lorsque la matrice $A \in \mathbb{C}^{n\times n}$ est hermitienne, les formules de la méthode d'Arnoldi se simplifient par symétrie car

$$H_{i,j} = (Av_j, v_i) = (v_j, A^* v_i) = (v_j, Av_i) = \overline{H}_{j,i}.$$

et comme $H_{i,j} = 0$ pour $i > j + 1$ on a donc $H_{j,i} = 0$ pour $j > i + 1$: la matrice H_m est donc une matrice tridiagonale hermitienne.

En tenant compte de cette propriété, le calcul des vecteurs v_j peut être simplifié, et on obtient la **méthode de Lanczos** :

1) **initialisation :**

$v_1 \in \mathbb{C}^n,\ \|v_1\|_2 = 1$ et $v_0 = 0$

2) **itérations : pour** $i = 1, 2, \ldots m$ **faire**

$H_{i,i} = (Av_i, v_i)$

$w_i = Av_i - H_{i,i}v_i - H_{i,i-1}v_{i-1}$

$H_{i+1,i} = \|w_i\|_2$

si $w_i \neq 0, \quad v_{i+1} = w_i/H_{i+1,i}$

sinon Arrêt des calculs

fin

Puisque la matrice H_m est tridiagonale hermitienne, on peut utiliser la méthode de bissection pour le calcul de ses valeurs propres, ce qui donne tout son intérêt à cette approche.

13.8 Lien avec la méthode du gradient conjugué

A partir de la définition des vecteurs r^k et d^k dans l'algorithme du gradient conjugué, utilisé pour résoudre le système linéaire $Ax = b$, lorsque la matrice A est symétrique définie positive, on peut èrire pour $k = 1, \ldots, n - 1$

$$r^k = d^k - \beta^k d^{k-1}$$
$$Ad^k = (r^k - r^{k+1})/\alpha^k$$

on en déduit la relation

$$Ar^k = -\frac{\beta^k}{\alpha^{k-1}} r^{k-1} + (\frac{1}{\alpha^k} + \frac{\beta^k}{\alpha^{k-1}}) r^k - \frac{1}{\alpha^k} r^{k+1}$$

En notant que $\beta^{k+1} = (r^{k+1}, r^{k+1})/r^k, r^k) > 0$, on introduit les matrices

$$R = \left[\frac{r^0}{\|r^0\|}, \frac{r^1}{\|r^1\|}, \ldots, \frac{r^{n-1}}{\|r^{n-1}\|} \right]$$

$$L = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ -\sqrt{\beta^1} & 1 & \ddots & \vdots & \vdots \\ \vdots & -\sqrt{\beta^2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & -\sqrt{\beta^{n-1}} & 1 \end{bmatrix}$$

puis

$$D = \begin{bmatrix} \frac{1}{\alpha^1} & 0 & \ldots & 0 \\ 0 & \frac{1}{\alpha^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & \frac{1}{\alpha^{n-1}} \end{bmatrix}.$$

La relation liant les résidus est alors résumée dans l'écriture matricielle

$$AR = R\left[LDL^T\right]$$

dans laquelle la matrice R est orthogonale $RR^T = I_n$ (voir la Partie 1) et la matrice LDL^T est tridiagonale par construction. Finalement on peut donc écrire

$$R^T AR = LDL^T$$

et l'on retrouve une relation semblable à celle de la méthode de Lanczos, dans laquelle $V_n = R$ et $H_n = LDL^T$. Le sous–espace de Krylov associé est dans ce cas

$$K_{n-1}(A, r^0) =< r^0, Ar^0, A^2r^0, \ldots, A^{n-1}r^0 >$$

et c'est le même sous–espace que celui engendré par les directions d^0, d^1, … d^{n-1} (voir la Partie 1).

Ce qu'il faut retenir

1. les méthodes de projection sont de algorithmes de calcul des valeurs propres et vecteurs propres d'une matrice A d'ordre n dans un sous-espace K de dimension $k \ll n$ à partir d'une matrice H d'ordre k représentant l'opérateur linéaire associé à la matrice A dans K.

2. si la matrice A est symétrique, la méthode de Lanczos permet de construire une matrice H tridiagonale symétrique.

3. si la matrice A n'est pas symétrique, la méthode d'Arnoldi permet de construire une matrice H Hessenberg supérieure.

Annexe

Chapitre 14

Quelques rappels de calcul différentiel

Dans ce chapitre, nous rappelons les fondements du calcul différentiel, en adoptant une approche relativement abstraite, qui ne repose que marginalement sur la notion de dérivée partielle.

14.1 Différentiabilité

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels normés sur $\mathbb{R}$, on note $\mathcal{L}_c(\mathbb{E},\mathbb{F})$ l'ensemble des applications linéaires et continues de $\mathbb{E}$ dans $\mathbb{F}$.
♠ Lorsque la dimension de $\mathbb{E}$ est *finie*, toutes les applications linéaires sont continues. C'est faux lorsque la dimension de $\mathbb{E}$ *infinie*!
Dans la suite, on notera Ω un ouvert de $\mathbb{E}$ contenant u, et f une application de $\Omega \subset \mathbb{E}$ dans $\mathbb{F}$; on dit que f est **continue** en un point $u \in \Omega$ si

$$\forall h \in \mathbb{E} \qquad f(u+h) = f(u) + \varepsilon_0(h), \tag{14.1}$$

où ε_0 est une application de $\mathbb{E}$ dans $\mathbb{F}$ telle que

$$\|\varepsilon_0(h)\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{E}} \to 0.$$

La notation $\forall h \in \mathbb{E}$ sous-entend : pour tout h de $\mathbb{E}$ tel que $u+h$ appartient à Ω, un ouvert contenant u et tel que f est définie sur celui-ci.
(En termes plus mathématiques, ceci signifie

$$\forall \epsilon > 0, \quad \exists \eta > 0, \quad \forall v \in \Omega, \quad \|v-u\|_{\mathbb{E}} < \eta \Longrightarrow \|f(v)-f(u)\|_{\mathbb{F}} < \epsilon.)$$

L'expression (14.1) est un **développement limité d'ordre 0** au voisinage de u.

Remarque 14.1.1 *(préliminaire) Certaines des définitions de ce chapitre sont données dans le contexte général d'espaces vectoriels normés ; on peut pour simplifier se limiter au cas $\mathbb{E} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^p$, où $n \geq 1$ et $p \geq 1$ sont deux entiers naturels.*

Définition 14.1.1 *On dit que l'application f est* **différentiable** *en un point $u \in \mathbb{E}$ s'il existe g appartenant à $\mathcal{L}_c(\mathbb{E},\mathbb{F})$, qui vérifie*

$$\forall h \in \mathbb{E} \qquad f(u+h) = f(u) + g(h) + \|h\|\,\varepsilon(h), \tag{14.2}$$

où ε est une application de $\mathbb{E}$ dans $\mathbb{F}$ telle que

$$\|\varepsilon(h)\|_{\mathbb{F}} \to 0 \quad \textit{quand} \quad \|h\|_{\mathbb{E}} \to 0.$$

L'application linéaire continue g est notée $df(u)$, et on l'appelle **différentielle de f en u.** *On note l'action de $df(u)$ sur h*

$$df(u)\cdot h$$

L'expression (14.2) est un **développement limité d'ordre 1** au voisinage de u.

Remarque 14.1.2 *On retrouve là des* **petites variations**, *telles que décrites précédemment. En effet, on écrit*

$$f(u+h) = f(u) + df(u)\cdot h + o(h),$$

avec la propriété $\dfrac{\|o(h)\|_{\mathbb{F}}}{\|h\|_{\mathbb{E}}} \to 0$ *lorsque* $\|h\|_{\mathbb{E}} \to 0$, *ce qui correspond à un développement limité du premier ordre de f au voisinage de u.*

Proposition 14.1.1 *Si la différentielle de f en u existe, elle est unique.*

Preuve : L'unicité de la différentielle en u est obtenue de la manière élémentaire suivante. Soient deux applications linéaires continues $df_1(u)$ et $df_2(u)$ satisfaisant la relation (14.2). Alors, pour tout vecteur non nul v, et pour tout réel λ strictement positif suffisamment petit pour que $u+\lambda v$ appartienne à Ω, on a l'égalité

$$df_1(u)\cdot(\lambda v) - df_2(u)\cdot(\lambda v) = \lambda(\varepsilon_1(\lambda v) - \varepsilon_2(\lambda v)).$$

Par linéarité de $df_1(u)$ et $df_2(u)$, on arrive à

$$df_1(u)\cdot v - df_2(u)\cdot v = (\varepsilon_1(\lambda v) - \varepsilon_2(\lambda v)).$$

Si on fait tendre λ vers 0, on obtient que l'application linéaire $df_1(u) - df_2(u)$ est nulle, soit finalement $df_1(u) = df_2(u)$. ∎

Exercice 14.1.1 *1. Soit $f : \mathbb{R} \to \mathbb{R}$ dérivable. Montrer que f est différentiable sur $\mathbb{R}$ et calculer $df(x)$, pour $x \in \mathbb{R}$.*
2. Soit $f : \mathbb{R}^n \to \mathbb{R}^p$, une application affine, $f(u) = Au + b$. Montrer que f est différentiable sur $\mathbb{R}^n$ et calculer $df(u)$, pour $u \in \mathbb{R}^n$.
3. Soit $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $f(A) = A^2$. Montrer que f est différentiable sur $\mathbb{R}^{n\times n}$ et calculer $df(A)$, pour $A \in \mathbb{R}^{n\times n}$.
4. Soit Ω_n l'ensemble des matrices inversibles de $\mathbb{R}^{n\times n}$, et $f : \Omega_n \to \Omega_n$, définie par $f(A) = A^{-1}$. Pourquoi Ω_n est-il ouvert ? Montrer que f est différentiable sur Ω_n et vérifier que

$$df(A)\cdot H = -A^{-1}H\,A^{-1},$$

pour $A \in \Omega_n$.
4. Soit $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$. Montrer que f n'est pas différentiable en $x = 0$, mais qu'elle l'est sur $\mathbb{R}^n \setminus \{0\}$, et calculer $df(x)$ pour $x \neq 0$.

Bien sûr, en toute généralité, toute application différentiable en un point est continue en ce point, et on retrouve la formule de limite de taux de variation ; c'est l'objet de la

Proposition 14.1.2 *Si l'application f de $\mathbb{E}$ dans $\mathbb{F}$ est différentiable en u, elle est continue en ce point et*

$$\forall h \in \mathbb{E} \qquad df(u)\cdot h = \lim_{\theta\to 0^+} \frac{f(u+\theta h) - f(u)}{\theta}. \tag{14.3}$$

Preuve : A partir de la définition de la différentiabilité, en utilisant notamment le fait que la différentielle en u est continue, on tire

$$\|f(u+h)-f(u)\|_{\mathbb{F}} \leq \|df(u)\|\ \|h\|_{\mathbb{E}} + \|h\|_{\mathbb{E}}\ \|\varepsilon(h)\|_{\mathbb{F}}$$

et ainsi

$$\|f(u+h)-f(u)\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{E}} \to 0.$$

De plus par linéarité de l'application $df(u)$,

$$\forall h \in \mathbb{E},\ \forall \theta > 0 \qquad f(u+\theta h) = f(u) + \theta df(u)\cdot h + \theta\ \|h\|\ \varepsilon(\theta h)$$

et finalement

$$\forall h \in \mathbb{E} \qquad df(u)\cdot h = \lim_{\theta \to 0^+} \frac{f(u+\theta h)-f(u)}{\theta}.$$

■

Définition 14.1.2 *On dit que l'application f est* **différentiable** *dans Ω, si elle est différentiable en tout point $u \in \Omega$. Dans ce cas, on peut définir une application df qui à tout point $u \in \Omega$ associe une application linéaire et continue $df(u)$ de $\mathbb{E}$ dans $\mathbb{F}$; on l'appelle* **différentielle** *de f dans Ω. Si la différentielle df est une application continue de $\mathbb{E}$ dans $\mathcal{L}_c(\mathbb{E},\mathbb{F})$, on dit que f est une application* **continûment différentiable**, *ou encore de* **classe $\mathcal{C}^1$**

La définition 14.1.1 introduit la notion de différentielle au sens de **Fréchet**. On peut également définir la différentielle de f en u, $h \mapsto df(u)\cdot h$, au sens de **Gateaux**. On parle souvent de **dérivée directionnelle**.

Définition 14.1.3 *On dit que l'application f, définie sur un voisinage de u, est différentiable au sens de Gateaux s'il existe $df(u)$ de $\mathcal{L}_c(\mathbb{E},\mathbb{F})$ telle que*

$$\forall h \in \mathbb{E} \qquad df(u)\cdot h = \lim_{\theta \to 0^+} \frac{f(u+\theta h)-f(u)}{\theta}.$$

On peut aussi écrire la définition équivalente, pour chaque h *dans* $\mathbb{E}$

$$f(u+\theta\, h) = f(u) + \theta\, df(u)\cdot h + o(\theta), \qquad \theta \geq 0, \tag{14.4}$$

avec la propriété $\dfrac{\|o(\theta)\|_{\mathbb{F}}}{\theta} \to 0$ *lorsque* $\theta \to 0^+$.

Proposition 14.1.3 *Si $\mathbb{E} = \mathbb{R}$, la Fréchet-différentiabilité et la Gateaux-différentiabilité coïncident.*

Supposons maintenant que $dim[\mathbb{E}] \geq 2$. La différence entre (14.2) écrite avec $\theta\, h$ au lieu de h et (14.4) se trouve dans l'expression du reste

$\theta\, \|h\| \varepsilon(\theta\, h)$ pour la Fréchet-différentiabilité.

$o(\theta)$ pour la Gateaux-différentiabilité.

En d'autres termes, elle est *uniforme* en h pour la première, ce qui n'est pas garanti pour la seconde. De manière plus imagée, la Gateaux-différentiabilité est la différentiabilité le long de toute *droite* passant par u, alors que la Fréchet-différentiabilité correspond à la différentiabilité le long de toute *courbe* passant par u. De façon générale, la proposition 14.1.2 montre qu'une application différentiable au sens de Fréchet est toujours différentiable au sens de Gateaux (et les différentielles sont égales !), alors que la réciproque est fausse. D'ailleurs, la Gateaux-différentiabilité n'implique même pas la continuité, comme le montre le contre-exemple qui suit.

Exercice 14.1.2 *On se place dans* $\mathbb{E} = \mathbb{R}^2$. *Soient* $q \geq p > 5$ *deux réels. Montrer que la fonctionnelle* f *définie par*

$$f(x,y) = \begin{cases} \dfrac{x^p}{(y-x^2)^2 + x^q} & si\ (x,y) \neq (0,0) \\ 0 & si\ (x,y) = (0,0) \end{cases} .$$

est différentiable au point $(0,0)$ *au sens de Gateaux, mais qu'elle n'est pas continue en ce point.*

Exercice 14.1.3 *Soit encore* $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$. *Vérifier que* f *n'est pas Gateaux-differentiable en* $x = 0$.

Plaçons nous maintenant dans le cas, important en pratique, où $\mathbb{E} = \mathbb{R}^n$ et $\mathbb{F} = \mathbb{R}$, c'est-à-dire que f est à valeurs réelles. On munit $\mathbb{R}^n$ d'un produit scalaire, noté $(\cdot,\cdot)$ dans la suite. Si f est différentiable en u, sa différentielle $df(u)$ est une forme linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$. A cette forme, on peut associer un unique vecteur de $\mathbb{R}^n$, appelé **gradient de** f **en** u, et noté $\nabla f(u)$, tel que

$$\forall h \in \mathbb{E} \qquad df(u)\cdot h = (\nabla f(u), h). \tag{14.5}$$

Dans ce cas particulier, la formule (14.2) prend la forme

$$\forall h \in \mathbb{E} \qquad f(u+h) = f(u) + (\nabla f(u), h) + \|h\|\, \varepsilon(h). \tag{14.6}$$

Par exemple, avec J_0 définie en (1.3), qu'obtient-on comme expressions de $dJ_0(u)\cdot h$ et $\nabla J_0(u)$ pour u et h éléments de $\mathbb{R}^n$? L'expression (1.6) nous fournit la réponse ; en effet, on a

$$J_0(u+h) - J_0(u) = (Au - b, h) + \frac{1}{2}(Ah, h).$$

D'après l'inégalité de Cauchy-Schwarz et par définition de la norme matricielle induite par la norme euclidienne, on trouve

$$|(Ah,h)| \leq \|Ah\|\,\|h\| \leq \|A\|\,\|h\|^2 \ ;$$

ainsi, par identification, on trouve que

$$\varepsilon(h) = \frac{1}{2}\frac{(Ah,h)}{\|h\|}, \quad et \quad |\varepsilon(h)| \leq \frac{1}{2}\|A\|\,\|h\| \to 0\ quand\ \|h\| \to 0.$$

On infère immédiatement que

$$dJ_0(u)\cdot h = (Au - b, h),\ \text{et}\ \nabla J_0(u) = Au - b.$$

♠ Lorsque la matrice A n'est pas symétrique, les expressions ci-dessus sont *fausses* ! En effet, on doit remplacer A par $\frac{1}{2}(A + A^{\mathsf{T}})$.

Le gradient dépend seulement du produit scalaire. En particulier, il est *indépendant* de la base de l'espace *euclidien* $\mathbb{R}^n$. Supposons maintenant que $\mathbb{R}^n$ est muni d'une base *orthonormale* $(e_k)_{1\leq k\leq n}$, et soit $(x_k)_{1\leq k\leq n}$ le système de coordonnées associé.
Dans la base $(e_k)_{1\leq k\leq n}$, le vecteur $\nabla f(u)$ a pour composantes

$$\nabla f(u) = \begin{pmatrix} \partial_1 f(u) \\ \partial_2 f(u) \\ \vdots \\ \partial_n f(u) \end{pmatrix}. \tag{14.7}$$

On note aussi $\dfrac{\partial f}{\partial x_k}(u)$ ses composantes ; $\partial_k f(u)$ est appelée $k^{ème}$ **dérivée partielle** de f en u.

Remarque 14.1.3 *pourquoi parle-t-on de dérivée partielle ? La raison en est simple. Si on choisit $h = \theta e_k$ dans (14.6), on arrive à*

$$f(u+\theta e_k) = f(u) + \theta(\nabla f(u), e_k) + |\theta|\ \varepsilon(\theta e_k) = f(u) + \theta \frac{\partial f}{\partial x_k}(u) + |\theta|\ \varepsilon(\theta e_k).$$

Par ailleurs, modulo un petit abus de notations, on peut réécrire $f(u)$ sous la forme $f(x_1, \cdots, x_n)$. En d'autres termes, $\dfrac{\partial f}{\partial x_k}(u)$ représente la dérivée de f en u dans la direction e_k, ce qui correspond à la dérivée de l'application

$$\theta \mapsto f(x_1, \cdots, x_{k-1}, x_k + \theta, x_{k+1}, \cdots, x_n) \text{ en } \theta = 0.$$

Exercice 14.1.4 *Vérifier que si f est différentiable en u, elle admet une dérivée partielle par rapport à chaque variable en ce point. Réciproquement, montrer que, si f admet des dérivées partielles sur Ω qui sont continues en u, alors f est différentiable en u et que, de plus, elle est de classe $\mathcal{C}^1$ sur un ouvert contenant u.*

Dans le cas où $\mathbb{F} = \mathbb{R}^p$, alors pour $u = \sum_{k=1}^n x_k\, e_k$, $f(u)$ correspond à un vecteur à p composantes

$$f(u) = \begin{pmatrix} f_1(u) \\ f_2(u) \\ \vdots \\ f_p(u) \end{pmatrix},$$

dès lors que l'on a choisi une base $(e'_l)_{1\le l\le p}$ de $\mathbb{F}$. On peut reprendre la construction ci-dessus, et différencier chaque composante de f. La différentielle de f en u (lorsqu'elle existe) peut alors être écrite composante par composante

$$\begin{aligned}
df_1(u)\cdot h &= (\nabla f_1(u), h) = \partial_1 f_1(u)h_1 + \partial_2 f_1(u)h_2 + \ldots + \partial_n f_1(u)h_n \\
df_2(u)\cdot h &= (\nabla f_2(u), h) = \partial_1 f_2(u)h_1 + \partial_2 f_2(u)h_2 + \ldots + \partial_n f_2(u)h_n \\
\vdots &= \vdots \\
df_p(u)\cdot h &= (\nabla f_p(u), h) = \partial_1 f_p(u)h_1 + \partial_2 f_p(u)h_2 + \ldots + \partial_n f_p(u)h_n
\end{aligned}$$

ou encore sous forme matricielle

$$\begin{pmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \ldots & \partial_n f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \ldots & \partial_n f_2(u) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_p(u) & \partial_2 f_p(u) & \ldots & \partial_n f_p(u) \end{pmatrix}.$$

La matrice associée à $df(u)$ dans les bases $(e_k)_{1\le k\le n}$ et $(e'_l)_{1\le l\le p}$ est appelée **matrice jacobienne de f en** u, et on la note $[df(u)]$. Lorsque $n = p$, son déterminant est appelé **jacobien** de f en u, égal à

$$J_{f(u)} = \begin{vmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \ldots & \partial_n f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \ldots & \partial_n f_2(u) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_n(u) & \partial_2 f_n(u) & \ldots & \partial_n f_n(u) \end{vmatrix}.$$

Remarque 14.1.4 *Revenons un instant au cas $\mathbb{F} = \mathbb{R}$, c'est-à-dire $p = 1$.*
On a $[df(u)] = (\,\partial_1 f(u) \quad \partial_2 f(u) \quad \ldots \quad \partial_n f(u)\,)$. Si on compare cette expression à (14.7), on en déduit que dans ce cas

$$\nabla f(u) = [df(u)]^{\mathsf{T}}.$$

14.2 Propriétés de la différentielle

Nous démontrons quelques résultats simples, concernant l'addition et la composition d'applications différentiables.

Proposition 14.2.1 *Soient f et g deux applications de $\mathbb{E}$ dans $\mathbb{F}$ Fréchet-différentiables en $u \in \mathbb{E}$, alors l'application $f+g$ est Fréchet-différentiable en u et $d(f+g)(u) = df(u) + dg(u)$.*

Preuve : Des relations

$$f(u+h) = f(u) + df(u)\cdot h + \|h\|\ \varepsilon_f(h)$$
$$g(u+h) = g(u) + dg(u)\cdot h + \|h\|\ \varepsilon_g(h)$$

on tire par addition

$$(f+g)(u+h) = (f+g)(u) + df(u)\cdot h + dg(u)\cdot h + \|h\|(\ \varepsilon_f(h) +\ \varepsilon_g(h)).$$

Comme

$$\|(\varepsilon_f(h) + \varepsilon_g(h))\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{E}} \to 0$$

on voit que l'application linéaire $d(f+g)(u)$ définie par

$$d(f+g)(u)\cdot h = df(u)\cdot h + dg(u)\cdot h$$

correspond à la définition de la différentiabilité en u de $f+g$. ∎

Remarque 14.2.1 *De la même façon, on peut prouver que la somme de deux applications Gateaux-différentiables en un point est Gateaux-différentiable.*

Proposition 14.2.2 *Soit f une application de $\mathbb{E}$ dans $\mathbb{F}$ Fréchet-différentiable en $u \in \mathbb{E}$, et soit g une application de $\mathbb{F}$ dans $\mathbb{G}$ Fréchet-différentiable en $f(u) \in \mathbb{F}$, alors l'application $g \circ f$ est Fréchet-différentiable en u et*

$$d(g\circ f)(u) = dg(f(u)) \circ df(u).$$

Preuve : Des relations

$$f(u+h) = f(u) + df(u)\cdot h + \|h\|\ \varepsilon_f(h)$$
$$g(f(u)+h') = g(f(u)) + dg(f(u))\cdot h' + \|h'\|\ \varepsilon_g(h'),$$

on tire

$$\begin{aligned} g\circ f(u+h) &= g(f(u+h)) \\ &= g\left(f(u) + df(u)\cdot h + \|h\|\ \varepsilon_f(h)\right) \\ &= g(f(u)+h') \quad \text{avec} \quad h' = df(u)\cdot h + \|h\|\ \varepsilon_f(h) \\ &= g(f(u)) + dg(f(u))\cdot h' + \|h'\|\ \varepsilon_g(h'). \end{aligned}$$

Mais l'application différentielle $dg(f(u))$ est linéaire par définition, d'où

$$dg(f(u))\cdot h' = dg(f(u))\cdot(df(u)\cdot h) + \|h\|\ dg(f(u))\cdot(\varepsilon_f(h)).$$

On arrive alors à l'expression

$$g\circ f(u+h) = g\circ f(u) + [dg(f(u))\circ df(u)]\cdot h + \|h\|\ dg(f(u))\cdot(\varepsilon_f(h)) + \|h'\|\ \varepsilon_g(h').$$

Il suffit maintenant de vérifier que les deux termes de droite peuvent être réécrits sous la forme $\|h\|\,\varepsilon_{g\circ f}(h)$, avec $\|\varepsilon_{g\circ f}(h)\| \to 0$ lorsque $\|h\| \to 0$. Or, on a d'une part

$$\|dg(f(u))\cdot(\varepsilon_f(h))\| \leq \|dg(f(u))\|\,\|\varepsilon_f(h))\| \to 0 \quad quand \quad \|h\| \to 0\ ;$$

et d'autre part

$$\|h'\| \leq \|df(u)\|\,\|h\| + \|h\|\,\|\varepsilon_f(h)\| = O(\|h\|).$$

On obtient finalement

$$g\circ f(u+h) = g\circ f(u) + [dg(f(u))\circ df(u)]\cdot h + \|h\|\,\varepsilon_{g\circ f}(h).$$

■

On a également le résultat suivant, si l'on affaiblit l'hypothèse sur f.

Proposition 14.2.3 *Soit f une application de $\mathbb{E}$ dans $\mathbb{F}$ Gateaux-différentiable en $u \in \mathbb{E}$, et soit g une application de $\mathbb{F}$ dans $\mathbb{G}$ Fréchet-différentiable en $f(u) \in \mathbb{F}$, alors l'application $g\circ f$ est Gateaux-différentiable en u et*

$$d(g\circ f)(u) = dg(f(u))\circ df(u).$$

Preuve : Soit $h \in \mathbb{E}$ donné : $f(u+\theta\,h) = f(u) + \theta\,df(u)\cdot h + o(\theta)$.
Si on note $h'_\theta = \theta\,df(u)\cdot h + o(\theta)$, on a en particulier $\dfrac{\|h'_\theta\|}{\theta}$ borné lorsque $\theta > 0$ est petit.

$$\begin{aligned} g\circ f(u+\theta\,h) - g\circ f(u) &= dg(f(u))\cdot h'_\theta + \|h'_\theta\|\varepsilon_g(h'_\theta) \\ &= \theta\,dg(f(u))\cdot(df(u))\cdot h + dg(f(u))\cdot o(\theta) + \|h'_\theta\|\varepsilon_g(h'_\theta), \text{d'où} \\ \frac{g\circ f(u+\theta\,h) - g\circ f(u)}{\theta} &= dg(f(u))\cdot(df(u))\cdot h + dg(f(u))\cdot o(1) + \frac{\|h'_\theta\|}{\theta}\varepsilon_g(h'_\theta). \end{aligned}$$

Comme $\|h'_\theta\| \to 0$ lorsque $\theta \to 0^+$, on a de même $\|\varepsilon_g(h'_\theta)\| \to 0^+$. Ainsi

$$\lim_{\theta\to 0^+} \frac{g\circ f(u+\theta\,h) - g\circ f(u)}{\theta} = [dg(f(u))\circ df(u)]\cdot h.$$

■

Remarque 14.2.2 *La Fréchet-différentiabilité de g est nécessaire pour pouvoir considérer la différentielle de la composée. Le résultat sur la composition est faux, si l'on suppose uniquement que g est Gateaux-différentiable, même si f est Fréchet-différentiable.*

Posons $v = f(u)$.
Lorsque $\mathbb{E} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^p$ et $\mathbb{G} = \mathbb{R}^m$, et que chacun de ces trois espaces est muni d'une base orthonormale, $df(u)$ est représentée par une matrice de $\mathbb{R}^{p\times n}$, $dg(v)$ par une matrice de $\mathbb{R}^{m\times p}$ et $d(g\circ f)(u)$ par une matrice de $\mathbb{R}^{m\times n}$:

$$\begin{pmatrix} \partial_1[g\circ f]_1(u) & \partial_2[g\circ f]_1(u) & \dots & \partial_n[g\circ f]_1(u) \\ \partial_1[g\circ f]_2(u) & \partial_2[g\circ f]_2(u) & \dots & \partial_n[g\circ f]_2(u) \\ \dots & \dots & \dots & \dots \\ \partial_1[g\circ f]_m(u) & \partial_2[g\circ f]_m(u) & \dots & \partial_n[g\circ f]_m(u) \end{pmatrix}.$$

D'après la proposition 14.2.2, $d(g\circ f)(u)$ est représentée par une matrice égale au produit des matrices associées à $dg(v)$ et $df(u)$:

$$\begin{aligned} [d(g\circ f)(u)] &= [dg(v)]\,[df(u)] \qquad (14.8) \\ &= \begin{pmatrix} \partial_1 g_1(v) & \partial_2 g_1(v) & \dots & \partial_p g_1(v) \\ \partial_1 g_2(v) & \partial_2 g_2(v) & \dots & \partial_p g_2(v) \\ \dots & \dots & \dots & \dots \\ \partial_1 g_m(v) & \partial_2 g_m(v) & \dots & \partial_p g_m(v) \end{pmatrix} \begin{pmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \dots & \partial_n f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \dots & \partial_n f_2(u) \\ \dots & \dots & \dots & \dots \\ \partial_1 f_p(u) & \partial_2 f_p(u) & \dots & \partial_n f_p(u) \end{pmatrix}, \end{aligned}$$

que l'on écrit composante par composante

$$\frac{\partial (g\circ f)_i}{\partial x_j}(u)=\sum_{k=1}^{p}\frac{\partial g_i}{\partial x_k}(v)\frac{\partial f_k}{\partial x_j}(u)\qquad 1\le i\le m,\quad 1\le j\le n. \tag{14.9}$$

Dans le cas où la fonctionnelle g est à valeurs dans $\mathbb{R}$ (soit $m=1$), on sait que $[dg(v)]=\nabla g(v)^{\mathsf{T}}$ (cf. remarque 14.1.4) ; $g\circ f$ est également à valeurs dans $\mathbb{R}$, et l'on a de même $[d(g\circ f)(u)]=\nabla(g\circ f)(u)^{\mathsf{T}}$. En transposant (14.8), on en déduit finalement que

$$\nabla(g\circ f)(u)=[df(u)]^{\mathsf{T}}\nabla g(v)\qquad avec\quad v=f(u).$$

Exercice 14.2.1 *Soit toujours $f:\mathbb{R}^n\to\mathbb{R}$, $f(x)=\|x\|$. En l'écrivant sous la forme $f=s\circ g$, s et g à déterminer, retrouver l'expression de $df(x)$ et de son gradient, lorsque x est non nul.*

14.3 Différentielles d'ordre supérieur et formules de Taylor

Dans cette section, on considère des applications différentiables au sens de Fréchet.

14.3.1 Différentielles d'ordre supérieur

On supose que f est différentiable en u. Si la différentielle df est elle-même différentiable en u, on définit l'application $d^2f(u)$, appelée **différentielle seconde** de l'application f en u, et on dit que f est deux fois différentiable au point u ; $d^2f(u)$ appartient à $\mathcal{L}_c(\mathbb{E},\mathcal{L}_c(\mathbb{E},\mathbb{F}))$. Si la différentielle d^2f est une application continue de $\mathbb{E}$ dans $\mathcal{L}_c(\mathbb{E},\mathcal{L}_c(\mathbb{E},\mathbb{F}))$, on dit que f est une application de classe $\mathcal{C}^2$.

Remarque 14.3.1 *(***importante***) Pour définir la différentielle seconde* **en** *u, on doit supposer que f est différentiable sur un* **voisinage de** *u.*

On peut *identifier*[1] $\mathcal{L}_c(\mathbb{E},\mathcal{L}_c(\mathbb{E},\mathbb{F}))$ à $\mathcal{L}_c(\mathbb{E}\times\mathbb{E},\mathbb{F})$, et on écrit donc :

$$(d^2f(u)\cdot h)\cdot h'=d^2f(u)\cdot(h,h'),\qquad (h,h')\in\mathbb{E}\times\mathbb{E}.$$

Si $h'=h$, on condense les notations en $d^2f(u)\cdot h^2$. On rappelle le

Théorème 14.3.1 *(de* **Schwarz***) Soit f une application deux fois différentiable en u. Alors $d^2f(u)$ est une application (bilinéaire, continue et) symétrique de $\mathbb{E}\times\mathbb{E}$ dans $\mathbb{F}$.*

1. Si $A\in\mathcal{L}_c(\mathbb{E}\times\mathbb{E},\mathbb{F})$, ceci signifie que A est bilinéaire en (x,y), et qu'il existe $C\in\mathbb{R}$ tel que

$$\sup_{\|x\|=1,\|y\|=1}\|A(x,y)\|\le C.$$

(i) Pour $x\in\mathbb{E}$, soit $A_x=A(x,\cdot):A_x$ est linéaire et $\sup_{\|y\|=1}\|A_x(y)\|=\sup_{\|y\|=1}\|A(x,y)\|\le C\|x\|$. Ainsi, $A_x\in\mathcal{L}_c(\mathbb{E},\mathbb{F})$, et $\|A_x\|\le C_x$, avec $C_x=C\|x\|$.
(ii) Soit maintenant $\tilde{A}:x\to A_x$. Comme A est linéaire en sa première variable, $\tilde{A}$ est un élément de $\mathcal{L}(\mathbb{E},\mathcal{L}_c(\mathbb{E},\mathbb{F}))$. Il reste à vérifier la continuité, or

$$\sup_{\|x\|=1}\|\tilde{A}(x)\|\le\sup_{\|x\|=1}C_x\le C.$$

Réciproquement, soit $\tilde{A}\in\mathcal{L}_c(\mathbb{E},\mathcal{L}_c(\mathbb{E},\mathbb{F}))$. On définit $A:(x,y)\to\tilde{A}(x)(y)$. Par construction, A est bilinéaire de $\mathbb{E}\times\mathbb{E}$ dans $\mathbb{F}$ et, par ailleurs, comme $\tilde{A}(x)\in\mathcal{L}_c(\mathbb{E},\mathbb{F})$ pour tout x,

$$\sup_{\|x\|=1,\|y\|=1}\|A(x,y)\|=\sup_{\|x\|=1,\|y\|=1}\|\tilde{A}(x)(y)\|\le\sup_{\|x\|=1}\|\tilde{A}(x)\|\le\|\tilde{A}\|.$$

Replaçons-nous maintenant dans le cadre qui nous a permis de définir les dérivées partielles (premières), c'est-à-dire $\mathbb{E} = \mathbb{R}^n$ et $\mathbb{F} = \mathbb{R}$: $d^2 f(u)$ est une forme bilinéaire et continue de $\mathbb{R}^{n\times n}$. D'après l'identification ci-dessus, il existe un **unique** élément $\nabla^2 f(u)$ de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ tel que

$$d^2 f(u)\cdot(h, h') = (\nabla^2 f(u)h, h'), \qquad h, h' \in \mathbb{R}^n.$$

Encore une fois, $\nabla^2 f(u)$ est indépendante de la base choisie.
Munissons $\mathbb{R}^n$ d'une base orthonormée ; on peut construire, de la même façon que les dérivées partielles premières, les dérivées partielles secondes, notées $\dfrac{\partial^2 f}{\partial x_k \partial x_l}(u)$ ou $\partial_k \partial_l f(u)$. La différentielle seconde $d^2 f(u)$ est formée des $n \times n$ dérivées partielles $\partial_l \partial_k f(u)$ de chacune des composantes $\partial_k f(u)$ du gradient. On se trouve donc dans le cas du calcul de la différentielle d'une application de $\mathbb{R}^n$ dans lui-même, et on représente $d^2 f(u)$ par la **matrice Hessienne** :

$$[\nabla^2 f(u)] = \begin{pmatrix} \partial_1\partial_1 f(u) & \partial_2\partial_1 f(u) & \dots & \partial_n\partial_1 f(u) \\ \partial_1\partial_2 f(u) & \partial_2\partial_2 f(u) & \dots & \partial_n\partial_2 f(u) \\ \dots & \dots & \dots & \dots \\ \partial_1\partial_n f(u) & \partial_2\partial_n f(u) & \dots & \partial_n\partial_n f(u) \end{pmatrix}.$$

En particulier, on peut écrire : $d^2 f(u)\cdot(h, h') = (\nabla^2 f(u)h, h') = \sum_{i,j=1}^n h_i\, h'_j \partial_i \partial_j f(u)$.
Enfin, on infère facilement du théorème de Schwarz le

Corollaire 14.3.1 *$[\nabla^2 f(u)]$ est une matrice symétrique.*

Exercice 14.3.1 *Soit J_0 définie en (1.3). Calculer $d^2 J_0(u)\cdot(h, h')$ et $[\nabla^2 J_0(u)]$ pour u, h et h' éléments de $\mathbb{R}^n$.*

Bien évidemment, il est loisible de définir, par récurrence, les différentielles d'ordre supérieur ($k \geq 3$), à partir ce qui est écrit ci-dessus :

$$d^k f(u)\cdot(h, h', \cdots, h^{(k-1)}), \qquad h, h', \cdots, h^{(k-1)} \in \mathbb{E} \times \mathbb{E} \times \cdots \times \mathbb{E}.$$

Lorsque tous les arguments sont identiques, on adopte la notation $d^k f(u)\cdot h^k$. Si la différentielle $d^k f$ est une application continue de $\mathbb{E}$ dans l'espace $\mathcal{L}_c(\mathbb{E} \times \mathbb{E} \times \cdots \times \mathbb{E}, \mathbb{F})$, on dit que f est une application de classe $\mathcal{C}^k(\Omega)$. Pour définir la différentielle d'ordre k **en** u, on doit supposer que f est $k-1$ fois différentiable sur un **voisinage de** u.

14.3.2 Formules de Taylor

Nous énonçons pour finir quelques résultats concernant les formules de Taylor des applications différentiables.
On suppose que l'application f de $\mathbb{E}$ dans $\mathbb{F}$ est k fois différentiable en u, avec $k \geq 0$ (si $k = 0$, ceci signifie simplement que f est continue en u). Pour h suffisamment petit, c'est-à-dire tel que $u + h \in \Omega$, on introduit le **reste de rang k de f en u**

$$r_k(h) = f(u + h) - f(u) - \sum_{m=1}^{k} \frac{1}{m!} d^m f(u)\cdot h^m.$$

En d'autres termes, on écrit le **développement limité d'ordre** k au voisinage de u

$$f(u + h) = f(u) + \sum_{m=1}^{k} d^m f(u)\cdot h^m + r_k(h). \tag{14.10}$$

Remarque 14.3.2 *Supposons par exemple que $\mathbb{E} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}$ et $k = 2$; dans ce cas particulier*

$$f(u + h) = f(u) + (\nabla f(u), h) + \frac{1}{2}(\nabla^2 f(u)h, h) + r_2(h).$$

L'objet des résultats ci-dessous (pour lesquels on indique brièvement l'idée de la démonstration, voir également [15]) est d'estimer le reste $r_k(h)$. Pour u et v deux éléments de $\mathbb{E}$, on appelle $[u,v]$ le segment défini par

$$[u,v] = \{w \in \mathbb{E} \ : \ \exists \lambda \in [0,1], \ w = \lambda u + (1-\lambda)v\}.$$

On suppose ici que f est un peu plus régulière.

Théorème 14.3.2 *(***inégalité de Taylor–Lagrange***) Supposons que f est de classe $\mathcal{C}^k$ sur Ω. On choisit h tel que le segment $[u, u+h]$ est inclus dans Ω. On suppose de plus que f admet en tout point de $]u, u+h[$ une différentielle d'ordre $k+1$, dont la norme est majorée par M uniformément sur $]u, u+h[$. Alors, le reste r_k vérifie*

$$\|r_k(h)\| \leq \frac{1}{(k+1)!}\, M\, \|h\|^{k+1}.$$

Preuve : On note $\gamma(t) = u + t\,h$ le chemin défini le long du segment $[u,v]$, pour $t \in [0,1]$, ce qui permet d'introduire la fonction $\mu : t \mapsto f \circ \gamma(t)$.
On applique ensuite l'inégalité de Taylor-Lagrange pour μ, fonction d'une variable réelle. ■

Lorsque $k = 0$, l'inégalité précédente est appelée **inégalité des accroissements finis**.

Théorème 14.3.3 *(***formule de Taylor–Mac Laurin***) On se place dans le cas d'une fonctionnelle à valeurs numériques, c'est-à-dire que $\mathbb{F} = \mathbb{R}$. Supposons que f est de classe $\mathcal{C}^k$ sur Ω. On choisit h tel que le segment $[u, u+h]$ est inclus dans Ω. On suppose de plus que f admet en tout point de $]u, u+h[$ une différentielle d'ordre $k+1$. Alors, il existe $\lambda \in]0,1[$ tel que*

$$r_k(h) = \frac{1}{(k+1)!} d^{k+1} f(u+\lambda h) \cdot h^{k+1}.$$

Preuve : On procède comme pour l'inégalité de Taylor-Lagrange. ■

Remarque 14.3.3 *Le théorème 14.3.3 est faux si $\mathbb{F} \neq \mathbb{R}$.*

Et, si f est encore un peu plus régulière, on obtient le

Théorème 14.3.4 *(***du reste intégral***) Supposons que f est de classe $\mathcal{C}^{k+1}$ sur Ω. On choisit h tel que le segment $[u, u+h]$ est inclus dans Ω. Alors, $r_k(h)$ est égal à*

$$r_k(h) = \int_0^1 \frac{(1-\theta)^k}{k!} d^{k+1} f(u+\theta h) \cdot h^{k+1}\, d\theta.$$

Preuve : On procède comme pour l'inégalité de Taylor-Lagrange. ■

Pour finir, si l'on en revient à la régularité initiale, on a le

Théorème 14.3.5 *(***de Taylor–Young***) Soit f une application k fois différentiable en u. Le reste r_k vérifie*

$$\|r_k(h)\| = o(\|h\|^k).$$

Preuve : La démonstration est faite par récurrence sur k.
Lorsque $k = 1$, on retrouve la définition de la différentiabilité de f en u.
Par récurrence, on différencie le reste r_{k+1}, et on utilise la formule des accroissements finis, en notant que la différentielle de $h \mapsto d^m f(u) \cdot h^m$ est

si $m = 1$: $df(u)$.

si $m > 1$: $h \mapsto m\, d^m f(u) \cdot h^{m-1}$.

■

Chapitre 15

La méthode des différences finies

15.1 Introduction

Dans ce chapitre, on se propose de présenter une méthode numérique qui permet de calculer des approximations de quantités liées à des problèmes issus de la physique (au sens large !), telles que

- déplacement, vitesse (en mécanique) ;
- potentiel, champ (en électromagnétisme) ;
- prix d'une option (en finance) ;
- et bien d'autres encore (voir [4])...

Dans la suite, on étudie plus en détail le calcul du déplacement transerval d'un fil, ainsi que celui du potentiel électrostatique.

15.2 Un problème monodimensionnel

Dans cette section, nous considérons un fil, *tendu* entre ses extrémités, situées en 0 et 1. On suppose qu'il est soumis à une force extérieure transverse (telle que son poids, si le fil est 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.

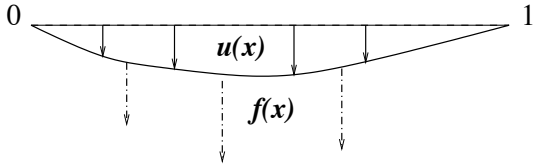


FIG. 15.1 – *Un fil pesant*

On admet que les équations de l'élasticité linéaire monodimensionnelles (1D) normalisées permettent de modéliser correctement le phénomène. Elles sont de la forme :

$$-u''(x) = f(x) \text{ sur }]0,1[, \quad u(0) = u(1) = 0. \tag{15.1}$$

On parle d'équation 1D, puisque la seule variable est x, qui varie de 0 à 1. Les conditions aux extrémités sont appelées **condition aux limites** : elles signifient simplement que les dites extrémités sont fixes. Notons que dans le cas particulier où $f \equiv 1$, la solution est égale à

$$u_0(x) = \frac{1}{2}x(1-x). \tag{15.2}$$

Dans 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 (15.1) (ça n'est pas la seule !), on utilise la

Proposition 15.2.1 *Soient $x \in]0,1[$ et h tel que $[x-h, x+h] \in [0,1]$. Alors*

$$\exists \theta \in]-1,1[\text{ tel que } -u''(x) = \frac{-u(x-h)+2u(x)-u(x+h)}{h^2} + \frac{h^2}{12}u^{(4)}(x+\theta\, h). \qquad (15.3)$$

Preuve : On utilise la formule de Taylor-Mac Laurin, rappelée au théorème 14.3.3.

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

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

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 15.2.1 *Le premier terme de (15.3) 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 (15.1) ; on parle souvent de **schéma numérique** de discrétisation. Le terme **différences finies** provient quant à lui de l'expression (15.3) : 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).

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**.

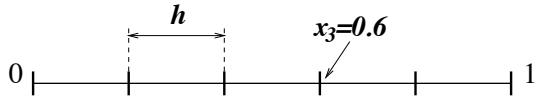


FIG. 15.2 – *Le segment découpé ($N = 4$, $h = 0.2$)*

Remarque 15.2.2 *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

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i,\ 1 \leq i \leq N, \text{ avec } u_0 = u_{N+1} = 0. \tag{15.4}$$

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

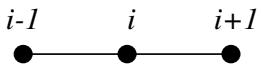


FIG. 15.3 – *Le schéma aux différences finies à trois points*

NB. Noter la similitude entre (15.4) d'une part, et (15.3) et (15.1) 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 (15.4) sous la forme matricielle *équivalente*

$$A\,\vec{u} = \vec{f},\ \text{avec } A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & \dots & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & \dots & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{15.5}$$

La matrice A est tridiagonale et symétrique par construction.

Il convient de vérifier qu'il existe une solution $\vec{u}$ unique de (15.5). Qui plus est, comment calculer et majorer l'erreur commise? C'est l'objet des résultats ci-dessous.
Pour commencer, nous allons vérifier que la matrice A 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. Ensuite, pour exploiter cette formule, c'est-à-dire pour majorer l'erreur, nous allons étudier les caractéristiques de l'inverse A^{-1}.

Proposition 15.2.2 *La matrice A est symétrique définie-positive.*

Preuve : On applique la définition usuelle, en formant le produit scalaire $h^2(v, Av)$, pour un vecteur v de $\mathbb{R}^N$ quelconque :

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

Ainsi, $(v, Av) \geq 0$. De plus,$(v, Av) = 0$ 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 $v = 0$. ■

Comme A est symétrique définie-positive, elle est en particulier inversible et, de plus, on peut résoudre le système linéaire (15.5) à l'aide de l'algorithme du Gradient Conjugué.

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 (15.1) (resp. (15.5)), on a alors, d'après (15.3), pour i compris entre 1 et N :

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

Dans l'esprit de la remarque 15.2.1, on introduit le vecteur $\vec{\varepsilon}$ de $\mathbb{R}^N$, dont les composantes sont égales à $\varepsilon_i = (A\vec{e})_i = \dfrac{h^2}{12}f''(x_i+\theta_i\,h)$, 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

$$\vec{e} = A^{-1}\vec{\varepsilon}. \tag{15.6}$$

Pour aller plus loin, nous allons démontrer deux propriétés concernant la matrice A^{-1} : la première sur le signe de ses éléments, et la seconde sur la somme de ses mêmes éléments, ligne par ligne. Avant de commencer, introduisons la

Définition 15.2.1 *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 15.2.3 *Une matrice A de $\mathbb{R}^{N\times N}$ est monotone si, et seulement si, on a l'inclusion*

$$\{v \in \mathbb{R}^N \ : \ Av \geq 0\} \subset \{v \in \mathbb{R}^N \ : \ v \geq 0\}. \tag{15.7}$$

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 (15.7) 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 (15.7) 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 15.2.4 *La matrice A correspondant à (15.5) est monotone.*

Preuve : Pour prouver que A est monotone, on reprend la proposition 15.2.3. Soit v tel que $Av \geq 0$, 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

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

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

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

En reprenant l'expression de l'erreur (15.6), on peut alors écrire

$$|e_i| = |\sum_j (A^{-1})_{i,j}\varepsilon_j| \leq \sum_j (A^{-1})_{i,j}|\varepsilon_j| \leq \sum_j (A^{-1})_{i,j}\|\vec{\varepsilon}\|_\infty \leq \frac{h^2}{12} C_{f,2} \sum_j (A^{-1})_{i,j}. \qquad (15.8)$$

Pour finalement arriver à une majoration de l'erreur commise, il suffit de majorer $\sum_j (A^{-1})_{i,j}$ dans (15.8), pour $1 \leq i \leq N$.

Proposition 15.2.5 *La somme des éléments de chaque ligne de A^{-1} est inférieure ou égale à 1/8.*

Preuve : On remarque que $\sum_j (A^{-1})_{i,j} = \sum_j (A^{-1})_{i,j}\delta_j$, sous réserve que $\delta_j = 1$, pour tout j. A quoi correspond le vecteur $\vec{\delta}$ ainsi construit ? On pose $\vec{u}_0 = A^{-1}\vec{\delta}$, soit $A\vec{u}_0 = \vec{\delta}$.
$\vec{\delta}$ joue le rôle d'un second membre de (15.5). Il correspond de fait à $f \equiv 1$, ce qui nous renvoie à la solution u_0 de (15.2). 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

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

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

$$\|\vec{e}\|_\infty \leq \frac{h^2}{96} \sup_{x\in[0,1]} |f''(x)|. \qquad (15.9)$$

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 et la donnée f sont moins régulières?

Théorème 15.2.2 *Lorsque la solution u est de classe $\mathcal{C}^2([0,1])$, l'erreur est telle que*

$$\lim_{h\to 0^+} \|\vec{e}\|_\infty = 0. \tag{15.10}$$

Preuve : On introduit à nouveau le vecteur $\vec{\varepsilon}$, de composantes $\varepsilon_i = (A\vec{e})_i$. D'après les résultats portant sur la matrice A (propositions 15.2.4 et 15.2.5), 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, A, $\vec{e}$ 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...

15.3 Un problème multidimensionnel

Dans cette section, on va proposer une méthode de calcul du potentiel électrostatique, toujours basée sur les différences finies. Dans $\mathbb{R}^3$, on considère une cavité cubique, $\Omega =]0,1[^3$, dans laquelle on a fait le vide. On suppose qu'elle est entourée d'un conducteur parfait, et que le potentiel électrostatique, sur son bord $\partial\Omega$, est nul. Enfin, on place des charges dans la cavité. Si on appelle ρ la densité de charge et ϵ_0 la permittivité électrique, le potentiel électrostatique V généré par les charges dans la cavité est solution de l'équation dite de Coulomb, qui s'écrit

$$-\Delta V = \frac{\rho}{\epsilon_0} \text{ dans } \Omega, \quad V = 0 \text{ sur } \partial\Omega. \tag{15.11}$$

(L'opérateur Δ, appelé Laplacien, est égal par définition à $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2} + \dfrac{\partial^2}{\partial z^2}$.)

On parle d'un problème 3D, puisqu'il dépend des variables x, y et z. La condition $V = 0$ sur $\partial\Omega$ est une **condition aux limites**.

Dans la suite, nous allons nous contenter de la version 2D du même type de problème... La généralisation au cas 3D, relativement aisée, sera laissée aux bons soins du lecteur ! (voir l'exercice 15.3.1 ci-dessous.)

Soit donc à résoudre, dans $\Omega =]0,1[^2$, le problème 2D

$$-\Delta u = f \text{ dans } \Omega, \quad u = u_d \text{ sur } \partial\Omega. \tag{15.12}$$

Ci-dessus, f et u_d sont des fonctions données (f dans Ω et u_d sur le bord $\partial\Omega$).
Quant au Laplacien bidimensionnel, il est défini par $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$.

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 (15.1). En effet, on remarque qu'en 1D on a les égalités $-u'' = -\dfrac{\partial^2}{\partial x^2}u = -\Delta u$, *id est* (15.1) 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 15.2.1.

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 $\mathcal{C}^2([0,1]^2)$, même si $u_d \equiv 0$...

Proposition 15.3.1 *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 trouve alors

$$-\Delta u(x,y) = \frac{-u(x-h_1,y) + 2u(x,y) - u(x+h_1,y)}{{h_1}^2} + \frac{-u(x,y-h_2) + 2u(x,y) - u(x,y+h_2)}{{h_2}^2}$$
$$+\frac{{h_1}^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\theta_1 h_1, y) + \frac{{h_2}^2}{12}\frac{\partial^4 u}{\partial y^4}(x, y+\theta_2 h_2). \tag{15.13}$$

Remarque 15.3.1 *Les deux premiers termes de (15.13) sont une bonne approximation de* $-\Delta u(x,y)$*, sous réserve que* h_1 *et* h_2 *sont* petits*. Le reste est en effet borné par*

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

Dans la suite, on va considérer un pas de discrétisation *identique* 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 $(u_{i,j})_{0\le i,j\le n+1}$, qui sont les *valeurs approchées* de la solution u aux points d'abscisse $x_i=ih$ et d'ordonnée $y_j=jh$. On note $(f_{i,j})_{1\le i,j\le n}$ les valeurs $f_{i,j}=f(x_i,y_j)$, pour i et j variant de 1 à n.

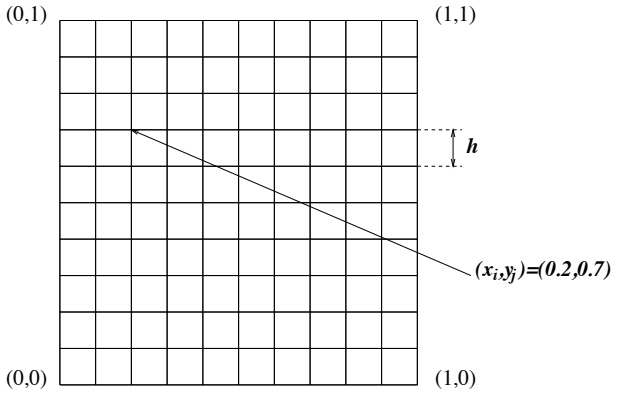


FIG. 15.4 – Ω *et les points de discrétisation (*$n=9$*,* $h=0.1$*)*

L'opérateur de Laplace est *approché* par une combinaison linéaire de valeurs $u_{i,j}$, selon le **schéma à cinq points**

$$-\Delta u(x_i,y_j)\approx\frac{-u_{i,j-1}-u_{i-1,j}+4u_{i,j}-u_{i+1,j}-u_{i,j+1}}{h^2},\quad 1\le i,j\le n. \tag{15.14}$$

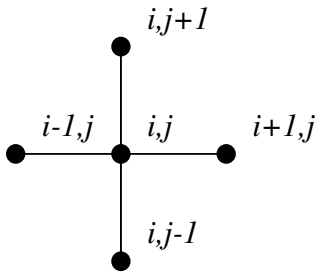


FIG. 15.5 – *Le schéma aux différences finies à cinq points*

Le problème (15.12) 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

$$\frac{-u_{i,j-1}-u_{i-1,j}+4u_{i,j}-u_{i+1,j}-u_{i,j+1}}{h^2}=f_{i,j},\quad 1\le i,j\le n. \tag{15.15}$$

Les valeurs de u sur le bord, ici $u(0,\cdot)$ et $u(1,\cdot)$, ainsi que $u(\cdot,0)$ et $u(\cdot,1)$ sont connues (car égales aux valeurs correspondantes de u_d). 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 $N = n^2$ 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

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

Au total, on a donc *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 (15.15). Ce système linéaire s'écrit encore

$$\frac{1}{h^2}\begin{pmatrix} B & T & & & \\ T & B & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & B & T \\ & & & T & B \end{pmatrix}\vec{u} = \vec{f} + \frac{1}{h^2}\begin{pmatrix} u_d(\cdot,0) \\ \cdot \\ \cdot \\ \cdot \\ u_d(\cdot,1) \end{pmatrix}. \tag{15.16}$$

Le dernier vecteur regroupe l'ensemble des valeurs prises par u sur le bord $\partial\Omega$, qui sont connues (car égales à la valeur prise par u_d 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 15.4 et 15.5, un des points voisins au moins se trouve sur le bord $\partial\Omega$. Au contraire, si i et j appartiennent tous les deux à $\{2, \cdots, n-1\}$, tous les voisins se trouvent dans Ω, et sa composante $_{i,j}$ est nécessairement nulle.

On écrit finalement (15.16) sous la forme condensée

$$A\vec{u} = \vec{f} + \frac{1}{h^2}\vec{u}_d, \tag{15.17}$$

avec A une matrice de $\mathbb{R}^{N\times N}$, $\vec{u}$, $\vec{f}$ et $\vec{u}_d$ trois vecteurs de $\mathbb{R}^N$.

Si l'on revient à la structure interne de A, on a $T = -I_n$, avec I_n la matrice identité d'ordre n, et $B \in \mathbb{R}^{n\times n}$ est la matrice tridiagonale définie par

$$B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}.$$

La matrice A des systèmes linéaires (15.16) et (15.17) est donc pentadiagonale (par points), 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 :

$$\text{composante }_I \equiv {}_{i,j} \quad \Longleftrightarrow \quad I = i + (j-1)n \text{ ou } \left\{ \begin{array}{l} i = (I-1) \mod n + 1 \\ j = \lfloor (I-1)/n \rfloor + 1 \end{array} \right. . \tag{15.18}$$

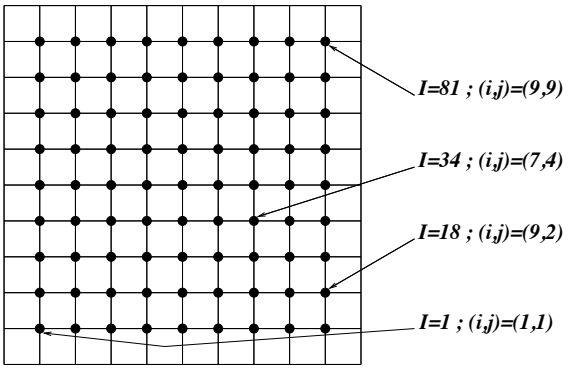


FIG. 15.6 – *Les deux numérotations : $I \in \{1, \cdots, 81\}$ et $i, j \in \{1, \cdots, 9\}$*

Nous allons maintenant étudier les propriétés de la matrice A.
Tout d'abord, A est symétrique par construction. Toutes ses valeurs propres sont donc réelles. Quel est leur signe ? La réponse est simple, si l'on se souvient du théorème 10.4.1 de Gerschgorin-Hadamard : soit $\lambda(A)$ une valeur propre de A, alors, dans le plan complexe $\mathbb{C}$,

$$\lambda(A) \in \bigcup_{I=1,\cdots,N} \{z \in \mathbb{C} \ : \ |z - A_{I,I}| \leq \sum_{J\neq I} |A_{I,J}|\}.$$

Dans le cas particulier où $\lambda(A) \in \mathbb{R}$,

$$\lambda(A) \in \bigcup_{I=1,\cdots,N} \{s \in \mathbb{R} \ : \ s \in \Big[A_{I,I} - \textstyle\sum_{J\neq I} |A_{I,J}|, A_{I,I} + \sum_{J\neq I} |A_{I,J}|\Big]\}.$$

Or, on a d'une part $A_{I,I} = 4$, et d'autre part $\sum_{J\neq I} |A_{I,J}| \leq 4$, pour tout I. Ainsi, les segments $\Big[A_{I,I} - \sum_{J\neq I} |A_{I,J}|, A_{I,I} + \sum_{J\neq I} |A_{I,J}|\Big]$ sont tous inclus dans $\mathbb{R}^+$, et $\lambda(A) \geq 0$. La matrice A est donc symétrique et positive.
Si on prouve en plus que A est inversible, aucune valeur propre n'est nulle, *id est* elles sont toutes strictement positives, et A est symétrique définie-positive... Ce résultat est un corollaire immédiat (voir le corollaire 15.3.1) du théorème ci-dessous.

Théorème 15.3.1 *La matrice A des systèmes linéaires (15.16) et (15.17) est monotone.*

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

$$4v_{1,1} - v_{2,1} - v_{1,2} \geq 0 \tag{15.19}$$
$$4v_{n,1} - v_{n-1,1} - v_{n,2} \geq 0 \tag{15.20}$$
$$4v_{1,n} - v_{1,n-1} - v_{2,n} \geq 0 \tag{15.21}$$
$$4v_{n,n} - v_{n,n-1} - v_{n-1,n} \geq 0 \tag{15.22}$$
$$4v_{i,1} - v_{i-1,1} - v_{i+1,1} - v_{i,2} \geq 0 \quad 2 \leq i \leq n-1 \tag{15.23}$$
$$4v_{1,j} - v_{1,j-1} - v_{2,j} - v_{1,j+1} \geq 0 \quad 2 \leq j \leq n-1 \tag{15.24}$$
$$4v_{n,j} - v_{n,j-1} - v_{n-1,j} - v_{n,j+1} \geq 0 \quad 2 \leq j \leq n-1 \tag{15.25}$$
$$4v_{i,n} - v_{i,n-1} - v_{i-1,n} - v_{i+1,n} \geq 0 \quad 2 \leq i \leq n-1 \tag{15.26}$$
$$4v_{i,j} - v_{i,j-1} - v_{i-1,j} - v_{i+1,j} - v_{i,j+1} \geq 0 \quad 2 \leq i, j \leq n-1. \tag{15.27}$$

Ci-dessus, on a isolé des lignes de la matrice A correspondant respectivement

1. En (15.19)-(15.22) :
   aux *coins* de la grille, d'indices (i,j) parmi $\{(1,1),(n,1),(1,n),(n,n)\}$.
2. En (15.23)-(15.26) :
   aux *côtés* de la grille, coins exclus, d'indices (i,j) avec $i \in \{1,n\}$ ou (exclusif) $j \in \{1,n\}$.
3. En (15.27) :
   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 15.2.3. Comme la grille comprend des points aux coins, sur les côtés et intérieurs, on traite les trois cas correspondants.

1. Si v_{min} correspond à un coin (par exemple $v_{min} = v_{1,1}$), (15.19) 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 d'un côté différent d'un coin (par exemple $v_{min} = v_{i,1}$, $i \in \{2,\cdots,n-1\}$ donné), (15.23) 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, (15.27) fournit l'inégalité
$$(v_{min} - v_{i,j-1}) + (v_{min} - v_{i-1,j}) + (v_{min} - v_{i+1,j}) + (v_{min} - v_{i,j+1}) \geq 0.$$
   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 15.2.4), 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 des côtés (hors coins) sont compris, et le cas 2. s'applique.

■

D'après la définition de la monotonie d'une matrice, on en déduit comme annoncé le

Corollaire 15.3.1 *La matrice A des systèmes linéaires (15.16) et (15.17) est symétrique définie-positive.*

On peut donc se servir de l'algorithme du Gradient Conjugué pour résoudre les systèmes linéaires (15.16) et (15.17), et calculer ainsi une approximation de la solution u du problème 2D (15.12)...

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[1]. Cependant, à l'aide d'autres techniques relativement lourdes à mettre en œuvre (cf. [7]), on peut malgré tout prouver le

1. En effet, si $f \equiv 1$, quelle est la solution du problème (15.12), quelle est sa régularité, avec par exemple $u_d \equiv 0$?

Théorème 15.3.2 *Lorsque la solution u est de classe $\mathcal{C}^4([0,1]^2)$, l'erreur est telle que*

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

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

Bref, *"tout est bien qui finit bien"* (librement traduit de [25]...).

Pour le cas 3D du problème (15.11), nous proposons pour finir l'exercice ci-dessous.

Exercice 15.3.1 *Suggérer une généralisation de l'approche du problème 2D au cas 3D, pour résoudre le problème du potentiel électrostatique (15.11).*
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 A ainsi obtenue :

- *A est-elle positive ?*
- *A est-elle monotone ?*
- *A est-elle définie-positive ?*

Chapitre 16

Mise en œuvre pratique

16.1 Introduction

Après la revue théorique de quelques algorithmes parmi les plus utilisés en calcul scientifique, il est temps de se demander comment les utiliser concrètement, et voir quels sont les avantages et inconvénients respectifs de ces méthodes quand on les implémente dans un ordinateur. Il est impossible à l'intérieur de ce polycopié de détailler toutes les variantes existantes de ces algorithmes, car pour en améliorer l'efficacité on doit en écrire une version spécifique non seulement à chaque classe de problèmes traités mais encore particulière à chaque type d'ordinateur (séquentiel, vectoriel, parallèle), à chaque architecture (mémoire locale ou partagée, réseau de processeurs en grille ou en hypercube, grappe de stations) et même à chaque langage de programmation (Fortran 90, C++) ou environnement (P.V.M., M.P.I.), la liste est infinie...

Nous allons donc montrer rapidement les principales difficultés qui apparaissent dès que l'on doit utiliser l'une quelconque de ces méthodes dans le but de simuler numériquement les grands problèmes physiques que doivent résoudre les ingénieurs, tels que :

- le calcul d'écoulements fluides autour d'obstacles (voitures, ailes d'avion), ou à l'intérieur de cavités (moteurs à explosion, circuits de refroidissement de réacteurs nucléaires, organes du corps humain) ;

- le calcul de contraintes à l'intérieur de structures (pièces mécaniques, gratte-ciel, prothèses médicales) ;

- l'étude des vibrations (ponts, fusées, acoustique) ;

  et bien d'autres encore !

Pour simplifier l'exposé, et également pour se ramener à un niveau de complexité adéquat, on se limitera dans ce chapitre aux difficultés liées à la structure des matrices.

16.2 Structure des matrices

Au chapitre 15 (voir (15.12)), on a introduit l'équation de Laplace dans le carré $\Omega =]0,1[^2$

$$\begin{cases} -\Delta u = f \text{ dans } \Omega \\ u = u_d \text{ sur } \partial\Omega \end{cases} \tag{16.1}$$

où f et u_d sont des données. Lorsque ce problème est approché par la méthode des différences finies, avec un pas de maillage constant $h = 1/(n+1)$ dans chaque direction, le système linéaire

résultant possède n^2 inconnues. En numérotant les inconnues de manière "naturelle", ligne par ligne, chacune des lignes du système linéaire résultant s'écrit

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

La matrice du système linéaire est donc tridiagonale par blocs, de la forme

$$A = \frac{1}{h^2} \begin{pmatrix} B & -I & \dots & \dots & 0 \\ -I & B & -I & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -I \\ 0 & \dots & \dots & -I & B \end{pmatrix}$$

où la matrice B de rang n est elle-même tridiagonale (par points) :

$$B = \begin{pmatrix} 4 & -1 & 0 & \ddots & 0 \\ -1 & 4 & -1 & \ddots & \ddots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots & -1 \\ 0 & \ddots & \ddots & -1 & 4 \end{pmatrix}.$$

Ce qu'il faut retenir de cet exemple, c'est le **caractère creux** de la matrice A : il y a au plus cinq coefficients non nuls dans une ligne, et cela quel que soit le nombre de points dans le carré ! Cette propriété est propre à la méthode d'approximation et on la retrouve dans toutes les matrices provenant de la méthode des différences finies ou des éléments finis (cf. [4]), quel que soit le problème traité.

Considérons par exemple un problème moins académique : le problème d'élasticité linéaire en dimension deux. Il s'agit de calculer la déformation verticale d'une plaque rectangulaire horizontale, soumise à des forces de gravité. Sans entrer dans le détail - inutile pour notre propos - de la formulation du problème physique, rappelons qu'il s'agit de calculer en *chacun des points* d'un maillage T_h, les deux composantes du vecteur déplacement $\mathbf{u}^h = u_1^h \vec{e}_1 + u_2^h \vec{e}_2$. Pour ce type de discrétisation, la matrice ainsi construite est *symétrique*. On l'appelle A dans la suite.

La figure 16.1 représente le maillage éléments finis T_h avec la numérotation initiale des nœuds.

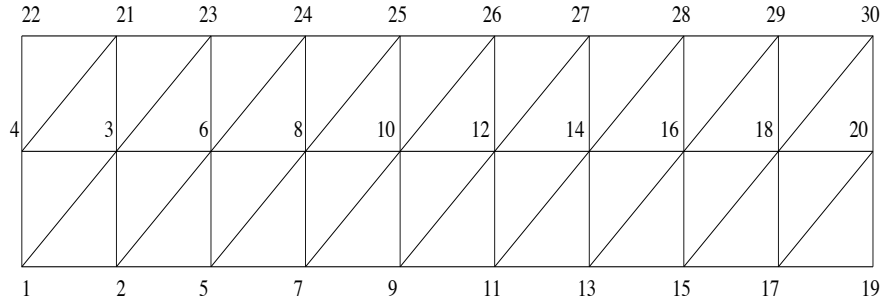


FIG. 16.1 – *La numérotation initiale*

Remarque 16.2.1 *Il y a deux fois plus d'inconnues que de points car à chaque nœud (point) de T_h sont associées les deux composantes du déplacement $\mathbf{u}^h$. Dans le système linéaire résultant,*

la numérotation des inconnues est liée à celle des nœuds du maillage de la façon suivante : au nœud N_k, de numéro k, sont associées les inconnues de numéros $2k-1$ et $2k$ représentant les deux composantes cherchées : $u_1^h(N_k)$ et $u_2^h(N_k)$. Le caractère creux de la matrice est bien visible sur la figure 16.2.

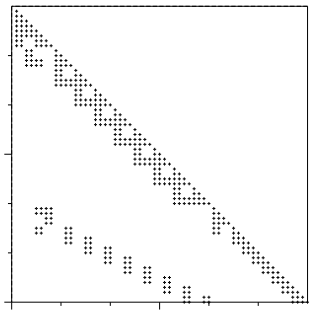


FIG. 16.2 – *La matrice associée*

Notons que l'on a sur cet exemple 60 inconnues (deux composantes de $\mathbf{u}^h$ sur 30 nœuds), ce qui correspond donc à une matrice symétrique de $\mathbb{R}^{60\times 60}$. Pour pouvoir reconstruire une telle matrice, il est nécessaire et suffisant de connaître sa diagonale, ainsi que sa partie triangulaire inférieure, soit un nombre total d'éléments égal à

$$1+2+\cdots+59+60=\frac{60\times 61}{2}=1830.$$

Après construction, on trouve effectivement 350 éléments non nuls dans cette partie de la matrice. Rappelons que le profil d'une matrice a été introduit à la proposition 6.19.1.

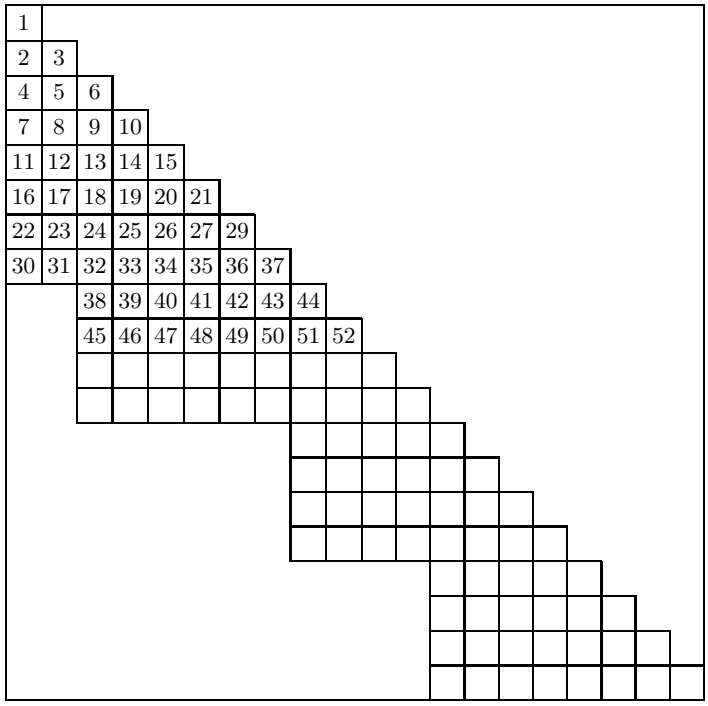
Définition 16.2.1 *La taille d'une ligne de la matrice A dans le profil est appelée* **largeur de bande** *de la ligne.*
La **longueur du profil** *est égale à la place mémoire nécessaire à la représentation de la matrice dans un ordinateur, optimisée au sens où on limite le stockage aux seuls éléments* internes *au profil. Elle est égale à la somme des largeurs de bande.*

Dans le cas qui nous intéresse, la longueur du profil de la matrice du système linéaire associé à la numérotation *initiale* de la la figure 16.1 est 926. D'après ce que l'on a vu, elle est comprise entre 350 et 1830 : 2,65 fois le minimum, et 50,6% du maximum (ou un gain de 49,4%).

16.3 Stockage profil

Pour préciser les idées montrons un exemple de rangement des coefficients de la matrice A dans un stockage profil (il s'agit du cas symétrique). La matrice A est représentée dans l'ordinateur par un tableau monodimensionnel noté PfA, de la façon suivante : pour chaque ligne i, on garde tous les coefficients compris entre $A_{i,il(i)}$ et $A_{i,i}$ (qu'ils soient nuls ou non) avec $il(i)$ le plus petit indice de colonne j tel que $A_{i,j}\neq 0$. Ces lignes sont alors rangées par ordre croissant dans PfA (figure 16.3).

Pour gérer la correspondance entre PfA et A, un tableau PdA est nécessaire, qui donne, en PdA$(i+1)$, l'adresse dans PfA du coefficient diagonal $A_{i,i}$ (avec la convention PdA(1)=0).



PdA : 0 1 3 6 10 15 21 29 37 44 52 ...

FIG. 16.3 – *Le stockage profil pour une matrice symétrique*

Exercice 16.3.1 *Montrer que l'adresse ia dans PfA d'un coefficient quelconque $A_{i,j}$ (avec $i \geq j$) est obtenue par la formule*

$$ia = \begin{cases} PdA(i+1) - i + j & si\ j \geq il(i) \\ \emptyset & sinon \end{cases} .$$

Ce mode de stockage appelé **stockage profil** (*skyline en anglais*), a été introduit par Jennings [14]. Il est particulièrement commode puisqu'il est utilisable pour stocker la matrice A ou la matrice L facteur de Cholesky de A ; en effet on peut stocker A et L dans le même tableau, en remplaçant les valeurs des coefficients de la matrice A par celles des coefficients de la matrice L, au fur et à mesure du calcul de ces derniers, en conservant le même profil (voir la Proposition 6.19.1). On dit que la matrice L "écrase" la matrice A.

16.4 Numérotation et stockage profil

La longueur du profil détermine la place mémoire occupée par la matrice A dans l'ordinateur ; afin de la diminuer, analysons plus précisément la structure de la matrice A. Comme A est une matrice issue d'une discrétisation par éléments finis, un coefficient $A_{k,l}$ ne peut être différent de zéro que si les nœuds associés N_k et N_l appartiennent à un même élément de T_h (cf. [4]) ; les nœuds $(N_l)_l$ ainsi associés au nœud N_k sont appelés **voisins** de N_k.

La largeur de bande de la ligne k dépend donc du nombre de nœuds $(N_l)_l$ voisins du nœud N_k, mais surtout de leurs numéros car plus grand est l'écart $\max_{N_l \text{voisin de } N_k} |k-l|$, plus grande est la largeur de bande

$$lp(k) = 2 \max_{N_l \text{voisin de } N_k} |k-l|.$$

(le facteur multiplicatif deux provient de la présence de deux inconnues par nœud, cf. la remarque 16.2.1.)

Pour diminuer la largeur de bande $lp(k)$ il convient donc de donner aux voisins du nœud N_k des numéros les plus proches possibles de k. En appliquant cette idée à tous les nœuds du maillage T_h, on va essayer de diminuer la longueur totale du profil.

Le procédé utilisé est appelé méthode de renumérotation de Cuthill et McKee [8]. Son principe est le suivant : partant d'un nœud donné, à qui l'on attribue le numéro 1, on numérote d'abord ses voisins (au sens précédent) puis on réitère le procédé en reprenant un par un chacun des nœuds nouvellement numérotés et en recherchant parmi ses voisins ceux qui ne possédent pas encore de numéro. L'ensemble des nœuds à numéroter à chaque étape constitue le front de renumérotation, qui parcourt le domaine au cours des itérations. Les nœuds du front sont renumérotés suivant leur nombre de voisins (non numérotés) croissant pour limiter l'accroissement du front.

La figure 16.4 montre l'application de cette méthode au maillage précédent ; le sommet de départ est le sommet inférieur gauche du rectangle.

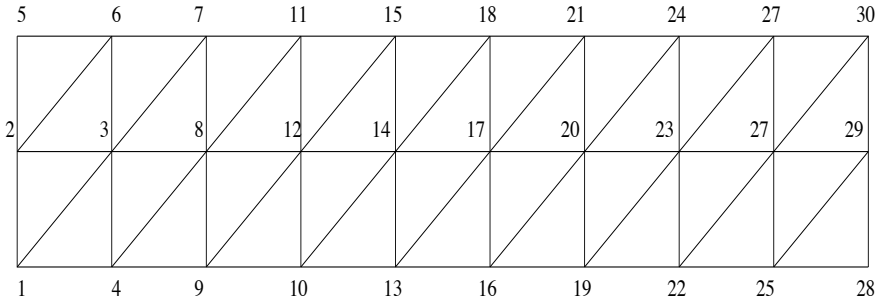


FIG. 16.4 – *La numérotation Cuthill-McKee*

Sur la figure 16.5, représentant la matrice A on constate bien, après renumérotation, une diminution de la largeur de bande, puisque la longueur du profil est maintenant 514 : 1,47 fois le minimum, et 28,1% du maximum (ou un gain de 71.9%).

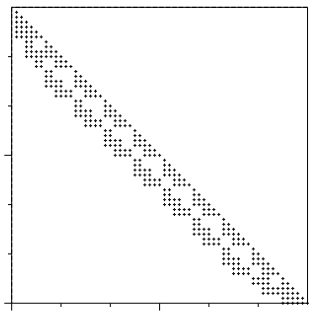


FIG. 16.5 – *La matrice associée*

Le principe de la renumérotation de Cuthill-McKee s'applique à toutes les matrices issues de la méthode des éléments finis, mais aussi dans un contexte plus général.

La recherche d'une longueur de profil minimum repose implicitement sur le fait que l'on représente la matrice dans l'ordinateur à l'aide du stockage profil, en vue de résoudre le système

linéaire associé par la méthode de Cholesky. Le stockage profil est bien adapté à une résolution par une méthode *directe*.

Si par contre on envisage de résoudre le système linéaire par une méthode *itérative*, comme par exemple la méthode du gradient conjugué, on doit avoir une autre approche.

16.5 Stockage condensé

Reprenons donc le cas d'un système linéaire à résoudre, écrit sous la forme $Ax = b$, avec A symétrique et *creuse*.

Dans toutes les méthodes itératives on doit calculer le résidu du système linéaire $r^k = b - Ax^k$, ce qui implique le calcul effectif d'un produit matrice-vecteur. Si on examine précisément les opérations nécessaires à ce calcul, on remarque que chaque composante du vecteur produit $p = Av$ est obtenue par la formule

$$p_i = \sum_{k=1}^{n} A_{i,k}\, v_k \quad \text{pour } 1 \leq i \leq n.$$

Etant donné le caractère creux de la matrice A, il est astucieux de restreindre la somme considérée aux seuls coefficients non nuls de la matrice A.

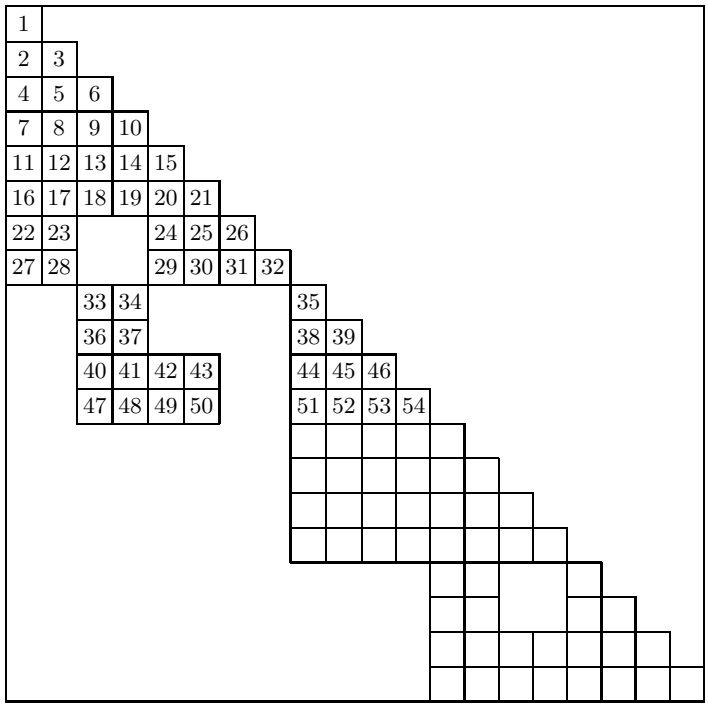
Il suffit donc de stocker la matrice A dans un tableau monodimensionnel CdA de la façon suivante : pour chaque ligne i, on ne garde que les coefficients $A_{i,j}$ non nuls (au nombre de 700 [1] dans notre exemple), rangés par indice de colonne j croissant. Les lignes de la matrice sont alors rangées par ordre croissant dans le tableau CdA (voir figure 16.6).

Pour gérer la correspondance entre CdA et A, un tableau PlA est nécessaire, qui donne pour chaque ligne i l'adresse dans CdA du dernier coefficient non nul $A_{i,j}$. Enfin un tableau PcA de même taille que CdA, et rangé de manière cohérente donne pour chaque coefficient $A_{i,j}$ l'indice de colonne j correspondant. On ne stocke ainsi que les coefficients non nuls de la matrice A. L'adresse ia dans CdA d'un coefficient quelconque $A_{i,j}$ n'est pas accessible directement, car il faut rechercher l'indice de colonne j dans PcA entre les adresses PlA(i)+1 et PlA(i+1) (avec la convention PlA(1)=0).

Ce mode de stockage, appelé **stockage condensé**, est donc bien adapté au calcul du produit d'une matrice par un vecteur, puisque l'on n'effectue pas de produits par des coefficients nuls, comme dans le cas du stockage profil. Par contre il n'est pas adapté à la factorisation de Cholesky puisqu'il n'y a pas de place réservée *a priori* aux coefficients de remplissage qui vont apparaitre en cours de factorisation : les matrices A et L n'ont pas la même représentation en stockage condensé !

Noter donc que chaque type de stockage impose une méthodologie de calcul, qui peut être très différente des formules algébriques associées à la théorie ; ainsi pour le stockage condensé symétrique le produit $p = Av$ est effectué à l'aide de l'algorithme suivant :

1. 350 coefficients et 350 positions...



PlA : 0 1 3 6 10 15 21 26 32 35 39

PcA : 1 1 2 1 2 3 1 2 3 4 1 2 3 4 5 1 2 3 4 5 6

1 2 5 6 7 1 2 5 6 7 8 3 4 9 3 4 9 10 . . .

FIG. 16.6 – *Le stockage condensé pour une matrice symétrique*

Boucle sur les inconnues :
Pour $i = 1$ à n faire
$p(i) = 0$
Fin de la boucle sur les inconnues
Boucle sur les inconnues :
Pour $i = 1$ à n faire
Boucle sur les colonnes de la ligne i :
Pour j = PlA(i) +1 à PlA(i+1) faire
$p(i) = p(i) + \text{CdA}(j) \times v(\text{PcA}(j))$
Fin de la boucle sur les colonnes
Boucle sur les lignes de la colonne i :
Pour j = PlA(i) +1 à PlA(i+1)-1 faire
$p(\text{PcA}(j)) = p(\text{PcA}(j)) + \text{CdA}(j) \times v(i)$
Fin de la boucle sur les lignes
Fin de la boucle sur les inconnues

Ce qu'il faut retenir

- les matrices obtenues dans le cadre de l'approximation de la solution d'une équation aux dérivées partielles par la méthode des différences finies ou des éléments finis sont creuses.

- cette propriété doit être exploitée lors de la résolution sur ordinateur des systèmes linéaires associés à ces problèmes, ce qui suppose un mode de stockage adapté.

- pour plus d'efficacité on distingue deux types de stockage, suivant que l'on utilise une méthode directe de résolution des systèmes linéaires (Gauss, Cholesky, Crout) ou une méthode itérative (SSOR, Gradient Conjugué symétrique et non symétrique).

  En conclusion, pour les méthodes directes on utilise le stockage profil, pour les méthodes itératives on utilise le stockage condensé.

Bibliographie

[1] **L. M. Adams, H. F. Jordan**, Is SOR color-blind?, *SIAM Journal on Scientific and Statistical Computing*, **7** (1986).

[2] **A. Björck**, *Solutions of equations in $\mathbb{R}^n$ (Part I): Least squares methods*, Handbook of numerical analysis, Volume I, éditeurs P. G. Ciarlet et J.–L. Lions, North Holland, Amsterdam (1990).

[3] **A.-S. Bonnet Ben Dhia, M. Lenoir**, *Outils élémentaires d'analyse pour les équations aux dérivées partielles*, Cours MA 102, ENSTA.

[4] **A.-S. Bonnet Ben Dhia, E. Luneville**, *Résolution numérique des équations aux dérivées partielles*, Cours MA 201, ENSTA.

[5] **F. Chatelin**, *Valeurs propres de matrices*, Masson, Paris (1988).

[6] **P. G. Ciarlet**, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris (1982).

[7] **P. G. Ciarlet, B. Miara, J.-M. Thomas**, *Exercices d'analyse numérique matricielle et d'optimisation*, Masson, Paris (1982).

[8] **E. Cuthill, J. McKee**, Reducing the bandwidth of sparse symmetric matrices, dans les *Proceedings of the ACM National Conference*, Academic Press, New-York (1969).

[9] **V. Faber, T. Manteuffel**, Necessary and sufficient conditions for the existence of a conjugate gradient method, *SIAM Journal on Numerical Analysis*, **21**, 352-362 (1984).

[10] **J.-C. Gilbert**, *Optimisation différentiable. Théorie et algorithmes*, Cours AO 201, ENSTA.

[11] **G. H. Golub, G. Meurant**, *Résolution numérique des grands systèmes linéaires*, Eyrolles, Paris (1983).

[12] **M. R. Hestenes, E. Stiefel**, Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards*, **49**, 409-436 (1952).

[13] **A. S. Householder**, *The theory of matrices in numerical analysis*, Blaisdell Publishing Company (1970).

[14] **A. Jennings**, A compact storage scheme for the solution of symmetric linear simultaneous equations, *Computing Journal*, **9** (1966).

[15] **R. Krikorian**, *Linéarisation et stabilité des équations différentielles*, Cours AO 102, ENSTA.

[16] **P. Lascaux, R. Théodor**, *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Masson, Paris (1984).

[17] **G. Meurant**, *Computer solution of large linear systems*, Studies in Mathematics and its Applications, **28**, North Holland, Amsterdam (1999).

[18] **J. Ortega, R. J. Plemmons**, Extension of the Ostrowski-Reich theorem for SOR iterations, *Linear Algebra and its Applications*, **28** (1987).

[19] **B. N. Parlett**, *The symmetric eigenvalue problem*, Prentice Hall, Englewood Cliffs (1980).

[20] **J. Pérez**, *Gravitation classique*, Enseignement thématique d'Astrophysique MAT 40, ENSTA.

[21] **E. Polak, G. Ribière**, Sur la convergence de la méthode des gradients conjugués, *Revue Française d'Informatique et de Recherche Opérationnelle*, **16**-R1 (1969).

[22] **A. Quarteroni, A. Valli**, *Domain decomposition methods for partial differential equations*, Oxford Science Publications, Oxford (1999).

[23] **Y. Saad**, *Numerical methods for large eigenvalue problems*, Manchester University Press, Manchester (1992).

[24] **Y. Saad, M. H. Schultz**, GMRES: a generalized minimum residual algorithm for solving nonsymmetric linear systems, *SIAM Journal on Scientific and Statistical Computing*, **7**, 856-869 (1986).

[25] **W. Shakespeare**, *Much ado about nothing* (ca. 1598).

[26] **B. F. Smith, P. E. Bjørstad, W. D. Gropp**, *Domain decomposition. Parallel multilevel methods for elliptic partial differential equations*, Cambridge University Press, New York (1996).

[27] **D. M. Young**, *Iterative solution of large linear systems*, Academic Press, New York (1971).

Index