ALGÈBRE LINÉAIRE ET ANALYSE NUMÉRIQUE MATRICIELLE

Yves Achdou[1]

23 novembre 2005

[1]UFR Mathématiques, Université Paris 7, Case 7012, 75251 PARIS Cedex 05, France and Laboratoire Jacques-Louis Lions, Université Paris 6. achdou@math.jussieu.fr

Table des matières

Chapitre 1

Rappels d'Algèbre Linéaire

On rappelle les résultats élémentaires (qu'il faut absolument connaître) en algèbre linéaire. On donne la majorité des énoncés sans démonstration : il est fortement conseillé au lecteur de faire toutes les démonstrations.

1.1 Espaces vectoriels

Définition 1.1 *Soit E et I deux ensembles : Une famille, notée $(x_i)_{i\in I}$, $x_i \in E$, d'éléments de E indexée par I est une application de I dans E.*

On désigne par $\mathbb{K}$ un corps commutatif. On prendra $\mathbb{K} = \mathbb{C}$ (corps des nombres complexes) ou $\mathbb{K} = \mathbb{R}$ (corps des nombres réels).

Définition 1.2 *Soit E un ensemble muni*
- *d'une loi de composition interne $+ : E \times E \to E$, $(u, v) \mapsto u + v$.*
- *d'une loi de composition externe $\cdot : \mathbb{K} \times E \to E$, $(\lambda, u) \mapsto \lambda \cdot u$.*

On dit que $(E, +, \cdot)$ est un espace vectoriel sur le corps $\mathbb{K}$ si

1. *$(E, +)$ est un groupe commutatif.*
2. *$\forall \lambda \in \mathbb{K}$, $\forall u, v \in E$, $\lambda \cdot (u + v) = \lambda \cdot u + \lambda \cdot v$.*
3. *$\forall \lambda, \mu \in \mathbb{K}$, $\forall u \in E$, $(\lambda + \mu) \cdot u = \lambda \cdot u + \mu \cdot u$.*
4. *$\forall \lambda, \mu \in \mathbb{K}$, $\forall u \in E$, $= \lambda \cdot (\mu \cdot u) = (\lambda\mu) \cdot u$.*
5. *$\forall u \in E$, $1 \cdot u = u$.*

On appelle vecteurs les éléments de E et scalaires les éléments de $\mathbb{K}$.

Dans la suite, on omettra la notation "$\cdot$" : $\lambda \cdot u = \lambda u$.

Définition 1.3 *Soient $(E, +, \cdot)$ un espace vectoriel sur le corps $\mathbb{K}$ et F un sous-ensemble de E. On dit que F est un sous-espace vectoriel de E si*

1. *F est non vide.*
2. *F est stable par $+ : \forall u, v \in F$, $u + v \in F$.*
3. *F est stable par $\cdot : \forall \lambda \in \mathbb{K}$, $\forall v \in F$, $\lambda v \in F$.*

Proposition 1.1 *Soient $(E, +, \cdot)$ un espace vectoriel sur le corps $\mathbb{K}$ et F un sous-ensemble de E. L'ensemble F est un sous-espace vectoriel de E si et seulement si $0 \in E$ et E est stable par combinaison linéaire, i.e. $\forall \lambda, \mu \in \mathbb{K}$, $\forall u, v \in F$, $\lambda u + \mu v \in F$.*

Proposition 1.2 *L'intersection d'une famille quelconque de sous-espaces vectoriels de E est un sous-espace vectoriel de E.*

Si F est un sous-espace vectoriel de E, alors $(F, +, \cdot)$ est un espace vectoriel.

Définition 1.4 *On appelle sous-espace vectoriel engendré par une famille (pas forcément finie) A de vecteurs de E et on note* $\mathrm{Vect}(A)$ *le sous-espace vectoriel formé des combinaisons linéaires (finies) des vecteurs de A :*

$$\mathrm{Vect}(A) = \left\{ x \in E : \text{ il existe } p \in \mathbb{N},\ x_1, \dots, x_p \in A, \quad \text{et } \lambda_1, \dots, \lambda_p \in \mathbb{K} \text{ tels que } x = \sum_{i=1}^{p} \lambda_i x_i \right\} \tag{1.1}$$

On dit qu'un vecteur de $\mathrm{Vect}(A)$ *est une combinaison linéaire des vecteurs de A .*

Exercice. Vérifier que $\mathrm{Vect}(A)$ est bien un sous-espace vectoriel de E.

Lemme 1.1 *Soit une famille A de vecteurs de E,* $\mathrm{Vect}(A)$ *est le plus petit sous-espace vectoriel qui contient A.*

Définition 1.5 *On dit qu'une famille A de vecteurs de E est génératrice si* $\mathrm{Vect}(A) = E$*, c'est à dire si pour tout $x \in E$, il existe $p \in \mathbb{N}$, $x_1, \dots, x_p \in A$, et $\lambda_1, \dots, \lambda_p \in \mathbb{K}$ tels que $x = \sum_{i=1}^{p} \lambda_i x_i$.*

Définition 1.6 *On dit qu'une famille A de vecteurs de E est libre si pour toute famille finie $(x_1, \dots, x_p)$ de vecteurs de A, distincts deux à deux, et pour toute famille de scalaires $(\lambda_1, \dots, \lambda_p)$, $\sum_{i=1}^{p} \lambda_i x_i = 0$ si et seulement si $\lambda_i = 0$ pour $i = 1, \dots, p$. On dit aussi que les vecteurs $x_1, \dots, x_p$ sont linéairement indépendants.*

Définition 1.7 *Une famille A de vecteurs de E à la fois libre et génératrice est appelée une base de E.*

Proposition 1.3 *Une famille A de vecteurs de E est une base si et seulement si, pour tout $x \in E$, il existe un unique $p \in \mathbb{N}$, une unique sous-famille de A avec p vecteurs $(x_1, \dots, x_p)$, et une unique famille $(\lambda_1, \dots, \lambda_p)$ de scalaire non nuls tels que $x = \sum_{i=1}^{p} \lambda_i x_i$.*

Exemple. Dans $\mathbb{K}^n$, on utilise très souvent la base dite *base canonique* : cette base est formée par les vecteurs e_i, $1 \leq i \leq n$:

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ \\ \dots i \\ \\ \\ \end{matrix} \begin{matrix} \dots 1 \\ \vdots \\ \dots i-1 \\ \\ \dots i+1 \\ \vdots \\ \dots n \end{matrix} \tag{1.2}$$

ou encore : la j-ème composante de e_i est δ_{ij}, (les δ_{ij} sont les symboles de Kronecker : $\delta_{ij} = 0$ si $j \neq i$ et $\delta_{ii} = 1$.)

Définition 1.8 *Soit k un entier positif. Soient $V_1, \ldots, V_k$ k sous-espaces vectoriels de E. On appelle somme des sous-espace $V_1, \ldots, V_k$ et on note $V_1 + \cdots + V_k$ le sous-espace vectoriel*

$$V_1 + \cdots + V_k = \text{Vect}\left(\bigcup_{i=1}^{k} V_i\right) \tag{1.3}$$

Proposition 1.4 *Soient $V_1, \ldots, V_k$ k sous-espaces vectoriels de E.*

$$V_1 + \cdots + V_k = \left\{ x \in E \ : \ \text{il existe } x_1 \in V_1, \ldots, x_k \in V_k \text{ tels que } x = \sum_{i=1}^{k} x_i \right\}. \tag{1.4}$$

Définition 1.9 *Si pour tout $x \in V_1 + \ldots V_k$ la décomposition $x = \sum_{i=1}^{k} x_i$, avec $x_i \in V_i$, est unique, alors on dit alors que les sous-espaces $V_1, \ldots, V_k$ sont en somme directe et on utilise la notation $\bigoplus_{i=1}^{k} V_i$ ou encore $V_1 \oplus \cdots \oplus V_k$ à la place de la notation $V_1 + \ldots V_k$.*

Proposition 1.5 *Pour que les sous-espaces V_1 et V_2 de E soient en somme directe, il faut et il suffit que*

$$V_1 \cap V_2 = \{0\}. \tag{1.5}$$

Définition 1.10 *On dit que les sous-espaces V_1 et V_2 de E sont supplémentaires si et seulement si $V_1 \oplus V_2 = E$ ou de manière équivalente si*

$$V_1 + V_2 = E \quad et \quad V_1 \cap V_2 = \{0\}. \tag{1.6}$$

Proposition 1.6 (fondamentale) *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ admettant une famille génératrice G de k vecteurs. Alors une famille libre F de vecteurs de E a un nombre fini ℓ de vecteurs avec $\ell \leq k$, et on peut obtenir une nouvelle famille génératrice de E en échangeant ℓ vecteurs de G avec ceux de F.*

Démonstration. Par récurrence sur le nombre de vecteurs de F quand ce nombre est $\leq k$. Après, on montre qu'on ne peut pas avoir une famille libre avec $k+1$ vecteurs. ■

Théorème 1.1 (de la base incomplète) *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ admettant une famille génératrice finie G. Soit L une famille libre de vecteurs de E. Alors il existe une sous-famille G' de G telle que $L \cup G'$ soit une base de E.*

Corollaire 1.1 *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ admettant une famille génératrice finie. Alors il existe au moins une base de E avec un nombre fini de vecteurs et toutes les bases de E ont le même nombre de vecteurs. Ce nombre noté $\dim(E)$, est appelé la dimension de E et on dit que E est de dimension finie.*

Proposition 1.7 *Si E est de dimension finie,*

- *toute famille génératrice de E a au moins $\dim(E)$ vecteurs.*
- *toute famille libre de E a au plus $\dim(E)$ vecteurs.*
- *une famille libre de E comportant $\dim(E)$ vecteurs est une base de E.*
- *une famille génératrice de E comportant $\dim(E)$ vecteurs est une base de E.*
- *tout sous-espace F de E est de dimension finie inférieure ou égale à $\dim(E)$. Si $\dim(F) = \dim(E)$ alors $F = E$.*

– *soient deux sous-espaces F et G de E. On a*

$$\dim(F+G) = \dim(F) + \dim(G) - \dim(F \cap G).$$

En particulier, soient deux sous-espaces F et G tels que $F \oplus G = E$. On a

$$\dim(E) = \dim(F) + \dim(G).$$

Soient k sous-espaces vectoriels $F_1, \ldots, F_k$ de E tels que $F_1 \oplus \cdots \oplus F_k = E$, on a

$$\dim(E) = \sum_{i=1}^{k} \dim(F_i).$$

Définition 1.11 *Si un espace vectoriel E sur $\mathbb{K}$ n'admet pas de famille génératrice finie ou de manière équivalente, si pour tout $p \in \mathbb{N}$, il existe une famille libre avec p vecteurs de E, alors on dit que E est dimension infinie.*

Dans ce cours qui est surtout consacré à l'analyse numérique matricielle, il sera peu question d'espaces vectoriels de dimension infinie.

1.2 Applications Linéaires

Définition 1.12 *Soient $(E,+,\cdot)$ et $(F,+,\cdot)$ deux espaces vectoriels sur $\mathbb{K}$. On dit que l'application ℓ de E dans F est une application linéaire de E dans F si pour tout $\lambda, \mu \in \mathbb{K}$, pour tout $u, v \in E$,*

$$\ell(\lambda u + \mu v) = \lambda \ell(u) + \mu \ell(v). \tag{1.7}$$

On note $\mathcal{L}(E,F)$ l'ensemble des applications linéaires de E dans F. On utilise la notation plus courte $\mathcal{L}(E)$ pour $\mathcal{L}(E,E)$.

Exemple. L'application identité de E noté id_E est une application linéaire de E dans E.

Proposition 1.8 *Soient $(E,+,\cdot)$, $(F,+,\cdot)$ et $(G,+,\cdot)$ trois espaces vectoriels sur $\mathbb{K}$, $\ell \in \mathcal{L}(E,F)$, et $\jmath \in \mathcal{L}(F,G)$. La composée $\jmath \circ \ell$ est une application linéaire de E dans F : $\jmath \circ \ell \in \mathcal{L}(E,G)$.*

Proposition 1.9 *L'ensemble des applications de E dans F, noté F^E est un espace vectoriel, quand on le munit de*
- *l'addition d'applications*
- *le produit par un scalaire.*

$\mathcal{L}(E,F)$ est un sous-espace vectoriel de $(F^E,+,\cdot)$, l'espace des applications de E dans F.

Définition 1.13 *Une application linéaire de E dans $\mathbb{K}$ est appelée une forme linéaire.*

Définition 1.14 *Soit ℓ une une application linéaire de E dans F. On appelle noyau de ℓ l'image réciproque de $\{0_F\}$ par ℓ. C'est un sous-espace vectoriel de E. Le noyau de ℓ est noté* $\ker(\ell)$, *(du mot anglais kernel (ou allemand kern) qui veut dire noyau).*

$$\ker(\ell) = \{x \in E :\ \ell(x) = 0_F\} \tag{1.8}$$

Définition 1.15 *Soit ℓ une une application linéaire de E dans F. L'image de ℓ (c'est à dire l'image de E par ℓ) est un sous-espace vectoriel de F. Elle est notée* $\mathrm{Im}(\ell)$.

$$\mathrm{Im}(\ell) = \{\ell(x),\ x \in E\} \tag{1.9}$$

Définition 1.16 *Soit ℓ une une application linéaire de E dans F. Le rang de ℓ, noté* $\mathrm{rang}(\ell)$ *est la dimension de* $\mathrm{Im}(\ell)$ *(éventuellement infinie).*

Soit ℓ une une application linéaire de E dans F. Il est clair que ℓ est injective si et seulement si $\ker(\ell) = \{0_E\}$.

Proposition 1.10 *Soit ℓ une une application linéaire de E dans F. L'application ℓ est un isomorphisme de E sur F, c'est à dire une application linéaire bijective de E sur F (à la fois injective et surjective) si et seulement si il existe une application linéaire notée ℓ^{-1} de F dans E telle que $\ell \circ \ell^{-1} = \mathrm{id}_F$ et $\ell^{-1} \circ \ell = \mathrm{id}_E$.*

Dans le cas où E est de dimension finie, on a le théorème du rang

Théorème 1.2 (du rang) *Soient $(E, +, \cdot)$ et $(F, +, \cdot)$ deux espaces vectoriels sur $\mathbb{K}$, E étant de dimension finie. Soit ℓ une application linéaire de E dans F. On a l'identité*

$$\mathrm{rang}(\ell) + \dim(\ker(\ell)) = \dim(E). \tag{1.10}$$

Du Théorème du rang, on déduit en particulier que si E et F sont de dimension finie, une condition nécessaire et suffisante pour qu'il existe un isomorphisme de E sur F est que $\dim(E) = \dim(F)$.

Corollaire 1.2 *Soient $(E, +, \cdot)$ et $(F, +, \cdot)$ des espaces vectoriels sur $\mathbb{K}$ de dimension finie et tels que* $\dim(E) = \dim(F)$. *Soit ℓ une une application linéaire de E dans F. Alors les assertions suivantes sont équivalentes*

- *ℓ est injective*
- *ℓ est surjective*
- *ℓ est un isomorphisme de E sur F.*
- *ℓ admet un inverse à gauche : il existe $\jmath \in \mathcal{L}(F, E)$ telle que $\jmath \circ \ell = \mathrm{id}_F$*
- *ℓ admet un inverse à droite : il existe $\jmath \in \mathcal{L}(F, E)$ telle que $\ell \circ \jmath = \mathrm{id}_E$.*

Si ℓ admet un inverse à droite (ou à gauche) $\jmath$, alors ℓ est inversible et $\jmath = \ell^{-1}$.

Le corollaire 1.2 s'applique particulièrement si $E = F$, et un endomorphisme de l'espace vectoriel E inversible est appelé un *automorphisme* de E.
Donnons quelques résultats sur les formes linéaires :

Lemme 1.2 *Soient $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ de dimension finie n. Soit ℓ une forme linéaire non nulle sur E. Le noyau de ℓ est un sous-espace de E de dimension $n-1$ (on dit un hyperplan de E).*
Réciproquement, pour tout hyperplan F de E, on peut trouver une forme linéaire ℓ sur E dont le noyau est F.

1.3 Matrices

Définition 1.17 *Soient n et m deux entiers strictement positifs. On appelle matrice à n lignes et m colonnes à coefficients dans $\mathbb{K}$, un tableau à double entrée (a_{ij}), $1 \leq i \leq n, 1 \leq j \leq m$ (un tableau à double entrée étant une application de $\{1,\dots,n\} \times \{1,\dots,m\}$ dans $\mathbb{K}$). Le premier indice est l'indice de ligne et le deuxième indice est l'indice de colonne. Une matrice A à n lignes et m colonnes à coefficients dans $\mathbb{K}$ est donc de la forme*

$$A = \begin{pmatrix} a_{11} & \dots & a_{1,j} & \dots & a_{1m} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{i,j} & \dots & a_{im} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{n,j} & \dots & a_{nm} \end{pmatrix}$$

On note $\mathcal{M}_{n,m}(\mathbb{K})$ l'ensemble des matrices à n lignes et m colonnes à coefficients dans $\mathbb{K}$. Pour abréger, on note $\mathcal{M}_n(\mathbb{K})$ l'ensemble des matrices carrées à n lignes et n colonnes à coefficients dans $\mathbb{K}$. Les matrices de $\mathcal{M}_n(\mathbb{K})$ sont appelées matrices carrées d'ordre n.

Convention 1.1 *On identifiera les matrices colonnes de $\mathcal{M}_{n,1}(\mathbb{K})$ aux vecteurs de $\mathbb{K}^n$.*

Remarque 1.1 *On pourra aussi employer la notation $\mathbb{K}^{n\times m}$ pour $\mathcal{M}_{n,m}(\mathbb{K})$.*

Définition 1.18 *Soient $A = (a_{ij})$ et $B = (b_{ij})$ deux matrices de $\mathcal{M}_{n,m}(\mathbb{K})$. On définit la somme des matrices A et B et on note $A + B$ la matrices dont le coefficient situé sur la i-ème ligne et la j-ème colonne est $a_{ij} + b_{ij}$, $i = 1,\dots,n$, $j = 1,\dots,m$.*
Soient $A = (a_{ij})$ une matrice de $\mathcal{M}_{n,m}(\mathbb{K})$ et $\lambda \in \mathbb{K}$ un scalaire. On définit le produit de la matrice A par le scalaire λ et on note λA la matrices dont le coefficient situé sur la i-ème ligne et la j-ème colonne est λa_{ij}, $i = 1,\dots,n$, $j = 1,\dots,m$.

Proposition 1.11 *$(\mathcal{M}_{n,m}(\mathbb{K}), +, \cdot)$ est un espace vectoriel sur $\mathbb{K}$ de dimension nm.*

Définition 1.19 *Soient m, n, p trois entiers naturels strictement positifs. Soient $A = (a_{ij})$ une matrice de $\mathcal{M}_{m,n}(\mathbb{K})$ et $B = (b_{ij})$ une matrice de $\mathcal{M}_{n,p}(\mathbb{K})$. On définit le produit AB comme la matrice de $\mathcal{M}_{m,p}(\mathbb{K})$ dont le coefficient situé sur la i-ème ligne et la j-ème colonne est $\sum_{k=1}^{n} a_{ik}b_{kj}$, $i = 1,\dots,m$, $j = 1,\dots,p$.*
On vient de définir une loi de composition appelée produit matriciel.

Proposition 1.12
- *Le produit matriciel est associatif : soient $A \in \mathcal{M}_{m,n}(\mathbb{K})$, $B \in \mathcal{M}_{n,p}(\mathbb{K})$, $C \in \mathcal{M}_{p,q}(\mathbb{K})$.*
$$A(BC) = (AB)C$$
- *Le produit matriciel est distributif à gauche sur l'addition : soient $A \in \mathcal{M}_{m,n}(\mathbb{K})$, $B_1, B_2 \in \mathcal{M}_{n,p}(\mathbb{K})$,*
$$A(B_1 + B_2) = AB_1 + AB_2$$
- *Le produit matriciel est distributif à droite sur l'addition : soient $B \in \mathcal{M}_{n,p}(\mathbb{K})$, $A_1, A_2 \in \mathcal{M}_{m,n}(\mathbb{K})$,*
$$(A_1 + A_2)B = A_1B + A_2B$$

– *Soient* $\lambda \in \mathbb{K}$, $A \in \mathcal{M}_{m,n}(\mathbb{K})$ *et* $B \in \mathcal{M}_{n,p}(\mathbb{K})$ *:*

$$(\lambda A)B = A(\lambda B) = \lambda(AB) = \lambda AB$$

– *L'élément neutre pour le produit dans* $\mathcal{M}_n(\mathbb{K})$ *est la matrice identité notée* I *ou* I_n *:*

$$I = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

Remarque 1.2 *Dans* $\mathcal{M}_n(\mathbb{K})$*, le produit matriciel n'est pas commutatif, i.e. pour des matrices* A *et* B *de* $\mathcal{M}_n(\mathbb{K})$*, on a en général* $AB \neq BA$*. Exercice : donner des exemples.*

Définition 1.20 *On dit que deux matrices* $A \in \mathcal{M}_n(\mathbb{K})$ *et* $B \in \mathcal{M}_n(\mathbb{K})$ *commutent si* $AB = BA$*.*

Exemple. Soit $A \in \mathcal{M}_n(\mathbb{K})$. Les puissances de A commutent entre elles :

$$\forall p, q \in \mathbb{N}, \quad A^p A^q = A^q A^p = A^{p+q}$$

où l'on pose $A^0 = I_n$, et $A^p = \underbrace{A \times \cdots \times A}_{p}$.

Définition 1.21 *Soient* $d_1, \dots, d_n$ *des scalaires. On notera* $\mathrm{Diag}(d_1, \dots, d_n)$ *la matrice de* $D \in \mathcal{M}_n(\mathbb{K})$*, dont les coefficients sont*

$$D_{ij} = d_i \delta_{ij}, \quad 1 \leq i, j \leq n.$$

Une telle matrice est appelée matrice diagonale.

Proposition 1.13 *Soit* $D = \mathrm{Diag}(d_1, \dots, d_n)$ *une matrice diagonale de* $\mathcal{M}_n(\mathbb{K})$*.*
– *Soit* A *une matrice de* $\mathcal{M}_{n,m}(\mathbb{K})$*. La matrice* $DA \in \mathcal{M}_{n,m}(\mathbb{K})$ *est obtenue en multipliant les coefficients de la* i*ème ligne de* A *par* d_i*, pour* $i = 1, \dots, n$*.*
– *Soit* A *une matrice de* $\mathcal{M}_{m,n}(\mathbb{K})$*. La matrice* $AD \in \mathcal{M}_{m,n}(\mathbb{K})$ *est obtenue en multipliant les coefficients de la* j*ème colonne de* A *par* d_j*, pour* $j = 1, \dots, n$*.*

Définition 1.22 *On appelle permutation d'ordre* n *une bijection de* $\{1, \dots, n\}$ *sur* $\{1, \dots, n\}$*. L'ensemble des permutations d'ordre* n *compte* $n!$ *éléments.*

Définition 1.23 *Soit* σ *une permutation d'ordre* n*, on associe à* σ *une matrice* $P^\sigma \in \mathcal{M}_n(\mathbb{K})$ *:*

$$P^\sigma_{ij} = \delta_{\sigma(i)j},$$

c'est à dire

$$\begin{array}{rcll} P^\sigma_{ij} & = & 1 & si \quad j = \sigma(i), \\ P^\sigma_{ij} & = & 0 & si \quad j \neq \sigma(i), \end{array}$$

On dit que P^σ *est une matrice de permutation d'ordre* n*.*

Lemme 1.3 *Soit P et Q deux matrices de permutation d'ordre n, alors PQ est une matrice de permutation d'ordre n.*

On peut définir les permutations élémentaires d'ordre n : pour $1 \leq k < l \leq n$, $\sigma^{k,l}$ est la permutation d'ordre n échangeant k et l en laissant les autres valeurs inchangées.

Proposition 1.14 *Pour une permutation σ d'ordre n, il existe un entier p et une famille finie $(\sigma^{k_i,l_i})_{1\leq i\leq p}$ de permutations élémentaires telle que $\sigma = \sigma^{k_p,l_p} \circ \cdots \circ \sigma^{k_1,l_1}$. De plus, pour tous les entiers p ayant la propriété précédente, -1^p est invariant et ne dépend que de σ : ce nombre est appelé la signature de la permutation σ, et noté $s(\sigma)$.*

Définition 1.24 *Soient $1 \leq k < l \leq n$, la matrice $P^{kl} \in \mathcal{M}_n(\mathbb{K})$ définie par*

$$\begin{array}{rcll} P^{kl}_{ij} & = & \delta_{ij} & si \quad i \neq k \ et \quad i \neq l \\ P^{kl}_{kj} & = & \delta_{lj} & \forall 1 \leq j \leq n \\ P^{kl}_{lj} & = & \delta_{kj} & \forall 1 \leq j \leq n \end{array}$$

est appelée matrice de permutation élémentaire d'ordre n.

De la Proposition 1.14, on déduit immédiatement le

Corollaire 1.3 *Toute matrice de permutation d'ordre n peut se décomposer en un produit de matrices de permutation élémentaire d'ordre n.*

Proposition 1.15 *Soient trois entiers positifs $k < l \leq n$ et $P^{kl} \in \mathcal{M}_n(\mathbb{K})$ une matrice de permutation élémentaire.*

- *Soit $A \in \mathcal{M}_{n,m}(\mathbb{K})$. Le produit à gauche de A par P^{kl} revient à échanger les lignes k et l de A, en laissant les autres lignes inchangées, i.e.*

$$B = P^{kl}A \Leftrightarrow \left| \begin{array}{rcll} B_{ij} & = & A_{ij} & si \quad i \neq k \ et \quad i \neq l \\ B_{kj} & = & A_{lj} & \forall 1 \leq j \leq N \\ B_{lj} & = & A_{kj} & \forall 1 \leq j \leq N \end{array} \right.$$

- *Soit $A \in \mathcal{M}_{m,n}(\mathbb{K})$. Le produit à droite de A par P^{kl} revient à échanger les colonnes k et l de A, en laissant les autres colonnes inchangées*

1.4 Représentation matricielle d'une application linéaire

Soient F un espace vectoriel sur $\mathbb{K}$ de dimension n et E un espace vectoriel sur $\mathbb{K}$ de dimension m. Soit $(f_1, \ldots, f_n)$ une base de F et $(e_1, \ldots, e_m)$ une base de E. Soit ℓ une application linéaire de E dans F. Pour chaque vecteur e_j, $1 \leq j \leq m$, on peut écrire $\ell(e_j)$ sur la base $(f_i)_{1\leq i\leq n}$: il existe un unique n-uplet de scalaires $(a_{1j}, \ldots, a_{nj})$ tels que

$$\ell(e_j) = \sum_{i=1}^{n} a_{ij} f_i \tag{1.11}$$

On a alors par linéarité de ℓ : pour tout $x \in E$, $x = \sum_{j=1}^{m} x_j e_j$,

$$\ell(x) = \sum_{i=1}^{n} (\sum_{j=1}^{m} a_{ij} x_j) f_i. \tag{1.12}$$

Les coefficients a_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$, forment une matrice de $\mathcal{M}_{n,m}(\mathbb{K})$:

Définition 1.25 *On appelle représentation matricielle (ou matrice) de l'application linéaire ℓ de E dans F relativement aux bases $(e_j)_{1\leq j\leq m}$ de E et $(f_j)_{1\leq j\leq n}$ de F la matrice $A=(a_{i,j})$ de $\mathcal{M}_{n,m}(\mathbb{K})$ dont les coefficients sont définis par (1.11).*

L'application ℓ est donc uniquement caractérisée par sa représentation matricielle relativement aux bases $(e_j)_{1\leq j\leq m}$ de E et $(f_i)_{1\leq i\leq n}$ de F. On montre de plus que la représentation matricielle relativement aux bases $(e_j)_{1\leq j\leq m}$ de E et $(f_i)_{1\leq i\leq n}$ de F définit un isomorphisme de $(\mathcal{L}(E,F),+,\cdot)$ dans $(\mathcal{M}_{n,m}(\mathbb{K}),+,\cdot)$, ce qui montre que la dimension de $\mathcal{L}(E,F)$ est mn. Cependant, cet isomorphisme n'est pas intrinsèque car il dépend des bases choisies.

On fait donc le choix d'une application linéaire *canonique* associée à une matrice A de $\mathcal{M}_{n,m}(\mathbb{K})$: on choisit $E=\mathbb{K}^m$, $F=\mathbb{K}^n$, $(e_j)_{j=1,\dots,m}$ la base canonique de $\mathbb{K}^m$, $(f_i)_{i=1,\dots,n}$ la base canonique de $\mathbb{K}^n$ et on considère l'application linéaire ℓ de $\mathcal{L}(\mathbb{K}^m,\mathbb{K}^n)$ dont la représentation matricielle relativement aux bases $(e_j)_{1\leq j\leq m}$ et $(f_i)_{1\leq i\leq n}$ est A (c'est à dire ℓ est définie par (1.12)).

On a donc un isomorphisme *canonique* entre $\mathcal{M}_{n,m}(\mathbb{K})$ et $\mathcal{L}(\mathbb{K}^m,\mathbb{K}^n)$. De plus, grâce à la Convention 1.1, on voit que si $\ell\in\mathcal{L}(\mathbb{K}^m,\mathbb{K}^n)$ et si A est sa matrice relativement aux bases canoniques, alors pour tout vecteur $v\in\mathbb{K}^m$, $\ell(v)$ est le vecteur $Av\in\mathcal{M}_{n,1}(\mathbb{K})=\mathbb{K}^n$ (obtenu comme le produit matriciel Av défini dans la Définition 1.19, en voyant v comme une matrice de $\mathcal{M}_{m,1}(\mathbb{K})$).

Cet isomorphisme canonique permet de définir pour les matrices toutes les notions présentées dans la Section 2, et d'obtenir pour les matrices les résultats analogues à ceux énoncés pour les applications linéaires. Par exemple,

- La matrice I_n est associée canoniquement à l'identité $\mathrm{id}_{\mathbb{K}^n}$.
- le noyau d'une matrice $A\in\mathcal{M}_{n,m}(\mathbb{K})$, noté $\ker(A)$, est le noyau de l'application linéaire canonique associée à A, et on a

$$\ker(A)=\{v\in\mathbb{K}^m : Av=0_{\mathbb{K}^n}\}. \tag{1.13}$$

- l'image d'une matrice $A\in\mathcal{M}_{n,m}(\mathbb{K})$, notée $\mathrm{Im}(A)$, est l'image de l'application linéaire canonique associée à A, et on a

$$\mathrm{Im}(A)=\{w\in\mathbb{K}^n : \exists v\in\mathbb{K}^m \text{ tel que } Av=w\}. \tag{1.14}$$

- $\mathrm{Im}(A)$ est le sous-espace vectoriel de $\mathbb{K}^n$ engendré par les colonnes de A.
- Le rang de $A\in\mathcal{M}_{n,m}(\mathbb{K})$ est la dimension de l'image de A. C'est le nombre maximal de colonnes de A formant une famille libre de $\mathbb{K}^n$ (donc en particulier $\mathrm{rang}(A)\leq\min(m,n)$).

Le produit matriciel correspond à la composition d'application linéaire : Soient $A\in\mathcal{M}_{m,n}(\mathbb{K})$ et $B\in\mathcal{M}_{n,p}(\mathbb{K})$ deux matrices et $\ell\in\mathcal{L}(\mathbb{K}^n,\mathbb{K}^m)$ et $\jmath\in\mathcal{L}(\mathbb{K}^p,\mathbb{K}^n)$ leur applications linéaires canoniquement associées : alors l'application $\ell\circ\jmath\in\mathcal{L}(\mathbb{K}^p,\mathbb{K}^m)$ est canoniquement associée au produit matriciel $AB\in\mathcal{M}_{m,p}(\mathbb{K})$. L'analogue matriciel de la Proposition 1.10 permet alors de définir la notion d'inverse d'une matrice carrée :

Définition 1.26 *Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$. On dit que A est inversible si et seulement si il existe une matrice de $\mathcal{M}_n(\mathbb{K})$, notée A^{-1} telle que*

$$AA^{-1}=A^{-1}A=I_n.$$

et du Corollaire 1.2, on déduit

Proposition 1.16 *Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$. Les assertions suivantes sont équivalentes*

- *A est inversible*
- $\text{rang}(A) = n$
- $\ker(A) = \{0_{\mathbb{K}^n}\}$
- *A admet un inverse à gauche : il existe une matrice $B \in \mathcal{M}_n(\mathbb{K})$ telle que $BA = I_n$*
- *A admet un inverse à droite : il existe une matrice $B \in \mathcal{M}_n(\mathbb{K})$ telle que $AB = I_n$*

Si $A \in \mathcal{M}_n(\mathbb{K})$ admet un inverse à droite (ou à gauche) B, alors A est inversible et $B = A^{-1}$.

Proposition 1.17 *Soient A et B deux matrices inversibles de $\mathcal{M}_n(\mathbb{K})$. La matrice AB est inversible et*

$$(AB)^{-1} = B^{-1}A^{-1}.$$

On peut vérifier que pour $n \geq 2$, le sous-ensemble de $\mathcal{M}_n(\mathbb{K})$ formé par les matrices inversibles muni du produit matriciel est un groupe non commutatif.

On peut maintenant démontrer de nouvelles propriétés concernant les formes linéaires

Proposition 1.18
- *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ de dimension finie n. Une famille de formes linéaires sur E $(\ell_i)_{i=1,\dots,m}$ est libre si et seulement si $m \leq n$ et $\cap_{i=1}^m \ker(\ell_i)$ est un sous-espace de E de dimension $n - m$.*
- *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ de dimension finie n et F un sous-espace de E de dimension $m \leq n$. Il existe $n - m$ formes linéaires indépendantes $(\ell_i)_{i=1,\dots,n-m}$ telles que $F = \cap_{i=1}^{n-m} \ker(\ell_i)$.*

Démonstration. On a vu que $\dim(\mathcal{L}(E, \mathbb{K})) = n$, donc si $(\ell_i)_{i=1,\dots,m}$ est libre, alors $m \leq n$. Démontrons l'implication $\Rightarrow$ dans le premier point, par récurrence.
La propriété est vraie pour $m = 1$, c'est le Lemme 1.2.
Prenons $m \leq n$ et supposons la propriété vraie jusqu'à $m - 1$. L'ensemble $\cap_{i=1}^{m-1} \ker(\ell_i)$ est un sous-espace de E de dimension $n-m+1$. Considérons l'application linéaire $f^{(m-1)} \in \mathcal{L}(E, \mathbb{K}^{m-1})$, $f^{(m-1)}(x) = (\ell_1(x), \dots, \ell_{m-1}(x))$. On a $\ker(f^{(m-1)}) = \cap_{i=1}^{m-1} \ker(\ell_i)$. Donc d'après l'hypothèse de récurrence et le théorème du rang, $\text{rang}(f^{(m-1)}) = m - 1$. Il existe donc $m - 1$ vecteurs de E, $e_1, \dots, e_{m-1}$ tels que $(f^{(m-1)}(e_1), \dots, f^{(m-1)}(e_{m-1}))$ soit une base de $\mathbb{K}^{m-1}$. Ceci est équivalent à dire que $\forall z \in \mathbb{K}^{m-1}$, le système linéaire : trouver $(x_i)_{i=1,\dots,m-1}$, $x_i \in \mathbb{K}$, tels que

$$\sum_{i=1}^{m-1} x_i f^{(m-1)}(e_i) = z$$

a une solution unique. Ce système s'écrit

$$\begin{aligned}
x_1\ell_1(e_1) + \cdots + x_{m-1}\ell_1(e_{m-1}) &= z_1, \\
&\vdots \\
x_1\ell_i(e_1) + \cdots + x_{m-1}\ell_i(e_{m-1}) &= z_i, \\
&\vdots \\
x_1\ell_{m-1}(e_1) + \cdots + x_{m-1}\ell_{m-1}(e_{m-1}) &= z_{m-1}.
\end{aligned}$$

Par suite, la matrice

$$B \in \mathcal{M}_{m-1}(\mathbb{K}), \quad B_{ij} = \ell_i(e_j),$$

est inversible.
Considérons maintenant l'application linéaire $f^{(m)} \in \mathcal{L}(E, \mathbb{K}^m)$, $f^{(m)}(x) = (\ell_1(x), \dots, \ell_m(x))$ On a $\ker(f^{(m)}) = \cap_{i=1}^m \ker(\ell_i)$ donc $\ker(f^{(m)}) \subset \ker(f^{(m-1)})$, ce qui montre que

$$m - 1 \leq \text{rang}(f^{(m)}) \leq m,$$

en utilisant le théorème du rang. Supposons que $\text{rang}(f^{(m)}) = m-1$. Comme $(f^{(m)}(e_1), \ldots, f^{(m)})(e_{m-1}))$ est une famille libre, c'est une base de $\text{Im}(f^{(m)})$. Pour tout $y \in E$, le système linéaire : trouver $(x_i)_{i=1,\ldots,m-1}$, $x_i \in \mathbb{K}$, tels que

$$\sum_{i=1}^{m-1} x_i f^{(m)}(e_i) = f^{(m)}(y)$$

a une solution unique. Ce système s'écrit

$$\begin{aligned} x_1\ell_1(e_1) + \cdots + x_{m-1}\ell_1(e_{m-1}) &= \ell_1(y), \\ &\vdots \\ x_1\ell_i(e_1) + \cdots + x_{m-1}\ell_i(e_{m-1}) &= \ell_i(y), \\ &\vdots \\ x_1\ell_{m-1}(e_1) + \cdots + x_{m-1}\ell_{m-1}(e_{m-1}) &= \ell_{m-1}(y), \\ x_1\ell_m(e_1) + \cdots + x_{m-1}\ell_m(e_{m-1}) &= \ell_m(y). \end{aligned}$$

Il est équivalent à :

$$\begin{aligned} \begin{pmatrix} x_1 \\ \vdots \\ x_{m-1} \end{pmatrix} &= B^{-1} \begin{pmatrix} \ell_1(y) \\ \vdots \\ \ell_{m-1}(y) \end{pmatrix}, \\ x_1\ell_m(e_1) + \cdots + x_{m-1}\ell_m(e_{m-1}) &= \ell_m(y). \end{aligned}$$

On en déduit qu'il existe $(\alpha_i)_{i=1,\ldots,m-1}$ tels que $\forall y \in E$,

$$\ell_m(y) = \sum_{i=1}^{m-1} \alpha_i \ell_i(y),$$

ce qui contredit l'hypothèse $(\ell_1, \ldots, \ell_m)$ libre. Donc $\text{rang}(f^{(m)}) = m$, ce qui prouve l'hypothèse de récurrence au rang m.
La démonstration de la réciproque dans le premier point est laissée au lecteur.
Pour démontrer le deuxième point, on choisit une base de F : $(e_1, \ldots, e_m)$ que l'on complète avec $(e_{m+1}, \ldots, e_m)$ pour former une base de E. Tout vecteur s'écrit $x = \sum_{i=1}^{n} x_i e_i$ et on choisit pour $i \leq n-m$: $\ell_i(x) = x_{i+m}$. Les formes linéaires $(\ell_i)_{i=1,\ldots n-m}$ sont linéairement indépendantes et $\cap_{i=1}^{n-m} \ker(\ell_i) = F$. ■

On a démontré au passage la proposition

Proposition 1.19 – *Soit $(E, +, \cdot)$ un espace vectoriel sur $\mathbb{K}$ de dimension finie n. Soit $m \leq n$ et $(\ell_i)_{i=1,\ldots,m}$ une famille de formes linéaires sur E, linéairement indépendantes. Alors si $f \in \mathcal{L}(E, \mathbb{K}^m)$ est l'application*

$$f(x) = \begin{pmatrix} \ell_1(x) \\ \vdots \\ \ell_m(x) \end{pmatrix},$$

alors $\text{rang}(f) = m$.

1.5 Changement de Bases, Matrices Semblables

De la Proposition 1.16, on tire facilement

Proposition 1.20 *Soit $P \in \mathcal{M}_n(\mathbb{K})$ une matrice inversible. Les colonnes de P forment une base de $\mathbb{K}^n$. Réciproquement, soit E un espace vectoriel sur $\mathbb{K}$ de dimension n, et deux bases $\mathcal{B}$ et $\mathcal{B}'$ de E : la matrice dont la j-ème colonne contient les coordonnées du j-ème vecteur de $\mathcal{B}'$ dans la base $\mathcal{B}$, pour $1 \leq j \leq n$, est inversible.*

Définition 1.27 *Soit $\mathcal{B}$ une base de $\mathbb{K}^n$. La matrice P dont les colonnes sont les vecteurs de la base $\mathcal{B}$ est appelée matrice de passage ou matrice de changement de base de la base canonique vers la base $\mathcal{B}$.*
Plus généralement, si E est un espace vectoriel de dimension n sur $\mathbb{K}$, et si $\mathcal{B}$ et $\mathcal{B}'$ sont deux bases de E, on appelle matrice de changement de base de $\mathcal{B}$ vers $\mathcal{B}'$ la matrice dont la j-ème colonne contient les coordonnées du j-ème vecteur de $\mathcal{B}'$ dans la base $\mathcal{B}$, pour $1 \leq j \leq n$.

Proposition 1.21 *Soit E un espace vectoriel sur $\mathbb{K}$ de dimension n. Si P est la matrice de passage d'une base $\mathcal{B}$ de E vers une base $\mathcal{B}'$, alors P^{-1} est la matrice de passage de $\mathcal{B}'$ vers $\mathcal{B}$. Si les coordonnées d'un vecteur dans la base $\mathcal{B}$ sont données par le vecteur $x \in \mathbb{K}^n$, alors ses coordonnées dans la base $\mathcal{B}'$ sont données par le vecteur $P^{-1}x$.*

En particulier, si une matrice $P \in \mathcal{M}_n(\mathbb{K})$ est inversible et transforme la base canonique de $\mathbb{K}^n$ en une nouvelle base $(f_1, \ldots, f_n)$ de $\mathbb{K}^n$ (formée par les colonnes de P), on a pour tout vecteur x, (dont les coordonnées dans la base canonique sont x_i, $1 \leq i \leq n$)

$$x = \sum_{i=1}^{n} y_i f_i \quad \text{avec} \quad y = P^{-1}x. \tag{1.15}$$

On déduit de la Proposition 1.21 la

Proposition 1.22 *Soit E un espace vectoriel sur $\mathbb{K}$ de dimension n. Soient $\mathcal{B}$ et $\mathcal{B}'$ deux bases de E et P la matrice de passage de $\mathcal{B}$ vers $\mathcal{B}'$. Si ℓ est une application linéaire de $\mathcal{L}(E)$ dont la matrice dans la base $\mathcal{B}$ est A, alors la matrice de ℓ dans la base $\mathcal{B}'$ est $B = P^{-1}AP$.*

Définition 1.28 *On dit que deux matrices A et B sont semblables si il existe une matrice inversible P telle que*

$$B = P^{-1}AP. \tag{1.16}$$

La Proposition 1.22 montre donc que toutes les représentations matricielles d'une même application linéaire $\ell \in \mathcal{L}(E)$ sont semblables. Des Propositions 1.22 et 1.20, on déduit que deux matrices semblables ont même rang.

1.6 Transposition

Définition 1.29 *Soit $A = (a_{ij})$ une matrice de $\mathcal{M}_{n,m}(\mathbb{K})$. On définit la matrice transposée de A et on note A^T la matrice de $\mathcal{M}_{m,n}(\mathbb{K})$ dont le coefficient situé sur la i-ème ligne et la j-ème colonne est a_{ji} pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$.*

Exemple. Le transposée de la matrice colonne $V \in \mathcal{M}_{n,1}(\mathbb{K})$ est une matrice ligne $V^T \in \mathcal{M}_{1,n}(\mathbb{K})$.
On démontre facilement la

Proposition 1.23 *Soient $A \in \mathcal{M}_{n,m}(\mathbb{K})$ et $B \in \mathcal{M}_{m,p}(\mathbb{K})$. On a*

$$(AB)^T = B^T A^T. \tag{1.17}$$

a) Rang et Inversibilité Soit A une matrice de $\mathcal{M}_{n,m}(\mathbb{K})$. Le rang de A^T est le nombre maximal de colonnes de A^T linéairement indépendantes, donc de lignes de A linéairement indépendantes. Par suite, si $ker(A) \neq \{0\}$, il existe une combinaison linéaire nulle des lignes de A avec des coefficients non tous nuls et $\text{rang}(A^T) < n$. En particulier, on vient de démontrer le

Lemme 1.4 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On a*

$$A \text{ inversible } \Leftrightarrow A^T \text{ inversible} \tag{1.18}$$

On peut aussi par un simple calcul utilisant (1.17), prouver la

Proposition 1.24 *Si $A \in \mathcal{M}_n(\mathbb{K})$ est inversible, alors*

$$(A^T)^{-1} = (A^{-1})^T. \tag{1.19}$$

Lemme 1.5 *Le rang d'une matrice $A \in \mathcal{M}_{m,n}$ est le nombre maximal de lignes de A linéairement indépendantes et on a rang(A)=rang(A^T).*

Démonstration. Soit q le nombre maximal de lignes de A formant un système libre (on sait que q=rang(A^T)) et r le rang de A, i.e. le nombre maximal de colonnes de A formant un système libre. Alors en supposant que les lignes de A : L_{i_1},..., L_{i_q} forment un système libre, on a $\forall i \neq i_1, \ldots, i_q$, $L_i = \sum_{k=1}^{q} \alpha_{i,k} L_{i_k}$. Un vecteur de Im($A$) s'écrit $y_i = \sum_{i=1}^{n} L_{i,j} x_j$, $i = 1, \ldots, m$ donc on a $y_i = \sum_{k=1}^{q} \alpha_{i,k} y_{i_k}$, $i \neq i_1, \ldots i_q$. Donc $\text{Im}(A) \subset \{y \in \mathbb{K}^m, y_i = \sum_{k=1}^{q} \alpha_{i,k} y_{i_k}, i \neq i_1, \ldots, i_q\}$ qui est un sous-espace de E de dimension q d'après la Proposition 1.18 (car les formes linéaires $y \mapsto y_i - \sum_{k=1}^{q} \alpha_{i,k} y_{i_k}$, $i \neq i_1, \ldots, i_q$ sont linéairement indépendantes). Mais on sait que $\dim(\text{Im}(A)) = r$ donc $r \leq q$. En faisant le même raisonnement sur la matrice $A^T \in \mathcal{M}_{n,m}$, on voit que $q \leq r$. Donc $q = r$. ■

On peut caractériser le rang d'une matrice en étudiant l'inversibilité de matrices extraites : extraire une matrice d'une matrice plus grande, c'est ne conserver de la matrice initiale que les coefficients situés sur certaines lignes et sur certaines colonnes. Plus précisément,

Définition 1.30 *Soient m, n, m', n' quatre entiers strictement positifs, avec $m' \leq m$ et $n' \leq n$. Soit σ une application strictement croissante de $\{1, \ldots, m'\}$ dans $\{1, \ldots, m\}$, et τ une application strictement croissante de $\{1, \ldots, n'\}$ dans $\{1, \ldots, n\}$. Soit $A \in \mathcal{M}_{m,n}(\mathbb{K})$. On définit la matrice extraite $A^{\sigma,\tau} \in \mathcal{M}_{m',n'}(\mathbb{K})$ comme la matrice dont les coefficients sont $b_{i,j} = a_{\sigma(i),\tau(j)}$, $1 \leq i \leq m'$, $1 \leq j \leq n'$.*

De la caractérisation du rang d'une matrice comme le nombre maximal de colonnes linéairement indépendantes, on déduit la

Proposition 1.25 *Soit $A \in \mathcal{M}_{n,m}(\mathbb{K})$. Le rang de A est le plus grand entier p tel qu'on puisse extraire de A une matrice carrée $B \in \mathcal{M}_p(\mathbb{K})$ qui soit inversible.*

Remarque 1.3 *On en déduit avec le Théorème du rang que si $A \in \mathcal{M}_{m,n}(\mathbb{K})$, alors $n - \dim(Ker(A)) = m - \dim(Ker(A^T))$.*

b) Matrices symétriques et antisymétriques.

Définition 1.31 *Une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est symétrique si $A^T = A$.*
Une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est antisymétrique si $A^T = -A$.

Les matrices symétriques forment un sous-espace de $\mathcal{M}_n(\mathbb{K})$ de dimension $n(n+1)/2$. Une base de ce sous-espace est formée par les matrices $(S^{ij})_{1\leq i\leq j\leq n}$ dont les coefficients sont $s^{ij}_{kl} = \delta_{ki}\times\delta_{lj} + \delta_{li}\times\delta_{kj}$, $k = 1,\dots,n$, $l = 1,\dots,n$.
Les matrices antisymétriques forment un sous-espace de $\mathcal{M}_n(\mathbb{K})$ de dimension $n(n-1)/2$, supplémentaire au sous-espace des matrices symétriques. Une base de ce sous-espace est formée par les matrices $(A^{ij})_{1\leq i<j\leq n}$ dont les coefficients sont $a^{ij}_{kl} = \delta_{ki}\times\delta_{lj} - \delta_{li}\times\delta_{kj}$, $k = 1,\dots,n$, $l = 1,\dots,n$.
Soit $A \in \mathcal{M}_n(\mathbb{K})$, la partie symétrique de A est $\frac{1}{2}(A + A^T)$, et sa partie antisymétrique est $\frac{1}{2}(A - A^T)$.

c) Dualité Il est naturel de poser la question : quelle est la notion correspondant à la transposition de matrices pour les applications linéaires ? La réponse à cette question fait intervenir la notion de dualité :

Définition 1.32 *On appelle dual d'un espace vectoriel E sur $\mathbb{K}$, l'espace E^* des formes linéaires sur E, c'est à dire $E^* = \mathcal{L}(E,\mathbb{K})$.*

Supposons maintenant E de dimension n. Si $(e_i)_{1\leq i\leq n}$ est une base de E, on peut montrer que la famille $(e_i^*)_{1\leq i\leq n}$, avec $e_i^*(e_j) = \delta_{ij}$ est une base de E^*, appelée base duale de la base $(e_i)_{1\leq i\leq n}$. La forme linéaire e_i^* n'est rien d'autre que l'application qui à un vecteur x de E associe sa i-ème coordonnée dans la base $(e_i)_{1\leq i\leq n}$.
L'application qui à un vecteur $x = \sum_{i=1}^n x_i e_i$ associe la forme linéaire $x^* = \sum_{i=1}^n x_i e_i^*$ est un isomorphisme de E sur E^*. Cet isomorphisme dépend de la base choisie ; ce n'est pas un isomorphisme intrinsèque.

Définition 1.33 *Soient E et F deux espaces vectoriels sur $\mathbb{K}$. Soit une application linéaire $\ell \in \mathcal{L}(E,F)$, on voit que pour tout $v \in F^*$, $v\circ\ell \in E^*$. On définit alors l'application linéaire $\ell^T \in \mathcal{L}(F^*,E^*)$ par $\ell^T(v) = v\circ\ell$.*

Soit $(e_i)_{1\leq i\leq m}$ une base de E et $(e_i^*)_{1\leq i\leq m}$ sa base duale. Soit $(f_i)_{1\leq i\leq n}$ une base de F et $(f_i^*)_{1\leq i\leq n}$ sa base duale. Soit $A \in \mathcal{M}_{n,m}(\mathbb{K})$, $A = (a_{i,j})_{1\leq i\leq n, 1\leq j\leq m}$ est la matrice de ℓ relativement aux base (e_i) et (f_j), i.e. $\ell(e_i) = \sum_{j=1}^n a_{ji} f_j$. On a $(f_k^*\circ\ell)(e_i) = \sum_{j=1}^n a_{ji} f_k^*(f_j) = a_{ki}$, ce qui veut dire que $\ell^T(f_k^*) = f_k^*\circ\ell = \sum_{i=1}^m a_{ki} e_i^*$. La matrice de ℓ^T relativement aux bases (f_j^*) de F^* et (e_i^*) de E^* est donc A^T. La notion de transposition de matrice est donc l'analogue de la notion de transposition d'applications linéaires.

1.7 Espaces Euclidiens

Définition 1.34 *Un espace vectoriel E sur $\mathbb{R}$, de dimension finie n, et muni d'un produit scalaire $(x,y) \mapsto \langle x,y\rangle$, c'est à dire d'une forme bilinéaire sur $E\times E$, symétrique et définie positive, est appelé espace Euclidien. Plus explicitement, on a*

- *symétrie :* $\forall x,y \in E, \quad \langle x,y\rangle = \langle y,x\rangle$.
- $\forall \lambda,\mu \in \mathbb{R}, \forall x,y,z \in E, \quad \langle x,\lambda y + \mu z\rangle = \lambda\langle x,y\rangle + \mu\langle x,z\rangle$.
- $\forall x \in E, \quad \langle x,x\rangle \geq 0$ *et* $\langle x,x\rangle = 0 \Longleftrightarrow x = 0$.

Exemple. L'espace $\mathbb{R}^n$ du produit scalaire canonique : $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$, est un espace Euclidien. Remarquons que $\langle x, y\rangle$ s'écrit aussi $x^T y = y^T x$ en utilisant le produit matriciel.

Lemme 1.6 *Soit E un espace Euclidien. L'application $\|.\|$, $E \to \mathbb{R}_+$, qui à un vecteur x associe $\|x\| = \sqrt{\langle x, x\rangle}$ est une norme, appelé norme euclidienne associée à $\langle.\rangle$: on a*
- $\forall \lambda \in \mathbb{R}$, $\forall v \in E$, $\|\lambda v\| = |\lambda|\|v\|$.
- $\forall v, w \in E$, $\|v + w\| \leq \|v\| + \|w\|$.
- $\forall v \in E$, $\|v\| = 0 \Rightarrow v = 0_E$.

Lemme 1.7 *Soit E un espace Euclidien. On a l'inégalité de Cauchy-Schwarz $\forall v, w \in E$, $|\langle v, w\rangle| \leq \|v\|\|w\|$.*

Définition 1.35
- *Deux vecteurs x et y d'un espace E Euclidien sont orthogonaux si $\langle x, y\rangle = 0$. On note alors $x \perp y$.*
- *Soit K une partie de E : $K^\perp$ est le sous-ensemble de E formés des vecteurs othogonaux aux vecteurs de K. $K^\perp$ est un sous-espace vectoriel de E.*
- *On dit qu'une famille de vecteurs de E est orthogonale si les vecteurs qui la composent sont deux à deux orthogonaux.*
- *On dit qu'une famille de vecteurs de E est orthonormée ou orthonormale si elle est orthogonale et si chacun de ses vecteurs v est tel que $\|v\| = 1$.*

Proposition 1.26 *Soit E un espace Euclidien.*
- *Soient F et G deux sous-espaces vectoriels de E tels que $F \subset G^\perp$. Alors $F \cap G = \{0\}$.*
- *Soit F un sous-espace vectoriel de E. On a $F \oplus F^\perp = E$. On appelle $F^\perp$ le supplémentaire orthogonal de F.*

Proposition 1.27 *Soit E un espace Euclidien. Une famille de n vecteurs non nuls de E, deux à deux orthogonaux est une base. On dit que la base est orthogonale. Si les vecteurs ont de plus norme 1, on parle de base orthonormale.*

Dans les espaces Euclidiens, le produit scalaire permet d'identifier E et E^* :

Proposition 1.28 *Soit E un espace Euclidien, alors pour tout $f^* \in E^*$, il existe un unique $f \in E$ tel que $f^*(v) = \langle f, v\rangle \quad \forall v \in E$. L'application qui à f^* associe f est un isomorphisme intrinsèque (il ne dépend que du produit scalaire). Il permet d'identifier E à E^*.*

Démonstration. Soit $(e_i)_{1\leq i\leq n}$ une base de E. On note $g_{ij} = \langle e_i, e_j\rangle$ et G la matrice de $\mathcal{M}_n(\mathbb{R})$ dont les coefficients sont les g_{ij}. La matrice G est inversible, car le produit scalaire est défini positif (démontrer en détail pourquoi G est inversible).
Définissons un vecteur e^i tel que, pour tout vecteur x,

$$\langle e^i, x\rangle = e_i^*(x) = x_i$$

où e_i^* est la forme linéaire qui à un vecteur x fait correspondre sa $i$ème composante x_i dans la base $(e_1, \ldots, e_n)$.
On cherche e^i sous la forme : $e^i = \sum_{j=1}^n \alpha^{ij} e_j$. Alors $\langle e^i, e_j\rangle = \sum_{k=1}^n \alpha^{ik} g_{kj} = e_i^*(e_j) = \delta_{ij}$. On doit résoudre $\sum_{k=1}^n \alpha^{ik} g_{kj} = \delta_{ij}$, $j = 1, \ldots, n$. À i fixé, c'est un système à n équations et n inconnues, dont la matrice est G. Le système a pour solution unique la $i$ème colonne de G^{-1}. On a donc montré l'existence et l'unicité de e^i. La famille $(e^i)_{i=1,\ldots,n}$ est une base de E : en effet, s'il existe $x_1, \ldots, x_n$ tels que $\sum_{j=1}^n x_j e^j = 0$, on prend le produit scalaire avec e_i, $i = 1, \ldots, n$,

ce qui conduit à $x_i = 0$, $i = 1, \ldots, n$.
Pour le vecteur f^* de E^*, le vecteur $f = \sum_{i=1}^n f^*(e_i)\, e^i$ est l'unique vecteur tel que $f^*(v) = \langle f, v\rangle \quad \forall v \in E$. ■

Remarque 1.4 *La base $(e^i)_{i=1,\ldots,n}$ de E construite est l'image de la base duale $(e_i^*)_{i=1,\ldots,n}$ de E^* construite ci dessus par l'isomorphisme intrinsèque de E^* sur E construit ci-dessus.*

Remarque 1.5 *Si la base $(e_i)_{i=1,\ldots,n}$ est orthonormée, alors $e^i = e_i$, $i = 1, \ldots, n$.*

Soient deux espaces Euclidiens E et F et ℓ une application linéaire de E dans F. On définit l'application adjointe de ℓ et on note ℓ^* l'application linéaire de F dans E définie par

$$\langle \ell(u), v\rangle = \langle u, \ell^*(v)\rangle. \tag{1.20}$$

Soient $(e_i)_{1\le i\le m}$ une base de E et $(f_i)_{1\le i\le n}$ une base de F. Soit $A \in \mathcal{M}_{n,m}(\mathbb{R})$ la matrice de ℓ relativement à ces deux bases. Soient $\left(e^i\right)_{1\le i\le m}$ et $\left(f^i\right)_{1\le i\le n}$ les bases de E et F définies ci-dessus. la matrice de ℓ^* relativement aux bases $\left(f^i\right)_{1\le i\le n}$ et $\left(e^i\right)_{1\le i\le m}$ est A^T.
En choisissant $E = \mathbb{R}^m$ et $F = \mathbb{R}^n$ et en prenant $(e_i)_{1\le i\le m}$ et $(f_i)_{1\le i\le n}$ les bases canoniques (donc orthonormées), on voit que $\forall A \in \mathcal{M}_{n,m}(\mathbb{R})$, $\forall u \in \mathbb{R}^m$, $\forall v \in \mathbb{R}^n$,

$$v^T A u = \langle Au, v\rangle_{\mathbb{R}^n} = \langle u, A^T v\rangle_{\mathbb{R}^m} = u^T A^T v,$$

ce qui est aussi une conséquence de (1.17). Par analogie à ℓ^*, on emploiera aussi la notation A^* pour la matrice A^T, $A \in \mathcal{M}_n(\mathbb{R})$.
On peut montrer la

Proposition 1.29 *Soit $A \in \mathcal{M}_n(\mathbb{R})$, on a*

$$\begin{aligned} \ker(A^T) &= (\mathrm{Im}(A))^\perp, \\ \mathrm{Im}(A^T) &= (\ker(A))^\perp. \end{aligned} \tag{1.21}$$

Définition 1.36
- *On dit que $A \in \mathcal{M}_n(\mathbb{R})$ est normale si $AA^T = A^T A$.*
- *On dit que $A \in \mathcal{M}_n(\mathbb{R})$ est orthogonale ou unitaire si $AA^T = A^T A = I$, ou de manière équivalente si $A^{-1} = A^T$.*

Remarque 1.6 *Il est clair qu'une matrice $A \in \mathcal{M}_n(\mathbb{R})$ symétrique est normale, et qu'une matrice orthogonale est normale.*

Proposition 1.30 *Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est orthogonale si et seulement si ses colonnes forment une base orthonormée de $\mathbb{R}^n$. Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est orthogonale si et seulement si ses lignes forment une base orthonormée de $\mathbb{R}^n$.*
Une matrice orthogonale $O \in \mathcal{M}_n(\mathbb{R})$ est donc la matrice de passage de la base canonique (donc orthonormale) de $\mathbb{R}^n$ vers une base orthonormale. Réciproquement, si $\mathcal{B}$ et $\mathcal{B}'$ sont deux bases orthonormales de $\mathbb{R}^n$, la matrice de passage O de $\mathcal{B}$ vers $\mathcal{B}'$ est orthogonale. Dans ce cas, si A (resp. B) est la représentation matricielle de $\ell \in \mathcal{L}(\mathbb{R}^n)$ dans $\mathcal{B}$, (resp. $\mathcal{B}'$), la formule de changement de base s'écrit aussi $B = O^T AO$.

1.8 Espaces Hermitiens

Définition 1.37 *Un espace vectoriel E sur $\mathbb{C}$, de dimension finie n, et muni d'un produit scalaire hermitien $(x,y) \mapsto \langle x,y\rangle$, c'est à dire d'une forme sesquilinéaire sur $E \times E$, hermitienne et définie positive, est appelé espace Hermitien. Plus explicitement, on a*

- $\forall x, y \in E, \quad \langle x,y\rangle = \overline{\langle y,x\rangle}$.
- $\forall \lambda, \mu \in \mathbb{C}, \forall x,y,z \in E, \quad \langle x, \lambda y + \mu z\rangle = \bar{\lambda}\langle x,y\rangle + \bar{\mu}\langle x,z\rangle$ *et* $\langle \lambda y + \mu z, x\rangle = \lambda\langle y,x\rangle + \mu\langle z,x\rangle$.
- $\forall x \in E, \quad \langle x,x\rangle \geq 0$ *et* $\langle x,x\rangle = 0 \Longleftrightarrow x = 0$.

Exemple. L'espace $\mathbb{C}^n$ du produit scalaire hermitien canonique $\langle x,y\rangle = \sum_{i=1}^n x_i \bar{y_i}$ est un espace hermitien.

Lemme 1.8 *Soit E un espace hermitien. L'application $\|.\|$, $E \to \mathbb{R}_+$, qui à un vecteur x associe $\|x\| = \sqrt{\langle x,x\rangle}$ est une norme : on a*

- $\forall \lambda \in \mathbb{C}$, $\forall v \in E$, $\|\lambda v\| = |\lambda|\|v\|$.
- $\forall v, w \in E$, $\|v+w\| \leq \|v\| + \|w\|$.
- $\forall v \in E$, $\|v\| = 0 \Rightarrow v = 0_E$.

et l'inégalité de Cauchy-Schwarz : $\forall v, w \in E$, $|\langle v,w\rangle| \leq \|v\|\|w\|$.

Comme pour les espaces Euclidiens, on peut aussi définir les notions d'orthogonalité pour le produit scalaire hermitien, de sous-espace orthogonal à une partie de E, de bases orthogonales et orthonormées. On peut aussi définir la notion d'application linéaire adjointe :

Définition 1.38 *Soient deux espaces hermitiens E et F et ℓ une application linéaire de E dans F. On définit l'application adjointe de ℓ et on note ℓ^* l'application linéaire de F dans E définie par*

$$\langle \ell(u), v\rangle = \langle u, \ell^*(v)\rangle. \tag{1.22}$$

Proposition 1.31 *Soient deux espaces hermitiens E et F, et $(e_1,\ldots,e_m)$, $(f_1,\ldots,f_n)$ des bases orthonormées de E et de F. Soit ℓ une application linéaire de E dans F, dont la matrice relativement aux bases $(e_1,\ldots,e_m)$ et $(f_1,\ldots,f_n)$ est $A \in \mathcal{M}_{n,m}(\mathbb{C})$. La matrice de ℓ^* dans les bases $(f_1,\ldots,f_n)$, $(e_1,\ldots,e_m)$ est*

$$A^* = (\bar{A})^T, \qquad i.e. \qquad a^*_{ij} = \overline{a_{ji}}, \quad i = 1,\ldots,m;\ j = 1,\ldots,n \tag{1.23}$$

La matrice A^* est appelée matrice adjointe, elle vérifie

$$\langle Au, v\rangle = \langle u, A^* v\rangle, \quad \forall u \in \mathbb{C}^m,\ \forall v \in \mathbb{C}^n.$$

On peut montrer la

Proposition 1.32 *Soit $A \in \mathcal{M}_n(\mathbb{C})$, on a*

$$\begin{array}{l} \mathrm{rang}(A^*) = \mathrm{rang}(A), \\ \ker(A^*) = (\mathrm{Im}(A))^{\perp}, \\ \mathrm{Im}(A^*) = (\ker(A))^{\perp}. \end{array} \tag{1.24}$$

où l'orthogonalité est prise par rapport au produit scalaire hermitien dans $\mathbb{C}^n$.

Définition 1.39

- *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est hermitienne si $A^* = A$.*
- *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est normale si $AA^* = A^*A$.*

– *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est unitaire si $AA^* = A^*A = I_n$.*

Remarque 1.7 *Il est clair qu'une matrice hermitienne est normale. Une matrice unitaire est normale.*

On remarquera que l'ensemble des matrices hermitiennes n'est pas un sous-espace de $\mathcal{M}_n(\mathbb{C})$ (comme espace sur $\mathbb{C}$). On a l'analogue de la Proposition 1.30 pour les matrices réelles :

Proposition 1.33 *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est unitaire si et seulement si ses colonnes forment une base orthonormée de $\mathbb{C}^n$. Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est unitaire si et seulement si ses lignes forment une base orthonormée de $\mathbb{C}^n$.*
Une matrice unitaire $A \in \mathcal{M}_n(\mathbb{C})$ est donc la matrice de passage de la base canonique (donc orthonormale) de $\mathbb{C}^n$ vers une base orthonormale. Réciproquement, si $\mathcal{B}$ et $\mathcal{B}'$ sont deux bases orthonormales de $\mathbb{C}^n$, la matrice de passage de $\mathcal{B}$ vers $\mathcal{B}'$ est unitaire.

1.9 Procédé de Gram-Schmidt

Dans cette section, E est soit un espace Euclidien (espace vectoriel sur $\mathbb{R}$, de dimension finie, muni d'un produit scalaire) ou un espace hermitien (espace vectoriel sur $\mathbb{C}$, de dimension finie, muni d'un produit scalaire hermitien). On note $\langle , \rangle$ le produit scalaire. Le but est ici de décrire le procédé (ou algorithme) de Gram-Schmidt, permettant de construire des bases orthonormales de E. Le procédé de Gram-Schmidt est en fait l'essence de la démonstration du résultat suivant

Théorème 1.3 *Soit $(f_1, \ldots, f_n)$ une famille libre de vecteurs de E. Il existe une famille orthonormée $(e_1, \ldots, e_n)$ de vecteurs de E, telle que,*

$$\mathrm{Vect}(e_1, \ldots, e_p) = \mathrm{Vect}(f_1, \ldots, f_p), \qquad \forall p \leq n. \tag{1.25}$$

De plus, si E est un espace euclidien, la famille $(e_1, \ldots, e_n)$ est unique à une multiplication près des vecteurs par -1. Si E est un espace hermitien, la famille $(e_1, \ldots, e_n)$ est unique à une multiplication près des vecteurs par un nombre complexe de module 1.

Démonstration. On procède par récurrence sur n : pour $n = 1$, $e_1 = \frac{1}{\|f_1\|} f_1$ est bien l'unique vecteur normé (à une multiplication près par -1 si E est euclidien, par un nombre complexe de module 1 si E est hermitien) tel que $\mathrm{Vect}(e_1) = \mathrm{Vect}(f_1)$.
Supposons la propriété vrai jusqu'à $n-1 < \dim(E)$. Soit $(f_1, \ldots, f_n)$ une famille libre de vecteurs de E. D'après l'hypothèse de récurrence, il existe une famille orthonormée $(e_1, \ldots, e_{n-1})$ de vecteurs de E telle que,

$$\mathrm{Vect}(e_1, \ldots, e_p) = \mathrm{Vect}(f_1, \ldots, f_p), \qquad \forall p \leq n-1, \tag{1.26}$$

et cette famille est unique aux multiplications mentionnées plus haut près.
La famille $(e_1, \ldots, e_{n-1}, f_n)$ est libre et engendre le même sous-espace vectoriel que $(f_1, \ldots, f_n)$. Cherchons e_n sous la forme

$$e_n = \beta(f_n + \sum_{p=1}^{n-1} a_p e_p) \tag{1.27}$$

pour que $e_n \in \mathrm{Vect}(e_1, \ldots, e_{n-1})^\perp$ et $\|e_n\| = 1$; on obtient aisément que

$$\begin{aligned} a_1 &= -\langle f_n, e_1 \rangle \\ &\vdots \\ a_{n-1} &= -\langle f_n, e_{n-1} \rangle \end{aligned}$$

et que

$$\|e_n\|^2 = |\beta|^2 \|f_n - \sum_{p=1}^{n-1} \langle f_n, e_p \rangle e_p\|^2$$

Mais $f_n - \sum_{p=1}^{n-1} \langle f_n, e_p \rangle e_p \neq 0$ car $f_n \notin \text{Vect}(e_1, \ldots, e_{n-1})$. Donc il existe un unique scalaire β répondant à la question, à une multiplication près par -1 si E est euclidien, par un nombre complexe de module 1 si E est hermitien.
On a alors construit une famille $(e_1, \ldots, e_n)$ répondant à la question. ■

Si $(f_1, \ldots, f_n)$ est une base de E, le procédé de Gram-Schmidt permet d'en déduire une base orthonormale.

1.10 Déterminants

Définition 1.40 *On note $\mathcal{S}_n$ l'ensemble des permutations d'ordre n (bijections de l'ensemble de $\{1, \ldots, n\}$). Soit $A \in \mathcal{M}_n(\mathbb{K})$. On définit le déterminant de A comme le scalaire*

$$\det(A) = \sum_{\sigma \in \mathcal{S}_n} s(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}. \tag{1.28}$$

où $\mathcal{S}_n$ est l'ensemble des permutations d'ordre n et pour tout $\sigma \in \mathcal{S}_n$, $s(\sigma)$ est la signature de σ.

Remarque 1.8 *On voit que pour calculer le déterminant de $A \in \mathcal{M}_n(\mathbb{K})$ par la formule (1.28), il faut effectuer $n!$ sommes et $n * n!$ produits. Ce nombre d'opérations devient très vite énorme et on verra dans ce cours d'autres façons beaucoup plus rapides de calculer le déterminant de A.*

On a les trois propriétés

1. L'application qui à une colonne v (resp. une ligne v^T) de A associe $\det(A)$ (toutes les autres colonnes (resp. lignes) étant fixées) est une application linéaire. En particulier, si une matrice A contient une colonne ou une ligne nulle, son déterminant est nul. L'application déterminant peut être vue comme une forme n-linéaire sur le produit d'espaces $\underbrace{\mathbb{K}^n \times \cdots \times \mathbb{K}^n}_{\text{n}}$.
2. Soit $A \in \mathcal{M}_n(\mathbb{K})$ et B une matrice obtenue en échangeant deux colonnes de A en laissant les autres inchangées. On a $\det(B) = -\det(A)$. On dit que la forme n-linéaire sur $\underbrace{\mathbb{K}^n \times \cdots \times \mathbb{K}^n}_{\text{n}}$ est alternée. Ceci implique que si A comporte au moins deux colonnes égales (ou au moins deux lignes égales) alors $\det(A) = 0$.
3. $\det(I_n) = 1$.

On peut en fait montrer le

Théorème 1.4 (fondamental) *L'ensemble des applications de $\mathcal{M}_n(\mathbb{K})$ dans $\mathbb{K}$ ayant les deux premières propriétés ci-dessus est un espace vectoriel de dimension un, engendré par l'application $A \mapsto \det(A)$.*
L'application $A \mapsto \det(A)$ est la seule application de $\mathcal{M}_n(\mathbb{K}^n)$ dans $\mathbb{K}$ ayant les trois propriétés ci-dessus.

Soit B une matrice de $\mathcal{M}_n(\mathbb{K})$. L'application qui à $A \in \mathcal{M}_n(\mathbb{K})$ associe $\det(BA)$ a les deux premières propriétés ci-dessus, et vérifie de surcroît $\det(BI_n) = \det(B)$. Donc, d'après le Théorème 1.4, cette application est en fait $A \mapsto \det(A)\det(B)$.

Proposition 1.34 *Soient A et B deux matrices de $\mathcal{M}_n(\mathbb{K})$. On a*

$$\det(AB) = \det(A)\det(B) = \det(BA). \tag{1.29}$$

On a aussi la

Proposition 1.35 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On a*

$$\det(A^T) = \det(A). \tag{1.30}$$

Démonstration.

$$\begin{aligned}
\det(A) &= \sum_{\sigma\in\mathcal{S}_n} s(\sigma)\prod_{i=1}^{n} a_{i,\sigma(i)} \\
&= \sum_{\sigma\in\mathcal{S}_n} s(\sigma)\prod_{i=1}^{n} a_{\sigma^{-1}(i),i} && \text{caractère bijectif de } \sigma \\
&= \sum_{\sigma\in\mathcal{S}_n} s(\sigma^{-1})\prod_{i=1}^{n} a_{\sigma^{-1}(i),i} && \text{car } s(\sigma^{-1}) = s(\sigma) \\
&= \sum_{\sigma\in\mathcal{S}_n} s(\sigma)\prod_{i=1}^{n} a_{\sigma(i),i} && \text{car } \mathcal{S}_n = \{\sigma^{-1}, \sigma \in \mathcal{S}_n\} \\
&= \det(A^T).
\end{aligned}$$

■

On déduit de la Proposition 1.34 et des propriétés de l'application $A \mapsto \det(A)$ les

Lemme 1.9 *Si une matrice $P \in \mathcal{M}_n(\mathbb{K})$ est inversible, alors $\det(P) \neq 0$ et on a $\det(P^{-1}) = (\det(P))^{-1}$.*

Lemme 1.10 *Soient A et B deux matrices de $\mathcal{M}_n(\mathbb{K})$, semblables : on a $\det(A) = \det(B)$.*

On déduit des deux Lemmes précédents la

Proposition 1.36 *Une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est inversible si et seulement si $\det(A) \neq 0$.*

Démonstration. Si A n'est pas inversible, A est semblable à une matrice dont une colonne au moins est nulle, et d'après le Lemme 1.10, $\det(A) = 0$. On conclut alors avec le Lemme 1.9. ■

Définition 1.41 *Soit $A \in \mathcal{M}_n(\mathbb{K})$, et $1 \leq i, j \leq n$. On appelle mineur associé au coefficient a_{ij} de A et on note m_{ij}, le déterminant de la matrice d'ordre $n-1$ extraite de A en supprimant la i-ème ligne et la j-ème colonne. On appelle cofacteur associé au coefficient a_{ij} de A le scalaire $\Delta_{ij} = (-1)^{i+j} m_{ij}$.*

Proposition 1.37 (développement suivant une ligne ou une colonne) *On a, pour $A \in \mathcal{M}_n(\mathbb{K})$,*

$$\begin{array}{rcll}
\det(A) &=& \sum_{j=1}^{n} \Delta_{lj} a_{lj}, & \forall l,\ 1 \leq l \leq n \\
\det(A) &=& \sum_{i=1}^{n} \Delta_{ik} a_{ik}, & \forall k,\ 1 \leq k \leq n
\end{array} \tag{1.31}$$

Proposition 1.38 *Soit $A \in \mathcal{M}_n(\mathbb{K})$, on note $\tilde{A}$ la matrice des cofacteurs de A. On a*

$$\tilde{A}^T A = A\tilde{A}^T = \det(A) I_n. \tag{1.32}$$

Si A est inversible ,

$$A^{-1} = \frac{1}{\det(A)} \tilde{A}^T. \tag{1.33}$$

1.11 Traces

Définition 1.42 *Soit $A = (a_{ij})_{1 \leq i,j \leq n}$ une matrice carrée d'ordre n à coefficents dans $\mathbb{K}$. La trace de A, notée* $\mathrm{trace}(A)$ *est le scalaire*

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii}. \tag{1.34}$$

L'application trace qui à une matice asssocie sa trace est une forme linéaire sur $\mathcal{M}_n(\mathbb{K})$. Il est clair que $\mathrm{trace}(A) = \mathrm{trace}(A^T)$

Proposition 1.39 *Soit A et B deux matrices carrées d'ordre n à coefficients dans $\mathbb{K}$. On a*

$$\mathrm{trace}(AB) = \mathrm{trace}(BA). \tag{1.35}$$

Corollaire 1.4 *Soit A et B deux matrices carrées d'ordre n à coefficients dans $\mathbb{K}$. Si A et B sont semblables, on a*

$$\mathrm{trace}(A) = \mathrm{trace}(B). \tag{1.36}$$

Chapitre 2

Réduction de matrices

Dans ce qui suit, on s'intéresse à des matrices carrées de $\mathcal{M}_n(\mathbb{K})$, où $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$. On note I la matrice identité de $\mathcal{M}_n(\mathbb{K})$.
Il existe des classes de matrices particulièrement simples : les matrices diagonales $A = (a_{ij})_{1 \leq i,j \leq n}$ avec $a_{ij} = 0$ si $i \neq j$ et les matrices triangulaires supérieures ($a_{i,j} = 0$ si $i < j$) ou triangulaires supérieures ($a_{i,j} = 0$ si $i > j$). Par exemple, pour de telles matrices, résoudre un système linéaire est particulièrement simple (c'est l'idée de base de la méthode de Gauss). Ces matrices sont aussi particulièrement utiles pour la résolution des systèmes d'équations différentielles linéaires et pour l'étude des propriétés qualitatives des systèmes d'équations différentielles non linéaires. Pour une matrice donnée A, il est donc souvent très utile de chercher un changement de base qui transforme A en une matrice semblable, diagonale ou à défaut triangulaire. On appelle cette démarche réduction de matrices.
Pour la réduction des matrices, les notions importantes sont celles de valeurs propres et vecteurs propres, polynôme caractéristique, polynôme de matrices, sous-espace stable. On montrera en particulier que si le polynôme caractéristique d'une matrice est scindé, on peut trouver un changement de base où la matrice devient triangulaire supérieure et un théorème fondamental, le théorème d'Hamilton-Cayley, permet de préciser la structure de la matrice triangulaire obtenue.

2.1 Valeurs propres, vecteurs propres, polynôme caractéristique

Définition 2.1 *Soit $A \in \mathcal{M}_n(\mathbb{K})$, on dit que $\lambda \in \mathbb{K}$ est une valeur propre de A s'il existe un vecteur $v \in \mathbb{K}^n$, $v \neq 0$ tel que $Av = \lambda v$. On dit que le vecteur $v \neq 0$ est un vecteur propre de A associé à la valeur propre λ.*

Autrement dit, λ est une valeur propre de A si et seulement si le noyau de $A - \lambda I$ n'est pas réduit à $\{0\}$, ce qui a lieu si et seulement si $\det(A - \lambda I) = 0$.

Remarque 2.1 *Un vecteur non nul v est un vecteur propre de A si et seulement si Av est colinéaire à v.*
L'union de $\{0\}$ et de l'ensemble des vecteurs propres associés à une valeur propre λ de A est un sous-espace vectoriel de $\mathbb{K}^n$.

Définition 2.2 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. Soit $\lambda \in \mathbb{K}$ est une valeur propre de A. Le sous-espace $V_\lambda = \{v \in \mathbb{K}^n : Av = \lambda v\}$ (non réduit à $\{0\}$) est appelé sous-espace propre associé à la valeur propre λ. Le sous-espace V_λ est le noyau de $A - \lambda I$:*

$$V_\lambda = \ker(A - \lambda I).$$

Définition 2.3 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. L'ensemble des valeurs propres de A est appelé spectre de A et est noté $\sigma(A)$.*

Lemme 2.1 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. Soit $\lambda_1, \ldots, \lambda_k$ des valeurs propres distinctes de A. On a*

$$V_{\lambda_1} + \cdots + V_{\lambda_k} = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_k}$$

Démonstration. On effectue une démonstration par récurrence sur k : pour $k = 1$, c'est immédiat.
Supposons le résultat vrai pour les entiers positifs inférieurs strictement à k. Soient $w = u_1 + \cdots + u_k = v_1 + \cdots + v_k$, avec $u_i, v_i \in V_{\lambda_i}$. On a $(u_1 - v_1) + \cdots + (u_k - v_k) = 0$. Ceci implique que $\lambda_1(u_1 - v_1) + \cdots + \lambda_k(u_k - v_k) = 0$, en multipliant par A. Supposons que $\lambda_1 \neq 0$, (ce qui est vrai quitte à permuter les indices car $k > 1$). En divisant par λ_1 on obtient que $(u_1 - v_1) + \frac{\lambda_2}{\lambda_1}(u_2 - v_2) \cdots + \frac{\lambda_k}{\lambda_1}(u_k - v_k) = 0$. On en déduit en retranchant la première identité que $(\frac{\lambda_2}{\lambda_1} - 1)(u_2 - v_2) \cdots + (\frac{\lambda_k}{\lambda_1} - 1)(u_k - v_k) = 0$, avec $\frac{\lambda_i}{\lambda_1} - 1 \neq 0 \ \forall 2 \leq i \leq k$, et $u_i - v_i \in V_{\lambda_i}$. L'hypothèse de récurrence implique que $u_i = v_i$, pour $i = 2, \ldots, k$. On obtient alors que $u_1 = v_1$ en revenant à la première égalité.
Le résultat est démontré par récurrence. ■

Lemme 2.2 *Soient $A \in \mathcal{M}_n(\mathbb{K})$ et $\lambda_1, \ldots, \lambda_k$ des valeurs propres distinctes de A, $v_1, \ldots, v_k$ des vecteurs propres associés. Le système $(v_1, \ldots, v_k)$ est libre.*

Démonstration. Conséquence du Lemme 2.1. ■

Définition 2.4 *Soit $A \in \mathcal{M}_n(\mathbb{K})$. Le polynôme caractéristique de A est le polynôme $P_A(\lambda)$ à coefficients dans $\mathbb{K}$ défini par*

$$P_A(\lambda) = \det(A - \lambda I)$$

C'est un polynôme de degré égal à n et dont le coefficient du monôme de degré n est égal à $(-1)^n$. On a vu que les racines de P_A sont les valeurs propres de A.
La multiplicité d'une valeur propre λ de A est sa multiplicité en tant que racine de P_A. Une valeur propre de A de multiplicité un (respectivement deux) est appelée valeur propre simple (respectivement double).

On vérifie aisément la Proposition suivante

Lemme 2.3 *Le polynôme caractéristique est invariant par changement de base. En d'autres termes, deux matrices semblables ont le même polynôme caractéristique et le même spectre.*

Démonstration. Si Q est une matrice inversible,

$$\det(QAQ^{-1} - \lambda I) = \det(Q(A - \lambda I)Q^{-1}) = \det(A - \lambda I).$$

■

Exemple. La matrice

$$A = \begin{pmatrix} 5. & -1. & -2. & -1. & -1. \\ 1. & 3. & -2. & -1. & -1. \\ 1. & -1. & 2. & -1. & -1. \\ 1. & -1. & 0. & 1. & -1. \\ 4. & -4. & 0. & 0. & 0. \end{pmatrix}$$

a pour polynôme caractéristique $P_A(\lambda) = -\lambda(4-\lambda)^2(2-\lambda)(1-\lambda)$ (le vérifier) et $\sigma(A) = \{4, 2, 1, 0\}$. La valeur propre 4 est double et les autres valeurs propres sont simples. On a $V_4 = \{(a+b, b, 0, 0, a)^T, a, b \in \mathbb{R}\}$, $V_0 = \{(a, a, a, a, a)^T, a \in \mathbb{R}\}$, $V_2 = \{(a, a, a, 0, 0)^T, a \in \mathbb{R}\}$, $V_1 = \{(a, a, a, a, 0)^T, a \in \mathbb{R}\}$.

Exemple. Soit $a \in \mathbb{K}$, et $A = aI \in \mathcal{M}_n(\mathbb{K})$ la matrice de l'homothétie de rapport a. Tout vecteur non nul de $\mathbb{K}^n$ est un vecteur propre de A associé à la valeur propre a. Le polynôme caractéristique de A est $P_A(\lambda) = (a-\lambda)^n$.

Exemple. On considère une réel θ et la matrice $A \in \mathcal{M}_2(\mathbb{R})$ de la rotation d'angle θ

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{2.1}$$

Son polynôme caractéristique est $P_A(\lambda) = \lambda^2 - 2\lambda\cos\theta + 1$. Si $\theta \notin \pi\mathbb{Z}$, le polynôme n'a pas de racines dans $\mathbb{R}$ et A n'a donc pas de valeur propre dans $\mathbb{R}$.

En revanche, si on considère A comme une matrice de $\mathcal{M}_2(\mathbb{C})$, on voit que les racines de P_A sont $\lambda_\pm = \cos\theta \pm i\sin\theta$. Si $\theta \notin \pi\mathbb{Z}$, les valeurs propres de A sont les complexes conjugués (non réels) $\lambda_\pm$ et des vecteurs propres correspondants sont $v_\pm = (1, \pm i)^T$. On voit que pour toute valeur de θ, on peut trouver une base de $\mathbb{C}^2$ dans laquelle la matrice A devient diagonale. On dira que A est diagonalisable dans $\mathcal{M}_2(\mathbb{C})$.

Définition 2.5 *On dit qu'un sous-espace vectoriel V de $\mathbb{K}^n$ est un sous-espace stable pour A si pour tout vecteur $v \in V$, $Av \in V$.*

Exemple. On considère la matrice

$$A = \begin{pmatrix} 2 & 1 & 0 & 2 \\ 3 & 3 & 0 & 1 \\ 0 & 0 & 4 & 0 \\ 1 & 2 & 0 & 5 \end{pmatrix}$$

Les sous-espaces $\mathrm{Vect}(e_1, e_2, e_4)$ et $\mathrm{Vect}(e_3)$ sont stables pour A.

Comment traduire de façon matricielle qu'un sous-espace est stable pour A ? Si V est un sous-espace stable pour A de dimension $m \leq n$, le théorème de la base incomplète nous dit que l'on peut trouver une base $(f_1, \ldots, f_n)$ de K^n dont les m premiers vecteurs forment une base de V. Soit Q la matrice de changement de base (dont les colonnes contiennent les coordonnées des vecteurs f_i dans la base canonique), la matrice $Q^{-1}AQ$ est de la forme

$$Q^{-1}AQ = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \tag{2.2}$$

où $B_{11} \in \mathcal{M}_m(\mathbb{K})$ et $B_{22} \in \mathcal{M}_{n-m}(\mathbb{K})$.

Réciproquement si il existe une matrice inversible Q telle que $Q^{-1}AQ$ est de la forme (2.2), alors le sous-espace V engendré par les m premières colonnes de Q est stable pour A.

Lemme 2.4 *Soient $A \in \mathcal{M}_n(\mathbb{K})$ et une matrice inversible Q telle que $Q^{-1}AQ$ soit de la forme (2.2), alors*

$$P_A = P_{B_{11}} P_{B_{22}}. \tag{2.3}$$

Démonstration. En notant $B = Q^{-1}AQ$, on a $P_A = P_B$ d'après le Lemme 2.3 et on voit immédiatement que $P_B = P_{B_{11}} P_{B_{22}}$. ■

2.2 Rappels sur les polynômes et polynômes de matrices

Dans la suite, on va avoir besoin de quelques résultats fondamentaux sur les polynômes à coefficients dans $\mathbb{K}$.
On rappelle que l'ensemble $\mathbb{K}[X]$ des polynômes à coefficients dans $\mathbb{K}$ est un anneau, c'est à dire que

- $(\mathbb{K}[X], +)$ est un groupe commutatif.
- La multiplication $\times$ qui à deux polynômes P et Q associe le produit PQ, est une loi de composition interne, commutative et associative dans $\mathbb{K}[X]$, avec pour élément neutre le polynôme 1.
- La multiplication $\times$ est distributive sur l'addition : soit P, Q, R trois polynômes de $\mathbb{K}[X]$, on a $P(Q + R) = PQ + PR$.

On résume ces propriétés en disant que $(\mathbb{K}[X], +, \times)$ est un anneau commutatif.

Définition 2.6 *On dit qu'un polynôme est unitaire si son coefficient de plus haut degré est* 1*.*

Définition 2.7 *Soit P et Q deux polynômes de $\mathbb{K}[X]$. On dit que Q est un diviseur de P (ou que P est un multiple de Q) s'il existe un polynôme R tel que $P = QR$.*

Définition 2.8 *Soit $P \in \mathbb{K}[X]$. On dit que $\lambda \in \mathbb{K}$ est racine de P de multiplicité m (m entier > 0) si le polynôme $(X - \lambda)^m$ est un diviseur de P.*

On admettra le résultat important suivant :

Théorème 2.1 *Le corps $\mathbb{C}$ est algébriquement clos, ce qui signifie que tout polynôme $P \in \mathbb{C}[X]$ non nul est scindé, c'est à dire que si le degré de P est n, P a exactement n racines dans $\mathbb{C}$ (comptées avec leur multiplicité).*

Il est clair que $\mathbb{R}$ n'est pas algébriquement clos, puisque le polynôme $X^2 + 1$ n'admet pas de racine dans $\mathbb{R}$.

Définition 2.9 *On dit que P et Q sont premiers entre eux si P et Q n'ont pas d'autre diviseur commun que les polynômes de degré* 0*.*

Remarque 2.2 *Comme $\mathbb{C}$ est algébriquement clos, $P \in \mathbb{C}[X]$ et $Q \in \mathbb{C}[X]$ sont premiers entre eux si et seulement P et Q n'ont pas de racines communes dans $\mathbb{C}$.*

Exemple. Les polynômes $P(X) = 3X + 5$ et $Q(X) = X^2 + 2x + 1$ sont premiers entre eux.

Comme pour les entiers, on peut définir une division euclidienne dans $\mathbb{K}[X]$:

Proposition 2.1 (division euclidienne) *Soient P_1 et P_2 deux polynômes : il existe un unique polynôme Q et un unique polynôme R de degré strictement inférieur à celui de P_2 tels que $P_1 = QP_2 + R$.*

Démonstration. La preuve consiste à écrire l'algorithme de la division euclidienne. ■

Exemple. On considère les polynômes $P_1(X) = 5X^5 + 6X^3 + 2X + 1$ et $P_2(X) = X^3 + X + 2$. On a $P_1 = QP_2 + R$, où $Q(X) = 5X^2 + 1$ et $R(X) = -10X^2 + X - 1$.

Proposition 2.2 (identité de Bézout) *Deux polynômes P_1 et P_2 sont premiers entre eux si et seulement si il existe deux polynômes Q_1 et Q_2 tels que $Q_1P_1 + Q_2P_2 = 1$.*

Démonstration. Commençons par supposer qu'il existe deux polynômes Q_1 et Q_2 tels que $Q_1P_1 + Q_2P_2 = 1$. Un diviseur commun à P_1 et P_2 est donc un diviseur de 1, donc P_1 et P_2 sont premiers entre eux.
Pour la réciproque, on procède par récurrence sur le degré maximal des deux polynômes P_1 et P_2, que l'on note n. Pour $n = 0$, on voit que deux polynômes de degré 0 sont premiers entre eux si l'un des deux polynômes est non nul. Le résultat souhaité est donc vrai si $n = 0$.
Soit $n > 0$, supposons la propriété vraie pour tous les entiers positifs ou nuls inférieur à n. Prenons P_1 et P_2 deux polynômes :
supposons dans un premier temps que le degré de P_1 est n et que le degré de P_2 est strictement inférieur à n. On effectue la division euclidienne de P_1 par P_2 : on a $P_1 = QP_2 + R$, où R est un polynôme de degré inférieur strictement à celui de P_2. On vérifie que R et P_2 sont premiers entre eux, car un diviseur commun à P_2 et R est un diviseur commun à P_1 et P_2. On applique l'hypothèse de récurrence à P_2 et R : il existe deux polynômes A et B tels que $AP_2 + BR = 1$. On en conclut que $AP_2 + B(P_1 - QP_2) = 1$ ce qui est la conclusion souhaitée.
Si P_1 et P_2 sont tous deux de degré n : on effectue la division euclidienne de P_1 par P_2 : on a $P_1 = QP_2 + R$, où R est un polynôme de degré inférieur strictement à celui de P_2, et R est premier avec P_2. On sait alors qu'il existe deux polynômes A et B tels que $AP_2 + BR = 1$ puisque R est de degré strictement inférieur à n. On en conclut que $AP_2 + B(P_1 - QP_2) = 1$. ■

Exemple. Les polynômes $P_1(X) = 3X + 5$ et $P_2(X) = X^2 + 2x + 1$ sont premiers entre eux. On a bien $Q_1P_1 + Q_2P_2 = 1$, avec $Q_1(X) = -\frac{3}{4}X - \frac{1}{4}$ et $Q_2(X) = \frac{9}{4}$.

Définition 2.10 *Soit $P(X)$ un polynôme à coefficients dans $\mathbb{K}$: $P(X) = \alpha_m X^m + \cdots + \alpha_0 = \sum_{i=0}^{m} \alpha_i X^i$. Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$, on note $P(A)$ la matrice $P(A) = \alpha_m A^m + \cdots + \alpha_0 I = \sum_{i=0}^{m} \alpha_i A^i$. On dit que $P(A)$ est un polynôme de la matrice A.*

On vérifie aisément les lemmes suivants

Lemme 2.5 *Soient P et Q deux polynômes à coefficients dans $\mathbb{K}$, et A une matrice de $\mathcal{M}_n(\mathbb{K})$, alors $P(A)$ et $Q(A)$ commutent :*

$$P(A)Q(A) = Q(A)P(A) = PQ(A) = QP(A)$$

Lemme 2.6 *Soient P un polynôme à coefficients dans $\mathbb{K}$, et A une matrice de $\mathcal{M}_n(\mathbb{K})$. Si λ est une valeur propre de A alors $P(\lambda)$ est une valeur de $P(A)$.*

Démonstration. Soit λ une valeur propre de A. Il existe un vecteur $v \neq 0$ tel que $Av = \lambda v$. On voit facilement que $P(A)(v) = P(\lambda)v$. On en déduit que $P(A)$ a pour valeur propre $P(\lambda)$. ■

Remarque 2.3 *On verra plus tard que si $A \in \mathcal{M}_n(\mathbb{C})$, $\sigma(P(A))$ est exactement l'image par P de $\sigma(A)$.*
Ceci n'est pas vrai pour $A \in \mathcal{M}_n(\mathbb{R})$, car le spectre de la matrice A donnée par (2.1) avec $\theta = \frac{\pi}{n}$ est vide si $n > 1$, alors que le spectre de A^n est $\{-1\}$.

Lemme 2.7 *Soient P un polynôme à coefficients dans $\mathbb{K}$, et A une matrice de $\mathcal{M}_n(\mathbb{K})$. Le sous-espace* $\ker(P(A))$ *est stable pour A.*

Démonstration. Soit $v \in \ker(P(A))$, on a d'après le lemme 2.5, $P(A)Av = A\,P(A)v = 0$. ■

Proposition 2.3 *Soient P_1 et P_2 deux polynômes premiers entre eux et $A \in \mathcal{M}_n(\mathbb{K})$. Les sous-espaces* $\ker(P_1(A))$ *et* $\ker(P_2(A))$ *sont en somme directe et*

$$\ker((P_1P_2)(A)) = \ker(P_1(A)) \oplus \ker(P_2(A)) \tag{2.4}$$

Démonstration. Pour montrer que $\ker(P_1(A)) \cap \ker(P_2(A)) = \{0\}$, on utilise l'identité de Bézout : il existe deux polynômes Q_1 et Q_2 tels que $Q_1P_1 + Q_2P_2 = 1$. Donc $Q_1(A)P_1(A) + Q_2(A)P_2(A) = I$. Soit $v \in \ker(P_1(A)) \cap \ker(P_2(A))$, on a $v = (Q_1(A)P_1(A) + Q_2(A)P_2(A))v = 0$. Pour $v_1 \in \ker(P_1(A))$ et $v_2 \in \ker(P_2(A))$, le Lemme 2.5 implique que $(P_1P_2)(A)(v_1 + v_2) = P_2(A)P_1(A)v_1 + P_1(A)P_2(A)v_2 = 0$. Donc $\ker(P_1(A)) \oplus \ker(P_2(A)) \subset \ker((P_1P_2)(A))$.
Réciproquement en utilisant l'identité de Bézout, si $v \in \ker((P_1P_2)(A))$, on a $v = Q_1(A)P_1(A)v + Q_2(A)P_2(A)v$ et $Q_1(A)P_1(A)v \in \ker(P_2(A))$, $Q_2(A)P_2(A)v \in \ker(P_1(A))$; on a bien montré que $\ker((P_1P_2)(A)) \subset \ker(P_1(A)) \oplus \ker(P_2(A))$. ■

2.3 Polynôme minimal

On va s'intéresser à trouver des polynômes P qui annulent A, c'est à dire tels que $P(A) = 0$. De tels polynômes existent : en effet, on sait que la dimension de $\mathcal{M}_n(\mathbb{K})$ est n^2. Par suite, la famille de matrices $(I, A, A^2, \ldots, A^{n^2})$ ne peut pas être une famille libre de $\mathcal{M}_n(\mathbb{K})$. En conséquence, il existe une famille de coefficients non tous nuls $(\alpha_i)_{i=0,\ldots,n^2}$ tels que $\sum_{i=0}^{n^2} \alpha_i A^i = 0$. Le polynôme $P(X) = \sum_{i=0}^{n^2} \alpha_i X^i$ annule donc A. On note $\mathcal{I}_A$ l'ensemble des polynômes qui annulent A.
On vérifie facilement que

- $\mathcal{I}_A$ est un sous-groupe de $(\mathbb{K}[X], +)$,
- quels que soient $P \in \mathbb{K}[X]$, $Q \in \mathcal{I}_A$, on a $PQ \in \mathcal{I}_A$.

On résume ces deux propriétés en disant que $\mathcal{I}_A$ est un idéal de $\mathbb{K}[X]$.

Proposition 2.4 (polynôme minimal) *Il existe un unique polynôme unitaire $\mu_A \in \mathcal{I}_A$, tel que tout polynôme de $\mathcal{I}_A$ soit un multiple de μ_A. On a*

$$\mathcal{I}_A = \{\mu_A P, P \in \mathbb{K}[X]\}. \tag{2.5}$$

On dit que μ_A est le polynôme minimal de A.

Démonstration. Il est clair que le polynôme 1 n'appartient pas à $\mathcal{I}_A$.
Soit $\mu_A \in \mathcal{I}_A$ un polynôme unitaire de $\mathcal{I}_A$ de degré minimal, (on vient de voir que son degré est non nul).
Soit donc P un polynôme de $\mathcal{I}_A$. On veut montrer que P est un multiple de μ_A. On effectue la division euclidienne de P par μ_A : il existe un unique polynôme Q et un unique polynôme R de degré strictement inférieur à celui de μ_A tels que $P = Q\mu_A + R$. On voit alors que

- $R \in \mathcal{I}_A$ car $\mathcal{I}_A$ est un idéal de $\mathbb{K}[X]$.
- Le degré de R est strictement inférieur à celui de μ_A.

Le polynôme R ne peut être que nul : si ce n'était pas le cas, en divisant R par son coefficient de plus haut degré, on obtiendrait un polynôme unitaire de $\mathcal{I}_A$ de degré strictement inférieur à celui de μ_A. Donc P est un multiple de μ_A.
On a donc trouvé un polynôme unitaire μ_A de degré ≥ 1 tel que tout polynôme de $\mathcal{I}_A$ soit un multiple de μ_A. L'unicité de μ_A et (2.5) sont laissés en exercice. ■

Remarque 2.4 *En fait, pour tout idéal $\mathcal{J}$ de $\mathbb{K}[X]$, il existe un unique polynôme $G \in \mathcal{J}$ tel que $\mathcal{J} = \{GP, P \in \mathbb{K}[X]\}$. Cette propriété est une propriété de l'anneau $\mathbb{K}[X]$: on dit que $\mathbb{K}[X]$ est un anneau principal.*

Remarque 2.5 *Le polynôme minimal est moins facile à calculer que le polynôme caractéristique, car ce dernier est défini par le calcul d'un déterminant. Le lien entre les deux polynômes est donné par le théorème d'Hamilton-Cayley qui nous dit que le polynôme minimal divise le polynôme caractéristique et qu'ils ont les mêmes racines, (mais les racines n'ont pas forcément les mêmes multiplicités).*

Exemple. Le polynôme minimal de la matrice $A = aI$ (homothétie de rapport a) est $\mu_A(X) = X - a$, tandis que le polynôme caractéristique est $P_A(X) = (a - X)^n$.

2.4 Théorème de Hamilton-Cayley

Théorème 2.2 (Hamilton-Cayley) *Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$ et P_A son polynôme caractéristique. On a $P_A \in \mathcal{I}_A$ (ou de manière équivalente $P_A(A) = 0$).*

Démonstration. Soit $v \in \mathbb{K}^n$.
- Si $v = 0$, alors $P_A(A)v = 0$.
- Si $v \neq 0$, on considère l'espace vectoriel $V = \text{Vect}(v, Av, \ldots, A^n v)$. Le sous-espace V est de dimension $m \leq n$ et $(v, Av, \ldots, A^{m-1}v)$ est une base de V. Il existe donc m coefficients $(\alpha_0, \ldots, \alpha_{m-1})$ tels que
$$A^m v = \sum_{i=0}^{m-1} \alpha_i A^i v \tag{2.6}$$
et V est donc un sous-espace stable pour A. On complète $(v, Av, \ldots, A^{m-1}v)$ par $(f_{m+1}, \ldots, f_n)$ pour former une base de $\mathbb{K}^n$ et on note Q la matrice de changement de base. On a (2.2) et
$$B_{11} = \begin{pmatrix} 0 & \ldots & & & \alpha_0 \\ 1 & 0 & \ldots & & \alpha_1 \\ 0 & 1 & 0 & & \alpha_2 \\ & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & \alpha_{m-1} \end{pmatrix}$$
D'après le Lemme 2.4, le polynôme $P_{B_{11}}$ divise P_A. On peut facilement calculer $P_{B_{11}}(\lambda)$ en développant
$$\begin{vmatrix} -\lambda & \ldots & & & \alpha_0 \\ 1 & -\lambda & \ldots & & \alpha_1 \\ 0 & 1 & -\lambda & & \alpha_2 \\ & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & \alpha_{m-1} - \lambda \end{vmatrix}$$

par rapport à la première ligne. On a que

$$\begin{vmatrix} -\lambda & \dots & & & \alpha_0 \\ 1 & -\lambda & \dots & & \alpha_1 \\ 0 & 1 & -\lambda & & \alpha_2 \\ & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & \alpha_{m-1}-\lambda \end{vmatrix} = (-1)^{m+1}\alpha_0 - \lambda \begin{vmatrix} -\lambda & \dots & & \alpha_1 \\ 1 & -\lambda & & \alpha_2 \\ & \ddots & \ddots & \vdots \\ & & 1 & \alpha_{m-1}-\lambda \end{vmatrix}$$

On démontre alors par récurrence que $P_{B_{11}}(\lambda) = (-1)^m(\lambda^m - \sum_{i=0}^{m-1}\alpha_i\lambda^i)$.
On en déduit que $P_A(A)v = P_{B_{22}}(A)P_{B_{11}}(A)v = (-1)^m P_{B_{22}}(A)(\sum_{i=0}^{m-1}\alpha_i A^i v - A^m v) = 0$, d'après (2.6).
On a donc démontré que $P_A(A)v = 0$ pour tout $v \in \mathbb{K}^n$, ce qui implique que $P_A(A) = 0$. ■

Remarque 2.6 *Pour un polynôme unitaire de degré m, $P(X) = \sum_{i=0}^{m-1}\alpha_i X^i + X^m$, la matrice*

$$\begin{pmatrix} 0 & \dots & & & -\alpha_0 \\ 1 & 0 & \dots & & -\alpha_1 \\ 0 & 1 & 0 & & -\alpha_2 \\ & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & -\alpha_{m-1} \end{pmatrix}$$

s'appelle la matrice compagnon du polynôme P. Son polynôme caractéristique est $(-1)^m P(X)$.

Théorème 2.3 (Hamilton Cayley, version complète) *Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$. Le polynôme minimal μ_A divise le polynôme caractéristique P_A. De plus, les racines du polynôme minimal μ_A sont exactement celles du polynômes caractéristiques (mais avec des multiplicités éventuellement différentes)*

Démonstration. La première assertion est une conséquence de la Proposition 2.4 et du Théorème 2.2 .
On sait que μ_A est un diviseur de P_A, donc une racine de μ_A est nécessairement une racine de P_A, donc une valeur propre de A. Réciproquement, soit λ une valeur propre et v un vecteur propre associé, on a $0 = \mu_A(A)v = \mu_A(\lambda)v$ ce qui implique que $\mu_A(\lambda) = 0$. ■

Exemple. Si $\theta \notin \pi\mathbb{Z}$, le polynôme minimal de la matrice

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

(rotation d'angle θ) est $\mu_A(X) = X^2 - 2X\cos\theta + 1$. En effet, on sait que A (vue comme une matrice à coefficients complexes) a ses deux valeurs propres distinctes (de multiplicité un) et le théorème de Hamilton-Cayley implique alors que $\mu_A = P_A$.
Exemple. Considérons les trois matrices

$$A = \begin{pmatrix} a & 1 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{pmatrix} \qquad B = \begin{pmatrix} a & 1 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{pmatrix} \qquad C = \begin{pmatrix} a & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{pmatrix} \tag{2.7}$$

On a $P_A(X) = P_B(X) = P_C(X) = (a - X)^4$. D'après le théorème d'Hamilton-Cayley, μ_A, μ_B et μ_C sont de la forme $(X - a)^m$. On peut vérifier que $\mu_A(X) = \mu_B(X) = (X - a)^2$ et que $\mu_C(X) = (X - a)^4$. Le sous-espace propre de A associé à a est de dimension 3. Le sous-espace propre de B associé à a est de dimension 2. Le sous-espace propre de C associé à a est de dimension 1.

Exercice. Cet exercice consiste à prouver le théorème d'Hamilton-Cayley d'une autre manière, par des arguments d'"analyse". Montrer que l'ensemble des matrices diagonalisables est dense dans $\mathcal{M}_n(\mathbb{C})$. En déduire le théorème de Hamilton Cayley.

2.5 Diagonalisation

Définition 2.11 *On dit qu'une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est diagonalisable si et seulement si il existe une base de $\mathbb{K}^n$ formée de vecteurs propres de A, ou de manière équivalente si A est semblable à une matrice diagonale.*

Une conséquence immédiate du Lemme 2.2 est la

Proposition 2.5 *Si A a n valeurs propres distinctes, alors A est diagonalisable.*

Une conséquence du Lemme 2.1 est la

Proposition 2.6 *Si $A \in \mathcal{M}_n(\mathbb{K})$ a exactement k valeurs propres distinctes $\lambda_1, \ldots, \lambda_k$, alors A est diagonalisable si et seulement si* $\dim(V_{\lambda_1}) + \cdots + \dim(V_{\lambda_k}) = n$.

Démonstration. On sait par le Lemme 2.1 que les sous-espaces propres de A sont en somme directe, donc si $\dim(V_{\lambda_1}) + \cdots + \dim(V_{\lambda_k}) = n$, on a $\mathbb{K}^n = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_k}$, donc A est diagonalisable. La réciproque est immédiate. ■

Lemme 2.8 *Soit $A \in \mathcal{M}_n(\mathbb{K})$ et soit α une valeur propre de la matrice A. Soit p la dimension du sous-espace propre V_α. Le polynôme $(\alpha - X)^p$ est un diviseur de P_A.*

Démonstration. D'après la définition de V_α, A induit sur V_α une homothétie de rapport α, dont le polynôme caractéristique est $(\alpha - X)^p$. Comme V_α est un sous-espace propre pour A, on déduit du Lemme 2.4 que le polynôme $(\alpha - X)^p$ divise P_A. ■

On en déduit de ce lemme une condition nécessaire et suffisante pour que A soit diagonalisable :

Théorème 2.4 *Si $A \in \mathcal{M}_n(\mathbb{K})$ a exactement k valeurs propres distinctes $\lambda_1, \ldots, \lambda_k$, alors A est diagonalisable si et seulement si*

$$P_A(X) = \prod_{i=1}^{k} (\lambda_i - X)^{m_i}, \tag{2.8}$$

où m_i est la dimension du sous-espace propre V_{λ_i}.

Démonstration. D'après la proposition 2.6, A est diagonalisable si et seulement si $\sum_{i=1}^{k} m_i = n$ et donc si et seulement si $\prod_{i=1}^{k}(\lambda_i - X)^{m_i}$ est un polynôme de degré n, et le lecteur vérifiera que ceci équivaut à (2.8) car on a vu au Lemme 2.8 que $\prod_{i=1}^{k}(\lambda_i - X)^{m_i}$ divise P_A. ■

Remarque 2.7 *La dimension du sous-espace propre associé à la valeur propre λ_i de A peut être strictement inférieure à la multiplicité m_i de λ_i. Dans ce cas, A n'est pas diagonalisable, cf. l'exemple des matrices données par (2.7).*

Le Théorème de Hamilton-Cayley permet de donner une condition nécessaire pour que A soit diagonalisable en fonction de la structure du polynôme minimal :

Théorème 2.5 *Soit $A \in \mathcal{M}_n(\mathbb{K})$, A est diagonalisable si et seulement si son polynôme minimal μ_A est de la forme*

$$\mu_A(X) = \prod_{i=1}^{k}(X - \lambda_i), \qquad \text{avec } \lambda_i \neq \lambda_j \text{ si } i \neq j,$$

En d'autres termes, A est diagonalisable si et seulement si μ_A est scindé avec des racines simples.

Démonstration. Supposons que A a k valeurs propres distinctes $\lambda_1, \ldots, \lambda_k$. Si A est diagonalisable, on voit facilement que le polynôme $\prod_{i=1}^{k}(X - \lambda_i)$ annule A et donc μ_A divise ce polynôme. Comme on sait d'après le Théorème 2.3 que les racines de μ_A sont exactement les valeurs propres de A, on en déduit que $\mu_A = \prod_{i=1}^{k}(X - \lambda_i)$, et on voit que les racines du polynôme minimal sont simples.
Réciproquement, si le polynôme minimal a k racines simples $\lambda_1, \ldots, \lambda_k$, ces racines sont exactement les valeurs propres de A, d'après le Théorème 2.3. Soit $V_{\lambda_i} = \ker(A - \lambda_i I)$. On a d'après la Proposition 2.3 que $\mathbb{K}^n = \ker(\mu_A(A)) = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_k}$ car les polynômes $X - \lambda_i$ sont premiers entre eux. On conclut en utilisant la Proposition 2.6. ■

Exemple. D'après le Théorème 2.5, aucune des trois matrices dans (2.7) n'est diagonalisable.

2.6 Triangulation

Définition 2.12 *On dit qu'une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est triangularisable si elle est semblable à une matrice triangulaire supérieure $U \in \mathcal{M}_n(\mathbb{K})$, c'est à dire s'il existe une matrice inversible $Q \in \mathcal{M}_n(\mathbb{K})$ et une matrice triangulaire supérieure $U \in \mathcal{M}_n(\mathbb{K})$ telles que*

$$Q^{-1}AQ = U. \tag{2.9}$$

Il est clair que si $A \in \mathcal{M}_n(\mathbb{K})$ vérifie (2.9), alors

$$P_A(\lambda) = \prod_{i=1}^{n}(u_{ii} - \lambda),$$

et

- les coefficients diagonaux de U sont les valeurs propres de A,
- le polynôme caractéristique de A a n racines dans $\mathbb{K}$, comptées avec leur multiplicité. On dit que le polynôme caractéristique P_A est *scindé* dans $\mathbb{K}[X]$.

Donc, une matrice de $\mathcal{M}_n(\mathbb{R})$ dont le polynôme caractéristique est divisible par $x^2+a^2, a \in \mathbb{R}\backslash\{0\}$ ne peut pas être triangularisable dans $\mathcal{M}_n(\mathbb{R})$. On verra que cette même matrice, vue comme une matrice de $\mathcal{M}_n(\mathbb{C})$ est diagonalisable.
Le caractère scindé du polynôme caractéristique est en fait une condition nécessaire et suffisante pour que A soit triangularisable :

Théorème 2.6 *Une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est triangularisable si et seulement si son polynôme caractéristique est scindé.*

Démonstration. Il ne reste plus qu'à démontrer l'implication : P_A scindé $\Rightarrow$ A triangularisable. On le fait par récurrence sur n. Pour $n = 1$, la propriété est évidente. Supposons la propriété vraie pour $n-1$. Soit $A \in \mathcal{M}_n(\mathbb{K})$ et v un vecteur propre de A associé à la valeur propre λ. On peut par le théorème de la base incomplète trouver une base $(v, f_2, \ldots, f_n)$ de $\mathbb{K}^n$. Si on note $Q^{(1)}$ la matrice dont les colonnes sont ces vecteurs de bases, on a

$$Q^{(1)^{-1}} A Q^{(1)} = \begin{pmatrix} \lambda & l \\ 0 & B \end{pmatrix}, \quad l \in \mathcal{M}_{1,n-1}(\mathbb{K}), \quad B \in \mathcal{M}_{n-1}(\mathbb{K}).$$

D'après le Lemme 2.4, $P_A(X) = (\lambda - X)P_B(X)$. Comme P_A est scindé, P_B est un polynôme de degré $n-1$ scindé. On applique alors l'hypothèse de récurrence à B : il existe une matrice inversible $Q^{(2)} \in \mathcal{M}_{n-1}(\mathbb{K})$ et une matrice $U^{(2)} \in \mathcal{M}_{n-1}(\mathbb{K})$ triangulaire supérieure telles que $Q^{(2)^{-1}} B Q^{(2)} = U^{(2)}$. En notant $U \in \mathcal{M}_n(\mathbb{K})$ et $Q \in \mathcal{M}_n(\mathbb{K})$ les matrices

$$U = \begin{pmatrix} \lambda & l \\ 0 & U^{(2)} \end{pmatrix}, \quad Q = Q^{(1)} \begin{pmatrix} 1 & 0 \\ 0 & Q^{(2)} \end{pmatrix},$$

on vérifie que Q est inversible, U triangulaire supérieure et on a (2.9). ■

On déduit alors du Théorème 2.6 le

Théorème 2.7 *Toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ est triangularisable.*

Remarque 2.8 *Insistons bien sûr le fait que toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ n'est pas forcément triangularisable. On prendra par exemple la matrice de rotation*

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

dont le polynôme caractéristique est X^2+1. Cette matrice n'est pas triangularisable dans $\mathcal{M}_2(\mathbb{R})$ alors qu'elle est diagonalisable dans $\mathcal{M}_2(\mathbb{C})$

Pour des matrices réelles, on a le résultat suivant :

Proposition 2.7 *Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$. Il existe une matrice $Q \in \mathcal{M}_n(\mathbb{R})$ inversible et une matrice $R \in \mathcal{M}_n(\mathbb{R})$ triangulaire supérieure par blocs*

$$R = \begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1k} \\ 0 & R_{22} & \ldots & R_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & R_{kk} \end{pmatrix} \tag{2.10}$$

où R_{ii} est soit une matrice 1×1 (un réel) ou une matrice de $\mathcal{M}_2(\mathbb{R})$ avec des valeurs propres complexes conjuguées et non réelles telles que

$$Q^{-1} A P = R.$$

Démonstration. Montrons le résultat par récurrence sur n. Supposons la propriété vraie pour des matrices d'ordre inférieur ou égal à $n-1$.
Soit $A \in \mathcal{M}_n(\mathbb{R})$, son polynôme caractéristique est à coefficients réels. Il a n racines dans $\mathbb{C}$, et les racines non réelles se présentent par paires de nombres complexes conjugués.
Si toutes les racines sont réelles, on peut trouver $Q \in \mathcal{M}_n(\mathbb{R})$ inversible telle que $Q^{-1}AQ$ soit triangulaire avec les valeurs propres sur la diagonale.
Supposons que P_A a deux racines complexes conjuguées : $\lambda = \gamma + i\mu$, $\mu \neq 0$ et $\bar{\lambda}$: il existe deux vecteurs y et z de $\mathbb{R}^n$ tels que $A(y \pm iz) = (\gamma \pm i\mu)(y \pm iz)$. On a donc

$$A(y,z) = (y,z)\begin{pmatrix} \gamma & \mu \\ -\mu & \gamma \end{pmatrix}$$

Donc y, z engendrent un sous-espace vectoriel stable pour A : Par suite, il existe une matrice inversible $Q_1 \in \mathcal{M}_n(\mathbb{R})$ telle que

$$Q_1^{-1}AQ_1 = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix},$$

où $T_{11} \in \mathbb{R}^{2\times 2}$ a pour valeurs propres de λ et $\bar{\lambda}$. On applique l'hypothèse de récurrence à T_{22} : il existe une matrice inversible $\tilde{Q}_2 \in \mathcal{M}_{n-2}(\mathbb{R})$ telle que $\tilde{Q}_2^{-1}T_{22}\tilde{Q}_2$ ait la forme donnée dans (2.10). On pose alors $Q = Q_1\text{Block-Diag}(I_2, \tilde{Q}_2)$, et on a $Q^TAQ = R$ où R a la forme donnée dans (2.10). ■

En fait, la connaissance du polynôme caractéristique (qu'on suppose ici scindé) donne des renseignements précis sur la structure de la matrice U dans (2.9) :

Définition 2.13 *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice dont le polynôme caractéristique P_A est scindé : $P_A(X) = \prod_{i=1}^k (\lambda_i - X)^{m_i}$, avec $m_1 + \cdots + m_k = n$. Notons P_i le polynôme $P_i(X) = (\lambda_i - X)^{m_i}$. On appelle sous-espace caractéristique associé à la valeur propre λ_i le sous-espace*

$$E_i = \ker(P_i(A)). \tag{2.11}$$

Théorème 2.8 *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice dont le polynôme caractéristique P_A est scindé : $P_A(X) = \prod_{i=1}^k (\lambda_i - X)^{m_i}$, avec $m_1 + \cdots + m_k = n$. Les sous-espaces caractéristiques E_i sont stables pour A et $\dim(E_i) = m_i$. On a*

$$\mathbb{K}^n = E_1 \oplus \cdots \oplus E_k \tag{2.12}$$

et pour chaque i, $1 \leq i \leq k$, il existe une base $q_{i,1}, \ldots q_{i,m_i}$ de E_i, telle que si on appelle Q la matrice de $\mathcal{M}_n(\mathbb{K})$ dont les colonnes sont $q_{1,1}, \ldots q_{1,m_1}, \ldots, q_{k,1}, \ldots q_{k,m_k}$, alors $Q^{-1}AQ$ est diagonale par blocs,

$$Q^{-1}AQ = \begin{pmatrix} T_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & T_k \end{pmatrix}, \quad T_i \in \mathcal{M}_{m_i}(\mathbb{K}), \tag{2.13}$$

et chaque bloc diagonal T_i est triangulaire supérieur avec ses coefficients diagonaux égaux à λ_i. On écrit aussi

$$Q^{-1}AQ = \textit{Bloc-Diag}(T_1, \ldots, T_k), \quad T_i \in \mathcal{M}_{m_i}(\mathbb{K}). \tag{2.14}$$

Démonstration. La stabilité des E_i pour A est une conséquence du Lemme 2.7. Le Théorème d'Hamilton-Cayley nous dit que $\ker(P(A)) = \mathbb{K}^n$. On obtient alors (2.12) en appliquant (plusieurs fois) la Proposition 2.3. Soit alors une base de $\mathbb{K}^n$ obtenue en assemblant des bases de $E_1, \ldots, E_k$ et $\tilde{Q}$ la matrice de changement de base : on obtient que

$$\tilde{Q}^{-1}A\tilde{Q} = \text{Bloc-Diag}(\tilde{T}_1, \ldots, \tilde{T}_k) = \begin{pmatrix} \tilde{T}_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \tilde{T}_k \end{pmatrix}, \quad \tilde{T}_i \in \mathcal{M}_{m_i}(\mathbb{K}).$$

On sait que $P_A = P_{\tilde{T}_1} \ldots P_{\tilde{T}_k}$, et donc que chaque $P_{\tilde{T}_i}$ est scindé. D'autre part, comme $(\tilde{T}_i - \lambda_i I)^{m_i} = 0$, le théorème d'Hamilton-Cayley nous dit que $P_{\tilde{T}_i}$ ne peut avoir d'autre racine que λ_i. Donc $P_{\tilde{T}_i} = P_i = (\lambda_i - X)^{m_i}$, ce qui montre que

- $\tilde{T}_i$ est triangularisable avec des λ_i sur la diagonale.
- la dimension de E_i est m_i.

On a prouvé (2.13). ■

Dans la décomposition (2.13), le bloc T_i est de la forme $D_i + N_i$ où $D_i = \lambda_i I_{m_i}$ et N_i est une matrice de $\mathcal{M}_{m_i}(\mathbb{K})$, triangulaire supérieure strictement. Il est clair que D_i et N_i commutent. On a donc $Q^{-1}AQ = D + N$ où $D = \text{Bloc-Diag}(D_1, \ldots, D_k)$ et $N = \text{Bloc-Diag}(N_1, \ldots, N_k)$, D est diagonale, N est strictement triangulaire supérieure, et D et N commutent.

Définition 2.14 *On dit qu'une matrice $N \in \mathcal{M}_n\mathbb{K}$ est nilpotente s'il existe un entier $m > 0$ tel que $N^m = 0$.*

On a $N^n = 0$: N est donc une matrice nilpotente. On a presque démontré la

Proposition 2.8 (décomposition de Dunford) *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice dont le polynôme caractéristique P_A est scindé : A peut s'écrire comme la somme d'une matrice diagonalisable et d'une matrice nilpotente N ($N^n = 0$), qui commutent.*

Démonstration, suite et fin. On a vu que $Q^{-1}AQ = D + N$, donc $A = QDQ^{-1} + QNQ^{-1}$. La matrice QDQ^{-1} est clairement diagonalisable. On a $(QNQ^{-1})^n = QN^nQ^{-1} = 0$, et $QDQ^{-1}QNQ^{-1} = QDNQ^{-1} = QNDQ^{-1} = QNQ^{-1}QDQ^{-1}$. ■

Les blocs N_i étant nilpotents, on va montrer qu'on peut en effectuant encore un changement de base les mettre sous une forme particulièrement simple.

2.7 Matrices nilpotentes

Soit $N \in \mathcal{M}_n(\mathbb{K})$ une matrice telle que, pour un entier $m > 0$, $N^{m+1} = 0$ et $N^m \neq 0$. On note $E_i = \ker(N^i)$. On a $E_{m+1} = \mathbb{K}^n$ et $E_0 = \{0\}$.

Lemme 2.9 *Pour $0 \leq i \leq m$*

$$v \in E_{i+1} \Leftrightarrow Nv \in E_i. \tag{2.15}$$

Démonstration.

$$v \in E_{i+1} \Leftrightarrow N^{i+1}v = 0 \Leftrightarrow Nv \in E_i.$$

■

Lemme 2.10 *Soit $E_i = \ker(N^i)$. On a la suite d'inclusions strictes*

$$E_0 = \{0\} \subsetneq E_1 \subsetneq \cdots \subsetneq E_m \subsetneq E_{m+1} = \mathbb{K}^n. \tag{2.16}$$

Démonstration. L'inclusion $E_i \subset E_{i+1}$ est claire. Si $E_{i+1} = E_i$ pour $i < m+1$, alors $v \in E_{i+2}$ si et seulement si $Nv \in E_{i+1} = E_i$, ce qui a lieu si et seulement si $v \in E_{i+1}$. Donc $E_i = E_{i+1} = E_{i+2}$ et en itérant on montre que $E_i = \cdots = E_{m+1} = \mathbb{K}^n$, ce qui contredit l'hypothèse $N^m \neq 0$. ■

Remarque 2.9 *On a montré au passage qu'on a nécessairement $m < n$, car les inclusions strictes précédentes montrent que* $\dim(E_{m+1}) \geq m + 1$.

Lemme 2.11 *Soit F un sous-espace vectoriel de $\mathbb{K}^n$. Pour tout i,*

$$\begin{array}{l} F \cap E_{i+1} = \{0\} \Leftrightarrow \{v \in F, Nv \in E_i\} = \{0\}, \\ F \cap E_{i+1} = \{0\} \Rightarrow F \cap \ker(N) = \{0\} \end{array} \tag{2.17}$$

Démonstration.
- La première ligne est une conséquence de (2.15).
- Comme $\ker(N) = E_1 \subset E_i$, $F \cap E_{i+1} = \{0\} \Rightarrow \ker(N) \cap F = \{0\}$.

■

On définit alors

1. un supplémentaire F_m de E_m dans $\mathbb{K}^n$,
2. puis, comme $\forall v \in F_m \backslash \{0\}$, on a
   - $Nv \notin E_{m-1}$ d'après le Lemme 2.11
   - $Nv \in E_m$ d'après (2.15),

   on peut choisir F_{m-1} un supplémentaire de E_{m-1} dans E_m contenant l'image de F_m par N.
3. puis, par récurrence, on construit une suite de sous-espace F_i tel que F_i soit un supplémentaire de E_i dans E_{i+1} contenant l'image de F_{i+1} par N.

On a clairement $\dim(F_i) = \dim(E_{i+1}) - \dim(E_i)$, $1 \leq i \leq m$.
En posant $F_0 = E_1$, on a

$$\mathbb{K}^n = F_0 \oplus F_1 \oplus \cdots \oplus F_m,$$

Les dimensions des F_i forment une suite décroissante (au sens large) :

$$\dim(F_i) \leq \dim(F_{i-1})$$

car on sait que $\ker(N) \cap F_i = \{0\}$ et que l'image de F_i par N est contenue dans F_{i-1} (théorème du rang).

On va montrer comment construire une base de $\mathbb{K}^n$ dans laquelle la matrice de N devient triangulaire supérieure avec comme seuls coefficients éventuellement non nuls ceux situés immédiatement au dessus de la diagonale.

a) Soit e_n un vecteur de F_m, on construit les vecteurs $e_{n-1} = Ne_n \in F_{m-1}, \dots, e_{n-m} = Ne_{n-m+1} \in F_0$. Si $\dim(F_m) > 1$ on choisit alors e_{n-m-1} un vecteur de F_m indépendant de e_n, et on construit alors $e_{n-m-2} = Ne_{n-m-1}, \dots, e_{n-2m-1} = Ne_{n-2m}$. On répète le procédé jusqu'à ce que les vecteurs $(e_n, e_{n-m-1}, \dots)$ forment une base de F_m. Pour $p, 0 < p < m$, la famille $(e_{n-p}, e_{n-m-1-p} \dots) = (N^p e_n, N^p e_{n-m-1}, \dots)$ est une famille libre de vecteurs de F_{m-p} : en effet, $\sum_{i=0} \alpha_i e_{n-im-i-p} = 0$ si et seulement si $N^p(\sum_{i=0} \alpha_i e_{n-im-i}) = 0$. Ceci implique que les coefficients α_i sont tous nuls car $\sum_{i=0} \alpha_i e_{n-im-i} \in F_m$ et $F_m \cap \ker(N^p) = \{0\}$.

b) Quand on a construit une base de F_m, si l'image de F_m par N est strictement contenue dans F_{m-1}, on enrichit la famille précédemment construite d'un vecteur dans $F_{m-1} \backslash NF_m$ puis de ses images successives par $N, \dots, N^{m-1}$. On répète le procédé jusqu'à avoir construit une base de F_{m-1}.

c) Si la famille ainsi construite ne contient pas une base de F_{m-2}, on l'enrichit d'un vecteur dans $F_{m-2} \backslash NF_{m-1}$ puis de ses images successives par $N, \dots, N^{m-2}$. On répète le procédé jusqu'à avoir construit une base de F_{m-2}.

d) ainsi de suite, jusqu'à avoir aussi une base de F_0.

On obtient par cette méthode une base de $\mathbb{K}^n$, dans laquelle la matrice N devient $U = (u_{ij})_{1 \le i,j \le n}$ où $u_{ij} = 0$ si $j \neq i+1$, et $u_{i\,i+1} = y_i$ avec y_i vaut 0 ou 1. Les coefficients y_i valant 1 sont rangés en séquences de plus en plus longues, chaque séquence étant séparée de ses voisines par un seul $y_i = 0$. On a

- $\dim(F_0)$ colonnes nulles.
- $\dim(F_1) - \dim(F_2)$ séquences de $y_i = 1$ de longueur 1.
- $\dim(F_2) - \dim(F_3)$ séquences de $y_i = 1$ de longueur 2.
- $\vdots$
- $\dim(F_{m-1}) - \dim(F_m)$ séquences de $y_i = 1$ de longueur $m-1$.
- $\dim(F_m)$ séquences de $y_i = 1$ de longueur m.
- Le nombre total de séquences de $y_i = 1$ est $\dim(F_1)$. Elles sont séparées par $\dim(F_1) - 1$ coefficients nuls. Les $\dim(F_0) - \dim(F_1) + 1$ premières colonnes sont donc nulles.

Les points ci-dessus montrent en fait que la matrice U est unique, alors que la matrice de changement de base ne l'est pas.
On a démontré le

Théorème 2.9 *Soit $N \in \mathcal{M}_n(\mathbb{K})$ nilpotente telle que $N^{m+1} = 0$ et $N^m \neq 0$. La matrice N est semblable à une matrice $U \in \mathcal{M}_n(\mathbb{K})$ telle que $u_{ij} = 0$ si $j \neq i+1$, et $u_{i\,i+1} = y_i$ avec y_i valant 0 ou 1. Les coefficients y_i valant 1 sont rangés en séquences de plus en plus longues, chaque séquence étant séparée de ses voisines par un unique $y_i = 0$, et la dernière séquence de $y_i = 1$ ayant une longueur m.*

2.8 Réduction de Jordan

2.8.1 Réduction de Jordan des matrices triangularisables

Définition 2.15 *On appelle matrice de Jordan une matrice M de la forme*

$$M = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \lambda \end{pmatrix} \tag{2.18}$$

Le Théorème 2.9 dit qu'une matrice nilpotente est semblable à une matrice diagonale par blocs dont les blocs diagonaux sont des matrices de Jordan ayant 0 sur leur diagonale.
Des Théorèmes 2.8 et 2.9, on déduit le

Théorème 2.10 (réduction de Jordan) *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice dont le polynôme caractéristique P_A est scindé : $P_A(X) = \prod_{i=1}^{k}(\lambda_i - X)^{m_i}$, avec $\lambda_i \neq \lambda_j$ si $i \neq j$. La matrice A est semblable à une matrice diagonale par blocs :*

$$Q^{-1}AQ = \textit{Bloc-Diag}(T_1, \dots, T_k)$$

chaque bloc $T_i \in \mathcal{M}_{m_i}(\mathbb{K})$ s'écrivant $T_i = \lambda_i I + U_i$ où U_i est une matrice nilpotente sous la forme décrite dans le Théorème 2.9.
On peut aussi écrire

$$Q^{-1}AQ = \textit{Bloc-Diag}(M_1, \dots, M_r) \tag{2.19}$$

où les blocs M_i sont des matrices de Jordan (les coefficients diagonaux des blocs M_i sont les valeurs propres de A, et deux blocs M_i, M_j peuvent avoir les mêmes coefficients diagonaux). La réduction (2.19) est unique aux permutations près sur les blocs de Jordan. Deux matrices dont les polynômes caractéristiques sont scindés sont semblables si et seulement si elles ont la même réduction de Jordan.

Comme conséquence du Théorème 2.8, on a la

Proposition 2.9 *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice dont le polynôme caractéristique P_A est scindé. Le polynôme minimal μ_A est $\mu_A(X) = \prod_{i=1}^{k}(X - \lambda_i)^{p_i}$ où p_i est la taille du plus grand bloc de Jordan M_k dans (2.19) ayant λ_i sur sa diagonale.*

Exemple. Considérons la matrice

$$A = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 0 \\ 0 & 2 & -4 & -3 & -4 & 0 \\ 3 & -1 & 4 & -1 & -1 & 1 \\ 0 & 0 & 1 & 5 & 1 & 0 \\ -3 & 1 & 1 & 1 & 6 & -1 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{pmatrix} \tag{2.20}$$

Son polynôme caractéristique est $P_A(\lambda) = (5-\lambda)^4(2-\lambda)^2$. Le polynôme minimal de A est de la forme $\mu_A(X) = (5-\lambda)^p(2-\lambda)^q$ avec $p \leq 4$ et $q \leq 2$. En cherchant les vecteurs propres de A, on trouve que les vecteurs propres associés à la valeur propre 5 sont $(0, a, b, -a, -b, 0)^T$,

$(a,b) \neq (0,0)$, et les vecteurs propres associés à la valeur propre 2 sont $(a,0,-a,0,a,0)^T$, $a \neq 0$. On peut alors facilement vérifier que le polynôme minimal de A est $\mu_A = (5-\lambda)^2(2-\lambda)^2$, et on peut donc pas avoir un bloc de Jordan associé à la valeur propre 5 d'ordre 3, i.e.

$$\begin{pmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{pmatrix}$$

car son polynôme minimal est $(X-5)^3$.
On en conclut que la réduction de Jordan de A est de la forme

$$B = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

Exercice. Dans l'exemple précédent, trouver une matrice de passage Q telle que $Q^{-1}AQ = B$.

2.8.2 Applications aux systèmes d'équations différentielles linéaires

Les puissances des matrices de Jordan sont particulièrement faciles à calculer :

Proposition 2.10 *Soit $M \in \mathcal{M}_n(\mathbb{K})$ une matrice de Jordan (de la forme (2.18)) et $p \in \mathbb{N}$: on a*

$$\begin{array}{rcll} M^p & = & (b_{ij})_{1\leq i,j\leq n} & \\ b_{ij} & = & 0 & si \quad j < i \text{ ou } si\ j-i > p, \\ b_{ij} & = & C_p^{j-i}\lambda^{p+i-j} & si \quad 0 \leq j-i \leq p, \end{array} \tag{2.21}$$

et pour $t \in \mathbb{K}$, notant par $\exp(tM) = \sum_{i=0}^{+\infty} \frac{1}{i!} t^i M^i$,

$$\begin{array}{rcll} \exp(tM) & = & (c_{ij})_{1\leq i,j\leq n} & \\ c_{ij} & = & 0 & si \quad j < i, \\ c_{ij} & = & \frac{t^{j-i}}{(j-i)!} e^{t\lambda} & si \quad i \leq j. \end{array} \tag{2.22}$$

Démonstration. Exercice. ■

Soit $A \in \mathcal{M}_n(\mathbb{K})$, on sait que la solution du problème de Cauchy

$$\begin{array}{rcll} \frac{dv}{dt}(t) & = & Av(t) & t > 0, \\ v(0) & = & w & \end{array} \tag{2.23}$$

est donnée par

$$v(t) = \exp(tA)w \tag{2.24}$$

En conséquence, si $M \in \mathcal{M}_n(\mathbb{K})$ est une matrice de Jordan (de la forme (2.18)), la solution du problème de Cauchy pour le système d'équations différentielles

$$\begin{array}{rcll} \frac{dv}{dt}(t) & = & Mv(t) & t > 0, \\ v(0) & = & w & \end{array} \tag{2.25}$$

est donnée par

$$v_i(t) = \sum_{j=i}^{n} \frac{t^{j-i}}{(j-i)!} e^{t\lambda} w_j \tag{2.26}$$

Pour le système d'équations $\frac{dv}{dt}(t) = Av(t)$ avec une matrice A plus générale mais triangularisable (dont le polynôme caractéristique est scindé), on procède de la manière suivante :

1. on calcule une réduction de Jordan de A donnée par le Théorème 2.10. À l'aide d'un changement de base, la matrice devient diagonale par blocs, chaque bloc étant une matrice de Jordan.
2. On applique la formule (2.22) pour chaque bloc de Jordan.

Exercice. Résoudre le problème de Cauchy : trouver $u : \mathbb{R}_+ \to \mathbb{R}^5$ tel que $u(t = 0) = (1, 1, 1, 1, 1)^T$ et $u'(t) = Au(t)$ pour tout $t > 0$, avec A donnée par (2.20).

Toujours pour A triangularisable, sans même connaître une réduction de Jordan de A, on a des informations sur la solutions de (2.23), dés qu'on connaît le polynôme minimal de A : $\mu_A(X) = \prod_{i=1}^{k}(\lambda_i - X)^{p_i}$, avec $\lambda_i \neq \lambda_j$ si $i \neq j$; on sait en effet d'après le Théorème 2.10 et la Proposition 2.9 que la solution de l'équation différentielle s'écrit sous la forme

$$\sum_{i=0}^{k} \sum_{j=0}^{p_i - 1} u^{i,j} t^j e^{t\lambda_i},$$

où $u^{i,j} \in \mathbb{K}^n$ sont des vecteurs à déterminer.

2.9 Réduction simultanée

Proposition 2.11 *Si deux matrices $A \in \mathcal{M}_n(\mathbb{K})$ et $B \in \mathcal{M}_n(\mathbb{K})$ commutent, c'est à dire si $AB = BA$, alors les sous-espaces propres de A sont stables pour B.*

Démonstration. Soit λ une valeur propre de A, et soit v un vecteur propre associé. On a $ABv = BAv = \lambda Bv$, donc Bv appartient au sous-espace propre de A correspondant à λ. ■

Proposition 2.12 *Soient deux matrices $A \in \mathcal{M}_n(\mathbb{K})$ et $B \in \mathcal{M}_n(\mathbb{K})$ diagonalisables. Elles sont diagonalisables dans la même base si et seulement si A et B commutent.*

Démonstration. Si A et B sont diagonalisables dans la même base, il est clair qu'elles commutent : $A = Q^{-1}DQ$ et $B = Q^{-1}D'Q$, donc $AB = Q^{-1}DD'Q = Q^{-1}D'DQ = BA$, car deux matrices diagonales commutent.
Réciproquement, montrons la propriété par récurrence sur n. Si A n'est pas une homothétie, notons $\lambda_i, i = 1 \ldots, k$ les valeurs propres de A, m_i leur multiplicité, et $V_i = \ker(A - \lambda_i I_n)$ pour $i = 1, \ldots, k$. On a $\mathbb{K}^n = V_1 \oplus \cdots \oplus V_k$.
Comme A est diagonalisable, il existe une matrice Q inversible, dont les colonnes sont $q_{1,1}, \ldots q_{1,m_1}, \ldots, q_{k,1}, \ldots q_{k,m_k}$, $q_{i,j} \in V_i$, telle que $Q^{-1}AQ$ soit diagonale, et plus précisément $Q^{-1}AQ = \text{Bloc-Diag}(\lambda_1 I_{m_1}, \ldots, \lambda_k I_{m_k})$.
D'après la Proposition 2.11, $Q^{-1}BQ = \text{Bloc-Diag}(B_1, \ldots, B_k)$, où $B_i \in \mathcal{M}_{m_i}(\mathbb{K})$. On peut alors appliquer l'hypothèse de récurrence aux matrices $\lambda_i I_i$ et B_i, car $m_i < n$ (car la matrice A n'est pas une matrice d'homothétie). ■

En fait la démonstration précédente permet de démontrer le résultat plus fort :

Proposition 2.13 *Soit $\mathcal{E}$ une famille (pas forcément finie) de matrices de $\mathcal{M}_n(\mathbb{K})$, diagonalisables. Elles sont toutes diagonalisables dans une même base si et seulement si elles commutent deux à deux.*

Exercice. Pour $v = (v_0, \ldots, v_{n-1})^T \in \mathbb{C}^n$, on appelle $T^{(v)}$ la matrice telle que

$$T^{(v)}_{ij} = v_{(j+i-1) \bmod n}.$$

Montrer que toutes les matrices $T^{(v)}$ commutent, en montrant qu'elles sont toutes diagonalisables dans la même base.
Indication : on pourra considérer les vecteurs $e_k = (e^{i\frac{2k\pi}{n}}, \ldots, e^{i\frac{2kn\pi}{n}})^T$, pour $k = 0, \ldots, n-1$, avec $i^2 = -1$.

2.10 Décomposition de Schur et Réduction de Matrices Normales, Hermitiennes, Symétriques Réelles.

Donnons d'abord une version plus précise du Théorème 2.6 :

Lemme 2.12 *Soit A une matrice de $\mathcal{M}_n(\mathbb{K})$ dont le polynôme caractéristique est scindé : soient $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A (non nécessairement distinctes deux à deux). Quelque soit une permutation $\sigma \in \mathcal{S}_n$, on peut choisir une matrice inversible P telle que $P^{-1}AP$ soit triangulaire supérieure et sa diagonale soit $D = \mathrm{Diag}(\lambda_{\sigma(1)}, \ldots, \lambda_{\sigma(n)})$.*

Démonstration. Par récurrence. ■

On peut facilement démontrer le théorème suivant :

Théorème 2.11 (décomposition de Schur dans $\mathcal{M}_n(\mathbb{C})$) *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. Il existe une matrice Q unitaire et T une matrice triangulaire supérieure telle que*

$$T = Q^* A Q.$$

La matrice diagonale $D = \mathrm{Diag}(T)$ contient exactement les valeurs propres de A, et Q peut être choisie de manière à ce que les valeurs propres apparaissent dans n'importe quel ordre dans D.

Démonstration.
Soit P une matrice inversible telle que $P^{-1}AP = T^{(1)}$ est triangulaire supérieure. Les colonnes de P forment une base de $\mathbb{K}^n$. On peut effectuer une orthonormalisation de Gram-Schmidt de cette base : en terme matriciel, cela revient à multiplier P à droite par une matrice triangulaire supérieure $\tilde{P}$ (**exercice** : le prouver) : On a $Q = P\tilde{P}$ est unitaire, et $T = Q^{-1}AQ = (\tilde{P})^{-1}T^{(1)}\tilde{P}$ est triangulaire supérieure comme produit de telles matrices. Remarquons que $\tilde{T}$ et T ont la même diagonale.
D'après le Lemme 2.12, on peut choisir P de manière à ce que les valeurs propres de A apparaissent dans n'importe quel ordre dans $\mathrm{Diag}(\tilde{T})$ et donc dans $\mathrm{Diag}(T)$. ■

Remarque 2.10 *Remarquons que si A est réelle ($A \in \mathcal{M}_n(\mathbb{R})$) et si son polynôme caractéristique est scindé dans $\mathbb{R}$, alors on peut prendre $T \in \mathcal{M}_n(\mathbb{R})$ triangulaire supérieure et $Q \in \mathcal{M}_n(\mathbb{R})$ orthogonale dans le Théorème 2.11.*

On a le résultat plus général pour des matrices réelles quelconques :

Théorème 2.12 (décomposition de Schur dans $\mathcal{M}_n(\mathbb{R})$) *Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$. Il existe une matrice Q orthogonale et $R \in \mathcal{M}_n(\mathbb{R})$ une matrice triangulaire supérieure par blocs*

$$R = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ 0 & R_{22} & \dots & R_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & R_{nn} \end{pmatrix} \tag{2.27}$$

où R_{ii} est soit une matrice 1×1 (un réel) ou une matrice réelle 2×2 avec deux valeurs propres complexes conjuguées et non réelles , telles que

$$Q^T A Q = R. \tag{2.28}$$

Démonstration. Exactement identique à celle du Théorème 2.11. Appliquer la Proposition 2.7 puis effectuer une orthogonalisation de Gram-Schmidt dans $\mathcal{M}_n(\mathbb{R})$. ■

Théorème 2.13 (matrices normales) *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est normale ($AA^* = A^*A$) si et seulement si il existe une matrice Q unitaire telle que Q^*AQ soit diagonale.*

Démonstration. On sait d'après le Théorème 2.11 qu'il existe Q unitaire et T triangulaire supérieure telle que $Q^*AQ = T$. Du caractère normal de A on tire que $TT^* = T^*T$. En regardant la première ligne de ces matrices on voit que

$$|t_{11}|^2 = \sum_{i=1}^{n} |t_{1i}|^2.$$

Ceci implique que $t_{1i} = 0$ pour $i > 1$. Après, on regarde le coefficient diagonal sur la deuxième ligne et on obtient de même que $t_{2i} = 0$ pour $i > 2$, et on continue en raisonnant par récurrence. La réciproque est évidente. ■

On a comme conséquence

Corollaire 2.1 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice normale. Si $u \in \mathbb{C}^n$ et $v \in \mathbb{C}^n$ sont deux vecteurs propres associés à des valeurs propres distinctes de A alors $u^*v = 0$.*

On aurait pu prouver ce résultat autrement comme conséquence du

Lemme 2.13 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice normale. Le vecteur $u \in \mathbb{C}^n$ est un vecteur propre de A de valeur propre λ si et seulement si u est vecteur propre de A^* de valeur propre $\bar{\lambda}$.*

Démonstration. On a $0 = \|Au - \lambda u\|^2 = u^*A^*Au - \lambda u^*A^*u - \bar{\lambda}u^*Au + |\lambda|^2 u^*u = u^*AA^*u - \lambda u^*A^*u - \bar{\lambda}u^*Au + |\lambda|^2 u^*u = \|A^*u - \bar{\lambda}u\|^2$. ■

Si A est hermitienne, on montre facilement que la matrice diagonale obtenue au Théorème 2.13 est réelle :

Théorème 2.14 (matrices hermitiennes) *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est hermitienne ($A^* = A$) si et seulement si il existe une matrice $Q \in \mathcal{M}_n(\mathbb{C})$ unitaire telle que Q^*AQ soit diagonale à coefficients réels.*

Théorème 2.15 (matrices symétriques réelles) *Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est symétrique ($A^T = A$) si et seulement si il existe une matrice $Q \in \mathcal{M}_n(\mathbb{R})$ orthogonale telle que $Q^*AQ \in \mathcal{M}_n(\mathbb{R})$ soit diagonale.*

Démonstration. On utilise le Théorème 2.12. Il existe $R \in \mathcal{M}_n(\mathbb{R})$ de la forme (2.27), avec des blocs diagonaux d'ordre un (réel) ou deux (ayant deux valeurs propres complexes conjuguées non réelles), et Q orthogonale telles que $Q^TAQ = R$. Montrons que R ne peut pas avoir de valeur propre non réelle. Comme A est symétrique, on a que $R^T = Q^TA^TQ = Q^TAQ = R$, donc R est symétrique. Si R avait un bloc diagonal avec deux valeurs propres complexes conjuguées non réelles, ce bloc serait aussi symétrique à coefficient réel : le simple calcul du polynôme caractéristique d'un matrice symétrique réelle d'ordre deux, montre que cette situation est impossible. La matrice R est donc diagonale.
La réciproque est évidente. ■

2.11 Valeurs singulières d'une matrice

Théorème 2.16 (Valeurs singulières d'une matrice) *Soit une matrice $A \in \mathcal{M}_{n,m}(\mathbb{C})$. Il existe $Q \in \mathcal{M}_n(\mathbb{C})$ et $P \in \mathcal{M}_m(\mathbb{C})$ unitaires telles que Q^*AP soit diagonale, à coefficients réels positifs ou nuls sur la diagonale. Les $\min(m,n)$ coefficients diagonaux $(d_i)_{1\leq i\leq \min(m,n)}$ sont uniques à une permutation près, et sont appelés les valeurs singulières de A.*

Démonstration. Supposons que $m \leq n$. On sait que A^*A est une matrice hermitienne de $\mathcal{M}_m(\mathbb{C})$. Elle est donc diagonalisables et ses valeurs propres sont donc réelles d'après le Théorème 2.14. Montrons que les valeurs propres de A^*A sont positives ou nulles : Soit $\lambda \in \sigma(A^*A)$ et $v \in \mathbb{C}^m$ un vecteur propre associé à λ. On a $\lambda\|v\|^2 = \lambda v^*v = v^*A^*Av = \|Av\|^2$, donc $\lambda = \frac{\|Av\|^2}{\|v\|^2} \geq 0$.
Il existe donc P unitaire, $P \in \mathcal{M}_m(\mathbb{C})$ telle que $P^*A^*AP = D^2$, où $D = \mathrm{Diag}(d_1, \dots d_m)$ est diagonale réelle. On suppose qu'on a choisi P de manière à ce que les valeurs propres nulles de A^*A soient rangées en premier dans D^2. Donc il existe $0 \leq r \leq m$ tel $d_i = 0$ si $i \leq r$ et $d_i > 0$ si $i > r$. Pour $i > r$, on note q_i la ième colonne de AP divisée par d_i. La famille $(q_i)_{r<i\leq m}$ est orthonormale dans $\mathbb{C}^n$. On peut la compléter par $(q_i)_{i\leq r}$ et $(q_i)_{i>m}$ pour former une base orthonormale de $\mathbb{C}^n$. Soit $Q \in \mathcal{M}_n(\mathbb{C})$ la matrice unitaire dont les colonnes sont les q_i. On a

$$Q^*AP = \begin{pmatrix} D \\ 0 \end{pmatrix}$$

Si $m \leq n$, on reproduit le raisonnement précédent pour A^*. On obtient comme précédemment qu'il existe $P \in \mathcal{M}_m(\mathbb{C})$, $Q \in \mathcal{M}_n(\mathbb{C})$, et $\tilde{D} \in \mathcal{M}_{m,n}(\mathbb{R})$, diagonale avec des coefficients diagonaux positifs ou nuls tels que $P^*A^*Q = \tilde{D}$, et on a immédiatement que $Q^*AP = (\tilde{D})^T$, et on pose $D = (\tilde{D})^T$. ■

Le Théorème 2.16 nous dit que si $m < n$, il existe $Q \in \mathcal{M}_n(\mathbb{C})$ et $P \in \mathcal{M}_m(\mathbb{C})$ unitaires

telles que

$$Q^*AP = \left(\begin{array}{cccc} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & d_m \\ 0 & & \dots & 0 \\ \vdots & & & \vdots \\ 0 & & \dots & 0 \end{array}\right) \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\} m \\ \left.\begin{array}{l} \\ \\ \end{array}\right\} n-m \end{array}$$

Si $n < m$, il existe $Q \in \mathcal{M}_n(\mathbb{C})$ et $P \in \mathcal{M}_m(\mathbb{C})$ unitaires telles que

$$Q^*AP = \left(\begin{array}{ccccccc} d_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & d_2 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & & \vdots \\ 0 & \dots & 0 & d_n & 0 & \dots & 0 \end{array}\right)$$

Définition 2.16 *Une décomposition de A sous la forme $A = Q\Xi P^*$, avec Q, P unitaires, et $\Xi \in \mathcal{M}_{n,m}(\mathbb{R})$, diagonale à coefficients diagonaux positifs ou nuls, données par le Théorème 2.16, est appelée décomposition de A en valeurs singulières, ou décomposition SVD (singular values decomposition).*

Remarque 2.11 *La décomposition SVD utilise en général deux matrices de changement de bases au lieu d'une pour la diagonalisation.*

Remarque 2.12 *La démonstration du Théorème 2.16 donne un algorithme pour construire une décomposition SVD d'une matrice A, (en supposant connu un algorithme de diagonalisation de matrices hermitiennes).*

Exercice. Écrire l'algorithme correspondant à la démonstration du Théorème 2.16.

Remarque 2.13 *Il est clair que s'il n'y a pas unicité d'une décomposition SVD d'une matrice A (car il n'y a pas unicité dans le Théorème 2.14, l'ensemble des valeurs singulières de A ne dépend pas du choix de la décomposition SVD.*

Remarque 2.14 *Si $A \in \mathcal{M}_n(\mathbb{R})$, alors $A^*A = A^TA$ et $AA^* = AA^T$ sont des matrices symétriques à coefficients réels, dont les valeurs propres sont positives ou nulles. Dans ce cas, en refaisant la démonstration du Théorème 2.16, on voit qu'on peut donc choisir P et Q orthogonales. Tous les coefficients d'une telle décomposition SVD de A sont réels.*

Exemple. On considère la matrice

$$A = \frac{1}{\sqrt{2}} \left(\begin{array}{ccc} -1 & -3 & -\sqrt{3} \\ 1 & -3 & \sqrt{3} \end{array}\right)$$

Montrer qu'on a la décomposition $A = Q\Xi P^*$, où

$$P = \left(\begin{array}{ccc} \frac{1}{2} & 0 & -\frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ \frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{array}\right), \quad Q = \left(\begin{array}{cc} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{array}\right), \quad \Xi = \left(\begin{array}{ccc} 2 & 0 & 0 \\ 0 & 3 & 0 \end{array}\right).$$

Exercice. Montrer que pour toute matrice $A \in \mathcal{M}_{n,m}(\mathbb{C})$, alors A et A^* ont les mêmes valeurs singulières.
Exercice. Montrer que pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$, alors A est inversible si et seulement si ses valeurs singulières sont non nulles.

Proposition 2.14 *Soit $A \in \mathcal{M}_{n,m}(\mathbb{K})$. Soient $P \in \mathcal{M}_n(\mathbb{K})$ et $Q \in \mathcal{M}_m(\mathbb{K})$ deux matrices inversibles. Le rang de A est égal au rang de $P^{-1}AQ$.*

Démonstration. Exercice ■

Corollaire 2.2 *Le rang de $A \in \mathcal{M}_{n,m}(\mathbb{C})$ est le nombre de ses valeurs singulières non nulles.*

Proposition 2.15 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice normale. Les valeurs singulières de A sont exactement les modules des valeurs propres de A.*

Démonstration. On sait d'après le Théorème 2.13 qu'il existe une matrice $Q \in \mathcal{M}_n(\mathbb{C})$ unitaire et une matrice diagonale $D \in \mathcal{M}_n(\mathbb{C})$, $D = \mathrm{Diag}(d_1, \ldots, d_n)$ telles que $Q^*AQ = D$. On appelle s la fonction de $\mathbb{C}$ dans le cercle unité de $\mathbb{C}$ telle que $s(z) = \frac{\bar{z}}{|z|}$ si $z \neq 0$ et $s(0) = 1$. On appelle S la matrice diagonale de $\mathcal{M}_n(\mathbb{C})$ définie par $S = \mathrm{Diag}(s(d_1), \ldots, s(d_n))$. La matrice S est clairement unitaire. La matrice $P = QS$ est unitaire, et on a $Q^*AP = \Xi$ où $\Xi = \mathrm{Diag}(|z_1|, \ldots, |z_n|)$. On vient de trouver une décomposition SVD de A. Les valeurs singulières de A sont donc les modules des valeurs propres de A. ■

On peut donner une interprétation géométrique des valeurs singulières d'une matrice $A \in \mathcal{M}_{n,m}(\mathbb{R})$. Supposons que $n \leq m$ et que le rang de A est n (A a n valeurs singulières non nulles).
Considérons l'image par A de la boule unité $\mathcal{B}_1$ de $\mathbb{R}^m$, i.e. de l'ensemble des vecteurs de $\mathbb{R}^m$ tels que $x^T x \leq 1$. On sait par le Théorème 2.16 qu'il existe une décomposition SVD, $Q^TAP = \Xi$ avec Q et P orthogonales, et $\Xi \in \mathcal{M}_{n,m}(\mathbb{R})$ est diagonale à coefficients diagonaux positifs ou nuls. On suppose qu'on a choisi P et Q pour que les n premiers coefficients diagonaux de Σ sont non nuls et que les suivants sont nuls. Notons $\Xi^\dagger$ la matrice diagonale de $\mathcal{M}_n(\mathbb{R})$ dont les coefficients diagonaux sont les inverses de ceux de Ξ. On a

$$\Xi^\dagger Q^T AP = \begin{pmatrix} I_n, 0 \end{pmatrix} \quad \text{et } P^T A^T Q (\Xi^\dagger)^T \Xi^\dagger Q^T AP = \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix}$$

On en déduit que $x \in \mathcal{B}_1 \Rightarrow \|\Xi^\dagger Q^T APx\| \leq 1$. Réciproquement, par définition de P Q et $\Xi^\dagger$, pour tout vecteur y de $\mathbb{R}^n$ tel que $\|\Xi^\dagger y\| \leq 1$, il existe un vecteur Px de $\mathrm{Vect}(p_1, \ldots, p_n)$ ($p_1, \ldots, p_m$ sont les colonnes de P) vérifiant $\|x\| \leq 1$ tel que $APx = Qy$.
L'image des vecteurs de $\mathcal{B}_1$, exprimée dans la base orthonormale formée par des colonnes de Q est

$$\mathcal{E} = \{\sum_{i=1}^{n} y_i q_i,\ y \in \mathbb{R}^n;\ \exists x \in \mathcal{B}_1 \text{ tel que } y = Q^T APx\}$$

On a donc $\mathcal{E} = \{\sum_{i=1}^n y_i q_i,\ \|\Xi^\dagger y\| \leq 1\}$. Il est donc clair que $\mathcal{E}$ est un hyper-ellipsoïde dont les demi-axes ont pour longueur les valeurs singulières de A. On a démontré la proposition

Proposition 2.16 *Soit deux entiers strictement positif $n \leq m$ et soit $A \in \mathcal{M}_{n,m}(\mathbb{R})$ de rang n. L'image par A de la boule unité $\mathcal{B}_1$ de $\mathbb{R}^m$ est un hyper-ellipsoïde dont les demi-axes ont pour longueur les valeurs singulières de A.*

Remarque 2.15 *On peut généraliser le résultat précédent si A a des valeurs singulières quelconques. On obtient alors un hyper-ellipsoïde dégénéré.*

Remarque 2.16 *Les valeurs singulières maximale et minimale de A sont souvent notées $\sigma_{\max}(A)$ et $\sigma_{\min}(A)$.*

Chapitre 3

Formes Quadratiques et Hermitiennes

3.1 Formes Bilinéaires sur des Espaces Vectoriels Réels

Définition 3.1 *Soit E et F deux espaces vectoriels sur $\mathbb{R}$. On appelle forme bilinéaire sur $E \times F$ une application b de $E \times F$ dans $\mathbb{R}$, telle que $\forall x, y \in E$, $\forall u, v \in F$, $\forall \alpha, \beta \in \mathbb{R}$,*

$$b(\alpha x + \beta y, u) = \alpha b(x, u) + \beta b(y, u)$$
$$b(x, \alpha u + \beta v) = \alpha b(x, u) + \beta b(x, v)$$

On note $\mathcal{B}(E, F)$ l'ensemble des formes bilinéaires sur $E \times F$. Il est clair que $\mathcal{B}(E, F)$ est un espace vectoriel sur $\mathbb{R}$.
Si $E = F$, on parle de forme bilinéaire sur E et on utilise la notation plus courte $\mathcal{B}(E)$ pour l'espace vectoriel des formes bilinéaires sur E.

Remarque 3.1 *Rien n'empêche de définir des formes bilinéaires sur des espaces vectoriels sur $\mathbb{C}$. C'est dans le but de simplifier l'exposition et d'éviter les confusions avec les formes sesquilinéaires (définies plus loin) que nous nous restreignons aux espaces vectoriels réels.*

Exemple. L'application $b : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}$, $(x, u) \mapsto 2x_1u_1 + 4x_1u_2 + 2x_2u_1 + 3x_2u_2 + 6x_3u_2$ est une forme bilinéaire sur $\mathbb{R}^3 \times \mathbb{R}^2$.

Remarque 3.2 *Soit $x \in E$ et $u \in F$. Il est clair que l'application qui à $v \in F$ associe $b(x, v)$ est une forme linéaire sur F (un élément de F'), et que l'application qui à $y \in E$ associe $b(y, u)$ est une forme linéaire sur E (un élément de E').*

Supposons que E est de dimension finie n, F est de dimension finie m, et soient $\mathcal{E} = (e_1, \ldots, e_n)$ et $\mathcal{F} = (f_1, \ldots, f_m)$ des bases de E et F respectivement. Si $x = \sum_{j=1}^{n} x_j e_j$ et si $u = \sum_{i=1}^{m} u_i f_i$, on vérifie facilement que pour toute forme bilinéaire b, on a

$$b(x, u) = \sum_{j=1}^{n} \sum_{i=1}^{m} b(e_j, f_i) u_i x_j.$$

On voit donc qu'une forme bilinéaire b est complètement déterminée par la connaissance des $m \times n$ coefficients $b_{ij} = b(e_j, f_i)$, ou encore de la matrice $B = (b_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}} \in \mathcal{M}_{m,n}(\mathbb{R})$.

Définition 3.2 *On dit que la matrice B construite ci-dessus est la matrice de la forme bilinéaire b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$.*
Si $m = n$, on appelle discriminant de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ le déterminant de la matrice de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$.

Il est clair que l'application qui à b associe sa matrice par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ est un isomorphisme de $\mathcal{B}(E,F)$ sur $\mathcal{M}_{m,n}(\mathbb{R})$. On a donc $\dim(\mathcal{B}(E,F)) = nm$. Cet isomorphisme n'est pas canonique car il dépend du choix des bases.
On a un isomorphisme canonique de $\mathcal{B}(\mathbb{R}^n,\mathbb{R}^m)$ sur $\mathcal{M}_{n,m}(\mathbb{R})$, en choisissant $\mathcal{E}$ (respectivement $\mathcal{F}$) la base canonique de $\mathbb{R}^n$ (respectivement $\mathbb{R}^m$). Si B est la matrice de b par rapport aux bases canoniques de $\mathbb{R}^n$ et $\mathbb{R}^m$, on a

$$b(x,u) = u^T B x. \tag{3.1}$$

Regardons l'effet de changements de bases :

Lemme 3.1 *Soient $\mathcal{E} = (e_1,\dots,e_n)$ et $\mathcal{E}' = (e'_1,\dots,e'_n)$ deux bases de E, et $\mathcal{F} = (f_1,\dots,f_m)$ et $\mathcal{F}' = (f'_1,\dots,f'_m)$ deux bases de F. On note $P \in \mathcal{M}_n(\mathbb{R})$ la matrice de passage de $\mathcal{E}$ dans $\mathcal{E}'$ et $Q \in \mathcal{M}_m(\mathbb{R})$ la matrice de passage de $\mathcal{F}$ dans $\mathcal{F}'$. Soit B la matrice de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$. La matrice de b par rapport aux bases $\mathcal{E}'$ et $\mathcal{F}'$ est $Q^T B P$.*

Démonstration. On a $b(e'_j, f'_i) = b(\sum_{k=1}^n p_{kj} e_k, \sum_{l=1}^m q_{li} f_l) = \sum_{k=1}^n \sum_{l=1}^m p_{kj} b_{kl} q_{li}$, qui est bien le coefficient de la matrice $Q^T B P$ sur la $i$ème ligne et la $j$ème colonne. ■

3.1.1 Rang d'une forme bilinéaire

Le Lemme 3.1 et la Proposition 2.14 nous disent aussi que le rang de la matrice de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ ne dépend pas du choix des bases $\mathcal{E}$ et $\mathcal{F}$.

Définition 3.3 *Supposons que E est de dimension finie n, F est de dimension finie m. On appelle rang d'une forme bilinéaire b sur $E \times F$ et on note* $\mathrm{rang}(b)$ *le rang de B, où B est la matrice de b par rapport à deux bases quelconques $\mathcal{E}$ de E et $\mathcal{F}$ de F.*

Définition 3.4 *Supposons que E est de dimension finie n, F est de dimension finie m. On dit que $b \in \mathcal{B}(E,F)$ est une forme non dégénérée si $m = n = \mathrm{rang}(b)$. Dans le cas contraire, on dit que b est dégénérée.*
En d'autres termes, soit B la matrice de b par rapport à deux bases quelconques $\mathcal{E}$ de E et $\mathcal{F}$ de F ; la forme bilinéaire b est non dégénérée si et seulement si B est carrée et inversible.

Lemme 3.2 *La forme bilinéaire b est non dégénérée si et seulement si*
- *l'application qui à $x \in E$ associe la forme linéaire sur F : $v \mapsto b(x,v)$, est un isomorphisme de E sur F'.*
- *l'application qui à $u \in F$ associe la forme linéaire sur E : $y \mapsto b(y,u)$, est un isomorphisme de F sur E'.*

3.1.2 Formes Bilinéaires Symétriques et Orthogonalité

Définition 3.5 *Soit b une forme bilinéaire sur E. On dit que b est est une forme bilinéaire symétrique si $b(x,y) = b(y,x)$, $\forall x, y \in E$.*

Lemme 3.3 *L'ensemble des formes bilinéaires symétriques sur E, noté $\mathcal{B}_{\text{sym}}(E)$ forme un sous-espace de $\mathcal{B}(E)$.*

Lemme 3.4 *Soit E un espace vectoriel de dimension n sur $\mathbb{R}$ et b est une forme bilinéaire symétrique sur E. Pour toute base $\mathcal{E}$ de E, la matrice de b par rapport à $\mathcal{E}$ est symétrique. L'application qui à une forme bilinéaire symétrique associe sa matrice par rapport à $\mathcal{E}$ est un isomorphisme de $\mathcal{B}_{\text{sym}}(E)$ sur l'espace des matrices symétriques d'ordre n. L'espace $\mathcal{B}_{\text{sym}}(E)$ est de dimension $\frac{n(n+1)}{2}$.*

Définition 3.6 *Soient E un espace vectoriel de dimension n sur $\mathbb{R}$ et $b \in \mathcal{B}_{\text{sym}}(E)$. On dit que deux vecteurs x et y de E sont orthogonaux pour b et on note $x \perp_b y$ si et seulement si $b(x, y) = 0$.*
Soit C une partie de E. On note $C^{\perp_b}$ l'ensemble

$$C^{\perp_b} = \{x \in E : \forall y \in C,\ x \perp_b y\}$$

Lemme 3.5 *Soit $b \in \mathcal{B}_{\text{sym}}(E)$ et C une partie de E. L'ensemble $C^{\perp_b}$ est un sous-espace vectoriel de E.*

Remarque 3.3 *Soit b une forme bilinéaire symétrique. On a l'équivalence : b non dégénérée si et seulement si $E^{\perp_b} = \{0\}$.*

Lemme 3.6 *Soient E un espace vectoriel de dimension n sur $\mathbb{R}$. Soit $b \in \mathcal{B}_{\text{sym}}(E)$, b non dégénérée, et G un sous-espace de E. On a $\dim(G^{\perp_b}) + \dim(G) = n$ et $(G^{\perp_b})^{\perp_b} = G$.*

Démonstration. Le cas $G = \{0\}$ est trivial. Supposons G de dimension $p > 0$ et soit $(e_1, \ldots, e_p)$ une base de G. On note $f_i' \in E'$ la forme linéaire : $f_i'(v) = b(e_i, v)$. D'après le Lemme 3.2, $(f_1', \ldots, f_p')$ est une famille libre de E'. L'application linéaire qui à v associe $(f_1'(v), \ldots, f_p'(v))$ est donc de rang p. On a $G^{\perp_b} = \cap_{i=1}^{p} \ker(f_i')$ et le théorème du rang nous dit que $\dim(G^{\perp_b}) = n - p$. D'autre part, il est clair que $G \subset (G^{\perp_b})^{\perp_b}$ et on vient de montrer que $\dim((G^{\perp_b})^{\perp_b}) = n - \dim(G^{\perp_b}) = p$, donc $G = (G^{\perp_b})^{\perp_b}$. ■

Remarque 3.4 *Dans le Lemme 3.6, si on n'avait pas l'hypothèse b non dégénérée, on aurait $\dim(G^{\perp_b}) + \dim(G) \geq n$.*

Remarque 3.5 *On peut avoir $G^{\perp_b} \cap G \neq \{0\}$ pour b symétrique non dégénérée, ceci équivaut à l'existence d'un élément x non nul de G tel que $b(x, x) = 0$. Par exemple, on pourra considérer la forme bilinéaire symétrique sur $\mathbb{R}^2$: $b(x, y) = x_1y_2 + x_2y_1$.*

Proposition 3.1 *Soit b une forme bilinéaire symétrique sur E. L'ensemble des vecteurs tels que $b(x, x) = 0$ est un cône, appelé cône d'isotropie de b.*

Définition 3.7 *Soit b une forme bilinéaire symétrique sur E. On dit qu'un sous-espace G de E est non isotrope pour b, s'il ne contient pas de vecteur x non nul tel que $b(x, x) = 0$.*

Lemme 3.7 *Soit b une forme bilinéaire symétrique sur E et un sous-espace G de E non isotrope pour b. On a $G \oplus G^{\perp_b} = E$.*

Démonstration. Soit p la dimension de G.
– on a $G \cap G^{\perp_b} = \{0\}$, ce qui implique que $\dim(G \oplus G^{\perp_b}) = p + \dim(G^{\perp_b})$

– En reprenant la démonstration du Lemme 3.6, c.f. Remarque 3.4, on voit que $\dim(G^{\perp_b}) \geq n - p$ (avec égalité dans le cas non dégénéré).

On déduit des deux points précédents que $\dim(G \oplus G^{\perp_b}) \geq n$, et donc $G \oplus G^{\perp_b} = E$. ■

Remarque 3.6 *Dans le Lemme 3.7, on n'a pas besoin de l'hypothèse b non dégénérée.*

Définition 3.8 *Soit $\mathcal{E}$ une famille de vecteurs de E. On dit que $\mathcal{E}$ est une famille orthogonale pour b si et seulement si ses vecteurs sont orthogonaux deux à deux pour b. Soit $\mathcal{E}$ une base de E. On dit que $\mathcal{E}$ est une base orthogonale pour b si c'est une famille orthogonale de vecteurs de E. On dit que $\mathcal{E}$ est une base orthonormale pour b si $\mathcal{E}$ est une base orthogonale pour b et si chaque vecteur e de $\mathcal{E}$ vérifie $b(e, e) = 1$.*

Proposition 3.2 *Si b est une forme bilinéaire symétrique sur E, il existe une base orthogonale de E pour b.*

Démonstration. Par récurrence sur la dimension n de E. Le cas $n = 1$ est trivial. Supposons la propriété est vraie pour des entiers positifs $< n$, et soit E de dimension n.
Si b est nulle, toute base convient.
Si b est non nul, on peut trouver un vecteur $e_1 \neq 0$, tel que $b(e_1, e_1) \neq 0$: en effet, dans le cas contraire, on aurait une base $(e_1, \ldots, e_n)$ de E telle que

– $b(e_i, e_i) = 0$, $i = 1, \ldots, n$;
– $b(e_i + e_j, e_i + e_j) = 0$, $i, j = 1, \ldots, n$;

D'après le premier point et la symétrie de b, le deuxième point implique que $b(e_i, e_j) = 0$, $i, j = 1, \ldots, n$. Mais alors $b = 0$, ce qui est contraire à l'hypothèse. Soit donc $e_1 \neq 0$ tel que $b(e_1, e_1) \neq 0$. On sait d'après le Lemme 3.7 que $\{e_1\}^{\perp_b}$ est de dimension $n - 1$ et que ce sous-espace est supplémentaire à la droite vectorielle engendrée par e_1. On applique l'hypothèse de récurrence à la restriction de b à $\{e_1\}^{\perp_b}$. Il existe $(e_2, \ldots, e_n)$ une base orthogonale de $\{e_1\}^{\perp_b}$ pour b, et $(e_1, \ldots, e_n)$ est une base orthogonale de E. ■

Remarque 3.7 *Une autre façon de démontrer la Proposition 3.2 est de considérer la matrice B de b dans une base $\mathcal{E}$ de E.*
Trouver une base orthogonale pour b, c'est trouver un changement de base de $\mathcal{E}$ à $\mathcal{E}'$, i.e. une matrice $P \in \mathcal{M}_n(\mathbb{R})$ inversible telle que $P^T BP$ soit diagonale (cf (3.1)). Mais on sait que B est une matrice symétrique, et donc il existe une matrice P, orthogonale telle que $P^T BP$ soit diagonale.

Remarque 3.8 *Soit b une forme bilinéaire et $\mathcal{E}$ une base orthogonale pour b, et soit D la matrice (diagonale) de b par rapport à cette base. Si b est dégénérée, alors* $\mathrm{rang}(D) = \mathrm{rang}(b) < n$, *i.e. D a des coefficients diagonaux nuls. On en déduit qu'une forme bilinéaire dégénérée n'admet pas de base orthonormale.*

Théorème 3.1 (procédé d'orthogonalisation de Schmidt) *Soit b une forme bilinéaire symétrique sur E telle que pour tout vecteur non nul $x \in E$, $b(x, x) \neq 0$. Soit $(f_1, \ldots, f_p)$ une famille libre de vecteurs de E. Il existe une famille libre orthogonale $(e_1, \ldots, e_p)$ pour b de vecteurs de E, telle que,*

$$Vect(e_1, \ldots, e_q) = Vect(f_1, \ldots, f_q), \qquad \forall q \leq p. \tag{3.2}$$

Démonstration. La démonstration utilise le procédé d'orthogonalisation de Schmidt déjà vu si b est un produit scalaire. ■

Remarque 3.9 *L'hypothèse du Théorème 3.1 est plus forte que b non dégénérée. On verra plus tard comment construire une base orthogonale pour une forme bilinéaire symétrique générale en utilisant la réduction de Gauss des formes quadratiques.*

3.2 Formes Quadratiques

Définition 3.9 *Soit E un espace vectoriel sur $\mathbb{R}$. Une forme quadratique q sur E est une application E dans $\mathbb{R}$ telle qu'il existe une forme bilinéaire b sur E avec $\forall x \in E$, $q(x) = b(x,x)$. L'ensemble des formes quadratiques est clairement un espace vectoriel sur $\mathbb{R}$, que l'on note $\mathcal{Q}(E)$*

À toute forme bilinéaire, on peut associer une forme quadratique. Réciproquement, si q est une forme quadratique sur E, on peut construire à partir de q une forme bilinéaire symétrique c par $c(x,y) = \frac{1}{2}(q(x+y) - q(x) - q(y))$: en effet, soit b une forme bilinéaire telle $q(x) = b(x,x)$, on a $c(x,y) = \frac{1}{2}(b(x+y,x+y) - b(x,x) - b(y,y)) = \frac{1}{2}(b(x,y) + b(y,x))$, ce qui définit bien une forme bilinéaire symétrique. On appelle c la forme polaire de q. L'application qui à une forme quadratique lui fait correspondre sa forme polaire est un isomorphisme d'espaces vectoriels de $\mathcal{Q}(E)$ sur $\mathcal{B}_{\text{sym}}(E)$. Si E est de dimension n, la dimension de $\mathcal{Q}(E)$ est $\frac{n(n+1)}{2}$.
Grâce à la bijection ainsi construite, on peut utiliser pour les formes quadratiques les notions définies pour les formes bilinéaires. La matrice d'une forme quadratique par rapport à une base $\mathcal{E}$ de E est la matrice de sa forme polaire par rapport à $\mathcal{E}$. En particulier, si $E = \mathbb{R}^n$et si $\mathcal{E}$ est la base canonique de $\mathbb{R}^n$ et si $C = (c_{ij}) \in \mathcal{M}_n(\mathbb{R})$ est la matrice de Q par rapport à $\mathcal{E}$, on a

$$q(x) = \sum_{i=1}^{n} c_{ii}x_i^2 + 2\sum_{1\le i<j\le n} c_{ij}x_ix_j = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_ix_j = x^TCx. \tag{3.3}$$

Le rang d'une forme quadratique q est le rang de sa forme polaire. Une forme quadratique est dégénérée si sa forme polaire l'est. On définit l'orthogonalité par rapport à une forme quadratique q comme l'orthogonalité par rapport à sa forme polaire et on utilise le symbole $\perp_q$. Une famille orthogonale pour q est une famille orthogonale pour sa forme polaire c, et la notion d'orthogonal d'une partie G de $\mathbb{R}^n$ pour q coïncide avec celle d'orthogonal de G pour c.
La Proposition 3.2 se récrit en terme de formes quadratiques sous la forme

Proposition 3.3 *Soit E un espace vectoriel réel de dimension n, et soit q une forme quadratique sur E. Il existe n réels $(\gamma_1, \dots, \gamma_n)$ et n formes linéaires linéairement indépendantes $(\ell_i)_{i=1,\dots,n}$ telles que*

$$q(x) = \sum_{i=1}^{n} \gamma_i(\ell_i(x))^2 \tag{3.4}$$

Démonstration. Soit c la forme polaire de q, et soit $\mathcal{E} = (e_1, \dots, e_n)$ une base de E orthogonale pour q (la Proposition 3.2 nous dit qu'une telle base existe). La matrice de c par rapport à $\mathcal{E}$ est $\text{Diag}(\gamma_1, \dots, \gamma_n)$. On a

$$q(x) = \sum_{i=1}^{n} \gamma_i(\ell_i(x))^2$$

où $(\ell_i)_{1\le i\le n}$ est la base duale de $\mathcal{E}$, i.e. $\ell_i(e_j) = \delta_{ij}$. ■

La proposition 3.3 nous dit que toute forme quadratique peut être décomposée en une combinaison de carrés de forme linéaires.

La réduction de Gauss en somme de carré permet de construire une telle décomposition, *i.e.* de construire une base orthogonale pour une forme quadratique q.

La méthode de Gauss Supposons que la matrice de q dans une base $\mathcal{E}$ s'écrive $C = (c_{ij})$, *i.*., si on note x_i les coordonnées d'un vecteur x dans $\mathcal{E}$, on a

$$q(x) = \sum_{i=1}^{n} c_{ii}x_i^2 + 2\sum_{1\leq i<j\leq n} c_{ij}x_ix_j.$$

On distingue deux cas :

1. L'un des coefficients diagonaux de C est non nul. Sans restriction, on peut supposer que $c_{11} \neq 0$. On a

$$\begin{aligned} q(x) =& c_{11}\left(x_1^2 + \frac{2}{c_{11}}\sum_{1<j\leq n} c_{1j}x_1x_j + \frac{1}{c_{11}^2}(\sum_{1<j\leq n} c_{1j}x_j)^2\right) \\ &+ \sum_{i=2}^{n} c_{ii}x_i^2 + 2\sum_{2\leq i<j\leq n} c_{ij}x_ix_j - \frac{1}{c_{11}}(\sum_{1<j\leq n} c_{1j}x_j)^2. \end{aligned}$$

   On définit la forme linéaire ℓ_1 et la forme quadratique q' par

$$\ell_1(x) = x_1 + \sum_{1<j\leq n}\frac{c_{1j}}{c_{11}}x_j, \qquad q'(x) = \sum_{i=2}^{n} c_{ii}x_i^2 + 2\sum_{2\leq i<j\leq n} c_{ij}x_ix_j - \frac{1}{c_{11}}(\sum_{1<j\leq n} c_{1j}x_j)^2$$

   et on peut récrire l'identité précédente

$$q(x) = c_{11}(\ell_1(x))^2 + q'(x)$$

   Il est clair que $\text{rang}(q') \leq n-1$, car $q'(x)$ ne dépend pas de x_1.

2. Tous les coefficients diagonaux de C sont nuls. On peut supposer qu'un des coefficients extra-diagonaux de C est non nul (sinon $q = 0$ et on a fini). Sans restriction, on peut supposer que $c_{12} \neq 0$. On pose $\psi_1(x) = c_{12}x_1 + \sum_{i=3}^{n} c_{2i}x_i$ et $\psi_2(x) = c_{12}x_2 + \sum_{i=3}^{n} c_{1i}x_i$. On a

$$\psi_1(x)\psi_2(x) = c_{12}\left(\sum_{i=2}^{n} c_{1i}x_1x_i + \sum_{i=3}^{n} c_{2i}x_2x_i + \sum_{i=3}^{n}\sum_{j=3}^{n}\frac{c_{2i}c_{1j}}{c_{12}}x_ix_j\right).$$

   On pose $\ell_1(x) = (\psi_1(x) + \psi_2(x))$ et $\ell_2(x) = (\psi_1(x) - \psi_2(x))$ et on appelle q' la forme quadratique

$$q'(x) = 2\sum_{3\leq i<j\leq n} c_{ij}x_ix_j - 2\sum_{i=3}^{n}\sum_{j=3}^{n}\frac{c_{2i}c_{1j}}{c_{12}}x_ix_j.$$

   On a

$$q(x) = \frac{1}{2c_{12}}\left((\ell_1(x))^2 - (\ell_2(x))^2\right) + q'(x)$$

   et il est clair que $\text{rang}(q') \leq n-2$, car $q'(x)$ ne dépend pas de x_1, ni de x_2. Les formes linéaires ℓ_1 et ℓ_2 sont linéairement indépendantes, car ψ_1 et ψ_2 sont linéairement indépendantes.

Dans les deux cas, on a construit une forme linéaire q' de rang strictement inférieur à celui de q. Si q' est nulle, on a terminé. Sinon, on peut appliquer le procédé ci dessus à q'. En un nombre fini d'étapes (plus petit que n), on construit ainsi r formes linéaires linéairement indépendantes $\ell_1, \ldots, \ell_r$ et r réels non nuls $\gamma_1, \ldots, \gamma_r$ tels que

$$q(x) = \sum_{i=1}^{r} \gamma_i (\ell_i(x))^2$$

Après, si $r < n$, on complète la famille des ℓ_i pour former une base de E' et on prend $\gamma_i = 0$ pour $i > r$. □

Exercice. Appliquer la méthode de Gauss à la forme quadratique sur $\mathbb{R}^4$:

$$q(x) = 4x_1^2 + 2x_2^2 + 3x_3^2 + x_4 + 6x_1x_2 + 12x_1x_3 + 2x_2x_4$$

Définition 3.10 *Soit E un espace vectoriel sur $\mathbb{R}$ et q une forme quadratique sur E.*

- *On dit que q est positive si $q(x) \geq 0$, $\forall x \in E$. On dit aussi parfois que q est semi-définie positive.*
- *On dit que q est négative si $q(x) \leq 0$, $\forall x \in E$. On dit aussi parfois que q est semi-définie négative.*
- *On dit que q est définie positive si q est positive et si $q(x) = 0 \Leftrightarrow x = 0$.*
- *On dit que q est définie négative si q est négative et si $q(x) = 0 \Leftrightarrow x = 0$.*

Définition 3.11 *Soit $A \in \mathcal{M}_n(\mathbb{R})$, A symétrique.*

- *On dit que A est positive (ou semi-définie positive) si la forme quadratique $x \mapsto x^T Ax$ est positive.*
- *On dit que A est négative (ou semi-définie négative) si la forme quadratique $x \mapsto x^T Ax$ est négative.*
- *On dit que A est définie positive si la forme quadratique $x \mapsto x^T Ax$ est définie positive.*
- *On dit que A est définie négative si la forme quadratique $x \mapsto x^T Ax$ est définie négative.*

Exercice. Montrer que la matrice tridiagonale dont les coefficients de la diagonale principale sont égaux à 2 et les coefficients des deux autres diagonales sont -1 est positive.

Proposition 3.4 (inégalité de Cauchy-Schwarz) *Soit q une forme quadratique positive sur E et c sa forme polaire. On a, pour tout x et y dans E,*

$$|c(x,y)|^2 \leq q(x)q(y).$$

Démonstration. On considère le polynôme de degré deux, $\mu \mapsto q(x+\mu y) = \mu^2 q(y) + 2\mu c(x,y) + q(x)$. Comme q est positive, ce polynôme a son discriminant négatif ou nul : $|c(x,y)|^2 - q(x)q(y) \leq 0$. ■

Théorème 3.2 (d'inertie de Sylvester) *Soit E un espace vectoriel réel de dimension n, et soit q une forme quadratique sur E. Il existe deux sous-espaces E_+ et E_- tels que*

1. *la restriction de q à E_+ soit définie positive.*

2. *la restriction de q à E_- soit définie négative.*
3. *on a la somme directe*

$$E = E_+ \oplus E_- \oplus E^{\perp_q}.$$

De plus, si deux sous-espaces F_+ et F_- ont les mêmes propriétés que E_+ et E_-, on

$$\dim(E_+) = \dim(F_+), \qquad \dim(E_-) = \dim(F_-).$$

Démonstration. Soit $\mathcal{E} = (e_1, \ldots, e_n)$ une base de E orthogonale pour q. On note $\mathrm{Diag}(\gamma_1, \ldots, \gamma_n)$ la matrice de q dans cette base. Quitte à permuter les vecteurs de base, on peut supposer que pour $0 \leq p_1 \leq p_2 \leq n$, on a

$$\begin{aligned} \gamma_i &> 0 \quad \text{pour } 1 \leq i \leq p_1, \\ \gamma_i &< 0 \quad \text{pour } p_1 < i \leq p_2, \\ \gamma_i &= 0 \quad \text{pour } p_2 < i \leq n. \end{aligned}$$

On prend alors $E_+ = Vect(e_i, 1 \leq i \leq p_1)$, $E_- = Vect(e_i, p_1 < i \leq p_2)$, et on a $E^{\perp_q} = Vect(e_i, p_2 < i \leq n)$. On vérifie facilement que E_+ et E_- ont les propriétés désirées.
Soient F_- et F_+ ayant les mêmes propriétés que E_- et E_+. Comme $p_2 = n - dim(E^{\perp_q})$, on a $\dim(F_+)+\dim(F_-) = p_2$. Soit p_1' la dimension de F_+. On a $F_+ \cap (E_- \oplus E^{\perp_q}) = \{0\}$ ce qui implique que $p_1' \leq p_1$. De même, on a $F_- \cap (E_+ \oplus E^{\perp_q}) = \{0\}$ ce qui implique que $p_2 - p_1' \leq p_2 - p_1$. Des deux dernières inégalités, on déduit $p_1' = p_1$. ■

Définition 3.12 *Soit E un espace vectoriel sur $\mathbb{R}$ de dimension finie et q une forme quadratique sur E. Soit E_+, E_- deux sous-espaces de E ayant les propriétés décrites dans le Théorème 3.2. On appelle signature de q la paire* $(\dim(E_+), \dim(E_-))$

3.3 Formes Sesquilinéaires

Définition 3.13 *Soit E et F deux espaces vectoriels sur $\mathbb{C}$. On appelle forme sesquilinéaire sur $E \times F$ une application s de $E \times F$ dans $\mathbb{C}$, telle que $\forall x, y \in E$, $\forall u, v \in F$, $\forall \alpha, \beta \in \mathbb{C}$,*

$$\begin{aligned} s(\alpha x + \beta y, u) &= \alpha s(x, u) + \beta s(y, u) \\ s(x, \alpha u + \beta v) &= \bar{\alpha} s(x, u) + \bar{\beta} s(x, v) \end{aligned}$$

On note $\mathcal{S}(E, F)$ l'ensemble des formes sesquilinéaires sur $E \times F$. Il est clair que $\mathcal{S}(E, F)$ est un espace vectoriel sur $\mathbb{C}$.
Si $E = F$, on parle de forme sesquilinéaire sur E et on utilise la notation plus courte $\mathcal{S}(E)$ pour l'espace vectoriel des formes sesquilinéaires sur E.

Remarque 3.10 *Soit $x \in E$ et $u \in F$. Il est clair que l'application qui à $v \in F$ associe $b(x, \bar{v})$ est une forme linéaire sur F (un élément de F'), et que l'application qui à $y \in E$ associe $b(y, u)$ est une forme linéaire sur E (un élément de E').*

Supposons que E est de dimension finie n, F est de dimension finie m, et soient $\mathcal{E} = (e_1, \ldots, e_n)$ et $\mathcal{F} = (f_1, \ldots, f_m)$ des bases de E et F respectivement. Si $x = \sum_{j=1}^n x_j e_j$ et si $u = \sum_{i=1}^m u_i f_i$, on vérifie facilement que pour toute forme sesquilinéaire s, on a

$$s(x, u) = \sum_{i=1}^{m} \sum_{j=1}^{n} s(e_j, f_i) x_j \bar{u}_i.$$

On voit donc qu'une forme sesquilinéaire s est complètement déterminée par la connaissance des $m \times n$ coefficients complexes $s_{ij} = s(e_j, f_i)$, ou encore de la matrice $S = (s_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}} \in \mathcal{M}_{m,n}(\mathbb{C})$.

Définition 3.14 *On dit que la matrice S construite ci-dessus est la matrice de la forme sesquilinéaire s par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$.*
Si $m = n$, on appelle discriminant de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ le déterminant de la matrice de b par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$.

Il est clair que l'application qui à s associe sa matrice par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ est un isomorphisme de $\mathcal{S}(E, F)$ sur $\mathcal{M}_{m,n}(\mathbb{C})$. On a donc $\dim(\mathcal{S}(E, F)) = nm$. Cet isomorphisme n'est pas canonique car il dépend du choix des bases.
On a un isomorphisme canonique de $\mathcal{S}(\mathbb{C}^n, \mathbb{C}^m)$ sur $\mathcal{M}_{m,n}(\mathbb{C})$, en choisissant $\mathcal{E}$ (respectivement $\mathcal{F}$) la base canonique de $\mathbb{C}^n$ (respectivement $\mathbb{C}^m$). Si S est la matrice de s par rapport aux bases canoniques de $\mathbb{C}^n$ et $\mathbb{C}^m$, on a

$$s(x, y) = y^* S x. \tag{3.5}$$

Regardons l'effet de changements de bases :

Lemme 3.8 *Soient $\mathcal{E} = (e_1, \ldots, e_n)$ et $\mathcal{E}' = (e'_1, \ldots, e'_n)$ deux bases de E, et $\mathcal{F} = (f_1, \ldots, f_m)$ et $\mathcal{F}' = (f'_1, \ldots, f'_m)$ deux bases de F. On note $P \in \mathcal{M}_n(\mathbb{C})$ la matrice de passage de $\mathcal{E}$ dans $\mathcal{E}'$ et $Q \in \mathcal{M}_m(\mathbb{C})$ la matrice de passage de $\mathcal{F}$ dans $\mathcal{F}'$. Soit S la matrice de s par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$. La matrice de s par rapport aux bases $\mathcal{E}'$ et $\mathcal{F}'$ est $Q^* S P$.*

Démonstration. exercice ■

On voit que le rang de la matrice de s par rapport aux bases $\mathcal{E}$ et $\mathcal{F}$ ne dépend pas du choix des bases $\mathcal{E}$ et $\mathcal{F}$.

Définition 3.15 *Supposons que E est de dimension finie n, F est de dimension finie m. On appelle rang d'une forme sesquilinéaire s sur $E \times F$ et on note* $\mathrm{rang}(s)$ *le rang de S, où S est la matrice de s par rapport à deux bases quelconques $\mathcal{E}$ de E et $\mathcal{F}$ de F.*

Définition 3.16 *Supposons que E est de dimension finie n, F est de dimension finie m. On dit que $s \in \mathcal{S}(E, F)$ est une forme sesquilinéaire non dégénérée si $m = n = \mathrm{rang}(b)$ ou encore si S (la matrice de s par rapport à deux bases quelconques $\mathcal{E}$ de E et $\mathcal{F}$ de F) est carrée et inversible. Dans le cas contraire, on dit que s est dégénérée.*

Lemme 3.9 *La forme sesquilinéaire s est non dégénérée si et seulement si*
- *l'application qui à $x \in E$ associe la forme linéaire sur F : $v \mapsto b(x, \bar{v})$, est un isomorphisme de E sur F'.*
- *l'application qui à $u \in F$ associe la forme linéaire sur F : $y \mapsto b(y, \bar{u})$, est un isomorphisme de F sur E'.*

Définition 3.17 *Soit s une forme sesquinéaire sur E. On dit que s est une forme sesquilinéaire hermitienne si $b(x, y) = \overline{b(y, x)}$, $\forall x, y \in E$.*

Exemple. La forme sesquilinéaire s définie sur $\mathbb{C}^3$

$$s(x, y) = 2x_1\bar{y}_1 + i(x_1\bar{y}_2 - x_2\bar{y}_1) + 3x_2\bar{y}_2 + (2+i)x_1\bar{y}_3 + (2-i)x_3\bar{y}_1 + 3(x_2\bar{y}_3 + x_3\bar{y}_2) + 4x_3\bar{y}_3$$

est une forme sesquilinéaire hermitienne.

Lemme 3.10 *Soit E un espace vectoriel de dimension n sur $\mathbb{C}$ et s une forme sesquilinéaire hermitienne sur E. Pour toute base $\mathcal{E}$ de E, la matrice de s par rapport à $\mathcal{E}$ est hermitienne.*

Définition 3.18 *Soient E un espace vectoriel de dimension n sur $\mathbb{C}$ et s une forme sesquilinéaire hermitienne sur E. On dit que deux vecteurs x et y de E sont orthogonaux pour s et on note $x \perp_s y$ si et seulement si $s(x,y) = 0$.*
Soit C une partie de E. On note $C^{\perp_s}$ l'ensemble

$$C^{\perp_s} = \{x \in E : \forall y \in C,\ x \perp_s y\}$$

Lemme 3.11 *Soit s une forme sesquilinéaire hermitienne sur E et C une partie de E. L'ensemble $C^{\perp_s}$ est un sous-espace vectoriel de E.*

On a l'équivalence : s non dégénérée si et seulement si $E^{\perp_s} = \{0\}$.

Lemme 3.12 *Soient E un espace vectoriel de dimension n sur $\mathbb{C}$. Soit s une forme sesquilinéaire hermitienne sur E, s non dégénérée, et G un sous-espace de E. On a $\dim(G^{\perp_s}) + \dim(G) = n$ et $(G^{\perp_s})^{\perp_s} = G$.*
Si s est dégénérée, on a $\dim(G^{\perp_s}) + \dim(G) \geq n$ et $G \subset (G^{\perp_s})^{\perp_s}$.

Démonstration. Le cas $G = \{0\}$ est trivial. Supposons G de dimension $p > 0$ et soit $(e_1, \ldots, e_p)$ une base de G. On note $f_i' \in E'$ la forme linéaire : $f_i'(v) = s(e_i, \bar{v})$. D'après le Lemme 3.9, $(f_1', \ldots, f_p')$ est une famille libre de E'. L'application linéaire qui à v associe $(f_1'(v), \ldots, f_p'(v))$ est donc de rang p. Donc le sous-espace $W = \cap_{i=1}^{p} \ker(f_i')$ vérifie $\dim(W) = n-p$, par le théorème du rang et on a $G^{\perp_s} = \{\bar{w}, w \in W\}$ donc $\dim(G^{\perp_s}) = n - p$.
D'autre part, il est clair que $G \subset (G^{\perp_s})^{\perp_s}$ et on vient de montrer que $\dim((G^{\perp_s})^{\perp_s}) = n - \dim(G^{\perp_s}) = p$, donc $G = (G^{\perp_s})^{\perp_s}$. ■

Définition 3.19 *Soit s une forme sesquilinéaire hermitienne sur E. On dit qu'un sous-espace G de E est non isotrope pour s, s'il ne contient pas de vecteur x non nul tel que $b(x,x) = 0$.*

Lemme 3.13 *Soit s une forme sesquilinéaire hermitienne sur E et un sous-espace G de E non isotrope pour s. On a $G \oplus G^{\perp_s} = E$.*

Démonstration. exercice ■

Définition 3.20 *Soit $\mathcal{E}$ une famille de vecteurs de E. On dit que $\mathcal{E}$ est une famille orthogonale pour s si et seulement si ses vecteurs sont orthogonaux deux à deux pour s. Soit $\mathcal{E}$ une base de E. On dit que $\mathcal{E}$ est une base orthogonale pour s si c'est une famille orthogonale de vecteurs de E. On dit que $\mathcal{E}$ est une base orthonormale pour s si $\mathcal{E}$ est une base orthogonale pour s et si chaque vecteur e de $\mathcal{E}$ vérifie $b(e,e) = 1$.*

Proposition 3.5 *Si s est une forme sesquilinéaire hermitienne sur E, il existe une base orthogonale de E pour s.*

Démonstration. Identique à celle de la Proposition 3.2. ■

Théorème 3.3 (procédé d'orthogonalisation de Schmidt) *Soit s une forme sesquilinéaire hermitienne sur E telle que pour tout vecteur non nul $x \in E$, $s(x,x) \neq 0$. Soit $(f_1, \ldots, f_p)$ une famille libre de vecteurs de E. Il existe une famille libre orthogonale $(e_1, \ldots, e_p)$ de vecteurs de E pour s, telle que,*

$$Vect(e_1, \ldots, e_q) = Vect(f_1, \ldots, f_q), \qquad \forall q \leq p. \tag{3.6}$$

3.4 Formes Hermitiennes

Définition 3.21 *Soit E un espace vectoriel sur $\mathbb{C}$. Une forme hermitienne q sur E est une application E dans $\mathbb{R}$ telle qu'il existe une forme sesquilinéaire hermitienne s sur E avec $\forall x \in E$, $q(x) = s(x,x)$.*

À toute forme sesquilinéaire s, on peut associer une forme hermitienne q. Réciproquement, si q est la forme hermitienne associée à s, on peut déduire s de q par

$$s(x,y) = \frac{1}{4}(q(x+y) - q(x-y) + iq(x+iy) - iq(x-iy))$$

Grâce à la bijection ainsi construite, on peut utiliser pour les formes quadratique hermitiennes les notions définies pour les formes sesquilinéaires hermitiennes.
La Proposition 3.5 se récrit en terme de formes quadratiques hermitiennes sous la forme

Proposition 3.6 *Soit E un espace vectoriel sur $\mathbb{C}$ de dimension n, et soit q une forme hermitienne sur E. Il existe n réels $(\gamma_1, \ldots, \gamma_n)$ et n formes linéaires linéairement indépendantes $(\ell_i)_{i=1,\ldots,n}$ telles que*

$$q(x) = \sum_{i=1}^{n} \gamma_i |\ell_i(x)|^2 \tag{3.7}$$

La méthode de Gauss Supposons que la matrice de q dans une base $\mathcal{E}$ s'écrive $C = (c_{ij})$, *i.e.*, si on note x_i les coordonnées d'un vecteur x dans $\mathcal{E}$, on a

$$q(x) = \sum_{i=1}^{n} c_{ii}|x|_i^2 + 2\Re(\sum_{1\le i<j\le n} c_{ij}\bar{x}_i x_j).$$

On distingue deux cas :

1. L'un des coefficients diagonaux de C est non nul. Sans restriction, on peut supposer que $c_{11} \neq 0$. On a

$$\begin{aligned} q(x) =& c_{11}\left(|x_1|^2 + \frac{2}{c_{11}}\Re(\sum_{1<j\le n} c_{1j}\bar{x}_1 x_j) + \frac{1}{c_{11}^2}|\sum_{1<j\le n} c_{1j}x_j|^2\right) \\ &+ \sum_{i=2}^{n} c_{ii}|x|_i^2 + 2\Re(\sum_{2\le i<j\le n} c_{ij}\bar{x}_i x_j) - \frac{1}{c_{11}}|\sum_{1<j\le n} c_{1j}x_j|^2 \end{aligned}$$

   On définit la forme linéaire ℓ_1 et la forme hermitienne q' par

$$\ell_1(x) = x_1 + \sum_{1<j\le n} \frac{c_{1j}}{c_{11}}x_j, \qquad q'(x) = \sum_{i=2}^{n} c_{ii}|x|_i^2 + 2\Re(\sum_{2\le i<j\le n} c_{ij}\bar{x}_i x_j) - \frac{1}{c_{11}}|\sum_{1<j\le n} c_{1j}x_j|^2$$

   et on peut récrire l'identité précédente

$$q(x) = c_{11}|\ell_1(x)|^2 + q'(x)$$

   Il est clair que $\text{rang}(q') \le n-1$, car $q'(x)$ ne dépend pas de x_1.

2. Tous les coefficients diagonaux de C sont nuls.

$$q'(x) = 2\Re(\sum_{1\le i<j\le n} c_{ij}\overline{x_i}x_j).$$

On peut supposer qu'un des coefficients extra-diagonaux de C est non nul (sinon $q=0$ et on a fini). Sans restriction, on peut supposer que $c_{12}\neq 0$. On pose $\psi_1(x) = x_1 + \sum_{i=3}^n \frac{c_{2i}}{\bar{c}_{12}}x_i$ et $\psi_2(x) = c_{12}x_2 + \sum_{i=3}^n c_{1i}x_i$. On a

$$\begin{aligned}
\overline{\psi_1(x)}\psi_2(x) &= \left(\sum_{i=2}^n c_{1i}\overline{x_1}x_i + \sum_{i=3}^n \overline{c_{2i}}x_2\overline{x_i} + \sum_{i=3}^n\sum_{j=3}^n \frac{\overline{c_{2i}}c_{1j}}{c_{12}}\overline{x_i}x_j\right).\\
\psi_1(x)\overline{\psi_2(x)} &= \left(\sum_{i=2}^n \overline{c_{1i}}x_1\overline{x_i} + \sum_{i=3}^n c_{2i}\overline{x_2}x_i + \sum_{i=3}^n\sum_{j=3}^n \frac{c_{2i}\overline{c_{1j}}}{\overline{c_{12}}}x_i\overline{x_j}\right).
\end{aligned}$$

Donc

$$q(x) = 2\Re\left(\overline{\psi_1(x)}\psi_2(x)\right) + q'(x) = \frac{1}{2}(|\psi_1(x)+\psi_2(x)|^2 - |\psi_1(x)-\psi_2(x)|^2) + q'(x)$$

où

$$q'(x) = 2\Re(\sum_{3\le i<j\le n} c_{ij}\overline{x_i}x_j) - 2\Re(\sum_{i=3}^n\sum_{j=3}^n \frac{\overline{c_{2i}}c_{1j}}{c_{12}}\overline{x_i}x_j)$$

et il est clair que $\mathrm{rang}(q') \le n-2$, car $q'(x)$ ne dépend pas de x_1, ni de x_2. Les formes linéaires $\ell_1 = \psi_1+\psi_2$ et $\ell_2 = \psi_1-\psi_2$ sont linéairement indépendantes, car ψ_1 et ψ_2 sont linéairement indépendantes.

Dans les deux cas, on a construit une forme linéaire q' de rang strictement inférieur à celui de q. Si q' est nulle, on a terminé. Sinon, on peut appliquer le procédé ci dessus à q'. En un nombre fini d'étapes (plus petit que n), on construit ainsi r formes linéaires linéairement indépendantes $\ell_1,\dots,\ell_r$ et r réels non nuls $\gamma_1,\dots,\gamma_r$ tels que

$$q(x) = \sum_{i=1}^r \gamma_i|\ell_i(x)|^2$$

Après, si $r<n$, on complète la famille des ℓ_i pour former une base de E' et on prend $\gamma_i = 0$ pour $i>r$. □On peut définir les formes hermitiennes positives, définies positives, négatives, définies négatives, comme pour les formes quadratiques réelles. De même, on définit les matrices hermitiennes positives, définies positives, négatives, définies négatives.

Proposition 3.7 (inégalité de Cauchy-Schwarz) *Soit q une forme hermitienne positive sur E et s sa forme sesquilinéaire associée. On a, pour tout x et y dans E,*

$$|s(x,y)|^2 \le q(x)q(y).$$

Démonstration. On considère l'application de $\mathbb{C}$ à valeurs dans $\mathbb{R}_+$, $\mu \mapsto q(x+\mu y) = |\mu|^2 q(y) + 2\Re(\mu s(y,x)) + q(x)$. Si $q(y)=0$, alors la positivité de cette application ne peut être obtenue que si $s(x,y)=0$.
Si $q(y)>0$, on prend $\mu = -\frac{s(x,y)}{q(y)}$, et on obtient

$$-\frac{|s(x,y)|^2}{q(y)} + q(x) \ge 0,$$

qui est le résultat désiré. ■

Théorème 3.4 (d'inertie de Sylvester) *Soit E un espace vectoriel sur $\mathbb{C}$ de dimension n, et soit q une forme hermitienne sur E. Il existe deux sous-espaces E_+ et E_- tels que*

1. *la restriction de q à E_+ soit définie positive.*
2. *la restriction de q à E_- soit définie négative.*
3. *on a la somme directe*
$$E = E_+ \oplus E_- \oplus E^{\perp_q}.$$

De plus, si deux sous-espaces F_+ et F_- ont les mêmes propriétés que E_+ et E_-, on

$$\dim(E_+) = \dim(F_+), \qquad \dim(E_-) = \dim(F_-).$$

Définition 3.22 *Soit E un espace vectoriel sur $\mathbb{C}$ de dimension finie et q une forme hermitienne sur E. Soit E_+, E_- deux sous-espaces de E ayant les propriétés décrites dans le Théorème 3.4. On appelle signature de q la paire* $(\dim(E_+), \dim(E_-))$

Chapitre 4

Analyse matricielle

Ici $\mathbb{K}$ désigne $\mathbb{C}$ ou $\mathbb{R}$.

4.1 Normes vectorielles et matricielles

4.1.1 Définitions

On rappelle qu'une norme $\| \ \|$ dans $\mathbb{K}^n$ est une application de $\mathbb{K}^n$ dans $\mathbb{R}_+$ qui à v associe $\|v\|$ avec les propriétés suivantes

1. $\|v\| = 0$ si et seulement si $v = 0$.
2. $\|\alpha v\| = |\alpha| \|v\|$.
3. $\|v + w\| \leq \|v\| + \|w\|$.

4.1.2 Les normes $\| \ \|_p$

Une famille de normes très utiles est celle des normes $\| \ \|_p$ où p est un réel plus grand que ou égal à 1 :

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}. \tag{4.1}$$

On définira aussi $\| \ \|_\infty$ par

$$\|v\|_\infty = \max_{i=1\ldots n} |v_i|. \tag{4.2}$$

Il est clair que $\| \ \|_\infty$ et $\| \ \|_1$ sont des normes. Pour montrer que $\| \ \|_p$ est une norme pour $1 < p < \infty$, on a les lemmes :

Lemme 4.1 *Si p est un réel tel que $1 < p < \infty$, et si x et y sont des réels non négatifs, alors*

$$xy \leq \frac{1}{p} x^p + \frac{1}{q} y^q \tag{4.3}$$

où q est le réel défini par $\frac{1}{q} + \frac{1}{p} = 1$.

Démonstration. On a $xy = \exp(\log(x) + \log(y)) = \exp(\frac{1}{p}\log(x^p) + \frac{1}{q}\log(y^q))$. On utilise la convexité de l'application exponentielle pour conclure ■

Lemme 4.2 (inégalité de Hölder) *Si p est un réel tel que $1 < p < \infty$, alors pour tous vecteurs $x \in \mathbb{K}^n$, $y \in \mathbb{K}^n$,*

$$|y^*x| \leq \|x\|_p \|y\|_q \tag{4.4}$$

Démonstration. Posons $\tilde{x} = \frac{1}{\|x\|_p} x$ et $\tilde{y} = \frac{1}{\|y\|_q} y$. On applique (4.3). On a $|\tilde{y}^*\tilde{x}| \leq \frac{1}{p} \sum_{i=1}^n \frac{1}{\|x\|_p^p} |x_i|^p + \frac{1}{q} \sum_{i=1}^n \frac{1}{\|y\|_q^q} |y_i|^q = \frac{1}{p} + \frac{1}{q} = 1$. On en déduit immédiatement (4.4). ■

Proposition 4.1 *Pour $1 < p < \infty$, $\| \|_p$ est une norme.*

Démonstration. La seule propriété non complètement immédiate est l'inégalité triangulaire : pour cela on écrit $\|x+y\|_p^p = \sum_{i=1}^n |x_i+y_i||x_i+y_i|^{p-1} \leq \sum_{i=1}^n |x_i||x_i+y_i|^{p-1} + \sum_{i=1}^n |y_i||x_i+y_i|^{p-1}$. On utilise l'inégalité de Hölder :

$$\sum_{i=1}^n |x_i||z_i| \leq \|x\|_p \|z\|_q \quad \text{et} \quad \sum_{i=1}^n |y_i||z_i| \leq \|y\|_p \|z\|_q,$$

où $z \in \mathbb{R}^n$ est le vecteur dont les composantes sont $z_i = |x_i + y_i|^{p-1}$. Mais $q = \frac{p}{p-1}$, donc $\|z\|_q = \|x + y\|_p^{p-1}$. On en conclut que

$$\|x + y\|_p^p \leq (\|x\|_p + \|y\|_p) \|x + y\|_p^{p-1}.$$

■

Exercice. Montrer en trouvant des contre-exemples que si $p < 1$, $\| \|_p$ n'est pas une norme dans $\mathbb{R}^2$.

La norme $\| \|_2$ est la norme euclidienne bien connue : elle est associée au produit scalaire usuel, et on a l'inégalité de Cauchy-Schwarz bien connue

$$|u^*v| \leq \|u\|_2 \|v\|_2. \tag{4.5}$$

En dimension finie, toutes les normes sont équivalentes, mais les constantes d'équivalence dépendent en général de la dimension n :

Proposition 4.2 *Soit u un vecteur de $\mathbb{K}^n$: on a*

$$\text{pour} \quad 1 \leq q < p < +\infty, \qquad \|u\|_p \leq \|u\|_q \leq n^{\frac{1}{q} - \frac{1}{p}} \|u\|_p. \tag{4.6}$$

$$\text{pour} \quad 1 \leq p < +\infty, \qquad \|u\|_\infty \leq \|u\|_p \leq n^{\frac{1}{p}} \|u\|_\infty. \tag{4.7}$$

Démonstration. Soit p et q tels que $1 \leq q < p < +\infty$, $\|u\|_p = (\sum_{i=1}^n |u_i|^p)^{\frac{1}{p}} = (\sum_{i=1}^n (|u_i|^q)^{\frac{p}{q}})^{\frac{1}{p}}$. Mais $\sum_{i=1}^n (|u_i|^q)^{\frac{p}{q}} \leq (\sum_{i=1}^n (|u_i|^q))^{\frac{p}{q}}$ car $q < p$.
C'est l'inégalité de Jensen : $\forall \alpha > 1, \forall x_i > 0, i = 1 \ldots n$, $\sum_{i=1}^n x_i^\alpha \leq (\sum_{i=1}^n x_i)^\alpha$, car

$$\sum_{i=1}^n \frac{x_i^\alpha}{(\sum_{j=1}^n x_j)^\alpha} \leq \sum_{i=1}^n \frac{x_i}{\sum_{j=1}^n x_j} = 1$$

car $\frac{x_i}{\sum_{j=1}^n x_j} \leq 1$.
On en déduit que $(\sum_{i=1}^n (|u_i|^q)^{\frac{p}{q}})^{\frac{1}{p}} \leq (\sum_{i=1}^n |u_i|^q)^{\frac{1}{q}} = \|u\|_q$. Donc $\|u\|_p \leq \|u\|_q$ et on a l'inégalité

de gauche dans (4.6).
Pour montrer l'inégalité de droite dans (4.6), on procède comme suit

$$\|u\|_q^q = \sum_{i=1}^n |u_i|^q = \sum_{i=1}^n 1.|u_i|^q \leq n^{1-\frac{q}{p}} (\sum_{i=1}^n |u_i|^p)^{\frac{q}{p}}.$$

C'est l'inégalité de Hölder avec les normes $\| \|_{\frac{p}{q}}$ et$\| \|_{\frac{p}{p-q}}$. Donc $\|u\|_q \leq n^{\frac{1}{q}-\frac{1}{p}} \|u\|_p$.
On laisse l'inégalité (4.7) en exercice. ■

4.1.3 Normes matricielles

Définition 4.1 *L'application $\| \|$ de $\mathcal{M}_n(\mathbb{K}) \to \mathbb{R}_+$ est une norme matricielle si et seulement si*

1. *$\forall A \in \mathcal{M}_n(\mathbb{K})$, $\|A\| = 0$ si et seulement si $A = 0$.*
2. *$\forall A \in \mathcal{M}_n(\mathbb{K})$, $\forall \alpha \in \mathbb{K}$, $\|\alpha A\| = |\alpha| \|A\|$.*
3. *$\forall A, B \in \mathcal{M}_n(\mathbb{K})$, $\|A + B\| \leq \|A\| + \|B\|$.*

et de plus

$$\forall A, B \in \mathcal{M}_n(\mathbb{K}), \quad \|AB\| \leq \|A\| \|B\|. \tag{4.8}$$

Remarque 4.1 *Il existe des normes sur $\mathcal{M}_n(\mathbb{K})$ qui ne sont pas des normes matricielles ; La norme $\|A\| = \max_{1 \leq i,j \leq 2} |a_{ij}|$ est une norme sur $\mathcal{M}_2(\mathbb{R})$, mais*

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad \text{et} \qquad \|A^2\| = 2 \geq \|A\|^2 = 1$$

Normes subordonnées à des normes vectorielles

Proposition 4.3 *À une norme $\| \|$ sur $\mathbb{K}^n$, on peut associer une norme matricielle sur $\mathcal{M}_n(\mathbb{K})$ qu'on note encore $\| \|$ par : pour tout $A \in \mathcal{M}_n(\mathbb{K})$,*

$$\|A\| = \sup_{0 \neq x \in \mathbb{K}^n} \frac{\|Ax\|}{\|x\|} = \max_{0 \neq x \in \mathbb{K}^n} \frac{\|Ax\|}{\|x\|} = \max_{x \in \mathbb{K}^n, \|x\|=1} \|Ax\| \tag{4.9}$$

On dit que la norme matricielle $\| \|$ est la norme matricielle subordonnée à $\| \|$.

Démonstration. Soit donc $\| \|$ une norme sur $\mathbb{K}^n$.
Exercice Montrer les égalités dans (4.9). Vérifier que la norme subordonnée est bien une norme sur $\mathcal{M}_n(\mathbb{K})$.
Il reste juste à démontrer que la norme subordonnée a la propriété (4.8). On a par définition des normes subordonnées que $\|Ax\| \leq \|A\| \|x\|$ pour tout $x \in \mathbb{K}^n$: donc $\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\|$ et on obtient bien (4.8) en divisant par $\|x\|$ et en prenant le sup. ■

Lemme 4.3 *Soit $\|.\|$ une norme subordonnée sur $\mathcal{M}_n(\mathbb{K})$. On a $\|I_n\| = 1$.*

Démonstration. exercice. ■

Pour $1 \leq p \leq \infty$, on va noter $\| \|_p$ la norme matricielle subordonnée à la norme $\| \|_p$ sur $\mathbb{K}^n$. On a les identités suivantes

Proposition 4.4

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}|. \tag{4.10}$$

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}| = \|A^*\|_\infty. \tag{4.11}$$

$$\|A\|_2 = \sigma_{\max}(A) = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \|A^*\|_2. \tag{4.12}$$

où $\rho(M)$ désigne le module maximal des valeurs propres de M.

Démonstration. On a, pour tout $x \in \mathbb{K}^n$,

$$\|Ax\|_\infty = \max_{1\le i\le n} |\sum_{j=1}^{n} a_{i,j}x_j| \le \left(\max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}|\right) \|x\|_\infty,$$

d'où l'inégalité $\|A\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}|$. Pour montrer l'égalité, on appelle i_0 l'indice réalisant le maximum : $\sum_{j=1}^{n} |a_{i_0,j}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}|$. On choisit alors x de la manière suivante : $x_j = \frac{\bar{a}_{i_0,j}}{|a_{i_0,j}|}$ si $a_{i_0,j} \neq 0$ et $x_j = 0$ sinon. Il est clair que $\|x\|_\infty = 1$ et on voit que la coordonnée d'indice i_0 de Ax est $\sum_{j=1}^{n} |a_{i_0,j}|$. On a donc trouvé x tel que $\|x\|_\infty = 1$ et $\|Ax\|_\infty \ge \max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}|$. On a prouvé (4.10).
On a, pour tout $x \in \mathbb{K}^n$,

$$\|Ax\|_1 = \sum_{1\le i\le n} |\sum_{j=1}^{n} a_{i,j}x_j| \le \left(\max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}|\right) \sum_{j=1}^{n} |x_j|,$$

d'où l'inégalité $\|A\|_1 \le \max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}|$. Pour montrer l'égalité, on appelle j_0 l'indice réalisant le maximum : $\sum_{i=1}^{n} |a_{i,j_0}| = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}|$. On choisit alors x de la manière suivante : $x_j = \delta_{j,j_0}$. Il est clair que $\|x\|_1 = 1$ et on voit que $\|Ax\|_1 = \sum_{i=1}^{n} |a_{i,j_0}|$. On a donc trouvé x tel que $\|x\|_1 = 1$ et $\|Ax\|_1 \ge \max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}|$. On a prouvé (4.11).
On a

$$\|A\|_2^2 = \sup_{0\neq x\in\mathbb{K}^n} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{0\neq x\in\mathbb{K}^n} \frac{x^*A^*Ax}{\|x\|_2^2}$$

Mais A^*A est hermitienne donc diagonalisable dans une base orthonormale et on a $\sup_{0\neq x\in\mathbb{K}^n} \frac{x^*A^*Ax}{\|x\|_2^2} = \rho(A^*A)$. De plus, si λ est une valeur propre non nulle de A^*A, il existe $v_\lambda \neq 0$ tel que $A^*Av_\lambda = \lambda v_\lambda$. Alors $AA^*Av_\lambda = \lambda Av_\lambda$. Donc Av_λ est un vecteur propre (car non nul) de AA^* avec la valeur propre λ. Si λ est une valeur propre non nulle de A^*A c'est aussi une valeur propre de AA^*. Donc si $\rho(A^*A) \neq 0$ alors $\rho(AA^*) = \rho(A^*A)$. Ceci est encore vrai si $\rho(A^*A) = 0$, car on a lors $A = 0$. On a donc prouvé (4.12). ■

Proposition 4.5 *Si U est une matrice unitaire de $\mathcal{M}_n(\mathbb{C})$, on a pour tout $A \in \mathcal{M}_n(\mathbb{C})$,*

$$\|A\|_2 = \|UA\|_2 = \|AU\|_2 = \|UAU^*\|_2. \tag{4.13}$$

Si A est normale,

$$\|A\|_2 = \rho(A). \tag{4.14}$$

Enfin, pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$,

$$\|A\|_2^2 = \|A^*A\|_2. \tag{4.15}$$

Démonstration. On laisse le premier point en exercice.
Si A est normale, A est semblable à une matrice diagonale D par une transformation unitaire : on a $\|A\|_2 = \|D\|_2$ d'après le premier point et il est facile vérifier que $\|D\|_2 = \rho(D) = \rho(A)$.
Enfin, on a, pour toute matrice A,

$$\|A\|_2^2 = \rho(A^*A) = \|A^*A\|_2.$$

■

Remarque 4.2 *L'égalité (4.14) est vraie en particulier si A est réelle symétrique ou hermitienne.*

Proposition 4.6 *On a les inégalités : pour $p:\ 1 \leq p < +\infty$,*

$$n^{-\frac{1}{p}}\|A\|_\infty \leq \|A\|_p \leq n^{\frac{1}{p}}\|A\|_\infty \tag{4.16}$$

pour $p, q:\ 1 \leq p < q < +\infty$,

$$n^{\frac{1}{q}-\frac{1}{p}}\|A\|_q \leq \|A\|_p \leq n^{\frac{1}{p}-\frac{1}{q}}\|A\|_q \tag{4.17}$$

$$\|A\|_2 \leq \sqrt{\|A\|_\infty \|A\|_1} \tag{4.18}$$

Démonstration. On laisse (4.16) et (4.17) en exercice. Pour prouver (4.18), on utilise (4.15) : $\|A\|_2^2 = \|A^*A\|_2 = \rho(A^*A)$. On verra par la suite, cf. Proposition 4.9 que pour toute norme matricielle $\|.\|$, on a $\rho(A^*A) \leq \|A^*A\|$: donc

$$\begin{aligned}\|A\|_2^2 = \rho(A^*A) &\leq \|A^*A\|_\infty \\ &\leq \|A^*\|_\infty \|A\|_\infty = \|A\|_1 \|A\|_\infty.\end{aligned}$$

■

Norme de Frobenius

On définit la norme de Frobenius d'une matrice M d'ordre n

$$\|M\|_F^2 = \sum_{1\leq i\leq n} \sum_{1\leq j\leq n} |m_{i,j}|^2. \tag{4.19}$$

On vérifie facilement que

$$\|M\|_F^2 = \mathrm{trace}(M^*M) = \mathrm{trace}(MM^*) = \|M^*\|_F^2. \tag{4.20}$$

Le carré de la norme de Frobenius de M est donc la somme du carré de ses valeurs singulières. On en déduit que

$$\|M\|_2 \leq \|M\|_F \leq \sqrt{n}\|M\|_2. \tag{4.21}$$

De (4.20), on déduit aussi que si P s'écrit $P = Q^*MQ$ avec Q unitaire, alors $\|P\|_F = \|M\|_F$, car la trace de P^*P est égale à celle de Q^*M^*MQ qui est égale à celle de M^*M.
Enfin, la norme de Frobenius a la propriété (4.8) : en effet, en appliquant l'inégalité de Cauchy-Schwarz dans $\mathbb{K}^n$, on a

$$\|AB\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n |\sum_{k=1}^n A_{ik}B_{kj}|^2 \leq \sum_{i=1}^n \sum_{j=1}^n \left(\sum_{k=1}^n |A_{ik}|^2 \sum_{l=1}^n |B_{lj}|^2 \right) = \|A\|_F^2 \|B\|_F^2$$

La norme de Frobenius est donc une norme matricielle.
Elle n'est pas subordonnée à une norme vectorielle, car $\|I_n\| = \sqrt{n} \neq 1$.

4.2 Nombre de conditionnement

On s'intéresse ici à la sensibilité d'un système linéaire $Ax = b$, $A \in \mathcal{M}_n(\mathbb{K})$, $x, b \in \mathbb{K}^n$ à des perturbations du second membre b ou de la matrice A.

Définition 4.2 *On considère une norme $\| \ \|$ sur $\mathbb{K}^n$ et on note encore $\| \ \|$ la norme subordonnée dans $\mathcal{M}_n(\mathbb{K})$: $\forall M \in \mathcal{M}_n(\mathbb{K})$, $\|M\| = \sup_{0 \neq x \in \mathbb{K}^n} \frac{\|Mx\|}{\|x\|}$. Pour une matrice $A \in \mathcal{M}_n(\mathbb{K})$ inversible, on définit le nombre de conditionnement de A par rapport à la norme $\| \ \|$ par* $\mathrm{cond}_{\| \ \|}(A) = \|A\|\|A^{-1}\|$.

On vérifie facilement avec (4.8) que $\mathrm{cond}_{\| \ \|}(A) \geq 1$.

Remarque 4.3 *Il est clair que si $\alpha \in \mathbb{K}\backslash\{0\}$ alors $\mathrm{cond}_{\|.\|}(\alpha A) = \mathrm{cond}_{\|.\|}(A)$, et que $\mathrm{cond}_{\|.\|}(A^{-1}) = \mathrm{cond}_{\|.\|}(A)$.*

Un exemple important est le conditionnement par rapport à la norme $\| \ \|_2$, qu'on note cond_2 : on peut le caractériser par

- $\mathrm{cond}_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$, où $\sigma_{\max}(A)$ et $\sigma_{\min}(A)$ sont respectivement la plus grande et la plus petite valeur singulière de A.
- si A est normale (en particulier hermitienne), on a la caractérisation plus simple : $\mathrm{cond}_2(A) = \frac{|\lambda|_{\max}(A)}{|\lambda|_{\min}(A)}$ ou $|\lambda|_{\max}(A)$ et $|\lambda|_{\min}(A)$ sont les plus grand et plus petit modules des valeurs propres de A.

4.2.1 Sensibilité de la solution d'un système linéaire

On veut étudier la sensibilité de la solution du système $Ax = b$ (A inversible) à une variation du second membre δb : on a $A(x + \delta x) = b + \delta b$, ce qui implique que $\delta x = A^{-1}\delta b$, donc $\|\delta x\| \leq \|A^{-1}\|\|\delta b\|$. D'autre part $\|b\| \leq \|A\|\|x\|$, par suite $\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|} = \mathrm{cond}_{\| \ \|}(A)\frac{\|\delta b\|}{\|b\|}$. On a le résultat

Proposition 4.7 *Si $Ax = b$ et $A(x + \delta x) = b + \delta b$ on a*

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}_{\| \ \|}(A)\frac{\|\delta b\|}{\|b\|}, \tag{4.22}$$

et on peut trouver x et δb tels que cette inégalité soit une égalité.

Démonstration. On a déjà prouvé la première assertion.
La dernière assertion est prouvée en prenant x réalisant le maximum : $\frac{\|Ax\|}{\|x\|} = \max_{0 \neq y} \frac{\|Ay\|}{\|y\|} = \|A\|$ et δb réalisant le maximum $\frac{\|A^{-1}\delta b\|}{\|\delta b\|} = \max_{0 \neq y} \frac{\|A^{-1}y\|}{\|y\|} = \|A^{-1}\|$. L'inégalité (4.22) est donc optimale. ■

On s'intéresse maintenant à la sensibilité aux perturbations de la matrice A : soit donc une perturbation δA de A et $x + \delta x$ la solution de $(A + \delta A)(x + \delta x) = b$. On a $\delta x = -A^{-1}\delta A(x + \delta x)$, d'où $\frac{\|\delta x\|}{\|x+\delta x\|} \leq \|A^{-1}\|\|\delta A\| = \mathrm{cond}_{\| \ \|}(A)\frac{\|\delta A\|}{\|A\|}$.

Proposition 4.8 *Si $Ax = b$ et $(A + \delta A)(x + \delta x) = b$ on a*

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \mathrm{cond}_{\| \ \|}(A)\frac{\|\delta A\|}{\|A\|}, \tag{4.23}$$

et on peut trouver A, x et δx tels que cette inégalité soit une égalité.

Démonstration. La dernière assertion est prouvée en prenant $\delta A = I$ et $y \in \mathbb{K}^n$ réalisant le maximum de $\frac{\|A^{-1}y\|}{\|y\|}$. On prend alors $b = (A + \delta A)y$ et x tel que $Ax = b$. On a $\delta x = y - x$. Pour ce choix de matrices et de vecteurs, l'inégalité (4.23) devient une égalité. ■

4.3 Rayon spectral d'une matrice

Dans cette section et la suivante, on considère des matrices de $\mathcal{M}_d(\mathbb{C})$.

4.3.1 Définition et propriétés

Définition 4.3 *Soit A une matrice de $\mathcal{M}_d(\mathbb{C})$. On sait que le polynôme caractéristique de A a d racines complexes. On appelle spectre de A et on note $sp(A)$ l'ensemble formé de ses d valeurs propres (comptées avec leur multiplicité). On appelle rayon spectral de A et on note $\rho(A)$ le réel positif ou nul défini par*

$$\rho(A) = \max_{\lambda \in sp(A)} |\lambda|. \tag{4.24}$$

Proposition 4.9 ;

1. *Soit $\|\ \|$ une norme matricielle sur $\mathcal{M}_d(\mathbb{C})$ (ayant donc la propriété multiplicative (4.8))*

$$\rho(A) \leq \|A\| \tag{4.25}$$

2. *De plus pour toute matrice A et pour tout réel positif ϵ, il existe une norme matricielle $\|\ \|$ subordonnée à une norme vectorielle telle que*

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

Démonstration. Démontrons le premier point. Soit $\|\ \|$ une norme matricielle sur $\mathcal{M}_d(\mathbb{C})$. On peut alors définir une norme vectorielle sur $\mathbb{C}^d$ qu'on note $|||\ |||$ par

$$|||v||| = \|M_v\|, \quad \text{où} \quad M_v = (v, \dots, v)$$

Exercice Vérifier qu'il s'agit bien d'une norme sur $\mathbb{C}^d$.
Soit λ une valeur propre de A telle que $|\lambda| = \rho(A)$ et v_λ un vecteur propre correspondant : on a

$$\rho(A)|||v||| = |||Av||| \leq \|(Av, \dots, Av)\| = \|A(v, \dots, v)\| \leq \|A\|||(v, \dots, v)\| = \|A\|||v|||,$$

l'inégalité venant de la propriété (4.8). On a donc

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

Pour le deuxième point, on sait qu'il existe une matrice unitaire U et une matrice triangulaire supérieure R telles que $A = URU^*$. La matrice triangulaire R s'écrit

$$R = \begin{pmatrix} \lambda_1 & r_{12} & & \dots & r_{1d} \\ & \lambda_2 & r_{23} & \dots & r_{2d} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_{d-1} & r_{d-1\,d} \\ & & & & \lambda_d \end{pmatrix}$$

Soit alors la matrice $D_\eta = \mathrm{Diag}(1, \eta, \dots \eta^{d-1})$. On considère la norme ||| ||| définie par

$$|||M||| = \|D_\eta^{-1} U^* M U D_\eta\|_\infty$$

On a

$$D_\eta^{-1} U^* A U D_\eta = D_\eta^{-1} R D_\eta = \begin{pmatrix} \lambda_1 & \eta r_{12} & \dots & & \eta^{d-1} r_{1d} \\ & \lambda_2 & \eta r_{23} & \dots & \eta^{d-2} r_{2d} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_{d-1} & \eta r_{d-1\,d} \\ & & & & \lambda_d \end{pmatrix}$$

Donc $|||A||| = \max_{1\le i\le d} \left(|\lambda_i| + \sum_{j>i} \eta^{j-i} |r_{ij}| \right)$. Pour $\epsilon > 0$ fixé, on peut choisir η assez petit pour que $|||A||| \le \rho(A) + \epsilon$.

Enfin on vérifiera facilement que ||| ||| est la norme subordonnée à la norme vectorielle

$$|||v||| = \|D_\eta^{-1} U^* v\|_\infty$$

■

Proposition 4.10 *1. Pour toute norme matricielle sur $\mathcal{M}_d(\mathbb{C})$ et pour toute matrice $A \in \mathcal{M}_d(\mathbb{C})$, on a l'implication*

$$\|A\| < 1 \Rightarrow I_d + A \text{ inversible}$$

2. Si la norme $\|.\|$ est de plus subordonnée à une norme sur $\mathbb{C}^n$ et si $\|A\| < 1$ alors

$$\|(I_d + A)^{-1}\| \le \frac{1}{1 - \|A\|}.$$

Démonstration.

1. Si $I_d + A$ n'est pas inversible, il existe $u \neq 0$ tel que $Au = -u$, et $\|A\| \ge \rho(A) \ge 1$. Donc, si $\|A\| < 1$ alors $I_d + A$ est inversible.
2. Si $\|.\|$ est une norme subordonnée, on a $\|I_d\| = 1$ et $(I_d + A)^{-1} = I_d - A(I_d + A)^{-1}$, et donc $\|(I_d + A)^{-1}\| \le 1 + \|A\| \|(I_d + A)^{-1}\|$ ce qui donne le résultat désiré.

■

Remarque 4.4 *On a aussi par contraposition : si $I_d + A$ n'est pas inversible alors $\|A\| \ge 1$, pour toute norme matricielle subordonnée ou non.*

4.3.2 Suite des puissances d'une matrice

On dit qu'une suite de matrices $(A_n)_{n\in\mathbb{N}}$ dans $\mathcal{M}_{d,d'}(\mathbb{C})$ tend vers 0, si pour une norme $\|\ \|$ définie sur $\mathcal{M}_{d,d'}(\mathbb{C})$, la suite $(\|A_n\|)_{n\in\mathbb{N}}$ tend vers 0 quand $n \to \infty$.
Comme toutes les normes sont équivalentes, cette convergence a alors lieu pour toutes les normes.
Si $d = d'$, en prenant les normes subordonnées aux normes vectorielles, on voit que

$$A_n \to 0 \Leftrightarrow \forall v \in \mathbb{C}^d, A_n v \to 0.$$

On déduit de la proposition 4.9 le théorème :

Théorème 4.1 *Soit $A \in \mathcal{M}_d(\mathbb{C})$. Une condition nécessaire et suffisante pour que la suite des puissances de A, $(A^n)_{n\in\mathbb{N}}$, converge vers 0 est que*

$$\rho(A) < 1. \tag{4.26}$$

Démonstration. Soit une matrice $A \in \mathcal{M}_d(\mathbb{C})$ telle que $A^n \to 0$ quand $n \to \infty$. Il est clair que $\rho(A^n) = \rho^n(A)$. Mais d'après la proposition 4.9, $\rho^n(A) \leq \|A^n\|$, pour toute norme matricielle. Donc $\rho^n(A) \to 0$, ce qui implique que $\rho(A) < 1$.
Réciproquement, soit une matrice telle que $\rho(A) < 1$. D'après la proposition 4.9, il existe une norme matricielle $\| \ \|$ subordonnée à une norme vectorielle, telle que $\|A\| < 1$. Comme $\| \ \|$ a la propriété (4.8), $\|A^n\| \leq \|A\|^n \to 0$ quand $n \to \infty$. ■

Le résultat suivant est une caractérisation du rayon spectral d'une matrice :

Théorème 4.2 *Pour toute norme matricielle sur $\mathcal{M}^n(\mathbb{C})$,*

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}}. \tag{4.27}$$

Démonstration. On a bien sûr :

$$\rho(A) \leq \|A^n\|^{\frac{1}{n}}, \quad \forall n > 0,$$

car $\rho(A)^n = \rho(A^n) \leq \|A^n\|$. Donc

$$\rho(A) \leq \liminf_{n\to\infty} \|A^n\|^{\frac{1}{n}}. \tag{4.28}$$

D'autre part, pour $\epsilon > 0$, considérons la matrice $A_\epsilon = \frac{1}{\rho(A)+\epsilon}A$. On a que $\rho(A_\epsilon) < 1$, donc $\lim_{n\to\infty} A_\epsilon^n = 0$: il existe donc $N > 0$ tel que pour tout $n > N$,

$$\|A_\epsilon^n\| \leq 1,$$

ce qu'on écrit aussi

$$\|A^n\| \leq (\rho(A) + \epsilon)^n.$$

Donc, pour tout $\epsilon > 0$,

$$\limsup_{n\to\infty} \|A^n\|^{\frac{1}{n}} \leq \rho(A) + \epsilon,$$

donc

$$\limsup_{n\to\infty} \|A^n\|^{\frac{1}{n}} \leq \rho(A).$$

■

Proposition 4.11 *La série $\sum_{n=0}^{\infty} A^n$ est convergente si et seulement si $\rho(A) < 1$. On a alors*

$$(I_d - A)^{-1} = \sum_{n=0}^{\infty} A^n.$$

Démonstration. Pour tout ϵ, il existe une norme subordonnée $\|.\|$ vérifiant $\|A\| \leq \rho(A) + \epsilon$, ce qui montre que si $\rho(A) < 1$, il existe une norme subordonnée $\|.\|$ vérifiant $\|A\| < 1$; la série $\sum_{n=0}^{\infty} A^n$ est absolument convergente pour cette norme et on vérifie que

$$(I_d - A)\sum_{n=0}^{\infty} A^n = \lim_{N\to\infty}(I_d - A)\sum_{n=0}^{N} A^n = I_d - \lim_{N\to\infty} A^{N+1} = I_d.$$

Réciproquement, si $\sum_{n=0}^{\infty} A^n$ est convergente, alors pour toute valeur propre λ (complexe) de A, on a $\sum_{n=0}^{\infty} \lambda^n$ converge, ce qui implique que $|\lambda| < 1$. On en déduit que $\rho(A) < 1$. ■

Remarque 4.5 *Sous les hypothèses de la Proposition 4.11 on a*

$$(I_d + A)^{-1} = \sum_{n=0}^{\infty} (-1)^n A^n.$$

4.4 Sensibilité d'un problème aux valeurs propres

Exemple. On considère la matrice de $\mathcal{M}_n(\mathbb{R})$

$$\begin{pmatrix} 0 & & \dots & & \epsilon \\ 1 & 0 & \dots & & 0 \\ 0 & \ddots & \ddots & & \\ \vdots & & & & \\ 0 & \dots & & 1 & 0 \end{pmatrix}$$

Son polynôme caractéristique est $x^n \pm \epsilon$: les valeurs propres sont nulles si $\epsilon = 0$ et ce sont les racines n ièmes de $\pm\epsilon$ si $\epsilon \neq 0$. Si $n = 10$ et si $\epsilon = 10^{-10}$, les valeurs propres ont pour module 0.1. On voit donc qu'une perturbation de 10^{-10} d'un coefficient de la matrice entraîne des variations importantes des valeurs propres.

On note $\sigma(A)$ le spectre d'une matrice A de $\mathcal{M}_n(\mathbb{C})$.
Pour tenter de comprendre la sensibilité des valeurs propres aux perturbations d'une matrice, on a le résultat

Théorème 4.3 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ diagonalisable, et P une matrice de changement de base telle que $P^{-1}AP = \mathrm{Diag}(\lambda_1, \dots \lambda_n)$. Soit $|||$ une norme sur $\mathcal{M}_n(\mathbb{C})$ telle que*

$$\|MN\| \leq \|M\|\|N\| \quad \text{et } \|\mathrm{Diag}(\mu_1 \dots \mu_n)\| = \max_{i=1\dots n} |\mu_i|$$

alors pour toute perturbation δA de A, on a

$$\sigma(A + \delta A) \subset \cup_{i=1}^{n} D_i \tag{4.29}$$

et

$$D_i = \{\mu \in \mathbb{C};\ |\mu - \lambda_i| \leq \mathrm{cond}_{\|\ \|}(P)\|\delta A\|\}. \tag{4.30}$$

Démonstration On sait que μ est valeur propre de $A + \delta A$ si et seulement si $A + \delta A - \mu I$ n'est pas inversible. En multipliant à droite par P et à gauche par P^{-1}, ceci est équivalent à dire que $\mathrm{Diag}(\lambda_1 - \mu, \dots, \lambda_n - \mu) + P^{-1}\delta AP$ est singulière. Deux cas se présentent :

1. si $\mathrm{Diag}(\lambda_1 - \mu, \ldots, \lambda_n - \mu)$ est singulière, alors $\mu \in D_i$ pour un indice $i \in \{1, \ldots, n\}$.
2. sinon, $I + \mathrm{Diag}((\lambda_1 - \mu)^{-1}, \ldots, (\lambda_n - \mu)^{-1})P^{-1}\delta AP$ est singulière. Ceci implique que $\|\mathrm{Diag}((\lambda_1 - \mu)^{-1}, \ldots, (\lambda_n - \mu)^{-1})P^{-1}\delta AP\| \geq 1$, autrement on pourrait construire un inverse avec une série. Les propriétés de la norme impliquent alors que $\|\mathrm{Diag}((\lambda_1 - \mu)^{-1}, \ldots, (\lambda_n - \mu)^{-1})\|\|P^{-1}\|\|\delta A\|\|P\| \geq 1$, soit encore $\max_i |\lambda_i - \mu|^{-1}\mathrm{cond}_{\|\ \|}(P)\|\delta A\| \geq 1$. On en déduit que
$$\min_i |\lambda_i - \mu| \leq \mathrm{cond}_{\|\ \|}(P)\|\delta A\|$$

□

Remarque 4.6 *La sensibilité d'un problème aux valeurs propres d'une matrice A diagonalisable par rapport aux perturbations dépend donc du conditionnement de la matrice de passage P et non pas de celui de A.*

Remarque 4.7 *Si A est normale, on sait que A est diagonalisable avec P unitaire. On a $\mathrm{cond}_2(P) = 1$. Dans ce cas les valeurs propres de A sont contenues dans des cercles centrés aux valeurs propres de A et de rayon $\|\delta A\|_2$.*

Remarque 4.8 *On a un résultat plus précis si A et δA sont hermitiennes. Dans ce cas, cf. [2], le théorème du min-max permet de dire que si on ordonne les valeurs propres par valeur croissante, les k ièmes valeurs propres de A et $A + \delta A$ sont distantes d'au plus $\|\delta A\|_2$.*

Chapitre 5

Méthodes Directes pour les Systèmes Linéaires

Le but est de résoudre le système linéaire dans $\mathbb{C}^n$: trouver $x \in \mathbb{C}^n$ tel que

$$Ax = b, \quad A \in \mathcal{M}_n(\mathbb{C}),$$

où $b \in \mathbb{C}^n$. On suppose que $A \in \mathcal{M}_n(\mathbb{C})$ est inversible.

Cas où A est triangulaire supérieure : remontée La matrice $A = (a_{ij})_{i,j\in\{1,\dots,n\}}$ est triangulaire supérieure :

$$a_{ij} = 0 \quad \text{si} \quad j < i$$

Comme A est inversible,

$$a_{ii} \neq 0 \quad \text{si} \quad 1 \leq i \leq n.$$

Le système s' écrit

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ & & & & \ddots & & & & \\ & & & & & & a_{nn}x_n & = & b_n \end{array}$$

On résout le système en remontant

$$\begin{array}{lcl} x_n & = & b_n/a_{nn} \\ x_{n-1} & = & (b_{n-1} - a_{n-1\,n}x_n)/a_{n-1\,n-1} \\ & \vdots & \\ x_1 & = & (b_1 - a_{12}x_2 - \cdots - a_{1n}x_n)/a_{11} \end{array}$$

Coût du calcul de x_k :

$$x_k = (b_k - a_{k\,k+1}x_{k+1} - \cdots - a_{kn}x_n)/a_{kk} \rightarrow$$

$$\begin{array}{l} (n-k) \quad \text{additions} \\ (n-k) \quad \text{multiplications} \\ 1 \quad \text{division} \end{array}$$

Coût total de la remontée :

$$\sim \sum_{k=1}^{n}(n-k) = \frac{1}{2}(n-1)n \sim \frac{n^2}{2} \quad \text{additions+multiplications}$$

Cas où A est triangulaire inférieure : descente La matrice $A = (a_{ij})_{i,j\in\{1,\dots,n\}}$ est triangulaire inférieure :

$$a_{ij} = 0 \quad si \quad i < j$$

Comme A est inversible,

$$a_{ii} \neq 0 \quad si \quad 1 \leq i \leq n.$$

Le système s' écrit

$$\begin{array}{lllllll} a_{11}x_1 & & & & & & = b_1 \\ a_{21}x_1 & + & a_{22}x_2 & & & & = b_2 \\ & & \ddots & & & & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n = b_n \end{array}$$

On résout le système en descendant

$$\begin{array}{lll} x_1 & = & b_1/a_{11} \\ x_2 & = & (b_2 - a_{21}x_1)/a_{22} \\ & \vdots & \\ x_n & = & (b_n - a_{n\,n-1}x_{n-1} - \cdots - a_{n1}x_1)/a_{nn} \end{array}$$

Coût total de la descente :

$$\sim \sum_{k=1}^{n} k = \frac{1}{2}(n-1)n \sim \frac{n^2}{2} \quad \text{additions+multiplications}$$

5.0.1 Algorithme d'élimination de Gauss sans recherche de pivot, et interprétation matricielle

On suppose que A est inversible. On veut résoudre le système : trouver $x \in \mathbb{C}^n$ tel que

$$Ax = b, \quad A \in \mathcal{M}_n(\mathbb{C}),$$

où $b \in \mathbb{C}^n$. On suppose que $A \in \mathcal{M}_n(\mathbb{C})$ est inversible.
Le but est de se ramener à un système triangulaire

Première étape : élimination de l'inconnue x_1 des lignes 2 à n. On va chercher un système équivalent (ayant la même solution x), où x_1 n'apparaît que dans la ligne 1.
Première Hypothèse :

$$a_{11} \neq 0 \rightarrow \text{ on appelle premier pivot } \pi^{(1)} = a_{11}.$$

On peut alors former une combinaison linéaire de la ligne $i > 1$ avec la ligne 1 pour éliminer l'inconnue x_1 dans la ligne i : la ligne i devient

$$0x_1 + (a_{i2} - \frac{a_{i1}}{\pi^{(1)}}a_{12})x_2 + \cdots + (a_{in} - \frac{a_{i1}}{\pi^{(1)}}a_{1n})x_n = b_i - \frac{a_{i1}}{\pi^{(1)}}b_1,$$

ce qu'on peut réécrire

$$a_{i2}^{(2)}x_2 + \cdots + a_{in}^{(2)}x_n = b_i^{(2)}, \quad \forall i > 1$$

en posant

$$\begin{aligned} a_{i2}^{(2)} &= a_{i2} - \tfrac{a_{i1}}{\pi^{(1)}} a_{12}, \\ &\vdots \\ a_{in}^{(2)} &= a_{in} - \tfrac{a_{i1}}{\pi^{(1)}} a_{1n}, \\ b_i^{(2)} &= b_i - \tfrac{a_{i1}}{\pi^{(1)}} b_1. \end{aligned}$$

Si on pose aussi, pour $1 \leq j \leq n$,

$$a_{1j}^{(2)} = a_{1j} \quad \forall 1 \leq j \leq n \quad \text{et} \quad b_1^{(2)} = b_1,$$

on a obtenu le nouveau système équivalent

$$A^{(2)} x = b^{(2)},$$

avec

$$A^{(2)} = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix}$$

On voit que

$$A^{(2)} = M^{(1)} A \quad \text{où} \quad M^{(1)} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\frac{a_{n1}}{a_{11}} & 0 & \dots & 0 & 1 \end{pmatrix}$$

Le système s'écrit :

$$A^{(2)} = M^{(1)} A, \quad \text{et} \quad b^{(2)} = M^{(1)} b.$$

$k$ième étape : élimination de l'inconnue x_k des lignes $k+1$ à n On suppose qu'on a pu itérer le procédé ci-dessus $k-1$ fois, c'est à dire que les pivots apparus jusqu'àl'étape k sont non nuls :

$$\Pi^{(i)} = a_{ii}^{(i)} \neq 0, \quad \text{pour } 1 \leq i \leq k-1.$$

On obtient alors le système équivalent

$$A^{(k)} x = b^{(k)}$$

où $A^{(k)}$ est de la forme

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & \dots & \dots & \dots & & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \ddots & & & & & \vdots \\ \vdots & \ddots & a_{33}^{(k)} & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & \dots & 0 & a_{k-1k-1}^{(k)} & \dots & \dots & a_{k-1n}^{(k)} \\ \vdots & & & \vdots & 0 & & & \\ \vdots & & & \vdots & \vdots & & \widetilde{A^{(k)}} & \\ 0 & \dots & \dots & 0 & 0 & & & \end{pmatrix}$$

et $\widetilde{A^{(k)}}$ est une matrice carrée d'ordre $n-k+1$:

$$\widetilde{A^{(k)}} = \begin{pmatrix} a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \dots & a_{nk}^{(k)} \end{pmatrix}.$$

On va chercher un système équivalent (ayant la même solution x), où x_k n'apparaît pas dans les lignes $k+1$ à n.

k-ième Hypothèse :

$$a_{kk}^{(k)} \neq 0 \rightarrow \text{ on appelle } k\text{ième pivot } \pi^{(k)} = a_{kk}^{(k)}.$$

On introduit

$$M^{(k)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+1\,k}^{(k)}}{\pi^{(k)}} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{a_{n\,k}^{(k)}}{\pi^{(k)}} & 0 & \dots & 0 & 1 \end{pmatrix}$$

Remarquons que l'inverse de $M^{(k)}$ est $L^{(k)}$:

$$L^{(k)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & \frac{a_{k+1\,k}^{(k)}}{\pi^{(k)}} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{a_{n\,k}^{(k)}}{\pi^{(k)}} & 0 & \dots & 0 & 1 \end{pmatrix}$$

Soit

$$A^{(k+1)} = M^{(k)} A^{(k)} \quad \text{et} \quad b^{(k+1)} = M^{(k)} b^{(k)},$$

alors

$$A^{(k+1)} x = b^{(k+1)},$$

et

$$A^{(k+1)} = \begin{pmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & \dots & \dots & \dots & \dots & & a_{1n}^{(k+1)} \\ 0 & a_{22}^{(k+1)} & \ddots & & & & & \vdots \\ \vdots & \ddots & a_{33}^{(k+1)} & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & \dots & 0 & a_{kk}^{(k+1)} & \dots & \dots & a_{kn}^{(k+1)} \\ \vdots & & & \vdots & 0 & & & \\ \vdots & & & \vdots & \vdots & & \widetilde{A^{(k+1)}} & \\ 0 & \dots & \dots & 0 & 0 & & & \end{pmatrix},$$

où $\widetilde{A^{(k+1)}}$ est une matrice carrée d'ordre $n-k$:

$$\widetilde{A^{(k+1)}} = \begin{pmatrix} a_{k+1k+1}^{(k+1)} & \dots & a_{k+1n}^{(k+1)} \\ \vdots & & \vdots \\ a_{nk+1}^{(k+1)} & \dots & a_{nk+1}^{(k+1)} \end{pmatrix}.$$

Après $n-1$ étapes Si on itère $n-1$ fois, et si les pivots apparus sont tous non nuls :
Hypothèses des $n-1$ étapes :

$$\pi^{(i)} = a_{ii}^{(i)} \neq 0, \quad \text{pour } 1 \leq i \leq n-1.$$

on arrive au système **triangulaire équivalent**

$$A^{(n)}x = b^{(n)},$$

avec

$$A^{(n)} = \begin{pmatrix} \pi^{(1)} & * & & \dots & * \\ 0 & \pi^{(2)} & * & \dots & * \\ 0 & 0 & \ddots & & \\ 0 & \dots & 0 & \ddots & * \\ 0 & & \dots & 0 & \pi^{(n)} \end{pmatrix}$$

qui est inversible si de plus
Hypothèse :

$$\pi^{(n)} \neq 0.$$

Si on appelle L la matrice triangulaire inférieure avec des 1 sur la diagonale

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

et U la matrice triangulaire supérieure

$$U = A^{(n)},$$

on a

$$A = LU.$$

On dit qu'on a effectué une factorisation de Gauss ou factorisation LU de A.

Remarque 5.1 *Il est aussi possible avec le même algorithme d'obtenir la factorisation LU d'une matrice de $\mathcal{M}_m(\mathbb{C})$, avec $m \geq n$, si les pivots qui apparaissent sont non nuls.*

5.0.2 Coût de la méthode d'élimination de Gauss.

On estime le coût de l'élimination de x_k. Rappelons qu'à l'étape $k-1$, on obtient la matrice A_k

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & \dots & \dots & \dots & & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \ddots & & & & & \vdots \\ \vdots & \ddots & a_{33}^{(k)} & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & \dots & 0 & a_{k-1k-1}^{(k)} & \dots & \dots & a_{k-1n}^{(k)} \\ \vdots & & & \vdots & 0 & & & \\ \vdots & & & \vdots & \vdots & & \widetilde{A^{(k)}} & \\ 0 & \dots & \dots & 0 & 0 & & & \end{pmatrix}$$

C'est sur le bloc $\widetilde{A^{(k)}}$ de $A^{(k)}$ qu'il faut travailler.
Construction de $M^{(k)}$: $n-k$ divisions par le pivot.

$$M^{(k)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+1\,k}^{(k)}}{\pi^{(k)}} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{a_{n\,k}^{(k)}}{\pi^{(k)}} & 0 & \dots & 0 & 1 \end{pmatrix}$$

Multiplication de $A^{(k)}$ par $M^{(k)}$: $(n-k)^2$ adds+mults.
Total

$$\sum_{k=1}^{n}(n-k)^2 = \frac{1}{3}n(n-1)(n-\frac{1}{2}) \sim \frac{n^3}{3} \text{ additions+multiplications.}$$
$$\sum_{k=1}^{n}(n-k) = \frac{1}{2}n(n-1) \text{ divisions.}$$

5.0.3 Utilisations de la factorisation LU

Calcul de déterminant Une utilisation de la factorisation LU est le calcul du déterminant de A : en effet, si A admet une factorisation LU, on a

$$\det(A) = \det(U) = \text{ produit des pivots}$$

car $\det(L) = 1$. La factorisation LU permet donc de calculer le déterminant de A avec une complexité de l'ordre de $\frac{n^3}{3}$ adds + mults, plutôt que $n!$ avec la formule du déterminant !

Remarque 5.2 *Calculer le déterminant d'une matrice d'ordre 100 avec la formule du déterminant prendrait un temps supérieur à l'age de l'univers sur le calculateur le plus puissant disponible aujourd'hui.*

Résolution de systèmes linéaires Si on a plusieurs systèmes (par exemple p) à résoudre avec la même matrice A, il est intéressant de calculer une fois pour toute et stocker la factorisation LU de $A = LU$, puis de résoudre les systèmes par descentes-remontées plutôt que d'utiliser l'algorithme d'élimination pour chaque système. On résout le système $Ax = b$ en résolvant d'abord

$$Ly = b,$$

soit une descente, puis

$$Ux = y,$$

soit une remontée. Si on a p systèmes à résoudre, la complexité est de l'ordre de $\frac{n^3}{3} + pn^2$ adds+mults plutôt que $p\frac{n^3}{3}$.

Remarque 5.3 *Comme à la remarque précédente, calculer la solution du système linéaire en appliquant les formules de Crammer nécessiterait un temps inimaginable dés que n vaut quelques dizaines.*

5.0.4 Algorithme

Voici un algorithme réalisant la factorisation LU d'une matrice A et stockant cette factorisation dans la place mémoire occuppée par A : (la matrice A est perdue, on dit qu'on écrase A).

```
for (int j=1;j<=n-1;j++)
  {
   for(int i=j+1;i<=n;i++)
      A(i,j)=A(i,j)/A(j,j);   //construction de la j eme colonne de L

   for(int i=j+1;j<=n;j++)
      for(int k:=j+1;k<=n;k++)
          A(i,k)=A(i,k)-A(i,j)*A(j,k);
                    //actualisation des n-j-1 derni{\‘e}res lignes de A
  }
```

Remarque 5.4 *La diagonale de L (qui ne contient que des 1) n'est pas stockée.*

Le programme pour la descente-remontée pour calculer la solution de $Ax = b$ est alors, en écrasant b :

```
 for (int i=1;i<=n;i++)  //descente pour calculer y tq Ly=b
    {
      sum=0.;
      for (int k=1; k<i;k++)
        sum += A(i,k)*b(k);
      b(i) = b(i) - sum ;
    }
  for (int i=n;i>=1;i--) //remontee pour calculer x tq Ux=y
    {
      sum=0.;
      for (int k=i+1; k<=n;k++)
        sum += A(i,k)*b(k);
      b(i) = (b(i)-sum)/A(i,i) ;
    }
```

5.0.5 Condition nécessaire et suffisante d'existence d'une factorisation LU, unicité

Théorème 5.1 *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. Pour $1 \leq p \leq n$, on note A_p le bloc*

$$A_p = \begin{pmatrix} a_{1\,1} & \dots & \dots & a_{1\,p} \\ a_{2\,1} & \dots & \dots & a_{2\,p} \\ \vdots & & & \vdots \\ a_{p\,1} & \dots & \dots & a_{p\,p} \end{pmatrix}$$

La matrice A admet une factorisation

$$A = LU$$

où L est triangulaire inférieure avec des 1 sur la diagonale et U est triangulaire supérieure et inversible si et seulement si tous les blocs A_p, $1 \leq p \leq n$, sont inversibles. De plus, cette factorisation est unique. De plus si A est réelle, alors L et U le sont aussi.

Démonstration. Si A admet une factorisation LU avec L et U inversibles, alors les blocs A_p sont bien inversibles.La démonstration de la réciproque demande un peu plus de travail :
Existence
La démonstration se fait par récurrence : la propriété est clairement vraie si $n = 1$. Supposons la propriété vraie pour $p \leq n-1$, et soit une matrice $A \in \mathcal{M}_n(\mathbb{C})$ dont tous les blocs A_p sont inversibles. L'hypothèse de récurrence implique l'existence d'une factorisation LU pour le bloc A_{n-1} : $A_{n-1} = L'U'$. On va chercher une factorisation LU de A sous la forme

$$A = \begin{pmatrix} L' & 0 \\ x^T & 1 \end{pmatrix} \begin{pmatrix} U' & y \\ 0 & z \end{pmatrix}$$

où $x, y \in \mathbb{C}^{n-1}$ et $z \in \mathbb{C}$, ou de manière équivalente

$$\begin{aligned} L'y &= (a_{1,n}, \dots, a_{n-1,n})^T \\ U'^T x &= (a_{n,1}, \dots a_{n,n-1})^T \\ x^T y + z &= a_{n,n} \end{aligned}$$

D'après l'hypothèse de récurrence, L' et U' sont inversibles, donc x et y existent, et z aussi. De plus, comme $\det(A) \neq 0$, z est non nul. L'existence d'une factorisation LU de A où L a des 1 sur la diagonale est démontrée.
Unicité
Soient deux factorisation LU de A : $A = LU = L'U'$ où L sont triangulaires inférieures avec des 1 sur la diagonale, et U sont triangulaires supérieures et inversibles. On a donc l'identité

$$L^{-1}L' = U'U^{-1}$$

Mais $L^{-1}L'$ est triangulaire inférieure avec des 1 sur la diagonale et $U'U^{-1}$ est triangulaire supérieure. On a donc

$$L^{-1}L' = U'U^{-1} = Id \Rightarrow \left| \begin{array}{l} L = L' \\ U = U' \end{array} \right.$$

■

5.0.6 Un deuxième algorithme

La démonstration précédente fournit un autre algorithme. Là encore, on stocke la factorisation LU dans la place mémoire occupée par A. L'idée est de faire une boucle sur p pour p allant de 1 à n en construisant à chaque étape la factorisation LU de A_p.

```
for (int p=1;p<=n;p++)
 {
   for(int i=1;i<p;i++)   //construction de la p eme colonne de U
     {                    //sauf le coeff diagonal
       sum = 0;
       for (int k=1;k<i;k++)
         sum += A(i,k)*A(k,p);
       A(i,p) = A(i,p)-sum;
     }
   for(int j=1;j<p;j++)  //construction de la p eme ligne de L
     {
       sum = 0;
       for (int k=1; k<j;k++)
         sum += A(k,j)*A(p,k);
       A(p,j) = (A(p,j) - sum)/A(j,j);
     }
   sum=0;               //nouveau pivot
   for(int i=1;i<p;i++)
      sum+=A(p,i)*A(i,p);
   A(p,p)-=sum;
 }
```

5.0.7 Cas de matrices bandes

Définition 5.1 *Soit A une matrice de $C^{n\times n}$. On appelle largeur de bande de A le plus petit entier $n_b \leq n$ tel que pour tout i, $1 \leq i \leq n$*

$$\begin{aligned} a_{i,j} &= 0 \quad \text{si } n \geq j > i + n_b, \\ a_{j,i} &= 0 \quad \text{si } 0 < j < i - n_b, \end{aligned}$$

Les coefficients non nuls de la matrices A sont contenus dans une bande de largeur $2n_b + 1$, centrée sur la diagonale.

Dans le cas où la largeur de bande n_b de A est très inférieure à n, on parle de matrice bande. Si A est une matrice bande, sa factorisation LU demande moins d'opérations et de place mémoire :

Proposition 5.1 *Si A, (de largeur de bande n_b) admet une factorisation LU, alors les largeurs de bandes de L et de U sont inférieures à n_b et la complexité nécessaire à effectuer la factorisation LU est inférieure à*

$$\sum_{k=1}^{n} (n_b)^2 = n(n_b)^2 \ \textit{additions+multiplications.}$$

$$\sum_{k=1}^{n} n_b = nn_b \ \textit{divisions.}$$

Démonstration. Montrons la propriété par récurrence sur les étapes de la factorisation LU de A.
La propriété est évidente à l'étape 0 puisque A a la structure bande.
Supposons la propriété vraie jusqu'à k : on a $A = L^{(1)} \dots L^{(k-1)} A^{(k)}$ où $A^{(k)}_{i,j} = 0$ si $j < i$ et $i < k$ ou si $|j - i| > n_b$. Regardons la k ème etape de la factorisation :

- la première chose que l'on fait est de construire $L^{(k)}$ avec $L^{(k)}_{ik} = A^{(k)}_{i,k}/A^{(k)}_{k,k}$ pour $i > k$. Mais on voit que $A^{(k)}_{i,k} = 0$ si $i > k + nb$, donc $L^{(k)}$ a elle aussi une largeur de bande inférieure ou égale à n_b.
- on construit alors $A^{(k+1)}$ en ne modifiant que le bloc $k+1 \to n$ de $A^{(k)}$: $A^{(k+1)}_{i,j} = A^{(k)}_{i,j} - L^{(k)}_{i,k} A^{(k)}_{k,j}$ pour $i > k$ et $j > k$. Mais on sait que $L^{(k)}_{i,k} = 0$ pour $i > k + n_b$ et que $A^{(k)}_{k,j} = 0$ pour $j > k + n_b$. Les seuls coefficients de $A^{(k)}$ que l'on va modifier sont donc les coefficients correspondants à $k < i \leq \min(k + n_b, n)$ et $k < j \leq \min(k + n_b, n)$. Des inégalités $k < i \leq \min(k + n_b, n)$ et $k < j \leq \min(k + n_b, n)$, on tire que $-n_b < j - i < n_b$: les seuls coefficients de $A^{(k)}$ modifiés seront compris dans la bande $|j - i| \leq n_b$: la matrice $A^{(k+1)}$ hérite donc de la structure bande de la matrice $A^{(k)}$.

■

Exercice. Écrire l'algorithme de factorisation LU pour une matrice bande de largeur de bande n_b.

5.1 Méthode de Cholesky

Dans le cas où A est hermitienne et définie positive, on peut toujours efectuer la factorisation décrite ci-dessus. De plus, on peut trouver une factorisation du type $A = LL^*$ moins gourmande en place mémoire.

5.1.1 Existence de la factorisation de Cholesky

Théorème 5.2 *Si A est hermitienne et définie positive, alors A admet une unique factorisation LU, où L est triangulaire inférieure avec des 1 sur la diagonale et U est triangulaire supérieure et inversible.*

Démonstration. Si A est hermitienne définie positive, ses blocs A_p $1 \leq p \leq n$, voir Théorème 5.1, le sont aussi, et on peut appliquer le Théorème 5.1. ■

Théorème 5.3 *Si A est hermitienne et définie positive, alors il existe une unique matrice L triangulaire inférieure et inversible, avec des coefficients réels positifs sur la diagonale, telle que*

$$A = LL^*$$

Cette factorisation porte le nom de factorisation de Cholesky (qui était un colonel de l'armée de Napoléon). Si A est réelle symétrique définie positive, L est réelle.

Démonstration.

On part de la factorisation LU de A. Il existe une unique factorisation $A = L'U'$ où la matrice L' est triangulaire inférieure avec des 1 sur la diagonale et U' est triangulaire supérieure inversible. Appelons D la diagonale de U'. Comme pour tout p, $1 \leq p \leq n$, $\det(D_p) = \det(A_p) >$

0, tous les coefficients de D sont strictement positifs. Notons $U = D^{-\frac{1}{2}}U'$ et $L = L'D^{\frac{1}{2}}$. On a $A = LU$. Montrons que $U = L^*$. Comme A est hermitienne, $U^*L^* = A^* = A = LU$. On a donc $L^{-1}U^* = U(L^*)^{-1}$. Mais $L^{-1}U^*$ est triangulaire inférieure, avec des 1 sur la diagonale tandis que $U(L^*)^{-1}$ est triangulaire supérieure, avec des 1 sur la diagonale. Donc $U = L^*$, et on a $A = LL^*$. Pour l'unicité, on procède comme dans la démonstation de l'unicité pour la factorisation LU. ■

Remarque 5.5 *Il est important de comprendre que les matrices L dans les factorisation LU et de Cholesky sont différentes.*

5.1.2 Algorithme

Considérons la k ième étape de la méthode de Cholesky : A ce point, on suppose que l'on a déjà construit la matrice $L^{(k-1)}$ et $A^{(k)}$ telles que

$$L^{(k-1)} = \begin{pmatrix} \sqrt{\pi_1} & 0 & & & \dots & & 0 \\ * & \sqrt{\pi_2} & 0 & & \dots & & 0 \\ & & \ddots & & & & \\ * & \dots & * & \sqrt{\pi_{k-1}} & 0 & \dots & 0 \\ * & \dots & * & * & 1 & \dots & 0 \\ * & \dots & * & \vdots & 0 & \ddots & 0 \\ * & \dots & * & * & 0 & \dots & 1 \end{pmatrix}$$

et

$$A^{(k)} = \begin{pmatrix} * & * & & & \dots & & * \\ 0 & * & & & \dots & & * \\ 0 & 0 & * & & \dots & & * \\ & & & \ddots & & & \\ 0 & 0 & \dots & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & & & & \vdots & \ddots & \\ 0 & 0 & \dots & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}$$

et $\pi_k = a_{k,k}^{(k)}$.

Remarque 5.6 *Le bloc carré $k \to n$ de $A^{(k)}$ est hermitien et défini positif.*

Pour effectuer la $k$ième étape, on construit d'abord $L^{(k)}$ en remplaçant la $k$ième colonne de $L^{(k-1)}$ par le vecteur $l_k = (0, \dots, 0, \sqrt{\pi_k}, \frac{a_{k+1\,k}^{(k)}}{\sqrt{\pi_k}}, \dots \frac{a_{n\,k}^{(k)}}{\sqrt{\pi_k}})^T$.

Observation 5.1 *D'après la Remarque 5.6,* $(0, \dots, 0, a_{kk}^{(k)}, \dots, a_{kn}^{(k)}) = \sqrt{\pi_k} l_k^*$.

On construit maintenant $A^{(k+1)}$ en effectuant les éliminations de Gauss et d'après l'Observation 5.1, on a

$$A^{(k+1)} = A^{(k)} - l_k l_k^*. \tag{5.1}$$

On voit en fait que l'on peut ne construire que la partie triangulaire inférieure de la matrices $A^{(k+1)}$, ce qui économise la moitié des opérations. On aboutit à l'algorithme suivant dans le cas d'une matrice réelle symétrique définie positive :

```
  for (k = 1; k <= n; k++)
  {
    A(k,k)=sqrt(A(k,k) );
    for (i = k+1; i <= n ; i++)
      A(i,k)= A(i,k) / A(k,k);
    for (i = k + 1; i <=n; i++)
      for (j = k+1 ; j <= i; j++)
        A(i,j)=A(i,j) - A(i,k)*A(j,k);
  }
```

5.1.3 Complexité de la factorisation de Cholesky

$\sim \frac{n^3}{6}$ additions+multiplications.
$\frac{1}{2}n(n-1)$ divisions.
n évaluations de racines carrées.

Exercice Montrer l'évaluation précédente.

Remarque 5.7 *L'intérêt de la factorisation de Cholesky est qu'elle demande une place mémoire deux fois inférieure à celles de la factorisation LU ci-dessus, car on ne doit stocker que L et que sa complexité arithmétique est deux fois moindre.*

5.2 Programmes Scilab pour les factorisation LU et de Cholesky

5.2.1 Factorisation LU

Construction de la factorisation LU

```
//LU factorization
function [A]= LUfact(A)
[m,n]=size(A);
for i=1:n,
  if (A(i,i)==0) then
    print(%io(2)," zero pivot")
  else
    A(i+1:m,i)=A(i+1:m,i)/A(i,i);
    for j=i+1:n,
      A(i+1:m,j)=A(i+1:m,j)-A(i+1:m,i)*A(i,j);
    end ;
  end ;
end;
```

Descente-Remontée pour la factorisation LU

```
// upward sweep : solves an upper triangular system
function  [y]=up_sweep_LU(A,x)
[m,n]=size(A);
if (m~=n) then
```

```
  print(%io(2), "error, not a square matrix");
else
  y=x;
  y(n)=y(n)/A(n,n);
  for i=n-1:-1:1,
    y(i)=(y(i)-A(i,i+1:n)*y(i+1:n))/A(i,i);
  end;
end;

// downward sweep : solves a lower triangular system with ones on the diagonal
function  [y]=down_sweep_LU(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=x;
  for i=2:n
    y(i)=y(i)-A(i,1:i-1)*y(1:i-1);
  end;
end;

//assumes A contains the LU-factorization
function [y]=SolveLU(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=down_sweep_LU(A,x);
  y=up_sweep_LU(A,y);
end;
```

5.2.2 Factorisation de Cholesky

Construction de la factorisation de Cholesky

```
//Cholesky factorization
//the matrix is stored in the lower part
function [A]= Cholesky_fact(A)
[m,n]=size(A);
for i=1:n,
  if (A(i,i)<=0) then
    print(%io(2)," non positive pivot")
  else
    A(i,i)=sqrt(A(i,i));
    A(i+1:m,i)=A(i+1:m,i)/A(i,i);
    for j=i+1:n,
      A(j:m,j)=A(j:m,j)-A(j:m,i)*A(j,i);
    end ;
  end ;
```

```
end ;
```

Descente-Remontée pour une factorisation de Cholesky

```
// up sweep : solves the upper triangular system
function  [y]=up_sweep_Cholesky(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=x;
  y(n)=y(n)/A(n,n);
  for i=n-1:-1:1,
    y(i)=(y(i)-(A(i+1:n,i))'*y(i+1:n))/A(i,i);
  end;
end;

// down sweep : solves the lower triangular system
function  [y]=down_sweep_Cholesky(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=x;
   y(1)=y(1)/A(1,1);
  for i=2:n
    y(i)=y(i)-A(i,1:i-1)*y(1:i-1);
    y(i)=y(i)/A(i,i)
  end;
end;

function [y]=SolveCholesky(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=down_sweep_Cholesky(A,x);
  y=up_sweep_Cholesky(A,y);
end;
```

5.3 Méthode de Gauss avec pivot partiel

La méthode de Gauss sans pivotage peut être bloquée si on tombe sur un pivot nul. Pour pallier cet inconvénient, on introduit une méthode qui permet d'éviter ce blocage en échangeant les lignes du système considéré. On obtient alors une méthode qui fonctionne pour toute matrice inversible : la méthode de Gauss avec pivot partiel.
On définit quelques notions utiles :

Définition 5.2 *On appelle permutation de* $\{1,\dots,n\}$ *une bijection de* $\{1,\dots,n\}$ *sur* $\{1,\dots,n\}$.

Définition 5.3 *Soit* σ *une permutation de* $\{1,\dots,n\}$, *on associe à* σ *une matrice* P *dite de permutation d'ordre* n *par*

$$P_{ij} = \delta_{\sigma(i)j},$$

c'est à dire

$$\begin{array}{rcll} P_{ij} & = & 1 & si \quad j = \sigma(i), \\ P_{ij} & = & 0 & si \quad j \neq \sigma(i), \end{array}$$

Lemme 5.1 *Soit* P *et* Q *deux matrices de permutation d'ordre* n, *alors* PQ *est une matrice de permutation d'ordre* n.

Définition 5.4 *Soient* $1 \leq k < l \leq n$, *on dit que la matrice de permutation* $P^{(kl)}$ *définie par*

$$\begin{array}{rcll} P^{(kl)}_{ij} & = & \delta_{ij} & si \quad i \neq k \ et \quad i \neq l \\ P^{(kl)}_{kj} & = & \delta_{lj} & \forall 1 \leq j \leq n \\ P^{(kl)}_{lj} & = & \delta_{kj} & \forall 1 \leq j \leq n \end{array}$$

est appelée matrice de permutation élémentaire.

Observation 5.2 *Le produit à gauche de* A *par* $P^{(kl)}$ *conduit à échanger les lignes* k *et* l *de* A, *en laissant les autres lignes inchangées, c'est à dire si*

$$B = P^{(kl)}A$$

$$\begin{array}{rcll} B_{ij} & = & A_{ij} & si \quad i \neq k \ et \quad i \neq l \\ B_{kj} & = & A_{lj} & \forall 1 \leq j \leq n \\ B_{lj} & = & A_{kj} & \forall 1 \leq j \leq n \end{array}$$

Observation 5.3 *Le déterminant d'une matrice de permutation élémentaire est* ± 1. *En effet,*

$$P^{(kl)}P^{(kl)} = I_n.$$

Lemme 5.2 *Toute matrice de permutation peut se décomposer en un produit de matrices élémentaires.*

Corollaire 5.1 *Le déterminant d'une matrice de permutation est* ± 1.

La méthode du pivot partiel est un algorithme dont l'application permet de prouver le théorème suivant. La méthode du pivot partiel a été présentée en cours sur un exemple et revue dans la preuve du théorème :

Théorème 5.4 *Soit* $A \in \mathcal{M}_n(\mathbb{C})$, *inversible. Alors il existe*
- *une matrice de permutation* P,
- *une matrice triangulaire inférieure* L, *avec des 1 sur la diagonale,*
- *une matrice triangulaire supérieure* U, *et inversible,*

telles que

$$PA = LU$$

Démonstration. Par récurrence sur les éliminations de Gauss : à la $k$ième étape, supposons que l'on ait prouvé l'égalité :

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

où $Q^{(k-1)}$ est une matrice de permutation et $\forall i < k$,

$$L^{(i)} = \overbrace{\begin{array}{c} 1 \quad \dots \quad i \qquad\qquad\qquad n \end{array}}$$
$$\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & * & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & * & 0 & \dots & 0 & 1 \end{pmatrix}$$

et

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \dots & \dots & \dots & \dots & & a_{1n}^{(k)} \\ 0 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & & & \vdots \\ \vdots & & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}$$

On effectue la $k$ième élimination de Gauss avec pivotage partiel : avant d'éliminer l'inconnue x_k, on échange les lignes k et $r_k > k$. On peut choisir $r_k \geq k$ tel que

$$|a_{r_k\, k}^{(k)}| = \max_{k \leq j \leq n} |a_{j\, k}^{(k)}|.$$

L'inversibilité de A assure que $a_{r_k\, k}^{(k)} \neq 0$. On peut donc s'en servir de pivot pour la prochaine élimination de Gauss.
Le pivotage puis l'élimination de Gauss s'écrivent matriciellement

$$P^{(kr_k)} A^{(k)} = L^{(k)} A^{(k+1)},$$

avec $P^{(kr_k)}$ matrice de permutation élémentaire entre les lignes k et r, $L^{(k)}$ triangulaire inférieure L, avec des 1 sur la diagonale, et

$$A^{(k+1)} = \begin{pmatrix} a_{11}^{(k+1)} & \dots & \dots & \dots & \dots & & a_{1n}^{(k+1)} \\ 0 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & & & \vdots \\ \vdots & & & 0 & a_{k+1k+1}^{(k+1)} & \dots & a_{k+1n}^{(k+1)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & a_{nk+1}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{pmatrix}.$$

Comme

$$(P^{(kr_k)})^{-1} = P^{(kr_k)},$$
$$A^{(k)} = P^{(kr_k)} L^{(k)} A^{(k+1)},$$

ce qui implique

$$Q^{(k-1)} A = L^{(1)} \dots L^{(k-1)} P^{(kr_k)} L^{(k)} A^{(k+1)}.$$

A ce point, on utilise le lemme :

Lemme 5.3 *Soit $P^{(kr_k)}$ la matrice de permutation élémentaire entre les lignes k et $r_k > k$, alors $\forall i < k$, si $L^{(i)}$ s'écrit*

$$L^{(i)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & C_i & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & & 0 & \dots & 0 & 1 \end{pmatrix} \quad et \quad C_i = \begin{pmatrix} \vdots \\ \alpha \\ \vdots \\ \beta \\ \vdots \end{pmatrix}$$

alors

$$L^{(i)} P^{(kr_k)} = P^{(kr_k)} L'^{(i)},$$

où

$$L'^{(i)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & C'_i & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & & 0 & \dots & 0 & 1 \end{pmatrix} \quad et \quad C'_i = \begin{pmatrix} \vdots \\ \beta \\ \vdots \\ \alpha \\ \vdots \end{pmatrix}.$$

Donc,

$$\begin{aligned} Q^{(k-1)}A &= L^{(1)}\dots L^{(k-2)}P^{(kr_k)}L'^{(k-1)}L^{(k)}A^{(k+1)} \\ &= P^{(kr_k)}L'^{(1)}\dots L'^{(k-1)}L^{(k)}A^{(k+1)}. \end{aligned}$$

Par suite,

$$P^{(kr_k)}Q^{(k-1)}A = L'^{(1)}\dots L'^{(k-1)}L^{(k)}A^{(k+1)},$$

où

- $Q^{(k)} = P^{(kr_k)}Q^{(k-1)}$ est une matrice de permutation,
- $L'^{(1)}\dots L'^{(k-1)}L^{(k)}$ est triangulaire inférieure avec des 1 sur la diagonale.

On a montré que la propriété est vraie pour $k+1$. ■

Remarque 5.8 *Dans la méthode de Gauss avec pivotage partiel, on ne doit pas stocker la matrice P mais seulement les r_k. Le coût mémoire et la complexité sont du même ordre que ceux de la factorisation LU. La résolution du système linéaire $Ax = b$ connaissant P L et U se fait de la manière suivante :*

$$\begin{aligned} y &= Pb \\ Lz &= y \\ Ux &= z \end{aligned}$$

Remarque 5.9 *La méthode de Gauss avec pivotage partiel ne conserve pas l'aspect bande : si A est une matrice bande dont la largeur de bande est n_b, les matrices L et U n'ont pas en général cete propriété. La complexité et le coût mémoire de la méthode de Gauss avec pivotage partiel deviennent alors largement supérieurs à ceux de la factorisation LU.*

Remarque 5.10 *La méthode du pivot total consiste à échanger non seulement les lignes mais aussi les colonnes de manière à choisir comme pivot le plus grand coefficient en module du bloc carré restant à factoriser. Cette méthode conduit à l'existence de deux matrices de permutations P et Q telles que $PAQ = LU$.*

Influence du pivotage sur la précision Le pivotage partiel ne sert pas seulement à garantir l'obtention d'une factorisation, il garantit aussi une bien meilleure stabilité que la factorisation LU, pour des matrices A mal conditionnées : prenons le système $Ax = b$ où

$$A = \begin{pmatrix} 10^{-9} & 1 \\ 1 & 1 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

dont la solution est $x = (\frac{1}{1-10^{-9}}, \frac{1-2\,10^{-9}}{1-10^{-9}})^T \sim (1,1)^T$. Supposons que la machine utilisée pour résoudre ce système ne garde que 8 chiffres significatifs : la factorisation LU calculée par la machine est

$$U = \begin{pmatrix} 10^{-9} & 1 \\ 0 & -10^{9} \end{pmatrix} \quad \text{et} \quad L = \begin{pmatrix} 1 & 0 \\ 10^{9} & 1 \end{pmatrix}$$

En utilisant cette factorisation, la machine retourne $x = (0,1)^T$!
Si on utilise le pivotage partiel, on obtient :

$$U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad L = \begin{pmatrix} 1 & 0 \\ 10^{-9} & 1 \end{pmatrix}$$

La machine retourne $x = (1,1)^T$.

5.4 Factorisations QR

5.4.1 Les réflexions de Householder

Définition 5.5 *Soit v un vecteur non nul de $\mathbb{C}^n$. On appelle réflexion de Householder ou matrice de Householder relative au vecteur v la matrice*

$$H_v = I - 2\frac{vv^*}{v^*v}. \tag{5.2}$$

Les réflexions de Householder ont les propriétés suivantes :

1. H_v est une matrice hermitienne.
2. H_v est une matrice unitaire.
3. $H_v - I$ est une matrice de rang un.
4. $H_{\lambda v} = H_v$, pour tout $\lambda \neq 0$.

Proposition 5.2 *Soit e un vecteur unitaire de $\mathbb{C}^n$ et x un vecteur non nul de $\mathbb{C}^n$. Il existe un vecteur non nul $v \in \mathbb{C}^n$ tel que $H_v x$ soit colinéaire à e.*

Démonstration. On cherche v sous la forme $v = x + \lambda e$. Dans le cas réel, ($x \in \mathbb{R}^n$ et $e \in \mathbb{R}^n$), les vecteurs $v = x \pm \|x\|_2 e$ (au moins l'un des deux est non nul) sont les seuls vecteurs de cette forme ayant la propriété demandée et on a $H_v x = \mp\|x\|_2 e$. Dans le cas général, on trouve aussi deux vecteurs v sous cette forme, et au moins l'un d'entre eux est non nul. ■

Remarque 5.11 *Si x est presque colinéaire à e, on a intérêt en pratique à choisir le vecteur v pour que v^*v, qui est au dénominateur de (5.2), ne soit pas petit. En effet, en précision finie, il faut éviter les divisions par des nombres petits. En particulier si $v \in \mathbb{R}^n$, on préférera le vecteur $v^+ = x + \text{signe}(x^*e)\|x\|_2 e$ au vecteur $v^- = x - \text{signe}(x^*e)\|x\|_2 e$, car $\|v^-\|_2$ est petit.*

5.4.2 Factorisations QR à l'aide des réflexions de Householder

Théorème 5.5 *Soit A une matrice de $\mathcal{M}_{m,n}(\mathbb{C})$ avec $m \geq n$. Il existe une matrice unitaire $Q \in \mathcal{M}_{m,m}(\mathbb{C})$et une matrice triangulaire supérieure $R \in \mathcal{M}_{m,n}(\mathbb{C})$, telles que*

$$A = QR.$$

Démonstration. On démontre ce résultat par récurrence : supposons qu'à l'aide de deux réflexions de Householder H_1 et H_2, on ait obtenu

$$H_2H_1A = \begin{pmatrix} * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & 0 & * & \dots & * \\ \vdots & \vdots & * & \dots & * \\ 0 & 0 & * & \dots & * \end{pmatrix}$$

On appelle x le vecteur de $\mathbb{C}^{m-2}$ obtenu en prenant les $m-2$ derniers coefficients de la troisième colonne de H_2H_1A. Si x est non nul, on sait trouver $v_3 \in \mathbb{C}^{m-2}$, tel que

$$H_{v_3}\, x \text{ soit colinéaire à } \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

On prend alors H_3

$$H_3 = \begin{pmatrix} I_2 & 0 \\ 0 & H_{v_3} \end{pmatrix}.$$

Si $x = 0$ on prend $H_3 = I$. Dans les deux cas H_3 est hermitienne et unitaire. On a

$$H_3 H_2 H_1 A = \begin{pmatrix} * & * & * & & \dots & * \\ 0 & * & * & & \dots & * \\ 0 & 0 & * & & \dots & * \\ 0 & 0 & 0 & * & \dots & * \\ \vdots & \vdots & \vdots & * & \dots & * \\ 0 & 0 & 0 & * & \dots & * \end{pmatrix}$$

Comme $m \geq n$, on peut itérer ce procédé et construire n matrices H_i hermitiennes et unitaires, telles que $H_n \dots H_1 A$ soit une matrice triangulaire supérieure R. On note Q la matrice unitaire (mais pas forcément hermitienne) $Q = H_1 \dots H_n$, on a

$$A = QR.$$

■

Les éléments fondamentaux de l'algorithme de la factorisation QR à l'aide de réflexions de Householder sont

1. Le choix des vecteurs v_i, de manière à ce que $\|v_i\|$ ne soit pas petit, cf. Remarque 5.11.
2. Le produit à gauche d'une matrice M par la matrice H_m : $H_m M$, doit être programmé intelligemment : on ne doit surtout pas stocker la matrice H_m. Il suffit de garder en mémoire le vecteur v_m, et d'opérer les réflexions de Householder à chaque colonne de M.
3. On ne calcule donc pas Q en général, mais on garde en mémoire les vecteurs $v_1, \dots, v_n$.
4. On peut écraser la matrice A que l'on factorise en stockant les vecteurs v_i dans la partie triangulaire inférieure stricte, et la matrice R dans la partie triangulaire supérieure à condition de normaliser les vecteurs v_i de manière à ce que leur premier coefficient soit 1 : ainsi on n'a pas besoin de stocker le premier coefficient des vecteurs v_i.
5. Si la matrice A est réelle, alors la matrice Q l'est aussi.

La factorisation QR a de nombreuses propriétés intéressantes :

1. Résoudre le système $Qy = b$ est facile car

   $$Q^{-1} = Q^* = H_n \dots H_1.$$

   La complexité de la résolution du système $Qy = b$ est donc $3\sum_i^n (m - i)$ adds+mults. Si $m = n$, la complexité est de l'ordre de $\frac{3}{2}n^2$. Si la matrice A est carrée d'ordre n inversible, pour résoudre $Ax = b$, on résout d'abord $Qy = b$ comme ci-dessus, puis $Rx = y$ par une remontée. La complexité totale est de l'ordre de $2n^2$.
2. Si $A \in \mathcal{M}_n(\mathbb{C})$ est inversible, alors

   $$\mathrm{cond}_2(A) = \mathrm{cond}_2(R),$$

   car Q est unitaire.

3. la méthode de Householder permet de calculer $|\det(A)|$. En effet,

$$|\det(R)| = |\text{produit des coefficients diagonaux de } R| = |\det(A)|.$$

4. Si $A \in \mathcal{M}_{m,n}(\mathbb{C})$, $m \geq n$, est factorisée sous la forme $A = QR$, alors résoudre une système de la forme $A^*Ax = b$ est facile si rang(A)= n, car $A^*A = R^*R$ et rang$(R) = n$, et on peut résoudre $A^*Ax = b$ par une descente(de R^*)-remontée (de R). La factorisation QR d'une matrice se prête donc bien aux problèmes de moindres carrés (voir §6).

Théorème 5.6 *Soit $A \in \mathcal{M}_m(\mathbb{C})$. On peut trouver une factorisation QR telle que tous les coefficients diagonaux de R sont réels positifs. Si A est inversible, cette factorisation est unique.*

Démonstration. On part de la factorisation QR de A obtenue par la méthode de Householder ci dessus : $A = Q'R'$ où la matrice Q' est unitaire et R' est triangulaire supérieure. Appelons D la matrice diagonale $D = diag(d_1, \dots d_m)$ où

$$d_i = \begin{cases} \frac{\overline{r_{ii}}}{|r_{ii}|} & \text{si } r_{ii} \neq 0, \\ 1 & \text{si } r_{ii} = 0, \end{cases}$$

La matrice D est clairement unitaire. Posons $Q = Q'D^{-1}$ et $R = DR'$. Ces matrices ont les propriétés désirées.
Supposons que A soit inversible et soient deux factorisation QR de A,

$$A = Q_1R_1 = Q_2R_2$$

telles que tous les coefficients diagonaux de R_1 et R_2 sont réels positifs. Ils sont strictement positifs car A est inversible. On a alors $Q_2^*Q_1 = R_2R_1^{-1}$. Mais $Q_2^*Q_1$ est unitaire tandis que $R_2R_1^{-1}$ est triangulaire supérieure avec des coefficients diagonaux réels positifs strictement. On peut vérifier que l'Identité est la seule matrice unitaire et triangulaire supérieure avec des coefficients diagonaux réels positifs. Donc $Q_2 = Q_1$ et $R_2 = R_1$.

■

5.4.3 Algorithme de factorisation QR avec des symétries de Householder

Recherche du Vecteur pour la Réflexion de Householder

```
// householder reflexion : find Householder vector
//normalized so that v(1)=1
function [v]= householder_vector(x)
delta=1;
if (x(1)<0) then
  delta=-1;
end;
v=x;
beta=v(1)+delta* sqrt(x'*x);
v=v/beta;
v(1)=1;
```

Réflexion de Householder

```
// performs Householder reflexion of vector v on a vector x
function [y]= householder_reflexion(v,x)
beta=2*(v'*x)/(v'*v);
y=x-beta*v;
```

Factorisation QR

```
//performs QR factorisation of a matrix
//the R part is stored in the upper part of A
//the Q part (normalized so that the upper line is one)
//is stored in the strict lower part of A
function[A]=QRfactorization(A)
[m,n]=size(A);
for i=1:n,
  v=householder_vector(A(i:m,i))
  delta=2/(v'*v);
  A(i,i)=A(i,i)-delta* (v'* A(i:m,i));
  for j=i+1:n,
    c=v'*A(i:m,j);
    A(i:m,j)=A(i:m,j)-c*delta*v;
  end ;
  if (i<m) then
    A(i+1:m,i)=v(2:m-i+1);
  end ;
end;
```

Résolution du système

```
function  [y]=up_sweep(A,x) //solves Ry=x
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
else
  y=x;
  y(n)=y(n)/A(n,n);
  for i=n-1:-1:1,
    y(i)=(y(i)-A(i,i+1:n)*y(i+1:n))/A(i,i);
  end;
end;



//assumes A contains a QR-factorization of A
function [y]=SolveQR(A,x)
[m,n]=size(A);
if (m~=n) then
  print(%io(2), "error, not a square matrix");
```

```
else
  y=x;
  for i=1:n-1,
    y(i:n)= householder_reflexion( [1;A(i+1:n,i)] , y(i:n));
  end ;
  y(n)=-y(n);
  y=up_sweep(A,y);
//  print(%io(2), y);
end;
```

5.4.4 Les rotations de Givens

Les réflexions de Householder sont très utiles quand il s'agit d'annuler tous les coefficients d'une colonne sous la diagonale. En revanche, pour annuler sélectivement certains coefficients (si la matrice a déjà beaucoup de coefficients nuls par exemple), les rotations de Givens constituent un outil important.
Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$: il existe deux réels $c = \cos(\theta)$ et $s = \sin(\theta)$ tels que si

$$Q = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} & & & & i & & & & j & & \end{array} \\ \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & & & & \vdots \\ \vdots & \ddots & 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ \vdots & & 0 & c & 0 & \dots & 0 & s & 0 & \dots & 0 \\ \vdots & & \vdots & 0 & 1 & 0 & \dots & 0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & 0 & \dots & 0 & 1 & 0 & \vdots & & \vdots \\ \vdots & & \vdots & -s & 0 & \dots & 0 & c & 0 & & \vdots \\ \vdots & & \vdots & 0 & \dots & \dots & \dots & 0 & 1 & \ddots & \vdots \\ \vdots & & \vdots & \vdots & & & & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0 & \dots & \dots & \dots & 0 & \dots & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ i \\ \\ \\ \\ \end{array} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \\ \\ j \\ \\ \\ \end{array}$$

alors les lignes de QA d'indices différents de i et j sont identiques à celles de A, et $(QA)_{j,i} = 0$. On définit $\theta \in]-\frac{\pi}{2}, \frac{\pi}{2}[$ tel $\tan(\theta) = t = \frac{a_{j,i}}{a_{i,i}}$ et $c = \cos(\theta) = \frac{1}{\sqrt{t^2+1}}$ et $s = \sin(\theta) = \frac{t}{\sqrt{t^2+1}}$. On a $-sa_{i,i} + ca_{j,i} = 0$.

5.5 Problèmes aux moindres carrés

On va travailler avec des systèmes réels mais toute la suite serait valable avec des systèmes dans $\mathbb{C}^n$, à condition de changer A^T en A^*.
On considère le problème de trouver $x \in \mathbb{R}^n$ tel que $Ax = b$ où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et où $b \in \mathbb{R}^m$, avec $m \geq n$. Il y a plus d'équations que d'inconnues, on dit que le système est sur-déterminé. Un système sur-déterminé n'a généralement pas de solution, sauf si $b \in \mathrm{Im}(A)$.
L'idée est alors de chercher à minimiser $\|Ax - b\|$ plutôt que de chercher à résoudre exactement le système. La solution du problème si elle existe dépend de la norme choisie. La méthode des

moindres carrés consiste à minimiser $\|Ax-b\|_2$ où $\|f\|_2^2 = \sum_{i=1}^m f_i^2$. On choisit cette norme car la fonction $J(x) = \|Ax-b\|_2^2$ est convexe et différentiable dans $\mathbb{R}^n$, et son gradient est $gradJ(x) = 2(A^TAx - A^Tb)$. On verra que minimiser J revient à résoudre le système linéaire $A^TAx = A^Tb$, qui a toujours une solution, et dont la solution est unique si $\text{rang}(A) = n$.

Lemme 5.4 *Soit $A \in \mathcal{M}_{m,n}(\mathbb{R})$ avec $m \geq n$: le problème de trouver $x \in \mathbb{R}^n$ minimisant la fonction $J : \mathbb{R}^n \to \mathbb{R}, \quad J(y) = \|Ay-b\|_2^2$ a au moins une solution. Le problème de trouver $x \in \mathbb{R}^n$ minimisant J est équivalent à trouver x tel que*

$$A^TAx - A^Tb = 0. \tag{5.3}$$

L'équation (5.3) est appelée équation normale.

Démonstration. Le problème de minimiser J revient à trouver $z \in \text{Im}(A)$ minimisant la distance $\|z-b\|_2$. On sait que $\text{Im}(A)^\perp = \ker(A^T)$, et donc que $\text{Im}(A) \oplus \ker(A^T) = \mathbb{R}^n$. Prendre pour z la projection de b sur $\text{Im}(A)$ parallèlement à $\ker(A^T)$ répond donc à la question, et après on prend x tel que $Ax = z$ (qui existe car $z \in \text{Im}(A)$ mais qui n'est pas unique si $\ker(A) \neq \{0\}$). On a donc trouvé x solution du problème au sens des moindres carrés.
La fonction J est différentiable sur $\mathbb{R}^n$ et son gradient vaut

$$grad(J)(x) = 2(A^TAx - A^Tb).$$

En effet, on vérifie facilement que

$$J(x+y) - J(x) = 2y^T(A^TAx - A^Tb) + \|Ay\|_2^2. \tag{5.4}$$

Si x minimise J, on a $grad(J)(x) = 0$ et x vérifie (5.3). Réciproquement, si x vérifie (5.3), on voit d'après (5.4) que $J(x+y) - J(x) = \|Ay\|_2^2 \geq 0$, $\forall y \in \mathbb{R}^n$, ce qui montre qye x minimise J. ■

Proposition 5.3 *Si $\text{rang}(A) = n$, le problème aux moindres carrés a une solution unique $x = (A^TA)^{-1}A^Tb$. Sinon, les solutions du problème aux moindres carrés forment un espace affine parallèle à $\ker(A)$.*

Démonstration. On a

$$y \in \ker(A^TA) \Rightarrow A^TAy = 0 \Rightarrow y^TA^TAy = 0 \Leftrightarrow \|Ay\|_2 = 0 \Leftrightarrow y \in \ker(A)$$

et $\ker(A) \subset \ker(A^TA)$. Donc $\ker(A) = \ker(A^TA)$.
Si $\text{rang}(A) = n$, $\ker(A) = \{0\}$ et l'équation normale (5.3) a une solution unique $x = (A^TA)^{-1}A^Tb$ et il en va de même pour le problème aux moindres carrés par le Lemme 5.4. Sinon, les solutions de (5.3) forment un espace affine. ■

On peut trouver une matrice orthogonale $Q \in \mathcal{M}_m(\mathbb{R})$ et une matrice triangulaire supérieure $R \in \mathcal{M}_{m,n}(\mathbb{R})$ telles que $A = QR$. Le système $A^TAx = A^Tb$ est équivalent à $R^TRx = R^TQ^Tb$, qui est facile à résoudre (une descente et une remontée), si $\text{rang}(R) = n$. De plus $\text{cond}_2(R^TR) = \text{cond}_2(A^TA)$.

Remarque 5.12 *Il n'est pas recommandé d'utiliser une méthode des moindres carrés pour résoudre un système carré car $\text{cond}_2(A^TA) = (\text{cond}_2(A))^2$ (exercice, le vérifier), et on a vu qu'il fallait éviter les grands nombres de conditionnement pour des raisons de précision.*

Exercice. Trouver la droite du plan la plus proche au sens des moindres carrés des points $(0,1)$ $(1,0)$ $(2,3)$.

Chapitre 6

Méthodes Itératives Stationnaires pour résoudre des systèmes linéaires

Pour des systèmes linéaires de grande taille, les méthodes directes (du type élimination de Gauss ou Cholesky), peuvent s'avérer trop coûteuses en temps de calcul ou en place mémoire. L'idée est alors de ne plus chercher à résoudre exactement le système linéaire mais d'approcher sa solution par une suite de vecteurs, construite à l'aide d'une formule de récurrence simple.

6.1 Principe et résultats généraux

6.1.1 Principe général

Soit un système linéaire

$$Ax = b, \tag{6.1}$$

où $A \in \mathcal{M}_d(\mathbb{C})$ est inversible, $x, b \in \mathbb{C}^d$. Le principe des méthodes itératives présentées ici est d'écrire A comme la différence de deux matrices :

$$A = M - N, \tag{6.2}$$

où

1. M est inversible
2. le système linéaire $My = c$ peut être résolu simplement, avec un coût de calcul faible : typiquement M sera diagonale ou triangulaire, mais on verra aussi le cas où M est diagonale ou triangulaire par blocs.

On va alors approcher la solution de (6.1) par la suite $(x^{(n)})$ définie par récurrence à partir de $x^{(0)}$ qu'on choisit, et de la formule

$$x^{(n+1)} = M^{-1}(b + Nx^{(n)}) \tag{6.3}$$

Remarque 6.1 *On n'a pas besoin de calculer M^{-1}, mais juste de savoir calculer la solution de $Mx^{(n+1)} = b + Nx^{(n)}$.*

Observation 6.1 *Si la suite converge vers y alors $y = x$. En effet, si la suite converge, on a à la limite, $My = b + Ny$ ou de manière équivalente $Ay = b$. Comme la solution de (6.1) est unique, $x = y$.*

Considérons l'erreur à l'étape n,

$$e^{(n)} = x - x^{(n)}.$$

On a

$$\left.\begin{array}{l} Mx^{(n+1)} = b + Nx^{(n)} \\ Mx = b + Nx \end{array}\right\} \Rightarrow e^{(n+1)} = M^{-1}Ne^{(n)} \Rightarrow e^{(n+1)} = (M^{-1}N)^{n+1}e^{(0)}$$

On appelle $M^{-1}N$ la matrice d'itération de la méthode. On a démontré le résultat

Proposition 6.1 *La suite donnée par $x^{(0)}$ et (6.3) converge vers x pour tout choix de $x^{(0)}$ si et seulement si la matrice d'itération vérifie*

$$\rho(M^{-1}N) < 1. \tag{6.4}$$

Démonstration. La suite donnée par $x^{(0)}$ et (6.3) converge vers x pour tout choix de $x^{(0)}$ si et seulement si $(M^{-1}N)^n e^{(0)} \to 0$ pour tout $e^{(0)}$, ce qui équivaut à dire que $(M^{-1}N)^n \to 0$. D'après le Théorème 4.1, ceci a lieu si et seulement si $\rho(M^{-1}N) < 1$. ■

6.1.2 Une condition suffisante dans le cas où A est hermitienne, définie positive

Théorème 6.1 *Soit A une matrice hermitienne, définie positive, et M, N deux matrices telles que $A = M - N$, M soit inversible et $M^* + N$ soit elle aussi hermitienne définie positive. Alors $\rho(M^{-1}N) < 1$.*

Démonstration. Comme A est symétrique définie positive, on peut considérer la norme vectorielle définie par

$$\|v\|_A^2 = v^*Av,$$

et la norme matricielle subordonnée encore notée $\|.\|_A$. On a pour tout vecteur v,

$$\begin{array}{rl} \|M^{-1}Nv\|_A^2 & = v^*(M-A)^*(M^*)^{-1}AM^{-1}(M-A)v \\ & = v^*Av + v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av \end{array}$$

Montrons que $v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av < 0$ si $v \neq 0$. En effet

$$\begin{array}{l} v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av \\ = v^*A^*(M^*)^{-1}AM^{-1}Av - v^*A^*(M^*)^{-1}MM^{-1}Av - v^*A^*(M^*)^{-1}M^*M^{-1}Av \\ = v^*A^*(M^*)^{-1}(A - M - M^*)M^{-1}Av \\ = -v^*A^*(M^*)^{-1}(M^* + N)M^{-1}Av \end{array}$$

qui est strictement négatif dès que $M^{-1}Av \neq 0 \to v \neq 0$, car $(M^* + N)$ est symétrique définie positive et A et M sont inversibles. Donc si $v \neq 0$, $\|M^{-1}Nv\|_A < \|v\|_A$. On en déduit que $\|M^{-1}N\|_A < 1$, ce qui achève la démonstration. ■

6.2 La méthode de Jacobi

On considère une matrice inversible A dont la diagonale D est inversible. La méthode de Jacobi consiste à choisir $M = D$ et $N = D - A$. La matrice d'itération $\mathcal{L}_J$ de la méthode de Jacobi s'écrit $\mathcal{L}_J = I - D^{-1}A$. On a les résultats suivants

Proposition 6.2 *Si A est à diagonale strictement dominante, i.e.*

$$\forall i, \quad |a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

alors $\rho(\mathcal{L}_J) < 1$, et la méthode de Jacobi converge pour tout choix de $x^{(0)}$.

Démonstration. Si A est à diagonale strictement dominante, on a $\|M^{-1}N\|_\infty = \|D^{-1}(D - A)\|_\infty < 1$. ■

Proposition 6.3 *Si A et $2D - A$ sont hermitiennes définies positives, alors $\rho(\mathcal{L}_J) < 1$ et la méthode de Jacobi converge pour tout choix de $x^{(0)}$.*

Démonstration. Si A est hermitienne définie positive, alors D l'est aussi, et on peut utiliser la méthode de Jacobi. De plus, $M + N = 2D - A$ est aussi hermitienne définie positive. On peut appliquer le Théorème 6.1. ■

Algorithme La ième coordonnée de $x^{(n+1)}$ est donnée par

$$x_i^{(n+1)} = \frac{b_i - \sum_{j \neq i} a_{i\,j} x_j^{(n)}}{a_{i\,i}}.$$

Voici la boucle de la méthode de Jacobi : le test d'arrêt est ici du type $\|x^{(n+1)} - x^{(n)}\| \leq \epsilon$, mais d'autres tests sont évidemment possibles.

```
while( err>eps)
{
  w=x;
  x=b;
  for(int i=0; i<x.size();i++)
    {
      for(int j=0; j<i;j++)
        x(i)-=a(i,j)*w(j);
      for(int j=i+1; j<x.size();j++)
        x(i)-=a(i,j)*w(j);
      x(i)=x(i)/a(i,i);
    }
  e=w-x;
  err=norm(e);
}
```

Remarque 6.2 *Remarquons que l'algorithme ci-dessus peut être aisément parallélisé : si on dispose d'une machine parallèle (comportant P processeurs ($P > 1$)), on peut facilement répartir le calcul sur les P processeurs : le k ème processeur mettant à jour les coordonnées d'indices*

$\frac{d(k-1)}{P}, \ldots, \frac{dk}{P} - 1$, *et les processeurs travaillant en même temps. Si la mémoire est de plus distribuée, il faut ensuite que les processeurs communiquent leur données actualisées aux autres processeurs : cette étape de transmission de l'information prend évidemment du temps, mais ce temps est souvent petit par rapport au temps de calcul.*

Quand la matrice admet une décomposition par bloc :

$$A = \begin{pmatrix} A_{11} & \ldots & A_{1P} \\ \vdots & & \vdots \\ A_{i1} & \ldots & A_{iP} \\ \vdots & & \vdots \\ A_{P1} & \ldots & A_{PP} \end{pmatrix}$$

où les blocs diagonaux A_{ii} sont des matrices carrées (les blocs non diagonaux ne doivent pas nécessairement l'être), et si les blocs diagonaux sont tous inversibles, une méthode dite de Jacobi par blocs consiste à prendre $D = \text{Block} - \text{Diag}(A_{11}, \ldots, A_{PP})$. Elle nécessite de savoir résoudre les systèmes avec les blocs A_{ii} avec une complexité algorithmique raisonnable :
Exemple Si on considère la discrétisation du problème de Poisson dans un carré unité

$$\begin{aligned} -\Delta u &= f \quad \text{dans } \Omega \\ u &= 0 \quad \text{sur } \partial\Omega \end{aligned}$$

par différences finies sur une grille uniforme de pas $h = \frac{1}{N+1}$, on obtient après numérotation lexicographique des inconnues le système linéaire

$$Ax = b$$

où A est une matrice de $\mathcal{M}_{N^2}(\mathbb{R})$ tridiagonale par blocs, les blocs étant tous des matrices de $\mathcal{M}_N(\mathbb{R})$, et

$$A_{ii} = D = (N+1)^2 \begin{pmatrix} 4 & -1 & 0 & \ldots & \ldots & \ldots & 0 \\ -1 & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & -1 \\ 0 & \ldots & \ldots & \ldots & 0 & -1 & 4 \end{pmatrix} \qquad A_{i,i+1} = A_{i+1,i} = -(N+1)^2 I_N \tag{6.5}$$

Il est facile de calculer la factorisation de Cholesky des blocs diagonaux de A $A_{ii} = LL^T$, car on sait que L n'a que deux diagonales non nulles. On peut donc résoudre les systèmes avec les blocs A_{ii} avec une complexité arithmétique de l'ordre de $2N$ pour chaque blocs ; la méthode de Jacobi par blocs s'écrit, en désignant par x_i (respectivement w_i) le vecteur de $\mathbb{R}^N$ contenant les coordonnées d'indices $(i-1)N, \ldots, iN-1$ de x (resp.w).

```
while( err>eps)
{
  w=x;
```

```
  x=b;
  for(int i=0; i<N-1;i++)
      x(i)+=(N+1)^2 * w(i+1);
  for(int i=1; i<N;i++)
      x(i)+=(N+1)^2 * w(i-1);
  for(int i=0; i<N;i++)
        solve (D,x(i)); //r{\'e}sout le syst{\`e}me D y=x(i)
                        // et range la solution dans x(i)
  e=w-x;
  err=norm(e);
}
```

6.3 La méthode de Gauss-Seidel

On considère une matrice inversible A dont la diagonale D est inversible. On note $A = D-E-F$, où $-E$ (respectivement $-F$) est la partie triangulaire inférieure strictement (respectivement supérieure) de A. La méthode de Gauss-Seidel consiste à choisir $M = D - E$ et $N = F$. La matrice d'itération $\mathcal{L}_{GS}$ de la méthode de Gauss-Seidel s'écrit $\mathcal{L}_{GS} = I - (D-E)^{-1}A$. On a les résultats suivants :

Proposition 6.4 *Si A est à diagonale strictement dominante, alors $\rho(\mathcal{L}_{GS}) < 1$ et la méthode de Gauss-Seidel converge pour tout choix de $x^{(0)}$.*

Démonstration. On s'intéresse à la solution de $My = Nx$ pour x donné : on a

$$a_{i\,i}y_i = \sum_{j<i} a_{i\,j}x_j + \sum_{j>i} a_{i\,j}y_j.$$

On considère i_0 tel que $|y_{i_0}| = \|y\|_\infty$: on a

$$|a_{i_0\,i_0}||y_{i_0}| \leq \sum_{j<i_0} |a_{i_0\,j}|\|y\|_\infty + \sum_{j>i_0} |a_{i_0\,j}|\|x\|_\infty.$$

Si A est à diagonale strictement dominante,

$$|a_{i_0\,i_0}| - \sum_{j<i_0} |a_{i_0\,j}| > \sum_{j>i_0} |a_{i_0\,j}|,$$

donc

$$\sup_{\substack{x\neq 0\\ My=Nx}} \frac{\|y\|_\infty}{\|x\|_\infty} < 1,$$

ce qui veut dire que $\|M^{-1}N\|_\infty < 1$. ■

Proposition 6.5 *Si A est hermitienne définie positive, alors $\rho(\mathcal{L}_{GS}) < 1$ et la méthode de Gauss-Seidel converge pour tout choix de $x^{(0)}$.*

Démonstration. Si A est hermitienne définie positive, alors D l'est aussi, et $E = F^*$.Donc $M^* + N = D - E^* + F = D$ est hermitienne définie positive. On peut appliquer le Théorème

6.1. ■

Algorithme La ième coordonnée de $x^{(n+1)}$ est donnée par

$$x_i^{(n+1)} = \frac{b_i - \sum_{j<i} a_{ij} x_j^{(n+1)} - \sum_{j>i} a_{ij} x_j^{(n)}}{a_{ii}}$$

Dans la méthode de Gauss-Seidel, dès que la ième coordonnée de $x^{(n+1)}$ est calculée, la ième coordonnée de $x^{(n)}$ devient inutile : on peut écraser la ième coordonnée de $x^{(n)}$ et la remplacer par la ième coordonnée de $x^{(n+1)}$ dès que celle ci est calculée.
Voici la boucle de la méthode de Gauss Seidel : le test d'arrêt est encore du type $\|x^{(n+1)} - x^{(n)}\| \leq \epsilon$.

```
while( err<eps)
{
  e=x;
  for(int i=0; i<x.size();i++)
    {
      x(i)=b(i);
      for(int j=0; j<i;j++)
        x(i)-=a(i,j)*x(j);
      for(int j=i+1; j<x.size();j++)
        x(i)-=a(i,j)*x(j);
      x(i)=x(i)/a(i,i);
    }
  e=e-x;
  err=norm(e);
}
```

Remarque 6.3 *La méthode de Gauss-Seidel nécessite de calculer $x_k^{(n+1)}$ avant de calculer $x_{k+1}^{(n+1)}$. Pour cette raison, il est beaucoup moins facile de paralléliser la méthode de Gauss-Seidel que la méthode de Jacobi.*

Remarque 6.4 *Comme pour la méthode de Jacobi, on peut généraliser la méthode de Gauss-Seidel à une matrice par blocs, dont les blocs diagonaux sont tous carrés et inversibles.*

Exercice. Proposer un programme pour résoudre le système de l'exemple ci dessus par une méthode de Gauss Seidel par blocs.

6.4 Méthodes SOR (successive over relaxation)

La méthode de Gauss-Seidel est très facile à programmer mais sa convergence peut être très lente pour certains systèmes : on la modifie en introduisant un paramètre $\omega \neq 0$ dit paramètre de relaxation et en choisissant $M = \frac{1}{\omega}D - E$ et $N = F + (1 - \frac{1}{\omega})D$. La matrice M est inversible si la diagonale D est inversible. Pour $\omega = 1$, on retrouve la méthode de Gauss-Seidel. Pour $\omega < 1$ on parle de sous-relaxation. Pour $\omega > 1$ on parle de sur-relaxation. Un calcul facile montre que la matrice d'itération de cette méthode est

$$\mathcal{L}_\omega = (I - \omega D^{-1}E)^{-1}((1-\omega)I + \omega D^{-1}F). \tag{6.6}$$

Proposition 6.6 *Si A est à diagonale strictement dominante, la méthode de relaxation avec paramètre ω converge pour tout $x^{(0)}$ si*

$$0 < \omega \leq 1. \tag{6.7}$$

Démonstration. On prend $0 < \omega \leq 1$, et on s'intéresse à la solution de $My = Nx$ pour x donné : on a

$$a_{i\,i}y_i = \omega \sum_{j<i} a_{i\,j}y_j + \omega \sum_{j>i} a_{i\,j}x_j + (1-\omega)a_{i\,i}x_i$$

On considère i_0 tel que $|y_{i_0}| = \|y\|_\infty$: on a

$$|a_{i_0\,i_0}||y_{i_0}| \leq \omega(\sum_{j<i_0} |a_{i_0\,j}|\|y\|_\infty + \sum_{j>i_0} |a_{i_0\,j}|\|x\|_\infty) + (1-\omega)|a_{i_0\,i_0}||x_{i_0}|.$$

Si A est à diagonale strictement dominante,

$$|a_{i_0\,i_0}| - \sum_{j<i_0} |a_{i_0\,j}| > \sum_{j>i_0} |a_{i_0\,j}|,$$

donc

$$\left((1-\omega)|a_{i_0\,i_0}| + \omega(|a_{i_0\,i_0}| - \sum_{j<i_0} |a_{i_0\,j}|)\right)|y_{i_0}| \leq \left((1-\omega)|a_{i_0\,i_0}| + \omega \sum_{j>i_0} |a_{i_0\,j}|\right)\|x\|_\infty$$

ce qui implique

$$\sup_{\substack{x \neq 0 \\ My = Nx}} \frac{\|y\|_\infty}{\|x\|_\infty} < 1,$$

ce qui veut dire que $\|M^{-1}N\|_\infty < 1$. ■

Proposition 6.7 *Si A est hermitienne définie positive, la méthode de relaxation avec paramètre ω converge pour tout x_0 si*

$$0 < \omega < 2. \tag{6.8}$$

Démonstration. Si A est hermitienne définie positive, alors D l'est aussi, et $E = F^*$. Donc $M^* + N = (\frac{2}{\omega} - 1)D - E^* + F = (\frac{2}{\omega} - 1)D$ est hermitienne définie positive dès que $0 < \omega < 2$. On peut appliquer le Théorème 6.1. ■

Remarque 6.5 *On peut aussi relaxer la méthode de Jacobi en prenant $M = \frac{1}{\omega}D$ et $N = \frac{1}{\omega}D - A$. Si A est hermitienne définie positive, sous quelles conditions sur ω $M^* + N$ est elle définie positive ?*

Théorème 6.2 *Si $0 \leq \omega$ ou si $\omega \geq 2$, La méthode SOR ne converge pas vers la solution x pour tout choix initial $x^{(0)}$. Si les inégalités sont strictes, on peut trouver des $x^{(0)}$ pour lesquels $\lim_{n\to\infty} \|x^{(n)}\| = +\infty$.*

Démonstration. D'après (6.6), le déterminant de $\mathcal{L}_\omega$ est $(1-\omega)^d$. Mais

$$|\rho(\mathcal{L}_\omega)| < 1 \Rightarrow |\det(\mathcal{L}_\omega)| < 1$$

car le déterminant est le produit des valeurs propres. Donc,

$$\omega \geq 2 \text{ ou } \omega \leq 0 \Rightarrow |(1-\omega)^d| \geq 1 \Rightarrow \rho(\mathcal{L}_\omega)| \geq 1$$

implique que la méthode SOR ne converge pas vers la solution pour tout choix initial. ■

Remarque 6.6 *La proposition 6.7 donne donc en fait une condition nécessaire et suffisante pour que la méthode SOR converge pour tout choix de* $x^{(0)}$.

Algorithme
Voici la boucle de la méthode SOR

```
while( err<eps)
{
  e=x;
  for(int i=0; i<x.size();i++)
    {
      x(i)=b(i)-(1-1/omega)*a(i,i)*x(i);
      for(int j=0; j<i;j++)
        x(i)-=a(i,j)*x(j);
      for(int j=i+1; j<x.size();j++)
        x(i)-=a(i,j)*x(j);
      x(i)=omega*x(i)/a(i,i);
    }
  e=e-x;
  err=norm(e);
}
```

6.5 Comparaisons des méthodes pour des matrices tridiagonales

Théorème 6.3 *Si A est tridiagonale, on a*

$$\rho(\mathcal{L}_{GS}) = \rho(\mathcal{L}_J)^2 \tag{6.9}$$

Démonstration. Soit A une matrice de $\mathcal{M}_d(\mathbb{C})$ tridiagonale :

$$A = \begin{pmatrix} a_1 & b_1 & 0 & \dots & \dots & 0 \\ c_1 & a_2 & b_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & c_{d-2} & a_{d-1} & b_{d-1} \\ 0 & \dots & \dots & 0 & c_{d-1} & a_d \end{pmatrix} \tag{6.10}$$

Le nombre complexe λ est valeur propre de $\mathcal{L}_J$ si et seulement si $\det(D^{-1}(D-A)-\lambda I)=0$ ou encore si $\det((1-\lambda)D-A)=0$, c'est à dire

$$\det\begin{pmatrix} \lambda a_1 & b_1 & 0 & \dots & \dots & 0 \\ c_1 & \lambda a_2 & b_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & c_{d-2} & \lambda a_{d-1} & b_{d-1} \\ 0 & \dots & \dots & 0 & c_{d-1} & \lambda a_d \end{pmatrix} = 0 \tag{6.11}$$

D'autre part μ est valeur propre de $\mathcal{L}_{GS}$ si et seulement si $\det((D-E)^{-1}(D-E-A)-\mu I)=0$ ou encore si $\det((1-\mu)(D-E)-A)=0$, c'est à dire

$$\det\begin{pmatrix} \mu a_1 & b_1 & 0 & \dots & \dots & 0 \\ \mu c_1 & \mu a_2 & b_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \mu c_{d-2} & \mu a_{d-1} & b_{d-1} \\ 0 & \dots & \dots & 0 & \mu c_{d-1} & \mu a_d \end{pmatrix} = 0 \tag{6.12}$$

Supposons $\lambda \neq 0$, (6.11) est équivalente à

$$\det\left[\mathrm{Diag}(\lambda,\lambda^2,\dots,\lambda^d)\begin{pmatrix} \lambda a_1 & b_1 & 0 & \dots & \dots & 0 \\ c_1 & \lambda a_2 & b_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & c_{d-2} & \lambda a_{d-1} & b_{d-1} \\ 0 & \dots & \dots & 0 & c_{d-1} & \lambda a_d \end{pmatrix}\mathrm{Diag}(\lambda^{-1},\lambda^{-2},\dots,\lambda^{-d})\right] = 0 \tag{6.13}$$

Ce produit de trois matrices vaut :

$$\lambda^{-1}\begin{pmatrix} \lambda^2 a_1 & b_1 & 0 & \dots & \dots & 0 \\ \lambda^2 c_1 & \lambda^2 a_2 & b_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda^2 c_{d-2} & \lambda^2 a_{d-1} & b_{d-1} \\ 0 & \dots & \dots & 0 & \lambda^2 c_{d-1} & \lambda^2 a_d \end{pmatrix}$$

ce qui veut dire d'après (6.12) que λ^2 est valeur propre de $\mathcal{E}_{GS}$. Donc si $\lambda \neq 0$, λ est valeur propre de $\mathcal{E}_J$ si et seulement si λ^2 est valeur propre de $\mathcal{E}_{GS}$, ce qui montre (6.9). ■

On a enfin un théorème plus précis dans le cas où A est de plus définie positive :

Théorème 6.4 *Si A est tridiagonale et hermitienne définie positive, les méthodes de Jacobi et de Gauss-Seidel convergent, et la méthode SOR converge si et seulement si $0 < \omega < 2$. Le paramètre optimal est*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathcal{L}_J)}} > 1, \tag{6.14}$$

et

$$\rho(\mathcal{L}_{\omega_{\text{opt}}}) = \frac{1}{\omega_{\text{opt}}}. \tag{6.15}$$

Démonstration. voir [2]. ■

6.6 Autres méthodes

Faute de temps, nous ne traitons pas les méthodes
- de Gauss-Seidel symétrisé
- des directions alternées

qui rentrent pourtant exactement dans le cadre décrit ci-dessous. Elles sont décrites dans [5], [3].
Les méthodes semi itératives de Chebyshev [3] permettent d'accélérer la convergence des méthodes ci-dessus. Nous ne les traitons pas non plus.
Les méthodes itératives que nous venons de décrire sont l'un des ingrédients des méthodes multigrilles, qui sont des méthodes souvent optimales pour résoudre certains systèmes linéaires issus de la simulation par éléments finis ou différences finies d'équations aux dérivées partielles, voir [1].

Chapitre 7

Méthodes de Descente pour des Systèmes Linéaires

7.1 Principe des méthodes de descente

7.1.1 Minimisation de fonctions quadratiques

Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice symétrique définie positive. Le but est de construire des méthodes itératives pour résoudre un système linéaire

$$Ax = b, \tag{7.1}$$

en lui associant un problème de minimisation équivalent et en construisant une suite minimisante : on va travailler avec des matrices réelles symétriques et définies positives, mais on pourrait tout généraliser au cas de matrices hermitiennes définies positives. Considérons la forme quadratique F sur $\mathbb{R}^d$ définie par

$$F(x) = \frac{1}{2}x^T Ax - x^T b. \tag{7.2}$$

La fonction F est continue et $\lim_{\|x\|\to\infty} F(x) = +\infty$. La fonction admet donc un minimum dans $\mathbb{R}^d$. La fonction F est de plus différentiable, et son gradient vaut

$$\forall y \in R^n, \quad DF(y) = Ay - b. \tag{7.3}$$

Exercice Démontrer l'assertion précédente.

Comme la solution de (7.1) est unique car A est inversible, le gradient de F ne s'annule qu'en un seul point qui réalise le minimum de F.

Exercice Montrer que F est strictement convexe, c'est à dire que

$$F(\alpha x + (1-\alpha)y) - \alpha F(x) - (1-\alpha)F(y) \leq -\frac{1}{2}\alpha(1-\alpha)\lambda_{min}(A)$$

où $\lambda_{min}(A)$ est la plus petite valeur propre de A.

On voit donc que le problème (7.1) est équivalent à la minimisation de F. L'idée va donc de construire des suites minimisantes de F pour approcher la solution de (7.1).

7.1.2 Méthodes de descente

Une méthode de descente consiste à construire une suite minimisante sous la forme

$$x_{n+1} = x_n + \alpha_n p_n, \tag{7.4}$$

où $p_n \in \mathbb{R}^d$, $p_n \neq 0$ et où le scalaire α_n est choisi pour que $F(x_{n+1}) < F(x_n)$. La convergence de la méthode dépend bien sûr des choix des p_n et α_n.

Définition 7.1 *L'erreur à l'étape n est le vecteur*

$$e_n = x - x_n. \tag{7.5}$$

On appelle résidu à l'étape n le vecteur

$$r_n = b - Ax_n = Ae_n. \tag{7.6}$$

Remarque 7.1 *Le résidu à la nième étape vérifie $r_n = -DF(x_n)$.*

7.1.3 Choix optimal de α_n pour p_n fixé

Supposons choisie la direction p_n : on peut choisir α_n de manière à minimiser la fonction ϕ de $\mathbb{R}^+$ dans $\mathbb{R}$ donnée par

$$\phi(t) = F(x_n + tp_n).$$

Cette fonction admet un minimum unique pour

$$\alpha_{opt} = \frac{r_n^T p_n}{p_n^T A p_n}, \tag{7.7}$$

et on a

$$x_{n+1} = x_n + \frac{r_n^T p_n}{p_n^T A p_n} p_n$$

Proposition 7.1 *Pour tout p_n, et si on choisit $\alpha_n = \alpha_{opt}$, on a*

$$r_{n+1}^T p_n = 0. \tag{7.8}$$

7.2 Méthodes de gradient

7.2.1 Principe des méthodes de gradient

L'idée des méthodes de gradient va être de choisir comme direction de descente le gradient de F en x_n, ou (voir Remarque 7.1), de manière équivalente, le résidu r_n :

$$\begin{aligned} x_{n+1} &= x_n - \alpha_n DF(x_n) \\ &= x_n + \alpha_n r_n. \end{aligned} \tag{7.9}$$

Pour tout $x \in \mathbb{R}^d$, il existe un nombre réel $\alpha_{\max}(x) > 0$, tel que

$$\rho \in]0, \quad \alpha_{\max}(x)[\Leftrightarrow F(x - \rho DF(x)) < F(x).$$

Exercice Démontrer cette assertion.
Le pas α_n doit donc être un réel positif choisi tel que $F(x_{n+1}) < F(x_n)$.

Remarque 7.2 *On peut généraliser ces méthodes à des fonctions strictement convexes et différentiables.*

7.2.2 Interprétation géométrique en dimension deux

Soit A une matrice d'ordre deux symétrique et définie positive. On sait que A est diagonalisable dans une base orthonormale. Quitte à changer les coordonnées on peut supposer que A est diagonale et que $b = 0$:

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \tag{7.10}$$

La fonction F s'écrit alors $F(x) = \lambda_1 x_1^2 + \lambda_2 x_2^2$, et les lignes de niveaux de F : $F(x) = r^2$ sont les ellipses concentriques

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = r^2. \tag{7.11}$$

Le gradient de F au point x est orthogonal à l'ellipse d'équation (7.11) passant par x. La Figure 7.1 donne un exemple des premiers itérés d'une méthode de gradient.

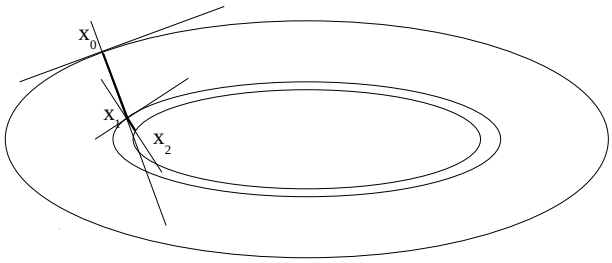

FIG. 7.1 – deux itérés d'une méthode de gradient

7.2.3 Méthodes du gradient à pas fixe

Si A est défini positif, on peut obtenir une méthode convergente en fixant α_n à une valeur bien choisie : la méthode du gradient à pas fixe consiste à construire la suite récurrente

$$x_{n+1} = x_n - \alpha(Ax_n - b).$$

L'erreur vérifie

$$e_{n+1} = (I - \alpha A)e_n = (I - \alpha A)^{n+1} e_0,$$

ce qui montre que la méthode du gradient à pas fixe converge si et seulement si

$$\rho(I - \alpha A) < 1$$

et on note $\tau(\alpha) = \rho(I - \alpha A)$ le rayon spectral de $I - \alpha A$. On appelle $\tau(\alpha)$ le taux de convergence de la méthode. On a donc

Théorème 7.1 *Si A est symétrique définie positive, la méthode du gradient à pas fixe converge vers la solution x de (7.1) si et seulement si*

$$\alpha < \frac{2}{\lambda_{\max}(A)}. \tag{7.12}$$

De plus la valeur de α minimisant le taux de convergence $\tau(\alpha)$ est

$$\alpha_{opt} = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}.$$

et le taux de convergence vaut alors

$$\tau_{opt} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}$$

Démonstration. La matrice d'itérations $(I - \alpha A)$ est diagonalisable et ses valeurs propres sont $1 - \alpha\lambda$ où λ est valeur propre de A. La condition nécessaire et suffisante s'obtient facilement en écrivant la condition $\rho(I - \alpha A) < 1$.
On a pour toute valeur propre λ de A,

$$1 - \alpha\lambda_{\max} \leq 1 - \alpha\lambda_{\max} \leq 1 - \alpha\lambda_{\min}$$

Donc $\rho(I - \alpha A) = \max(|1 - \alpha\lambda_{\max}(A)|, |1 - \alpha\lambda_{\min}(A)|)$. En tracant le graphe des deux fonctions

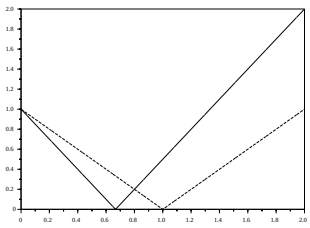

FIG. 7.2 – graphes des fonctions $\alpha \to |1 - \alpha\lambda_{\max}(A)|$ et $\alpha \to |1 - \alpha\lambda_{\min}(A)|$

$\alpha \to |1 - \alpha\lambda_{\max}(A)|$ et $\alpha \to |1 - \alpha\lambda_{\min}(A)|$, (voir Figure 7.2), on voit que

$$\rho(I - \alpha A) = \begin{cases} 1 - \alpha\lambda_{\min}(A) & \text{pour } \alpha \leq \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}, \\ \alpha\lambda_{\max}(A) - 1 & \text{pour } \alpha \geq \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}, \end{cases}$$

On voit aussi que le minimum de $\rho(I - \alpha A)$ est atteint pour $\alpha_{opt} = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$ et qu'il vaut

$$\tau_{opt} = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}$$

■

Remarque 7.3 *On voit donc que la méthode du gradient à pas fixe converge d'autant plus lentement que le conditionnement de A est grand. En effet, avec le choix optimal pour ρ, il faut de l'ordre de*

$$\frac{|\log(\epsilon)|}{|\log(\frac{\mathrm{cond}_2(A)-1}{\mathrm{cond}_2(A)+1})|}$$

itérations pour réduire l'erreur d'un facteur ε ; quand $\mathrm{cond}_2(A) \gg 1$*, le nombre d' itérations est donc de l'ordre de*

$$|\log(\epsilon)| \frac{\mathrm{cond}_2(A)}{2}.$$

7.2.4 Méthode du gradient à pas optimal

Comme dans § 7.1.3, on peut construire

$$x_{n+1} = x_n - \alpha_n(Ax_n - b),$$

en choisissant α_n de manière à minimiser la fonction ϕ de $\mathbb{R}^+$ dans $\mathbb{R}$ donnée par

$$\phi(t) = F(x_n - t(Ax_n - b)).$$

Cette fonction à un minimum unique pour

$$\alpha_{opt} = \frac{\|Ax_n - b\|_2^2}{\|Ax_n - b\|_A^2} = \frac{\|r_n\|_2^2}{\|r_n\|_A^2}, \tag{7.13}$$

où

$$\|y\|_A^2 = y^T Ay.$$

Il est important de noter que dans la méthode du gradient à pas optimal, d'après (7.8), les résidus successifs sont orthogonaux :

$$r_{n+1}^T r_n = 0. \tag{7.14}$$

Théorème 7.2 *Si A est symétrique et définie positive, la méthode du gradient à pas optimal converge vers la solution x de (7.1).*

Démonstration. La suite des $J(x_n)$ est décroissante par construction, et bornée inférieurement par $J(x)$, donc converge. On en déduit que

$$J(x_{n+1}) - J(x_n) \to 0,$$

ce qui implique que

$$-\alpha_n\|r_n\|_2^2 + \frac{1}{2}\alpha_n^2 r_n^T Ar_n \to 0,$$

et comme $\alpha_n = \frac{\|r_n\|_2^2}{r_n^T Ar_n}$, on trouve finalement que

$$\frac{\|r_n\|_2^4}{r_n^T Ar_n} \to 0.$$

Comme A est définie positive, ceci implique que $r_n \to 0$. Toujours grâce au caractère défini positif de A, on en déduit que $e_n \to 0$. ■

Remarque 7.4 *Cette démonstration se généralise à la méthode du gradient à pas optimal appliqué à la minimisation d'une fonction F fortement convexe et de gradient Lipchitzien.*

Pour déterminer la vitesse de convergence de la méthode de gradient à pas optimal, on utilise l'inégalité de Kantorovitch

Lemme 7.1 *Soient d réels strictement positifs,*

$$0 < \ell_1 < \cdots < \ell_i < \cdots < \ell_d$$

et d réels positifs β_i tels que $\sum_1^d \beta_i = 1$. On note $\ell = \sum_1^d \beta_i \ell_i$. On a

$$\sum_1^d \frac{\beta_i}{\ell_i} \leq \frac{\ell_1 + \ell_d - \ell}{\ell_1 \ell_d}, \tag{7.15}$$

et

$$\frac{1}{\ell \sum_1^d \frac{\beta_i}{\ell_i}} \geq \frac{4\ell_1 \ell_d}{(\ell_1 + \ell_d)^2}. \tag{7.16}$$

Démonstration. Pour prouver (7.15), on doit montrer que

$$\sum_1^d \beta_i (\frac{1}{\ell_i} + \frac{\ell_i}{\ell_1 \ell_d}) \leq \frac{\ell_1 + \ell_d}{\ell_1 \ell_d}$$

Pour cela, on voit que la fonction qui à $x \in [\ell_1, \ell_d]$ associe $\frac{1}{x} + \frac{x}{\ell_1 \ell_d}$ atteint son maximum en $x = \ell_1$ et en $x = \ell_d$ et le maximum vaut $\frac{1}{\ell_1} + \frac{1}{\ell_d}$. On conclut en utilisant le fait que $\sum_{i=1}^d \beta_i = 1$. Après, (7.16) s'obtient facilement en cherchant le minimum de $\ell \frac{\ell_1 + \ell_d - \ell}{\ell_1 \ell_d}$ sur l'intervalle $[\ell_1, \ell_d]$. ■

Proposition 7.2 *(Kantorovitch) Soit $A \in \mathcal{M}_d(\mathbb{R})$ symétrique définie positive dont les valeurs propres vérifient $0 < \lambda_{\min} = \lambda_1 < \cdots < \lambda_i < \cdots < \lambda_d = \lambda_{\max}$. On a*

$$\inf_{y \neq 0} \frac{\|y\|_2^4}{(y^T A y)\ (y^T A^{-1} y)} = \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}.$$

Démonstration. On note $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ les valeurs propres de A et $(v_i)_{1 \leq i \leq d}$ une base orthonormale de vecteurs propres : $Av_i = \lambda_i v_i$. Pour $y \in \mathbb{R}^d$, on note $\beta_i = \frac{(y^T v_i)^2}{\|y\|_2^2}$:

$$\sum_{i=1}^d \beta_i = 1.$$

On a aussi

$$\frac{(Ay, y)}{\|y\|_2^2} = \sum_{i=1}^d \beta_i \lambda_i \quad \text{et} \quad \frac{(A^{-1}y, y)}{\|y\|_2^2} = \sum_{i=1}^d \frac{1}{\lambda_i} \beta_i.$$

et on applique le lemme précédent, et on obtient

$$\inf_{y \neq 0} \frac{\|y\|_2^4}{(y^T A y)\ (y^T A^{-1} y)} \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}.$$

Enfin, cet infimum est atteint par $y = v_1 + v_d$. ■

Théorème 7.3 *Pour la méthode du gradient à pas optimal, on a l'estimation*

$$\|e_n\|_A \leq \left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1} \right)^n \|e_0\|_A$$

Démonstration. On a $\alpha_n = \frac{\|r_n\|_2^2}{r_n^T A r_n}$, $e_{n+1} = e_n - \frac{\|r_n\|_2^2}{r_n^T A r_n} r_n$, et $r_{n+1} = r_n - \frac{\|r_n\|_2^2}{r_n^T A r_n} A r_n$. Donc

$$\begin{aligned}
\|e_{n+1}\|_A^2 = e_{n+1}^T A e_{n+1} &= e_{n+1}^T r_{n+1} \\
&= e_n^T r_n - \frac{\|r_n\|_2^4}{r_n^T A r_n} \\
&= (1 - \frac{\|r_n\|_2^4}{(r_n^T A r_n)(r_n^T A^{-1} r_n)}) e_n^T A e_n \\
&\leq \frac{(\lambda_{\max} - \lambda_{\min})^2}{(\lambda_{\max} + \lambda_{\min})^2} e_n^T A e_n \\
&= \left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1} \right)^2 \|e_n\|_A^2
\end{aligned}$$

■

Interprétation géométrique en dimension deux

On reprend la matrice A donnée par (7.10). D'après (7.14), on peut construire graphiquement la suite des itérés, car le point x_{n+1} est à la fois sur la droite de direction r_n passant par x_n, et sur l'ellipse d'équation (7.11) tangentée par cette droite. La Figure 7.3 donne un exemple des premiers itérés d'une méthode de gradient à pas optimal. Dire que la matrice A est mal

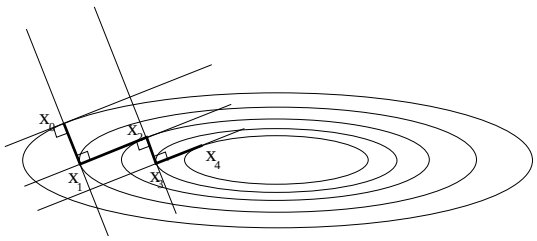

FIG. 7.3 – méthode de gradient à pas optimal

conditionnée, c'est dire que les lignes de niveaux de F sont des ellipses très allongées, ou encore à fort rapport d'aspect. Dans ce cas, on voit que la suite des x_n se rapproche de sa limite x en zigzaguant beaucoup, et on comprend que la convergence est lente.

7.3 La méthode du Gradient Conjugué

Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice symétrique définie positive. Pour $v \in \mathbb{R}^d$, on note $\|v\|_2$ la norme euclidienne de v : $\|v\|_2^2 = v^T v$ et $\|v\|_A$ la norme définie par

$$\|v\|_A^2 = v^T A v.$$

On veut résoudre le système

$$Au = b \tag{7.17}$$

par une méthode de descente à pas optimal.
On a vu précédemment que la méthode du gradient à pas optimal conduit à des trajectoires très oscillantes lorsque le conditionnement de la matrice est grand. Afin d'éviter ces phénomènes d'oscillations, l'idée est de chercher une méthode à pas optimal où la direction de descente n'est plus le gradient mais est construite en fonction des directions précédentes.
On rappelle que résoudre (7.17) est équivalent à minimiser la fonction $J(v) = \frac{1}{2}v^T Av, v - v^T b$, et que le gradient de J en v est donné par

$$\text{grad}J(v) = Av - b$$

On construit une suite minimisante pour J de la façon suivante :
L'initialisation de la méthode se fait en choisissant u_0 et en prenant $p_{-1} = 0$.
Connaissant u_k et p_{k-1}, $k > 0$, (on note le résidu $r_k = b - Au_k = -\text{grad}J(u_k)$), on choisit en même temps la nouvelle direction de descente p_k et le pas $\alpha_k > 0$. La direction de descente p_k est cherchée sous la forme

$$p_k = -\text{grad}J(u_k) + \beta_k p_{k-1} = r_k + \beta_k p_{k-1} \tag{7.18}$$

On cherche β_k et α_k de manière à minimiser la fonction

$$f(\beta_k, \alpha_k) = J(u_k + \alpha_k p_k) = J(u_k + \alpha_k(-\text{grad}J(u_k) + \beta_k p_{k-1})) = J(u_k + \alpha_k(r_k + \beta_k p_{k-1})). \tag{7.19}$$

Soulignons qu'il s'agit d'un problème d'optimisation à deux variables β_k et α_k.
On prend alors

$$u_{k+1} = u_k + \alpha_k p_k.$$

Cette méthode qui s'appelle l'algorithme du gradient conjugué, a été inventée en 1952 par Hestenes et Stieffel.

On écrit les conditions d'optimalité pour ce problème d'optimisation

$$\frac{\partial f}{\partial \alpha_k} = 0 \Leftrightarrow p_k^T \text{grad}J(u_k + \alpha_k p_k) = 0 \tag{7.20}$$

$$\frac{\partial f}{\partial \beta_k} = 0 \Leftrightarrow p_{k-1}^T \text{grad}J(u_k + \alpha_k p_k) = 0, \tag{7.21}$$

où p_k est donnée par (7.18). L'équation (7.20) est équivalente à $p_k)^T(b - Au_k - \alpha_k Ap_k) = 0$ ou encore à $p_k)^T(r_k - \alpha_k Ap_k) = 0$ et on a donc

$$\alpha_k = \frac{p_k^T r_k}{p_k^T Ap_k}. \tag{7.22}$$

qui est la formule déjà trouvée pour toute méthode de descente à pas optimal. Comme le pas est optimal, on a la relation d'orthogonalité entre la direction de descente p_{k-1} et le résidu r_k :

$$p_{k-1}^T r_k = 0, \quad k \geq 0. \tag{7.23}$$

On déduit alors immédiatement de (7.18) et de (7.23) que p_k donné par (7.18) n'est pas nul si $r_k \neq 0$.
De la relation (7.18), on tire que α_k s'écrit aussi

$$\alpha_k = \frac{r_k^T r_k}{p_k^T Ap_k} = \frac{\|r_k\|_2^2}{\|p_k\|_A^2}. \tag{7.24}$$

La condition d'optimalité vérifiée par β_k s'écrit

$$p_{k-1}^T(r_k - \alpha_k A p_k) = 0 \tag{7.25}$$

ce qui montre que

$$p_{k-1}^T r_k - \alpha_k p_{k-1}^T A(r_k + \beta_k p_{k-1}) = 0 \tag{7.26}$$

On déduit de (7.26),(7.23)

$$\beta_k = -\frac{p_{k-1}^T A r_k}{p_{k-1}^T A p_{k-1}} \tag{7.27}$$

ce qui implique que

$$p_k = r_k - \frac{p_{k-1}^T A r_k}{p_{k-1}^T A p_{k-1}} p_{k-1}, \tag{7.28}$$

et que

$$p_{k-1}^T A p_k = 0. \tag{7.29}$$

Il reste à montrer que les valeurs de α_k et β_k correspondent à un minimum local de f : pour cela, on doit montrer que $D^2 f(\alpha_k, \beta_k)$ est symétrique définie positive : on a

$$D^2 f(\alpha_k, \beta_k) = \begin{pmatrix} 2p_k^T A p_k & 0 \\ 0 & 2\alpha_k^2 p_{k-1}^T A p_{k-1} \end{pmatrix},$$

qui est semi-définie positive, et définie positive si $p_{k-1} \neq 0$,$p_k \neq 0$ et $r_k \neq 0$.
Les formules (7.24) et (7.27) suffisent à définir complètement cet algorithme, mais il s'avère qu'on peut prouver un grand nombre de propriétés qui expliquent pourquoi cet algorithme converge si bien, en particulier des propriétés de **conjugaison** entre les directions de descente :

Théorème 7.4 *L'algorithme défini ci-dessus a les propriétés suivantes*

$$p_l^T A p_k = 0, \quad \forall l, k,\ l \neq k \tag{7.30}$$

On dit que les directions de descentes sont deux à deux conjuguées.

$$r_l^T r_k = 0, \quad \forall l, k,\ l \neq k \tag{7.31}$$

$$r_l^T p_k = 0, \quad \forall l, k,\ k < l \tag{7.32}$$

$$p_k \in Vect(r_0, \ldots, r_k) \tag{7.33}$$

Le sous-espace $Vect(r_0, \ldots, r_k)$ est appelé $k$ème espace de Krylov.

Démonstration. La propriété (7.33) est une conséquence immédiate de (7.18) et de $p_{-1} = 0$ (se démontre par récurrence).
La relation de conjugaison

$$p_{k-1}^T A p_k = 0, \quad k \geq 0. \tag{7.34}$$

est une conséquence directe de (7.28). On déduit alors que $p_k^T A p_k = r_k^T A p_k$ et on tire de (7.24) que

$$\alpha_k = \frac{r_k^T r_k}{r_k^T A p_k}$$

Comme $r_{k+1} = r_k - \alpha_k A p_k$, on voit que

$$r_k^T r_{k+1} = 0. \tag{7.35}$$

On va maintenant démontrer par récurrence les relations (7.30), (7.31), (7.32). On sait que $p_0^T r_1 = 0$ cf. (7.23), et on vient de démontrer que $p_0^T A p_1 = 0$ et que $r_0^T r_1 = 0$.
supposons que la propriété de récurrence est vraie jusqu'à $l-1$. i.e. $\forall k_1, k_2,\ 0 \leq k_1 < k_2 < l$,

$$\begin{aligned} p_{k_1}^T r_{k_2} &= 0, \\ r_{k_1}^T r_{k_2} &= 0, \\ p_{k_1}^T A p_{k_2} &= 0. \end{aligned} \tag{7.36}$$

1. On tire de (7.23) que $p_{l-1}^T r_l = 0$. Il reste à montrer que $p_k^T r_l = 0$, pour $k < l-1$. Mais $r_l = r_{l-1} - \alpha_{l-1} A p_{l-1}$ et (7.36) implique que $p_k^T r_{l-1} = p_k^T A p_{l-1} = 0$. On a donc bien que $p_k^T r_l = 0$, pour $k < l-1$.
2. On tire de (7.35) que $r_{l-1}^T r_l = 0$. Il reste à montrer que $r_k^T r_l = 0$, pour $k < l-1$. Mais $r_k = p_k - \beta_k p_{k-1}$: $r_k^T r_l = (p_k - \beta_k p_{k-1})^T r_l = 0$ d'après le point 1.
3. On tire de (7.34) que $p_{l-1}^T A p_l = 0$. Il reste à montrer que $p_k^T A p_l = 0$, pour $k < l-1$. On utilise alors que $p_l = r_l + \beta_l p_{l-1}$, donc $p_k^T A p_l = p_k^T A r_l + \beta_l p_k^T A p_{l-1} = p_k^T A r_l$, d'après (7.36). Mais $p_k^T A r_l = r_l^T A p_k$ et $p_k \in Vect(r_0, \ldots, r_k)$ d'après (7.33), et $A p_k \in Vect(r_0, \ldots, r_k, r_{k+1}) \subset Vect(r_0, \ldots, r_{l-1})$. Du point 2, on déduit le résultat désiré.

On a montré (7.30), (7.31), (7.32) par récurrence. ■

Corollaire 7.1 *L'algorithme du gradient conjugué dans $\mathbb{R}^d$ converge en au plus d itérations.*

Démonstration. Les résidus sont deux à deux orthogonaux. Si pour tout $k < d$, $r_k \neq 0$, $(r_0, \ldots, r_d)$ est une famille orthogonale de vecteurs de $\mathbb{R}^d$ avec $d+1$ vecteurs : ceci ne peut avoir lieu que si $r_d = 0$. ■

Remarque 7.5 *On vient de voir que la méthode du gradient conjuguée qui est une méthode itérative peut aussi être vue comme une méthode directe, puisque qu'elle donne le résultat exact (en arithmétique exacte) en au plus n itérations. Le gradient conjugué, implémenté sur un ordinateur, ne peut pas être considéré comme une méthode directe, car l'arithmétique n'est pas exacte, et les directions de descentes ne sont pas exactement conjuguées.*

Corollaire 7.2 *L'itéré u_{k+1} minimise la fonction J sur l'espace de Krylov $Vect(r_0, \ldots, r_k)$.*

Démonstration. En effet, (7.33) implique que $u_{k+1} \in Vect(r_0, \ldots, r_k)$ (par récurrence) pour $l \leq k$, $r_l^T \mathrm{grad} J(u_{k+1}) = r_l^T r_{k+1} = 0$. ■

Remarque 7.6 *L'algorithme du gradient conjugué appartient donc à la classe des méthodes dites de Krylov, où l'on minimise à chaque itération une norme du résidu dans l'espace de Krylov. Pour en savoir plus sur les méthodes de Krylov, on pourra lire Y. Saad [8].*

Remarque 7.7 *Pour en savoir plus sur la méthode du gradient conjugué, on pourra lire [3].*

L'algorithme du gradient conjugué s'écrit

$$\begin{aligned}
\alpha_k &= \frac{p_k^T r_k}{p_k^T A p_k} \\
x_{k+1} &= x_k + \alpha_k p_k \\
r_{k+1} &= r_k - \alpha_k A p_k \\
\beta_{k+1} &= -\frac{r_{k+1}^T A p_k}{p_k^T A p_k} \\
p_{k+1} &= r_{k+1} + \beta_{k+1} p_k
\end{aligned}$$

L'algorithme du gradient conjugué a d'autres propriétés très intéressantes qui en font la meilleure méthode itérative pour des systèmes où A est symétrique et définie positive. On peut démontrer en particulier le résultat, voir [4, 5].

Théorème 7.5 *Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice symétrique définie positive. On a*

$$\|e_k\|_A \leq 2 \left(\frac{\sqrt{\mathrm{cond}_2(A)} - 1}{\sqrt{\mathrm{cond}_2(A)} + 1} \right)^k \|e_0\|_A \tag{7.37}$$

Démonstration. On voit que dans la méthode du gradient conjugué, l'erreur à l'étape k est de la forme

$$e^{(k)} = (I + \sum_{\ell=1}^{k} \gamma_\ell A^k) e^{(0)}$$

où les $(\gamma_\ell)_{1 \leq \ell \leq k}$ sont choisis pour minimiser $\|e^{(k)}\|_A$. On peut donc écrire l'erreur

$$e^{(k)} = P^{(k)}(A) e^{(0)},$$

où $P^{(k)}$ est le polynôme de degré k réalisant le minimum de la fonctionnelle

$$Q \mapsto \|Q(A) e^{(0)}\|_A$$

sur l'ensemble des polynômes de degré k prenant la valeur 1 en 0.
Soit t_k le k ième le polynôme de Chebyshev de première espèce :

$$\begin{aligned}
&t_k(x) = \cos(k \arccos(x)), \quad x \in [-1, 1] \\
&t_k(x) = \cosh(k \arg\cosh(x)) = \frac{1}{2} \left((x + \sqrt{x^2 - 1})^k + (x + \sqrt{x^2 - 1})^{-k} \right), \quad x > 1 \\
&t_k(x) = t_k(-x) \quad x < 1.
\end{aligned}$$

On peut montrer par récurrence que t_k est un polynôme de degré k et que son coefficient directeur est 2^{k-1}
Soient $\lambda_{\min}$ la plus petite valeur propre de A et $\lambda_{\max}$ la plus grande valeur propre de A. On prend

$$Q(x) = \frac{t_k(\frac{2x - \lambda_{\min} - \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}})}{t_k(\frac{-\lambda_{\min} - \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}})}.$$

On voit que $Q(0) = 1$ et que

$$\max_{x \in [\lambda_{\min}, \lambda_{\max}]} |Q(x)| = \frac{1}{t_k(\frac{-\lambda_{\min} - \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}})} = \frac{1}{t_k(\frac{\mathrm{cond}_2(A) + 1}{\mathrm{cond}_2(A) - 1})}$$

Un calcul facile montre que

$$t_k(\frac{\mathrm{cond}_2(A)+1}{\mathrm{cond}_2(A)-1}) = \frac{1}{2}\left(\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^k + \left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^{-k}\right) \geq \frac{1}{2}\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^{-k}$$

On déduit que

$$\|e^{(k)}\|_A = \|P^{(k)}(A)e^{(0)}\|_A \leq \|Q(A)e^{(0)}\|_A \leq \max_{\lambda\in\sigma(A)}|Q(\lambda)|\|e^{(0)}\|_A \leq 2\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^k \|e^{(0)}\|_A,$$

ce qui est bien le résultat désiré. ■

Si on a des informations supplémentaires sur le spectre de A, on peut obtenir une meilleure estimation : supposons par exemple que les valeurs propres de A soient

$$\lambda_1 \ll \lambda_2 \leq \cdots \leq \ldots \lambda_d.$$

On a donc

$$\frac{\lambda_d}{\lambda_2} \ll \mathrm{cond}_2(A).$$

On introduit alors le polynôme Q de degré k, prenant la valeur 1 en 0 :

$$Q(x) = \frac{t_{k-1}(\frac{2x-\lambda_2-\lambda_d}{\lambda_d-\lambda_2})}{t_{k-1}(\frac{-\lambda_2-\lambda_d}{\lambda_d-\lambda_2})}\frac{\lambda_1-x}{\lambda_1}.$$

On voit que

$$\max_{\lambda\in\sigma(A)}|Q(\lambda)| = \max_{\lambda\in\{\lambda_2,\ldots,\lambda_k\}}|Q(\lambda)| \leq \max_{\lambda\in[\lambda_2,\Lambda_d]}\left|\frac{t_{k-1}(\frac{2\lambda-\lambda_2-\lambda_d}{\lambda_d-\lambda_2})}{t_{k-1}(\frac{-\lambda_2-\lambda_d}{\lambda_d-\lambda_2})}\right|\frac{\lambda_d-\lambda_1}{\lambda_1} \leq 2\left(\frac{\sqrt{\frac{\lambda_d}{\lambda_2}}-1}{\sqrt{\frac{\lambda_d}{\lambda_2}}+1}\right)^{k-1} (\mathrm{cond}_2(A)-1)$$

qui est asymptotiquement beaucoup plus petit que $2\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^k$.

De même, si le spectre de A a ses ℓ plus petites valeurs propres isolées dur reste du spectre, pour $\ell \ll d$:

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_\ell \ll \lambda_{\ell+1}\cdots \leq \ldots \lambda_d.$$

$$\frac{\lambda_d}{\lambda_{\ell+1}} \ll \mathrm{cond}_2(A).$$

On suppose que $\ell \ll d$ et pour $k \geq \ell$, on introduit le polynôme Q de degré k, prenant la valeur 1 en 0 :

$$Q(\lambda) = \prod_{i=1}^{\ell}\frac{\lambda_i-\lambda}{\lambda_i}\frac{t_{k-\ell}(\frac{2x-\lambda_d-\lambda_{\ell+1}}{\lambda_d-\lambda_{\ell+1}})}{t}{}_{k-\ell}(\frac{-\lambda_d-\lambda_{\ell+1}}{\lambda_d-\lambda_{\ell+1}})$$

On voit que

$$\max_{\lambda\in\sigma(A)}|Q(\lambda)| \leq 2\left(\frac{\sqrt{\frac{\lambda_d}{\lambda_{\ell+1}}}-1}{\sqrt{\frac{\lambda_d}{\lambda_{\ell+1}}}+1}\right)^{k-\ell}\prod_{i=1}^{\ell}\frac{\lambda_d-\lambda_i}{\lambda_i} \leq 2\left(\frac{\sqrt{\frac{\lambda_d}{\lambda_{\ell+1}}}-1}{\sqrt{\frac{\lambda_d}{\lambda_{\ell+1}}}+1}\right)^{k-\ell}(\mathrm{cond}_2(A))^\ell$$

qui est asymptotiquement beaucoup plus petit que $2\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^k$.

Chapitre 8

Recherche de valeurs propres

Les notes ci-dessous s'inspirent de deux livres sur le calcul matriciel : le livre de P.G. Ciarlet [2] est une introduction à la fois très abordable et très complète. Le livre de G. Golub et al [3] est la référence dans le domaine de l'analyse numérique matricielle.

Trouver les valeurs propres d'une matrice carrée d'ordre n en factorisant son polynôme caractéristique devient impossible en général quand $n > 4$. Il faut donc utiliser des méthodes itératives. Il existe deux types de méthodes de recherche de valeurs propres :

- Les méthodes partielles qui visent à rechercher la plus grande ou la plus petite valeur propre ou encore la valeur propre la plus proche d'une valeur donnée. Ces méthodes marchent si la valeur propre recherchée est séparée des autres valeurs propres en module. Ces méthodes permettent de trouver le vecteur propre correspondant.
- Les méthodes globales qui donnent le spectre entier mais qui en général ne permettent pas d'obtenir directement les vecteurs propres.

8.1 Généralités

8.1.1 Décomposition de Schur

voir § 2.10.

8.1.2 Sensibilité d'un problème aux valeurs propres

On considère la matrice de $\mathcal{M}_n(\mathbb{R})$

$$\begin{pmatrix} 0 & & \dots & & \epsilon \\ 1 & 0 & \dots & & 0 \\ 0 & \ddots & \ddots & & \\ \vdots & & & & \\ 0 & \dots & & 1 & 0 \end{pmatrix}$$

Son polynôme caractéristique est $x^n \pm \epsilon$: les valeurs propres sont nulles si $\epsilon = 0$ et ce sont les racines n ièmes de $\pm\epsilon$ si $\epsilon \neq 0$. Si $n = 10$ et si $\epsilon = 10^{-10}$, les valeurs propres ont pour module 0.1. On voit donc qu'une perturbation de 10^{-10} d'un coefficient de la matrice entraîne des variations importantes des valeurs propres.
Pour tenter de comprendre la sensibilité des valeurs propres aux perturbations d'une matrice, on a le résultat

Théorème 8.1 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ diagonalisable, et P une matrice de changement de base telle que $P^{-1}AP = \mathrm{Diag}(\lambda_1, \dots \lambda_n)$. Soit $|||$ une norme sur $\mathcal{M}_n(\mathbb{C})$ telle que*

$$\|MN\| \leq \|M\|\|N\| \quad et \ \|\mathrm{Diag}(\mu_1 \dots \mu_n)\| = \max_{i=1\dots n} |\mu_i|$$

alors pour toute perturbation δA de A, on a

$$sp(A + \delta A) \subset \cup_{i=1}^{n} D_i \tag{8.1}$$

et

$$D_i = \{\mu \in \mathbb{C};\ |\mu - \lambda_i| \leq \mathrm{cond}_{\|\ \|}(P)\|\delta A\|\}. \tag{8.2}$$

Démonstration.

On sait que μ est valeur propre de $A + \delta A$ si et seulement si $A + \delta A - \mu I$ n'est pas inversible. En multipliant à droite par P et à gauche par P^{-1}, ceci est équivalent à dire que $\mathrm{Diag}(\lambda_1 - \mu, \dots, \lambda_n - \mu) + P^{-1}\delta AP$ est singulière. Deux cas se présentent :

1. si $\mathrm{Diag}(\lambda_1 - \mu, \dots, \lambda_n - \mu)$ est singulière, alors $\mu \in D_i$ pour un indice $i \in \{1, \dots, n\}$.
2. sinon, $I + \mathrm{Diag}((\lambda_1 - \mu)^{-1}, \dots, (\lambda_n - \mu)^{-1})P^{-1}\delta AP$ est singulière. Ceci implique que $\|\mathrm{Diag}((\lambda_1 - \mu)^{-1}, \dots, (\lambda_n - \mu)^{-1})P^{-1}\delta AP\| \geq 1$, autrement on pourrait construire un inverse avec une série. Les propriétés de la norme impliquent alors que $\|\mathrm{Diag}((\lambda_1 - \mu)^{-1}, \dots, (\lambda_n - \mu)^{-1})\|\|P^{-1}\|\|\delta A\|\|P\| \geq 1$, soit encore $\max_i |\lambda_i - \mu|^{-1}\mathrm{cond}_{\|\ \|}(P)\|\delta A\| \geq 1$. On en déduit que
$$\min_i |\lambda_i - \mu| \leq \mathrm{cond}_{\|\ \|}(P)\|\delta A\|$$

■

Remarque 8.1 *La sensibilité d'un problème aux valeurs propres d'une matrice A diagonalisable par rapport aux perturbations dépend donc du conditionnement de la matrice de passage P et non pas de celui de A.*

Remarque 8.2 *Si A est normale, on sait que A est diagonalisable avec P unitaire. On a $\mathrm{cond}_2(P) = 1$. Dans ce cas les valeurs propres de δA sont contenues dans des cercles centrés aux valeurs propres de A et de rayon $\|\delta A\|_2$.*

Remarque 8.3 *On a un résultat plus précis si A et δA sont hermitiennes. Dans ce cas, cf. [2], le théorème du min-max permet de dire que si on ordonne les valeurs propres par valeur croissante, les k ièmes valeurs propres de A et $A + \delta A$ sont distantes d'au plus $\|\delta A\|_2$.*

8.1.3 Valeurs singulières d'une matrice

Voir §2.11.

8.1.4 Norme de Frobenius

Voir §4.1.3

8.2 Méthodes partielles de recherche de valeurs propres

8.2.1 La méthode de la puissance

Soit $A \in \mathcal{M}_n(\mathbb{C})$ diagonalisable. On note $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A et $x_1, \ldots x_n$ des vecteurs propres unitaires associés. Dans la suite on va supposer que $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$.
On veut construire une méthode pour approcher numériquement λ_1. Pour cela, on se donne un vecteur $q^{(0)}$ tel que $\|q^{(0)}\|_2 = 1$ et on construit la suite $\lambda^{(k-1)}, q^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned} z^{(k)} &= Aq^{(k-1)}, \\ q^{(k)} &= \frac{z^{(k)}}{\|z^{(k)}\|_2}, \\ \lambda^{(k-1)} &= q^{(k-1)^*} z^{(k)}, \end{aligned}$$

Cette méthode itérative s'appelle méthode de la puissance, car $q^{(k)}$ et $z^{(k)}$ sont proportionnels à $A^k q^{(0)}$. On a le résultat de convergence

Théorème 8.2 (convergence de la méthode de la puissance) *Soit $A \in \mathcal{M}_n(\mathbb{C})$ diagonalisable, $\lambda_1, \ldots, \lambda_n$ ses valeurs propres, et $x_1, \ldots x_n$ des vecteurs propres unitaires associés. Si*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \tag{8.3}$$

et si $q^{(0)} = a_1 x_1 + \ldots a_n x_n$ avec $a_1 \neq 0$, alors $\lambda^{(k)}$ converge vers λ_1 et il existe $C \geq 0$ indépendant de k et pour tout k, $\epsilon^{(k)} \in \mathbb{C}$, $|\epsilon^{(k)}| = 1$ tels que

$$\begin{aligned} |\lambda^{(k)} - \lambda_1| &\leq C \left(\frac{|\lambda_2|}{|\lambda_1|} \right)^k, \\ \|\epsilon^{(k)} q^{(k)} - x_1\| &\leq C \left(\frac{|\lambda_2|}{|\lambda_1|} \right)^k, \end{aligned} \tag{8.4}$$

Démonstration. On vérifie facilement que

$$A^k q^{(0)} = \lambda_1^k \left(a_1 x_1 + a_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k x_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1} \right)^k x_n \right).$$

Comme $q^{(k)}$ est colinéaire à $A^k q^{(0)}$

$$q^{(k)} = \alpha^{(k)} \left(a_1 x_1 + a_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k x_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1} \right)^k x_n \right) \tag{8.5}$$

et en utilisant (8.3) et $\|q^{(k)}\|_2 = 1$, on voit que

$$\left| |\alpha^{(k)}| - \frac{1}{|a_1|} \right| \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k. \tag{8.6}$$

Comme $\lambda^{(k)} = q^{(k)^*} A q^{(k)}$, on voit facilement en utilisant (8.5) et (8.6) que

$$|\lambda^{(k+1)} - \lambda_1| \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k.$$

Le résultat est démontré. ■

Remarque 8.4 *Un des intérêts de cette méthode est qu'elle nécessite juste de savoir effectuer le produit matrice-vecteur par A.*

8.2.2 Test d'arrêt pour la méthode de la puissance

Un problème qui se pose est de choisir le test-d'arrêt de la méthode : pour cela, on introduit le résidu à l'étape k (qui tend vers 0 d'après le Théorème 1) :

$$r^{(k)} = Aq^{(k)} - \lambda^{(k)} q^{(k)}, \tag{8.7}$$

et la matrice $E^{(k)}$ de rang un définie par $\|r^{(k)}\|_2 E^{(k)} = -r^{(k)} q^{(k)^*}$. On a clairement $\|E^{(k)}\|_2 = 1$ et $E^{(k)} q^{(k)} = -\frac{r^{(k)}}{\|r^{(k)}\|_2}$ car $q^{(k)}$ est un vecteur unitaire. Donc d'après (8.7), $(A + \|r^{(k)}\|_2 E^{(k)}) q^{(k)} = \lambda^{(k)} q^{(k)}$, ce qui montre que $q^{(k)}$ est un vecteur propre de la matrice perturbée $A + \|r^{(k)}\|_2 E^{(k)}$, de valeur propre $\lambda^{(k)}$. On utilise alors le résultat suivant

Théorème 8.3 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice diagonalisable et λ une valeur propre simple de A. Il existe donc deux vecteurs unitaires x et y tels que $Ax = \lambda x$ et $A^* y = \lambda^* y$. Pour $\epsilon > 0$, on définit $A(\epsilon) = A + \epsilon E$ où $E \in \mathcal{M}_n(\mathbb{C})$ est une matrice telle que $\|E\|_2 = 1$. Il existe une fonction de classe C^1 : $\epsilon \mapsto (\lambda(\epsilon), x(\epsilon))$ d'un voisinage de 0 dans $\mathbb{C} \times \mathbb{C}^n$ telle que $\lambda(0) = \lambda$, $x(0) = x$, $A(\epsilon)x(\epsilon) = \lambda(\epsilon)x(\epsilon)$ et*

$$\left| \frac{\partial \lambda}{\partial \epsilon}(0) \right| \leq \frac{1}{|y^* x|}. \tag{8.8}$$

Avec ce théorème, on voit que $|\lambda^{(k)} - \lambda_1| \leq \frac{1}{|y_1^* x_1|} \|r^{(k)}\|_2$, ce qui donnerait un test d'arrêt si on connaissait x_1 et y_1.

On ne connaît pas x_1 et y_1, mais on sait que $q^{(k)}$ converge vers un vecteur qui est colinéaire à x_1. En modifiant légèrement la méthode de la puissance, on peut approcher le produit scalaire $y_1^* x_1$: on obtient le nouvel algorithme : On se donne deux vecteurs $z^{(0)}$ et $w^{(0)}$, et on construit la suite $\lambda^{(k)}, q^{(k)}, p^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned}
q^{(k-1)} &= \frac{z^{(k-1)}}{\|z^{(k-1)}\|_2}, & p^{(k-1)} &= \frac{w^{(k-1)}}{\|w^{(k-1)}\|_2}, \\
z^{(k)} &= Aq^{(k-1)}, & w^{(k)} &= A^* p^{(k-1)}, \\
\lambda^{(k-1)} &= q^{(k-1)^*} z^{(k)}, & & \\
e^{(k)} &= \frac{\|z^{(k)} - \lambda^{(k-1)} q^{(k-1)}\|_2}{|p^{(k-1)^*} q^{(k-1)}|} & &
\end{aligned}$$

Le réel $e^{(k)}$ fournit alors une estimation de $|\lambda_1 - \lambda^{(k-1)}|$.

Démonstration du Théorème 8.3 On peut trouver une base $(x, e_1, \dots e_{n-1})$ de vecteurs propres de A. L'existence des fonctions $(x(.), \lambda(.))$ vient du théorème des fonctions implicites appliqué dans l'espace engendré par les vecteurs propres $\mathbb{C} \times \text{Vect}(e_1, \dots, e_{n-1})$, et du fait que λ est une valeur propre simple : exercice.

On peut dériver par rapport à ϵ l'équation $(A + \epsilon E)x(\epsilon) = \lambda(\epsilon)x(\epsilon)$, ce qui donne $Ex(\epsilon) + (A + \epsilon E - \lambda(\epsilon)I)\frac{\partial x}{\partial \epsilon}(\epsilon) = \frac{\partial \lambda}{\partial \epsilon}(\epsilon)x(\epsilon)$. En $\epsilon = 0$, on a : $Ex + (A - \lambda I)\frac{\partial x}{\partial \epsilon}(0) = \frac{\partial \lambda}{\partial \epsilon}(0)x$. On multiplie à gauche par y^* : on obtient $y^* Ex = \frac{\partial \lambda}{\partial \epsilon}(0) y^* x$.

On doit montrer que $y^* x \neq 0$. Pour cela, si on note X la matrice dont les colonnes sont les vecteurs propres (à droite) de A, on a $AX = XD$ où D est une matrice diagonale. On en déduit

$X^{-1}A = DX^{-1}$ ou encore $A^*(X^*)^{-1} = (X^*)^{-1}D^*$. Pour $i \neq j$, la i ème colonne de $(X^*)^{-1}$ est orthogonale à la j ème colonne de X. On construit Y en divisant chaque colonne de $(X^*)^{-1}$ par sa norme. La i ème colonne de Y est un vecteur propre unitaire de A^* de valeur propre λ_i. Si λ_k est une valeur propre simple, alors $y_k^* x_k \neq 0$.
On conclut en majorant $|y^*Ex|$ par 1 d'après les hypothèses sur E, x et y. □

8.2.3 Méthode de la puissance inverse

Supposons que A est inversible et diagonalisable. On note $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A et $x_1, \ldots x_n$ des vecteurs propres normaux associés. On suppose cette fois que les valeurs propres de A vérifient

$$0 < |\lambda_1| < |\lambda_2| \leq |\lambda_3| \leq \cdots \leq |\lambda_n|. \tag{8.9}$$

On voit que A^{-1} est diagonalisable et que ses valeurs propres vérifient :

$$|\lambda_1^{-1}| > |\lambda_2^{-1}| \geq |\lambda_3^{-1}| \geq \cdots \geq |\lambda_n^{-1}|.$$

On peut donc appliquer la méthode de la puissance à A^{-1} pour approcher λ_1^{-1} : on se donne un vecteur $q^{(0)}$ tel que $\|q^{(0)}\|_2 = 1$ et on construit la suite $\mu^{(k-1)}, q^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned} \text{trouver } z^{(k)} \text{ t.q.} \quad Az^{(k)} &= q^{(k-1)}, \\ q^{(k)} &= \frac{z^{(k)}}{\|z^{(k)}\|_2}, \\ \mu^{(k-1)} &= q^{(k-1)^*} z^{(k)}, \end{aligned}$$

Cet algorithme est appelé méthode de la puissance inverse : on a le résultat suivant

Théorème 8.4 (convergence de la méthode de la puissance inverse) *Soit $A \in \mathcal{M}_n(\mathbb{C})$ diagonalisable et inversible vérifiant (8.9). Si $q^{(0)} = a_1 x_1 + \ldots a_n x_n$ avec $a_1 \neq 0$, alors $\mu^{(k)}$ converge vers λ_1^{-1} et il existe $C \geq 0$ indépendant de k et pour tout k, $\epsilon^{(k)} \in \mathbb{C}$, $|\epsilon^{(k)}| = 1$ tels que*

$$\begin{aligned} |\mu^{(k)} - \lambda_1^{-1}| &\leq C \left(\frac{|\lambda_1|}{|\lambda_2|} \right)^k, \\ \|\epsilon^{(k)} q^{(k)} - x_1\| &\leq C \left(\frac{|\lambda_1|}{|\lambda_2|} \right)^k, \end{aligned} \tag{8.10}$$

Remarque 8.5 *Comme $\lambda_1 \neq 0$, on a aussi*

$$|\lambda_1 - \frac{1}{\mu^{(k)}}| \leq C \left(\frac{|\lambda_1|}{|\lambda_2|} \right)^k,$$

où C est une constante indépendante de k.

Remarque 8.6 *La méthode de la puissance inverse nécessite de savoir résoudre des systèmes linéaires du type $Ax = b$. Elle est donc plus complexe et plus coûteuse en temps que la méthode de la puissance. Elle ne nécessite cependant pas la connaissance explicite de la matrice A^{-1}.*

8.2.4 Méthode de la puissance inverse avec translation

Pour une matrice diagonalisable A, une modification de la méthode de la méthode de la puissance inverse permet d'approcher numériquement la valeur propre λ_i la plus proche d'un nombre complexe η donné, à condition qu'il n'y ait pas d'autre valeur propre sur le cercle dans le plan complexe centré en η et passant par λ_i : on suppose que

$$0 < |\lambda_i - \eta| < |\lambda_j - \eta| \leq |\lambda_l - \eta|, \quad \forall l \neq i, j, 1 \leq l \leq n. \tag{8.11}$$

On applique alors la méthode de la puissance inverse à la matrice diagonalisable et inversible $A - \eta I$: on se donne un vecteur $q^{(0)}$ tel que $\|q^{(0)}\|_2 = 1$ et on construit la suite $\mu^{(k)}, q^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned} \text{trouver } z^{(k)} \text{ t.q.} \quad (A - \eta I) z^{(k)} &= q^{(k-1)}, \\ q^{(k)} &= \frac{z^{(k)}}{\|z^{(k)}\|_2}, \\ \mu^{(k-1)} &= q^{(k-1)^*} z^{(k)}, \end{aligned}$$

La suite $(\mu^{(k)})$ converge vers $(\lambda_i - \eta)^{-1}$ comme $\left|\frac{\lambda_i - \eta}{\lambda_j - \eta}\right|^k$, et on a une convergence similaire vers un vecteur propre x_i.

8.3 Une méthode globale pour des matrices symétriques : la méthode de Jacobi

On va se concentrer sur les matrices symétriques de $\mathcal{M}_n(\mathbb{R})$. La méthode de Jacobi permet de construire à l'aide des rotations de Givens une suite de matrices $A^{(k)}$ orthogonalement semblables à A qui converge vers une matrice diagonale : les coefficients diagonaux de cette matrice limite sont les valeurs propres de A. L'idée de la méthode est simple : pour tout couple (k, l), $0 \leq k < l \leq n$, et pour $\theta \in [0, 2\pi[$, on définit la matrice de rotation de Givens $Q_{k,l}(\theta)$ par

$$Q_{k,l}(\theta) = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} & & & & k & & & l & \\ \end{array} \\ \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & c & 0 & \dots & & s & & \\ & & 0 & 1 & 0 & \dots & & & \\ & & \vdots & & 1 & & \vdots & & \\ & & & & & \ddots & & & \\ & & -s & 0 & \dots & & c & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \\ k \\ \\ \\ l \\ \\ \end{array} \qquad c = \cos(\theta), \quad s = \sin(\theta) \tag{8.12}$$

Pour une matrice symétrique $B \in \mathcal{M}_n(\mathbb{R})$, pour $k < l$, on peut trouver une valeur de θ dans $]-\frac{\pi}{4}, \frac{\pi}{4}]$ telle que $M = Q_{k,l}(\theta)^T B Q_{k,l}(\theta)$ vérifie $m_{k,l} = 0$ et $m_{l,k} = 0$. Si $b_{kl} \neq 0$ cette valeur est unique. En effet,

– si $b_{l,k} = 0$, on prend $\theta = 0$.
– sinon, on voit que $c \neq 0$, et que si $t = \frac{s}{c}$, t doit vérifier

$$t^2 + 2\eta t - 1 = 0, \qquad \eta = \frac{b_{ll} - b_{kk}}{2b_{kl}} \tag{8.13}$$

qui a deux racines $-\eta \pm \sqrt{\eta^2+1}$. Pour que $\theta \in]-\frac{\pi}{4}, \frac{\pi}{4}]$, on doit choisir pour t la racine de plus petit module : $(-1 < t = -\eta - \sqrt{\eta^2+1} < 0$ si $\eta < 0$ et $0 < t = -\eta + \sqrt{\eta^2+1} \leq 1$ si $\eta \geq 0)$ puis $c = \frac{1}{\sqrt{1+t^2}}$ et $s = \frac{t}{\sqrt{1+t^2}}$. On remarque qu'on a aussi $\frac{1}{\tan(2\theta)} = \eta$.
Des calculs élémentaires montrent que la diagonale de B est transformée de la manière suivante : les coefficients d'indices différents de p, p et de q, q ne sont pas changés ; le coefficient d'indice p, p est transformé en $b_{pp} - b_{pq}\tan(\theta)$ et le coefficient d'indice q, q est transformé en $b_{qq} + b_{pq}\tan(\theta)$.

Les méthodes de Jacobi consistent à construire une suite de matrices symétriques $A^{(k)}$, avec $A^{(0)} = A$, en choisissant pour tout k un couple (p, q) $1 \leq p < q \leq n$ et en construisant $A^{(k+1)}$ par la relation $A^{(k+1)} = Q_{pq}(\theta)^T A^{(k)} Q_{pq}(\theta)$, où $\theta \in]-\frac{\pi}{4}, \frac{\pi}{4}]$ est choisi comme ci dessus de manière à annuler $a_{pq}^{(k+1)} = a_{qp}^{(k+1)}$.
Il y a plusieurs façon de choisir quels coefficients annuler à l'itération k :

1. dans la méthode de Jacobi classique, on choisit (p, q) tel que $a_{pq}^{(k)}$ soit le coefficient extra-diagonal de $A^{(k)}$ de plus grand module.

$$|a_{pq}^{(k)}| \geq |a_{ij}^{(k)}| \quad \forall i, j : \ i < j \ (i, j) \neq (p, q)$$

Cette méthode a de bonnes propriétés de convergence mais exige une recherche du plus grand coefficient extra-diagonal. Ceci nécessite de l'ordre de $\frac{n^2}{2}$ tests, et peut devenir coûteux pour n grand.

2. dans la méthode de Jacobi cyclique, on annule successivement tous les coefficients extra-diagonaux par un balayage cyclique : par exemple on choisira les couples (p, q) dans l'ordre lexicographique suivant

$$(1,2)\ (1,3)\ \ldots\ (1,n)\ (2,3)\ \ldots\ (2,n) \ldots (n-1,n)$$

On annulera $a_{12}^{(1)}$ pour construire $A^{(1)}$, $a_{23}^{(2)}$ pour construire $A^{(2)}$, etc... Chaque cycle nécessite donc de l'ordre de $\frac{n(n-1)}{2}$ transformations de jacobi.

3. On peut modifier légèrement l'algorithme ci-dessus en ne cherchant pas annuler les coefficients qui sont déjà très petits.

On va donner un résultat de convergence pour la méthode de Jacobi classique :

Théorème 8.5 (convergence de la méthode de Jacobi classique) *La suite $A^{(k)}$ construite par la méthode de Jacobi classique converge vers une matrice diagonale* $\mathrm{Diag}(\lambda_1, \ldots \lambda_n)$ *et $(\lambda_i)_{i=1\ldots n}$ est la famille des valeurs propres de A, (les valeurs propres multiples sont répétées avec leur multiplicité).*

Démonstration. i) On montre que la partie extradiagonale de $A^{(k)}$ converge vers 0.
On note $|M|_E$ la norme de Frobenius de la partie extradiagonale de M :

$$|M|_E^2 = 2 \sum_{1 \leq p < q \leq n} m_{p,q}^2. \tag{8.14}$$

On a l'inégalité :

$$\max_{i<j} |m_{i,j}| \geq \sqrt{\frac{1}{n(n-1)}} |M|_E. \tag{8.15}$$

Si p, q est tel que $|a_{p,q}^{(k)}| = \max_{i<j} |a_{i,j}^{(k)}|$, on a donc

1. $|a_{p,q}^{(k)}| \geq \sqrt{\frac{1}{n(n-1)}} |A^{(k)}|_E$.
2. $|a_{p,p}^{(k)}|^2 + 2|a_{p,q}^{(k)}|^2 + |a_{q,q}^{(k)}|^2 = |a_{p,p}^{(k+1)}|^2 + |a_{q,q}^{(k+1)}|^2$. car les matrices

$$\begin{pmatrix} a_{p,p}^{(k)} & a_{p,q}^{(k)} \\ a_{p,q}^{(k)} & a_{q,q}^{(k)} \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} a_{p,p}^{(k+1)} & 0 \\ 0 & a_{q,q}^{(k+1)} \end{pmatrix}$$

sont orthogonalement semblables.

Comme on a pour $i \neq p, q$, $a_{i,i}^{(k+1)} = a_{i,i}^{(k)}$, on en déduit

$$\|\mathrm{Diag}(A^{(k+1)})\|_F^2 \geq \|\mathrm{Diag}(A^{(k)})\|_F^2 + \frac{2}{n(n-1)} |A^{(k)}|_E^2. \tag{8.16}$$

On sait aussi que la norme de Frobenius est invariante par transformation orthogonale :

$$\|A^{(k+1)}\|_F^2 = \|A^{(k)}\|_F^2. \tag{8.17}$$

De (8.16) et (8.17), on déduit que

$$|A^{(k+1)}|_E^2 \leq (1 - \frac{2}{n(n-1)}) |A^{(k)}|_E^2, \tag{8.18}$$

ce qui montre que $A^{(k)}$ tend vers une matrice diagonale.

ii) En notant $D^{(k)}$ la diagonale de $A^{(k)}$, montrons que $D^{(k+1)} - D^{(k)}$ tend vers 0. On a $|a_{ii}^{(k+1)} - a_{ii}^{(k)}| = |a_{pq}^{(k)}||\tan(\theta)|$ ($(|\tan(\theta)| \leq 1)$ si $i = p$ ou $i = q$ et $|a_{ii}^{(k+1)} - a_{ii}^{(k)}| = 0$ sinon. Comme $a_{pq}^{(k)}$ tend vers 0 et $|\tan(\theta)| \leq 1$, on en déduit le résultat.

iii) Soit D une valeur d'adhérence de la suite $A^{(k)}$: pour tout réel λ, $\det(D - \lambda I)$ est une valeur d'adhérence de $\det(A^{(k)} - \lambda I)$. D'autre part, toute les matrices $A^{(k)}$ sont semblables à A : donc, $\det(D - \lambda I) = \det(A - \lambda I)$, ce qui implique que D a le même polynôme caractéristique et les mêmes valeurs propres que A (répétées avec leur multiplicité). De là, on déduit aussi que la suite $A^{(k)}$ a un nombre fini de valeurs d'adhérence.

iv) Il faut montrer que toute la suite $A^{(k)}$ converge. C'est une conséquence de ii) et iii).

1. On montre dans un premier temps par l'absurde que si $(L_1, \ldots L_p)$ sont les valeurs d'adhérence de la suite $A^{(k)}$,(on sait qu'il y en a un nombre fini) alors pour tout ϵ, il existe $k(\epsilon)$ tel que pour tout $k > k(\epsilon)$, il existe $i \in \{1, \ldots, p\}$ tel que $\|A^{(k)} - L_i\|_F \leq \epsilon$: exercice.
2. Après on prend $\epsilon = \frac{1}{3} \min_{i \neq j} \|L_i - L_j\|$. On sait qu'il existe K tel que pour tout $k > K$, $\|A^{(k+1)} - A^{(k)}\|_F \leq \epsilon$, et il existe $i \in \{1, \ldots, p\}$ tel que $\|A^{(k)} - L_i\|_F \leq \epsilon$. On conclut en montrant qu'il existe i_0 tel que pour tout $k > K$, $\|A^{(k)} - L_{i_0}\|_F \leq \epsilon$: exercice.

■

Remarque 8.7 *La méthode de Jacobi construit $A^{(k)}$ orthogonalement semblable à A : on a $A^{(k)} = (Q^{(k)})^T A Q^{(k)}$, où $Q^{(k)}$ est le produit de k matrices de Givens. On peut alors montrer que si toutes les valeurs propres de A sont simples, alors $Q^{(k)}$ converge vers Q et les colonnes de Q constituent une base orthonormale de vecteurs propres de A. La démonstration est semblable à celle du Théorème 8.5, cf. [2].*

8.4 Une méthode globale : la méthode QR

La méthode QR est une méthode permettant d'obtenir tout le spectre d'une matrice. C'est une méthode très populaire. On la décrit et on analyse ses propriétés. L'analyse de la méthode est difficile et n'est pas au programme de l'examen.

8.4.1 Factorisations QR d'une matrice

Les réflexions de Householder

Définition 8.1 *Soit v un vecteur non nul de $\mathbb{C}^n$. On appelle réflexion de Householder ou matrice de Householder relative au vecteur v la matrice*

$$H_v = I - 2\frac{vv^*}{v^*v}. \tag{8.19}$$

Les réflexions de Householder ont les propriétés suivantes :

1. H_v est une matrice hermitienne.
2. H_v est une matrice unitaire.
3. $H_v - I$ est une matrice de rang un.
4. $H_{\lambda v} = H_v$, pour tout $\lambda \neq 0$.

Proposition 8.1 *Soit e un vecteur unitaire de $\mathbb{C}^N$ et x un vecteur non nul de $\mathbb{C}^N$. Il existe un vecteur non nul $v \in \mathbb{C}^N$ tel que $H_v x$ soit colinéaire à e.*

Démonstration. On cherche v sous la forme $v = x + \lambda e$. Dans le cas réel, ($x \in \mathbb{R}^N$ et $e \in \mathbb{R}^N$), les vecteurs $v = x \pm \|x\|_2 e$ (au moins l'un des deux est non nul) sont les seuls vecteurs de cette forme ayant la propriété demandée et on a $H_v x = \mp\|x\|_2 e$. Dans le cas général, on trouve aussi deux vecteurs v sous cette forme, et au moins l'un d'entre eux est non nul. ■

Remarque 8.8 *Si x est presque colinéaire à e, on a intérêt en pratique à choisir le vecteur v pour que v^*v, qui est au dénominateur de (8.19), ne soit pas petit. En effet, en précision finie, il faut éviter les divisions par des nombres petits. En particulier si $v \in \mathbb{R}^n$, on préférera le vecteur $v^+ = x + \text{signe}(x^*e)\|x\|_2 e$ au vecteur $v^- = x - \text{signe}(x^*e)\|x\|_2 e$, car $\|v^-\|_2$ est petit.*

Factorisations QR à l'aide des réflexions de Householder

Théorème 8.6 *Soit A une matrice de $\mathcal{M}_{m,n}(\mathbb{C})$ avec $m \geq n$. Il existe une matrice unitaire $Q \in \mathcal{M}_m(\mathbb{C})$ et une matrice triangulaire supérieure $R \in \mathcal{M}_{m,n}(\mathbb{C})$, telles que*

$$A = QR.$$

Démonstration. On démontre ce résultat par récurrence : supposons qu'à l'aide de deux réflexions de Householder H_1 et H_2, on ait obtenu

$$H_2H_1A = \begin{pmatrix} * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & 0 & * & \dots & * \\ \vdots & \vdots & * & \dots & * \\ 0 & 0 & * & \dots & * \end{pmatrix}$$

On appelle x le vecteur de $\mathbb{C}^{m-2}$ obtenu en prenant les $m-2$ derniers coefficients de la troisième colonne de H_2H_1A. Si x est non nul, on sait trouver $v_3 \in \mathbb{C}^{m-2}$, tel que

$$H_{v_3}\, x \text{ soit colinéaire à } \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

On prend alors H_3

$$H_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & H_{v_3} \end{pmatrix}.$$

Si $x = 0$ on prend $H_3 = I$. Dans les deux cas H_3 est hermitienne et unitaire. On a

$$H_3H_2H_1A = \begin{pmatrix} * & * & * & & \dots & * \\ 0 & * & * & & \dots & * \\ 0 & 0 & * & & \dots & * \\ 0 & 0 & 0 & * & \dots & * \\ \vdots & \vdots & \vdots & * & \dots & * \\ 0 & 0 & 0 & * & \dots & * \end{pmatrix}$$

Comme $m \geq n$, on peut itérer ce procédé et construire n matrices H_i hermitiennes et unitaires, telles que $H_n \dots H_1A$ soit une matrice triangulaire supérieure R. On note Q la matrice unitaire (mais pas forcément hermitienne) $Q = H_1 \dots H_n$, on a

$$A = QR.$$

■

Les éléments fondamentaux de l'algorithme de la factorisation QR à l'aide de réflexions de Householder sont

1. Le choix des vecteurs v_i, de manière à ce que $\|v_i\|$ ne soit pas petit, cf. Remarque 8.8.
2. Le produit à gauche d'une matrice M par la matrice H_m : H_mM, doit être programmé intelligemment : on ne doit surtout pas stocker la matrice H_m. Il suffit de garder en mémoire le vecteur v_m, et d'opérer les réflexions de Householder à chaque colonne de M.
3. On ne calcule donc pas Q en général, mais on garde en mémoire les vecteurs $v_1, \dots, v_n$.
4. On peut écraser la matrice A que l'on factorise en stockant les vecteurs v_i dans la partie triangulaire inférieure stricte, et la matrice R dans la partie triangulaire supérieure à condition de normaliser les vecteurs v_i de manière à ce que leur premier coefficient soit 1.

5. Si la matrice A est réelle, alors la matrice Q l'est aussi.

La factorisation QR a de nombreuses propriétés intéressantes :

1. Résoudre le système $Qy = b$ est facile car

$$Q^{-1} = Q^* = H_n \dots H_1.$$

La complexité de la résolution du système $Qy = b$ est donc $3\sum_i^n (m-i)$ adds+mults. Si $m = n$, la complexité est de l'ordre de $\frac{3}{2}n^2$. Si la matrice A est carrée d'ordre n inversible, pour résoudre $Ax = b$, on résout d'abord $Qy = b$ comme ci-dessus, puis $Rx = y$ par une remontée. La complexité totale est de l'ordre de $2n^2$.

2. Si $A \in \mathcal{M}_n(\mathbb{C})$ est inversible, alors

$$\mathrm{cond}_2(A) = \mathrm{cond}_2(R),$$

car Q est unitaire.

3. la méthode de Householder permet de calculer $|\det(A)|$. En effet,

$$|\det(R)| = |\text{produit des coefficients diagonaux de } R| = |\det(A)|.$$

4. Si $A \in \mathcal{M}_{m,n}(\mathbb{C})$, $m \geq n$, est factorisée sous la forme $A = QR$, alors résoudre une système de la forme $A^*Ax = b$ est facile si $\mathrm{rang}(A) = n$, car $A^*A = R^*R$ et $\mathrm{rang}(R) = n$, et on peut résoudre $A^*Ax = b$ par une descente(de R^*)-remontée (de R). La factorisation QR d'une matrice se prête donc bien aux problèmes de moindres carrés.

Théorème 8.7 *Soit $A \in \mathbb{C}^{m\times m}$. On peut trouver une factorisation QR telle que tous les coefficients diagonaux de R sont réels positifs. Si A est inversible, cette factorisation est unique.*

Démonstration. On part de la factorisation QR de A obtenue par la méthode de Householder ci dessus : $A = Q'R'$ où la matrice Q' est unitaire et R' est triangulaire supérieure. Appelons D la matrice diagonale $D = \mathrm{Diag}(d_1, \dots d_m)$ où

$$d_i = \begin{cases} \frac{\overline{r_{ii}}}{|r_{ii}|} & \text{si } r_{ii} \neq 0, \\ 1 & \text{si } r_{ii} = 0, \end{cases}$$

La matrice D est clairement unitaire. Posons $Q = Q'D^{-1}$ et $R = DR'$. Ces matrices ont les propriétés désirées.

Supposons que A soit inversible et soient deux factorisation QR de A,

$$A = Q_1R_1 = Q_2R_2$$

telles que tous les coefficients diagonaux de R_1 et R_2 sont réels positifs. Ils sont strictement positifs car A est inversible. On a alors $Q_2^*Q_1 = R_2R_1^{-1}$. Mais $Q_2^*Q_1$ est unitaire tandis que $R_2R_1^{-1}$ est triangulaire supérieure avec des coefficients diagonaux réels positifs strictement. On peut vérifier que l'Identité est la seule matrice unitaire et triangulaire supérieure avec des coefficients diagonaux réels positifs. Donc $Q_2 = Q_1$ et $R_2 = R_1$. ■

Les rotations de Givens

Les réflexions de Householder sont très utiles quand il s'agit d'annuler tous les coefficients d'une colonne sous la diagonale. En revanche, pour annuler sélectivement certains coefficients (si la matrice a déjà beaucoup de coefficients nuls par exemple), les rotations de Givens constituent un outil important.
Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$: il existe deux réels $c = \cos(\theta)$ et $s = \sin(\theta)$ tels que si

$$Q = \begin{array}{c} \\ \\ \\ \\ i \\ \\ \\ \\ j \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} & & & & i & & & & j \end{array} \\ \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & c & 0 & \dots & & s & & \\ & & 0 & 1 & 0 & \dots & & & \\ & & \vdots & & 1 & & \vdots & & \\ & & & & & \ddots & & & \\ & & -s & 0 & \dots & & c & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \end{pmatrix} \end{array}$$

alors les lignes de QA d'indices différents de i et j sont identiques à celles de A, et $(QA)_{j,i} = 0$. On définit $\theta \in]-\frac{\pi}{2}, \frac{\pi}{2}[$ tel $\tan(\theta) = t = \frac{a_{j,i}}{a_{i,i}}$ et $c = \cos(\theta) = \frac{1}{\sqrt{t^2+1}}$ et $s = \sin(\theta) = \frac{t}{\sqrt{t^2+1}}$. On a $-sa_{i,i} + ca_{j,i} = 0$.

8.4.2 Généralisation de la méthode de la puissance

Soit $A \in \mathcal{M}_n(\mathbb{C})$. On note $\lambda_1, \dots, \lambda_n$ les valeurs propres de A et $x_1, \dots x_n$ les vecteurs propres associés. Dans la suite on va supposer que pour p, $p < n$,

$$|\lambda_1| \geq |\lambda_2| \geq \dots |\lambda_p| > |\lambda_{p+1}| \geq \dots \geq |\lambda_n|. \tag{8.20}$$

Soit $T = Q^*AQ$ une décomposition de Schur de A telle que $D = \mathrm{Diag}(T) = (\lambda_1, \dots, \lambda_n)$, vérifiant (8.20). On note $N = T - D$ et

$$T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \quad \text{et} \quad \mathrm{N} = \begin{pmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{pmatrix} \tag{8.21}$$

où T_{11} et N_{11} sont d'ordre p.
On notera

$$Q = (Q_1, Q_2) \tag{8.22}$$

où Q_1 est la matrice de $\mathbb{C}^{p\times n}$ obtenue en prenant les p premières colonnes de Q. On note $D_p(A)$ le sous-espace engendré par les colonnes de Q_1. On va chercher à approcher le sous-espace $D_p(A)$. Pour cela, on généralise la méthode de la puissance de la manière suivante : on se donne une matrice $Q^{(0)} \in \mathbb{C}^{n\times p}$ dont les colonnes sont orthonormales et on construit la suite $R^{(k)}, Q^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned} Z^{(k)} &= AQ^{(k-1)}, \\ Z^{(k)} &= Q^{(k)}R^{(k)}, \quad \text{(factorisation QR de } Z^{(k)}) \end{aligned}$$

Ici $Q^{(k)} \in \mathbb{C}^{n\times p}$ a ses colonnes orthonormales et $R^{(k)} \in \mathbb{C}^{p\times p}$ est une matrice triangulaire supérieure.

Remarque 8.9 *Pour $p = 1$, c'est exactement la méthode de la puissance et $R^{(k)} = \lambda^{(k)}$.*

On veut montrer que le sous-espace engendré par les colonnes de $Q^{(k)}$, noté noté $\mathrm{Im}(Q^{(k)})$, converge vers le sous-espace $D_p(A)$.

Théorème 8.8 (Convergence vers $D_p(A)$) *Sous l'hypothèse (8.20) et si de plus*

$$\mathrm{dist}(D_p(A^*), \mathrm{Im}(Q^{(0)})) < 1, \tag{8.23}$$

alors

$$\mathrm{dist}(D_p(A), \mathrm{Im}(Q^{(k)})) = O\left(\frac{|\lambda_{p+1}|^k}{|\lambda_p|^k}\right). \tag{8.24}$$

Démonstration. La preuve est assez complexe. Elle est donnée dans la section 8.4.5. ■

Remarque 8.10 *L'hypothèse (8.23) est de la même nature que l'hypothèse sur le choix initial dans la méthode de la puissance, cf. Théorème 8.2.*

8.4.3 La méthode QR

Si on prend $p = n$ dans l'algorithme précédent, et si on suppose de plus que les valeurs propres de A vérifient $|\lambda_1| < |\lambda_2| < \cdots < |\lambda_n|$ (ceci implique d'ailleurs que A est diagonalisable), on voit que l'algorithme calcule une décomposition de Schur de A. En effet, pour tout $i < n$ et si $\mathrm{dist}(D_i(A^*), vect(q_1^{(0)}, \ldots q_i^{(0)})) < 1$, on voit d'après le Théorème 8.11 que $\mathrm{dist}(D_i(A), vect(q_1^{(k)}, \ldots q_i^{(k)}))$ tend vers 0. Ceci implique que $T^{(k)} = (Q^{(k)})^* A Q^{(k)}$ converge vers une décomposition de Schur de A, (la limite de $T^{(k)}$ est triangulaire supérieure avec les valeurs propres sur la diagonale).

Remarque 8.11 *Si A est diagonalisable, les colonnes de $Q^{(k)}$ ne convergent pas en général vers les vecteurs propres de A, qui ne sont pas nécessairement orthogonaux.*

Remarque 8.12 *Cette méthode ne donne donc pas directement les vecteurs propres, mais une fois les valeurs propres approchées précisément, on peut utiliser une méthode de la puissance inverse translatée pour calculer les vecteurs propres.*

Dans le cas général, si plusieurs valeurs propres ont même module, $T^{(k)}$ ne converge plus vers une matrice triangulaire supérieure mais vers une matrice triangulaire supérieure par blocs, les blocs diagonaux correspondant aux valeurs propres d'un même module. On retrouve les valeurs propres de A en prenant les valeurs propres des blocs diagonaux.
En notant $T^{(k)} = (Q^{(k)})^* A Q^{(k)}$, on voit que $T^{(k-1)} = (Q^{(k-1)})^* Q^{(k)} R^{(k)} = U^{(k)} R^{(k)}$ si on note $U^{(k)} = (Q^{(k-1)})^* Q^{(k)}$. De même, $T^{(k)} = (Q^{(k)})^* A Q^{(k)} = (Q^{(k)})^* A Q^{(k-1)} U^{(k)} = (Q^{(k)})^* Q^{(k)} R^{(k)} U^{(k)} = R^{(k)} U^{(k)}$. On peut donc réécrire l'algorithme précédent sous sa forme la plus connue : on se donne une matrice $Q^{(0)} \in \mathbb{C}^{n\times p}$ dont les colonnes sont orthonormales, on pose $T^{(0)} = (Q^{(0)})^* A Q^{(0)}$ et on construit la suite $T^{(k)}, R^{(k)}, U^{(k)}$ pour $k \geq 1$, par la récurrence suivante

$$\begin{aligned} T^{(k-1)} &= U^{(k)} R^{(k)}, \quad \text{(factorisation QR de } T^{(k-1)}) \\ T^{(k)} &= R^{(k)} U^{(k)}, \end{aligned}$$

8.4.4 Accélération de la méthode QR pour $A \in \mathcal{M}_n(\mathbb{R})$

On voit qu'une étape de la méthode QR ci-dessus demande de l'ordre de n^3 opérations, donc cette méthode est coûteuse si n est grand.
On peut grandement accélérer la méthode QR en transformant au préalable A en une matrice orthogonalement semblable $T^{(0)}$ de type Hessenberg supérieure ($t^{(0)}_{i,j} = 0$ si $i > j+1$) à l'aide de symétries de Householder. Après, on applique l'algorithme décrit ci-dessus à $T^{(0)}$ en effectuant les factorisations QR par des rotations de Givens. L'intérêt de mettre $T^{(0)}$ sous la forme d'une matrice Hessenberg supérieure vient du fait que tous les $T^{(k)}$, $k > 0$ sont alors aussi de la forme Hessenberg supérieure et les factorisations QR des matrices Hessenberg ($T^{(k-1)} = U^{(k)}R^{(k)}$) par une succession de rotations de Givens demandent de l'ordre de n^2 opérations au lieu de n^3. De même le calcul de $T^{(k)} = R^{(k)}U^{(k)}$ demande de l'ordre de n^2 opérations.
Pour justifier cette modification de l'algorithme, on donne les résultats suivants :

Théorème 8.9 *Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$. Il existe une matrice orthogonale $Q \in \mathcal{M}_n(\mathbb{R})$ et une matrice de type Hessenberg supérieure $H \in \mathcal{M}_n(\mathbb{R})$ ($h_{i,j} = 0$ si $i > j+1$), telles que*

$$Q^T AQ = H.$$

Idée de la Démonstration On démontre ce résultat par récurrence : à l'aide de d'une réflexion de Householder H_1 , on peut obtenir

$$H_1^T A = \begin{pmatrix} * & * & * & \dots & * \\ * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ \vdots & \vdots & * & \dots & * \\ 0 & * & * & \dots & * \end{pmatrix}$$

Si on note c_1 le vecteur $c_1 = (a_{21}, \dots a_{n1})^T$, H_1 s'écrit $\mathrm{Diag}(1, \tilde{H}_1)$ où $\tilde{H}_1$ est une matrice de Householder d'ordre $n-1$ telle que $\tilde{H}_1 c_1 = (1, 0, \dots, 0)^T$ Alors on vérifie que $H_1^T AH_1$ est du même type que $H_1^T A$, i.e. ses coefficients d'indices $i, 1$ sont nuls pour $i > 2$.
Si après $k-1$ itérations, on a construit $k-1$ matrices H_k telles que

$$(H_1 \dots H_{k-1})^T A(H_1 \dots H_{k-1}) = \begin{array}{c} \begin{array}{ccc} k-1 & 1 & n-k \end{array} \\ \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{pmatrix} \end{array} \begin{array}{l} k-1 \\ 1 \\ n-k \end{array}$$

soit de type Hessenberg pour ses $k-1$ premières colonnes, on choisit alors $\tilde{H}_k$ une matrice Householder telle que $\tilde{H}_k^T B_{32}$ soit colinéaire à $(1, 0, \dots 0)^T \in \mathbb{R}^{n-k}$. On prend $H_k = \mathrm{Diag}(I_k, \tilde{H}_k)$, et on voit que

$$(H_1 \dots H_{k-1}H_k)^T A(H_1 \dots H_{k-1}H_k) = \begin{pmatrix} B_{11} & B_{12} & B_{13}\tilde{H}_k \\ B_{21} & B_{22} & B_{23}\tilde{H}_k \\ 0 & \tilde{H}_k^T B_{32} & \tilde{H}_k^T B_{33}\tilde{H}_k \end{pmatrix}$$

est de type Hessenberg sur ses k premières colonnes.
Le résultat est démontré. □

Théorème 8.10 *Soit H une matrice de $\mathcal{M}_n(\mathbb{R})$ de type Hessenberg supérieure. Il existe $n-1$ matrices de rotations de Givens et une matrice triangulaire supérieure R telles que $G_{n-1}^T \dots G_1^T H = R$. De plus $RG_1 \dots G_{n-1}$ est de type Hessenberg supérieure.*

Démonstration. Exercice ■

Remarque 8.13 *Si A est symétrique, alors le Théorème 8.9 dit qu'on peut construire à l'aide de symétries de Householder une matrice orthogonale $Q \in \mathcal{M}_n(\mathbb{R})$ telle que $Q^T AQ$ soit tridiagonale. On est alors ramené au problème de trouver les valeurs propres d'une matrice tridiagonale symétrique à coefficients réels : pour cela, on peut remplacer la méthode QR par la méthode des suites de Sturm encore appelée méthode de Givens, cf. [2].*

8.4.5 Démonstration du Théorème 8.8

En première lecture, on pourra lire le paragraphe § 8.4.5 et sauter les paragraphes § 8.4.5 et § 8.4.5.

Le lemme fondamental

Lemme 8.1 *Pour tout $k > 0$, $\theta \geq 0$, on a*

$$\|T_{22}^k\|_2 \leq (1+\theta)^{n-p-1} \left(|\lambda_{p+1}| + \frac{\|N\|_F}{1+\theta} \right)^k \tag{8.25}$$

et

$$\|T_{11}^{-k}\|_2 \leq (1+\theta)^{p-1} \left(|\lambda_p| - \frac{\|N\|_F}{1+\theta} \right)^{-k} \tag{8.26}$$

Démonstration. Soit $\Delta = \mathrm{Diag}(1, 1+\theta, \dots (1+\theta)^{n-p-1})$. On vérifie facilement que $\|\Delta N \Delta^{-1}\|_F \leq \frac{\|N\|_F}{1+\theta}$. On a

$$\begin{aligned}
\|T_{22}^k\|_2 &= \|\Delta^{-1}(\Delta N_{22}\Delta^{-1} + D_{22})^k \Delta\|_2 \leq \|\Delta^{-1}\|_2 (\|\Delta N_{22}\Delta^{-1}\|_2 + \|D_{22}\|_2)^k \|\Delta\|_2 \\
&\leq \mathrm{cond}_2(\Delta)(|\lambda_{p+1}| + \frac{\|N\|_F}{1+\theta})^k \\
&= (1+\theta)^{n-p-1}(|\lambda_{p+1}| + \frac{\|N\|_F}{1+\theta})^k
\end{aligned}$$

et

$$\begin{aligned}
\|T_{11}^{-k}\|_2 &= \|\Delta^{-1} D_{11}^{-k} (\Delta D_{11}^{-1} N_{11} \Delta^{-1} + I)^{-k} \Delta\|_2 \leq \|\Delta^{-1}\|_2 \|D_{11}^{-k}\|_2 \|(\Delta D_{11}^{-1} N_{11}\Delta^{-1} + I)^{-k}\|_2 \|\Delta\|_2 \\
&\leq \mathrm{cond}_2(\Delta) |\lambda_p|^{-k} \frac{1}{(1 - \|\Delta D_{11}^{-1} N_{11} \Delta^{-1}\|_2)^k} \\
&\leq (1+\theta)^{p-1} \frac{|\lambda_p|^{-k}}{(1 - \frac{|\lambda_p|^{-1}\|N\|_F}{1+\theta})^k} \\
&\leq (1+\theta)^{p-1} \frac{1}{(|\lambda_p| - \frac{\|N\|_F}{1+\theta})^k}
\end{aligned}$$

■

Un résultat préliminaire de changement de bases

Théorème 8.11 *Soit T une matrice triangulaire par blocs :*

$$T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$$

et $T_{11} \in \mathbb{C}^{p\times p}$ $T_{22} \in \mathbb{C}^{(n-p)\times(n-p)}$. *la transformation linéaire* $\Phi : \mathbb{C}^{(p)\times(n-p)} \to \mathbb{C}^{(p)\times(n-p)}$: $\Phi(X) = T_{11}X - XT_{22}$ *est non singulière si et seulement si* $sp(T_{11}) \cap sp(T_{22}) = \emptyset$. *Dans ce cas, il existe une unique matrice* $X \in \mathbb{C}^{(p)\times(n-p)}$ *telle que*

$$\begin{pmatrix} I_p & X \\ 0 & I_{n-p} \end{pmatrix}^{-1} T \begin{pmatrix} I_p & X \\ 0 & I_{n-p} \end{pmatrix} = \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix} \tag{8.27}$$

Démonstration. Supposons qu'il existe X tel que $\Phi(X) = 0$. On sait qu'il existe deux matrices unitaires $U \in \mathbb{C}^{p\times p}$ et $V \in \mathbb{C}^{(n-p)\times(n-p)}$ telles que

$$U^*XV = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$$

où Σ_r est une matrice diagonale d'ordre r, sans zéro sur la diagonale. On a

$$T_{11}U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^* = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*T_{22} \Leftrightarrow U^*T_{11}U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*T_{22}V. \tag{8.28}$$

On note $M = U^*T_{11}U$ et $N = V^*T_{22}V$,

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \quad N = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}$$

avec M_{11} et N_{11} d'ordre r. D'après (8.28), on a $M_{11}\Sigma_r = \Sigma_r N_{11}$ et $M_{21} = N_{21} = 0$. Les matrices M_{11} et N_{11} sont semblables, et ont donc les mêmes valeurs propres. Comme $M = U^*T_{11}U$ et $N = V^*T_{22}V$, T_{11} et T_{22} ont des valeurs propres communes.
Réciproquement si il existe λ $x \neq 0$ et $y \neq 0$ tels que $T_{11}x = \lambda x$ et $T_{22}y = \lambda y$. On vérifie que $\Phi(xy^*) = 0$.
Le dernier point du Théorème est facile. ■

Démonstration du Théorème 8.8

Avec les notations de § 8.4.2, on vérifie aisément par récurrence que

$$Q^{(k)}R^{(k)} \dots R^{(1)} = A^kQ^{(0)} \tag{8.29}$$

On multiplie (8.29) à gauche par Q^* : $Q^*Q^{(k)}R^{(k)} \dots R^{(1)} = T^kQ^*Q^{(0)}$ En notant

$$Q^*Q^{(k)} = \begin{pmatrix} V^{(k)} \\ W^{(k)} \end{pmatrix}, \quad \text{on a} \quad \begin{pmatrix} V^{(k)} \\ W^{(k)} \end{pmatrix} R^{(k)} \dots R^{(1)} = T^k \begin{pmatrix} V^{(0)} \\ W^{(0)} \end{pmatrix} \tag{8.30}$$

D'autre part, comme $|\lambda_p| > |\lambda_{p+1}|$, et d'après le Théorème 8.11, il existe $X \in \mathbb{C}^{p\times(n-p)}$ vérifiant (8.27). De (8.30), on déduit

$$\begin{pmatrix} I_p & X \\ 0 & I_{n-p} \end{pmatrix} \begin{pmatrix} T_{11}^k & 0 \\ 0 & T_{22}^k \end{pmatrix} \begin{pmatrix} I_p & -X \\ 0 & I_{n-p} \end{pmatrix} \begin{pmatrix} V^{(0)} \\ W^{(0)} \end{pmatrix} = \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix}^k \begin{pmatrix} V^{(0)} \\ W^{(0)} \end{pmatrix} = \begin{pmatrix} V^{(k)} \\ W^{(k)} \end{pmatrix} R^{(k)} \dots R^{(1)} \tag{8.31}$$

ou encore

$$\begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} V^{(0)} - XW^{(0)} \\ W^{(0)} \end{pmatrix} = \begin{pmatrix} V^{(k)} - XW^{(k)} \\ W^{(k)} \end{pmatrix} R^{(k)} \dots R^{(1)} \tag{8.32}$$

On suppose que $V^{(0)} - XW^{(0)}$ est inversible, ce qu'on démontrera plus tard. Alors (8.32) implique que

$$\begin{aligned}(V^{(0)} - XW^{(0)})^{-1}T_{11}^{-k}(V^{(k)} - XW^{(k)}) &= (R^{(k)} \dots R^{(1)})^{-1} \\ W^{(k)} &= T_{22}^k W^{(0)}(V^{(0)} - XW^{(0)})^{-1}T_{11}^{-k}(V^{(k)} - XW^{(k)})\end{aligned}$$

Du Théorème 8.12, on déduit que

$$\begin{aligned}\mathrm{dist}(D_p(A), \mathrm{Im}(Q^{(k)})) &= \|Q_2^* Q^{(k)}\|_2 = \|W^{(k)}\|_2 \\ &\leq \|T_{22}^k\|_2 \|W^{(0)}\|_2 \|(V^{(0)} - XW^{(0)})^{-1}\|_2 \|T_{11}^{-k}\|_2 \|V^{(k)} - XW^{(k)}\|_2 \\ &\leq \|T_{22}^k\|_2 \|W^{(0)}\|_2 \|(V^{(0)} - XW^{(0)})^{-1}\|_2 \|T_{11}^{-k}\|_2 (1 + \|X\|_F)\end{aligned} \tag{8.33}$$

Le résultat (8.24) vient alors du Lemme 8.1, en prenant θ assez grand.

Il ne reste plus qu'à démontrer que $V^{(0)} - XW^{(0)}$ est inversible. On utilise l'équation $A^*Q = QT^*$ ainsi que (8.27), qui s'écrit encore

$$T^* = \begin{pmatrix} I_p & 0 \\ -X^* & I_{n-p} \end{pmatrix} \begin{pmatrix} T_{11}^* & 0 \\ 0 & T_{22}^* \end{pmatrix} \begin{pmatrix} I_p & 0 \\ X^* & I_{n-p} \end{pmatrix} \tag{8.34}$$

On en déduit que

$$\begin{aligned}A^*(Q_1, Q_2) \begin{pmatrix} I_p & 0 \\ -X^* & I_{n-p} \end{pmatrix} &= (Q_1, Q_2) \begin{pmatrix} I_p & 0 \\ -X^* & I_{n-p} \end{pmatrix} \begin{pmatrix} T_{11}^* & 0 \\ 0 & T_{22}^* \end{pmatrix} \\ \Rightarrow A^*(Q_1 - Q_2X^*) &= (Q_1 - Q_2X^*)T_{11}^*.\end{aligned}$$

Ceci montre que les colonnes de $(Q_1 - Q_2X^*)$ engendrent l'espace vectoriel $D_p(A^*)$. En posant $Z = (Q_1 - Q_2X^*)(I_p + XX^*)^{-\frac{1}{2}}$, on voit facilement que les colonnes de Z sont orthonormales (en effet, $Z^*Z = I_p$). Les colonnes de Z forment donc une base orthonormale de $D_p(A^*)$.
D'autre part, $V^{(0)} - XW^{(0)} = (Q_1^* - XQ_2^*)Q^{(0)} = (I_p + XX^*)^{\frac{1}{2}} Z^* Q^{(0)}$, ce qui implique que $\sigma_{\min}(V^{(0)} - XW^{(0)}) \geq \sigma_{\min}(Z^*Q^{(0)})$.
D'après le Théorème 8.12,

$$\sigma_{\min}(Z^*Q^{(0)}) = \sqrt{1 - \mathrm{dist}^2(D_p(A^*), \mathrm{Im}(Q^{(0)}))}$$

qui est strictement positive d'après l'hypothèse du Théorème 8.8. Donc $V^{(0)} - XW^{(0)}$ est bien inversible. □

8.5 Annexe : Distance entre deux sous-espaces de $\mathbb{C}^n$

Dans l'analyse de la méthode QR ci-dessus, on a besoin de la notion de distance entre sous-espaces. Ce paragraphe peut être sauté en première lecture.

Définition 8.2 *Soient V_1 et V_2 deux sous-espaces de $\mathbb{C}^n$. On définit*

$$\mathrm{dist}(V_1, V_2) = \|P_{V_1} - P_{V_2}\|_2, \tag{8.35}$$

où P_V est la projection orthogonale sur V.

Remarque 8.14 *On peut définir de même la distance de deux sous-espaces de $\mathbb{R}^n$.*

Remarque 8.15

$$\mathrm{dist}(V_1, V_2) = \mathrm{dist}(V_1^{\perp}, V_2^{\perp}). \tag{8.36}$$

Remarque 8.16 *Que vaut la distance dans $\mathbb{R}^2$ de deux droites vectorielles engendrées respectivement par $x = (\cos\theta_1, \sin\theta_1)^T$ et $y = (\cos\theta_2, \sin\theta_2)^T$? on a*

$$\begin{aligned}\mathrm{dist}(V_1, V_2) &= \|xx^T - yy^T\|_2 = \left\| \begin{pmatrix} \cos^2\theta_1 - \cos^2\theta_2 & \cos\theta_1\sin\theta_1 - \sin\theta_2\cos\theta_2 \\ \cos\theta_1\sin\theta_1 - \sin\theta_2\cos\theta_2 & \sin^2\theta_1 - \sin^2\theta_2 \end{pmatrix} \right\|_2 \\ &= |\sin(\theta_2 - \theta_1)| \left\| \begin{pmatrix} -\sin(\theta_2+\theta_1) & \cos(\theta_2+\theta_1) \\ \cos(\theta_2+\theta_1) & \sin(\theta_2+\theta_1) \end{pmatrix} \right\|_2 = |\sin(\theta_2 - \theta_1)|\end{aligned}$$

En dimension supérieure, on le résultat

Théorème 8.12 *Soient W et Z deux matrices unitaires de $\mathcal{M}_n(\mathbb{C})$. On a $W = (W_1, W_2)$ et $Z = (Z_1, Z_2)$, avec $W_1, Z_1 \in \mathbb{C}^{n\times p}$ et $W_2, Z_2 \in \mathbb{C}^{n\times(n-p)}$. On appelle S le sous-espace engendré par les colonnes de W_1 et S' le sous-espace engendré par les colonnes de Z_1. On a*

$$\mathrm{dist}(S, S') = \sqrt{1 - \sigma_{\min}^2(W_1^* Z_1)} = \sqrt{1 - \sigma_{\min}^2(W_2^* Z_2)} \tag{8.37}$$

où $\sigma_{\min}(M)$ est la plus petite valeur singulière de M.

Démonstration.

$$\begin{aligned}\mathrm{dist}(S, S') &= \|W_1W_1^* - Z_1Z_1^*\|_2 = \|W^*(W_1W_1^* - Z_1Z_1^*)Z\|_2 = \left\| \begin{pmatrix} 0 & W_1^*Z_2 \\ -W_2^*Z_1 & 0 \end{pmatrix} \right\|_2 \\ &= \left\| \begin{pmatrix} 0 & W_1^*Z_2 \\ -W_2^*Z_1 & 0 \end{pmatrix}^* \begin{pmatrix} 0 & W_1^*Z_2 \\ -W_2^*Z_1 & 0 \end{pmatrix} \right\|_2^{\frac{1}{2}} = \left\| \begin{pmatrix} Z_1^*W_2W_2^*Z_1 & 0 \\ 0 & Z_2^*W_1W_1^*Z_2 \end{pmatrix} \right\|_2^{\frac{1}{2}}\end{aligned}$$

Mais $W_2W_2^* = I - W_1W_1^*$,(car pour tout sous-espace V, $P_{V^\perp} = I - P_V$). Donc $Z_1^*W_2W_2^*Z_1 = Z_1^*Z_1 - Z_1^*W_1W_1^*Z_1 = I - Z_1^*W_1W_1^*Z_1$. Mais $\|I - Z_1^*W_1W_1^*Z_1\|_2 = 1 - \sigma_{\min}^2(W_1^*Z_1)$.
Si on note $B = W_1^*Z_2$, on sait que $\|B^*B\|_2 = \|BB^*\|_2$, ce qui implique que $\|Z_2^*W_1W_1^*Z_2\|_2 = \|W_1^*Z_2Z_2^*W_1\|_2 = \|I - W_1^*Z_1Z_1^*W_1\|_2 = 1 - \sigma_{\min}^2(W_1^*Z_1)$.
On a démontré le résultat. ■

Chapitre 9

Problèmes

9.1 Partiel du 25 Novembre 2002

Exercice 1

1) a) Calculer une réduction de Jordan de la matrice (et la matrice de passage correspondante)

$$A = \begin{pmatrix} 2-\frac{\sqrt{10}}{25} & \frac{\sqrt{10}}{50} & 0 & -\frac{3\sqrt{2}}{10} \\ -2\frac{\sqrt{10}}{25} & 2+\frac{\sqrt{10}}{25} & 0 & -\frac{3\sqrt{2}}{5} \\ -\frac{3\sqrt{2}}{5} & \frac{3\sqrt{2}}{10} & 2 & \frac{1}{\sqrt{10}} \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

en calculant d'abord les valeurs propres de A puis son polynôme minimal.
b) En déduire la décomposition de Dunford-Jordan de A.
Indication On peut obtenir cette décomposition de manière simple, sans effectuer de produits matriciels.
c) Calculer A^{10}.

2) Calculer une réduction de Schur de la matrice

$$M = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 4 & 3 \\ 0 & -4 & -3 \end{pmatrix}$$

3) a) Expliquer comment on construit une décomposition en valeurs singulières d'une $B \in \mathcal{M}_{n,m}(\mathbb{R})$.
b) On choisit pour B la matrice

$$B = \begin{pmatrix} \frac{1}{\sqrt{2}} & 1 & 1 \\ -\frac{1}{\sqrt{2}} & 1 & 1 \\ 0 & -\frac{3}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{pmatrix}$$

Calculer une décomposition en valeurs singulières de la matrice B.
c) Donner $\|B\|_1$,$\|B\|_2$, $\|B\|_\infty$ et $\|B\|_F$.

4) Calculer une réduction de Gauss et la signature de la forme quadratique q sur $\mathbb{R}^4$

$$q(x) = 4x_1^2 + 4x_1x_2 + 4x_1x_3 + 8x_1x_4 + x_2^2 + x_3^2 + 4x_4^2 18x_3x_4 + 4x_4^2$$

Exercice 2

1) a) Montrer que pour tout matrice A de $\mathcal{M}_n(\mathbb{R})$ symétrique et définie positive, il existe une matrice B de $\mathcal{M}_n(\mathbb{R})$ symétrique et définie positive, telle que

$$B^2 = A. \tag{9.1}$$

Indication On diagonalisera A en justifiant que c'est possible.

2) Soient B une matrice symétrique définie positive et $\nu_1, \ldots, \nu_n$ ses valeurs propres.
a) Montrer qu'il existe un polynôme P de degré $\leq n-1$ tel que $P(\nu_i^2) = \nu_i$ pour $1 \leq i \leq n$.
b) Montrer que $P(B^2) = B$.
c) En déduire que B symétrique et définie positive vérifiant (9.1) est unique.

3) Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice inversible. Montrer que A peut s'écrire sous la forme

$$A = UB, \tag{9.2}$$

où U est unitaire et B est hermitienne définie positive.
Indication Pour montrer l'existence d'une telle décomposition, on partira d'une décomposition SVD de A.

4) Soient B_1 et B_2 deux matrices hermitiennes définies positives et U_1 et U_2 deux matrices unitaires telles que $A = U_1B_1 = U_2B_2$.
a) Montrer que $B_1^2 = B_2^2$.
b) En déduire que $B_1 = B_2$ puis que $U_1 = U_2$.
c) Que peut-on dire de B et U satisfaisant (9.2) ?

Exercice 3

Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$ antisymétrique, c'est à dire telle que

$$A^T = -A.$$

a) Montrer que A (vue comme une matrice de $\mathcal{M}_n(\mathbb{C})$) est diagonalisable.
b) Montrer que toutes ses valeurs propres sont imaginaires pures.
c) En déduire que pour tout réel α, $I_n + \alpha A$ est inversible.
d) Montrer que les matrices $(I_n - A)^{-1}$ et $(I_n + A)$ commutent, puis que $(I_n - A)^{-1}(I_n + A)$ est orthogonale.

Exercice 4

Soit Q une matrice orthogonale de $\mathcal{M}_3(\mathbb{R})$
a) Montrer que

$$\det(Q) = \pm 1$$

b)Expliquer pourquoi Q (vue comme matrice de $\mathcal{M}_3(\mathbb{C})$) est diagonalisable et pourquoi ses valeurs propres sont de module 1.
c) En déduire qu'il existe $p \in \{0, 1\}$ et $\theta \in [0, 2\pi[$ tel que Q est semblable à la matrice

$$\begin{pmatrix} (-1)^p & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

Exercice 5

Montrer que $A \in \mathbb{C}^{n\times n}$ est normale si et seulement si

$$\forall x \in \mathbb{C}^n, \quad \|Ax\|_2 = \|A^* x\|_2.$$

9.2 Partiel du 25 Novembre 2003.

Exercice 1 : Formes Quadratiques

1) Calculer la décomposition de Gauss de la forme quadratique sur $\mathbb{R}^4$.

$$q(x) = 2(x_1x_2 - 4x_1x_3 + x_2x_4 + 4x_3x_4) \tag{9.3}$$

Calculer la signature de q.

2) Donner une base orthogonale pour q.

3) (question indépendante des précédentes) On considère un entier $n > 4$. Démontrer que la forme quadratique sur R^n :

$$q(x) = 6\sum_{i=1}^{n} x_i^2 - 8\sum_{i=1}^{n-1} x_ix_{i+1} + 2\sum_{i=1}^{n-2} x_ix_{i+2}$$

est définie positive.
Suggestion On pourra calculer $(x_i - 2x_{i+1} + x_{i+2})^2$ et sommer de $i = 1$ à $n - 2$.

Exercice 2 : Décomposition en Valeurs Singulières

1) Donner une décomposition en valeurs singulières $P^TAQ = D$ de la matrice

$$A = \frac{1}{5}\begin{pmatrix} 3\sqrt{3} & -12 & 3 \\ 4\sqrt{3} & 9 & 4 \end{pmatrix}, \tag{9.4}$$

on calculera D, P et Q, et on donnera les valeurs singulières de A.

Exercice 3 : Décomposition de Jordan

1) Donner le polynôme minimal et une décomposition de Jordan de la matrice

$$A = \begin{pmatrix} 2 & \frac{3}{5} & 0 \\ -\frac{4}{5} & 2 & \frac{3}{5} \\ 0 & \frac{4}{5} & 2 \end{pmatrix}. \tag{9.5}$$

2) Résoudre l'équation différentielle dans $\mathbb{R}^3$:

$$\begin{array}{rcl} \frac{dU}{dt}(t) & = & AU(t) \quad t > 0, \\ U(0) & = & (1,1,1)^T. \end{array} \tag{9.6}$$

Exercice 4 : Inverses Généralisés

Soit A une matrice de $\mathcal{M}_{m,n}(\mathbb{C})$, le but de ce problème est l'étude des "inverses généralisés" de A.

1. (a) On désigne par $\text{rang}(A)$ le rang de A, montrer que
$$\ker(A^*A) = \ker(A), \quad \text{rang}(A^*A) = \text{rang}(A), \quad \text{et } \text{Im}(A^*A) = \text{Im}(A^*).$$
   (b) Montrer que A^*A est inversible lorsque $\text{rang}(A) = n$. Montrer que dans ce cas, la matrice $A_L = (A^*A)^{-1}A^*$ vérifie $A_L A = I_n$, matrice identité de $\mathcal{M}_n(\mathbb{C})$. On dit que A_L est une *inverse à gauche* de A. Montrer que pour toute matrice $V \in \mathcal{M}_{m,n}(\mathbb{C})$ telle que $\text{rang}(A^*VA) = n$, la matrice $A_{V,L} = (A^*VA)^{-1}A^*V$ est aussi une inverse à gauche de A.
   (c) Formuler et démontrer une proposition analogue quand $\text{rang}(A) = m$.
2. On se place maintenant dans le cas général où $\text{rang}(A) \leq \min(m, n)$. Montrer que pour toute matrice $B \in \mathcal{M}_{n,m}(\mathbb{C})$, les propriétés suivantes sont équivalentes :
   (a) Pour tout $c \in \text{Im}(A)$, $x = Bc$ est une solution de $Ax = c$.
   (b) $ABA = A$.
   (c) La matrice $C = BA$ est idempotente (*i.e.* $C^2 = C$) et $\text{rang}(C) = \text{rang}(A)$. **Suggestions.** On rappelle qu'une matrice $C \in \mathcal{M}_n(\mathbb{C})$ idempotente vérifie $\text{Im}(C) \oplus \ker(C) = \mathbb{C}^n$. On pourra comparer $\ker(BA)$ à $\ker(A)$ pour A vérifiant 2b.
   (d) La matrice $D = AB$ est idempotente et $\text{rang}(D) = \text{rang}(A)$. **Suggestions.** On pourra utiliser le fait que $ABA = A \Leftrightarrow A^*B^*A^* = A^*$ et utiliser le point 2c.

   Une matrice qui vérifie une de ces propriétés est appelée *inverse généralisé* de A.
3. Soit B une inverse généralisée de A, montrer que
   (a) $\text{rang}(B) \geq \text{rang}(A)$.
   (b) Les solutions x de l'équation $Ax = 0$ sont les vecteurs $x = (I_n - BA)z$ avec $z \in \mathbb{C}^n$. **Suggestions.** On pourra montrer que $\text{Im}(I_n - BA) \oplus \ker(I_n - BA) = \mathbb{C}^n$, $\text{Im}(I_n - BA) \subset \ker(A)$, et que $\ker(I_n - BA) \subset \text{Im}(BA)$.
   Quel est le rang de $I_n - BA$?
   (c) Une condition nécessaire et suffisante pour que l'équation $Ax = c$ admette une solution x est que $ABc = c$. Trouver alors toutes les solutions x de $Ax = c$.
4. On pose $\text{rang}(A) = r$. On sait que A a r valeurs singulières non nulles, donc on peut trouver une décomposition SVD de A : $V^*AU = \Sigma$, telle que
$$\Sigma = \begin{array}{c} \left[\begin{array}{cc} S & 0 \\ 0 & 0 \end{array}\right] \\ \begin{array}{cc} \leftrightarrow & \leftrightarrow \\ r & n-r \end{array} \end{array} \begin{array}{l} \updownarrow\ r \\ \updownarrow\ m-r \end{array} \tag{9.7}$$
   où $S \in \mathcal{M}_{r,r}(\mathbb{R})$ est diagonale inversible. Soit Σ^+ la matrice
$$\Sigma^+ = \begin{array}{c} \left[\begin{array}{cc} S^{-1} & 0 \\ 0 & 0 \end{array}\right] \\ \begin{array}{cc} \leftrightarrow & \leftrightarrow \\ r & m-r \end{array} \end{array} \begin{array}{l} \updownarrow\ r \\ \updownarrow\ n-r \end{array}$$
   Montrer que $U\Sigma^+V^*$ est un inverse généralisé de A.

9.3 Partiel 2004

Exercice 1 : Décomposition en Valeurs Singulières

Donner une décomposition en valeurs singulières $Q^T AP = D$ de la matrice

$$A = \begin{pmatrix} \frac{12}{5} & \frac{12}{5} \\ \frac{9}{5} & \frac{9}{5} \\ -5 & 5 \end{pmatrix}; \tag{9.8}$$

on calculera D, P et Q, et on donnera les valeurs singulières de A.

Exercice 2 : Réduction de Schur

Donner une décomposition de Schur de la matrice

$$A = \begin{pmatrix} 54 & 15 & -3 \\ 0 & 50 & 0 \\ -28 & 20 & 71 \end{pmatrix}. \tag{9.9}$$

Exercice 3 : Théorème de Hamilton-Cayley

Soit p matrices $A_i \in \mathcal{M}_n(\mathbb{R})$, $i = 1, \ldots, p$ telles que

$$\sum_{i=1}^{p} A_i = I_n, \tag{9.10}$$

et

$$A_i A_j = 0 \quad \text{si} \quad \text{i} \neq \text{j}. \tag{9.11}$$

a) Montrer que le polynôme $X^2 - X$ annule les matrices A_i , $i = 1, \ldots, p$.
b) En déduire que les matrices A_i sont diagonalisables dans une même base.

Exercice 4

Soit deux réels λ et μ non tous nuls. Soit $A \in \mathcal{M}_n(\mathbb{R})$ telle qu'il existe $B \in \mathcal{M}_n(\mathbb{R})$, $C \in \mathcal{M}_n(\mathbb{R})$ telles que

$$A = \lambda B + \mu C, \tag{9.12}$$
$$A^2 = \lambda^2 B + \mu^2 C, \tag{9.13}$$
$$A^3 = \lambda^3 B + \mu^3 C, \tag{9.14}$$

1) Soit P un polynôme de degré inférieur ou égal à 3 et tel que $P(0) = 0$. Montrer que

$$P(A) = P(\lambda)B + P(\mu)C.$$

2) En déduire un polynôme scindé avec des racines simples qui annule A. On distinguera les cas
- $0, \lambda, \mu$ sont deux à deux distincts
- deux des trois réels $0, \lambda, \mu$ coincident.

3) En déduire que A est diagonalisable.

4) En supposant que $\lambda\mu(\lambda - \mu) \neq 0$,
a) Exprimer A^3 en fonction de A^2 et A (on pourra utiliser la question 2)), puis A^4 en fonction de A et A^2.
b) Exprimer B et C en fonction de A et A^2.
c) En déduire que $B^2 = B$ et $C^2 = C$, puis que $BC = CB = 0$.
d) Montrer que $\ker(B) \oplus \mathrm{Im}(B) = \mathbb{R}^n$ et que $\ker(C) \oplus \mathrm{Im}(C) = \mathbb{R}^n$.
e) Montrer que $\mathrm{Im}(B) \oplus \mathrm{Im}(C) \oplus (\ker(B) \cap \ker(C)) = \mathbb{R}^n$.

Exercice 5 : Scilab

1) M étant une matrice, écrire une intruction scilab permettant de construire la matrice diagonale D dont la diagonale est celle de M.

2) On considère les lignes suivantes en langage scilab

```
for i=1:5
  for j=1:10
    t(i,j)=i/j;
  end
end
```

a) Modifier ces lignes en utilisant une opération vectorielle de manière à supprimer la boucle interne.
b) Obtenir la même matrice en utilisant uniquement des opérations vectorielles et la multiplication matricielle.

3) Écrire un programme scilab pour trouver la décomposition en valeurs singulières de la matrice A donnée par (9.8). **Indication** En annexe, on donne le résultat de la commande

```
help svd
```

Annexe

```
     Scilab Function
svd - singular value decomposition
Calling Sequence

s=svd(X)

[U,S,V]=svd(X)

[U,S,V]=svd(X,0) (obsolete)

[U,S,V]=svd(X,"e")

[U,S,V,rk]=svd(X [,tol])

Parameters

        0        X  : a real or complex matrix
        0        s  : real vector (singular values)
        0        S  : real diagonal matrix (singular values)
        0        U,V  : orthogonal or unitary square matrices (singular vectors).
        0        tol  : real number

Description

 [U,S,V] = svd(X)  produces a diagonal matrix  S  , of the same dimension as  X  and
 with nonnegative diagonal elements in decreasing order,
 and unitary matrices  U  and  V  so that  X = U*S*V' .

 [U,S,V] = svd(X,0)  produces the "economy size" decomposition.
 If  X  is m-by-n with m > n, then only the first n columns of  U  are computed
 and  S  is n-by-n.

 s = svd(X)  by itself, returns a vector  s  containing the singular values.

 [U,S,V,rk]=svd(X,tol)  gives in addition  rk , the numerical rank of  X
 i.e. the number of singular values larger than  tol .
 The default value of  tol  is the same as in  rank .
```

9.4 Partiel du 18 Novembre 2005

Exercice 1 : Valeurs singulières d'une matrice

1) Trouver la décomposition en valeurs singulières de la matrice A :

$$A = \begin{pmatrix} 12 & -60 & 16 \\ 9 & 80 & 12 \end{pmatrix}$$

On donnera la décomposition complète, c'est-à-dire deux matrices unitaire et une matrice diagonale. On indiquera les valeurs singulières de A.
Indication Les valeurs propres de la matrice

$$\begin{pmatrix} 32 & -36 \\ -36 & 53 \end{pmatrix}$$

sont 5 et 80. Pour travailler avec les plus petits entiers possibles, on aura le plus souvent intérêt à factoriser au maximum les expressions (en facteurs premiers).

2) On note

$$B = \begin{pmatrix} 11 & -2 \\ -5 & 10 \end{pmatrix}.$$

Calculer $\|B\|_1$, $\|B\|_2$, $\|B\|_\infty$, $\|B\|_F$.

Exercice 2 : Programmation en Scilab

Écrire en langage Scilab une fonction effectuant l'orthonormalisation au moyen de l'algorithme de Gram-Schmidt d'un ensemble de n vecteurs de $\mathbb{R}^n$. La fonction prendra comme argument la matrice formée par les n vecteurs mis en colonnes et renverra la matrice formée par les n vecteurs orthonormalisés mis en colonnes. On veillera à afficher à l'écran un message dans le cas où l'orthonormalisation est impossible.

Exercice 3 : Rang numérique d'une matrice

Soit $A \in \mathcal{M}_d(\mathbb{C})$. Le rang numérique de A est l'ensemble des nombres complexes

$$\mathcal{R}(A) = \left\{ \frac{x^* A x}{x^* x}, \quad x \in \mathbb{C}^d \backslash \{0\} \right\}.$$

1)
a) Montrer que $\mathcal{R}(A)$ contient les valeurs propres de A.
b) Montrer que pour tout $\alpha \in \mathbb{C}$

$$\mathcal{R}(A + \alpha I_d) = \{\mu + \alpha,\ \mu \in \mathcal{R}(A)\},$$

et que

$$\mathcal{R}(\alpha A) = \alpha \mathcal{R}(A).$$

c) Montrer que pour toute matrice unitaire U,

$$\mathcal{R}(U^* A U) = \mathcal{R}(A).$$

2) On suppose que le rang numérique de la matrice A est inclus dans le disque de centre c et de rayon r, i.e.

$$\mathcal{R}(A) \subset \{z \in \mathbb{C};\ \ |z-c| \leq r\},$$

avec $0 < r < |c|$.
a) Montrer en utilisant les résultats de 1.b) que

$$\mathcal{R}\left(I_d - \frac{1}{c}A\right) \subset \left\{z \in \mathbb{C}, |z| \leq \frac{r}{|c|}\right\}$$

b)En déduire que

$$\lim_{n\to\infty}\left(I_d - \frac{1}{c}A\right)^n = 0.$$

3) On suppose que A est une matrice normale. Soient $\lambda_1, \ldots, \lambda_d$ les valeurs propres de A, (les λ_i ne sont pas forcément deux à deux distincts).
Montrer que

$$\mathcal{R}(A) = \left\{ \sum_{i=1}^{d} m_i\lambda_i; \quad \left| \begin{array}{l} m_i \in [0,1]\ \ \forall i, 1 \leq i \leq d, \\ \\ \sum_{i=1}^{d} m_i = 1. \end{array} \right. \right\}.$$

Exercice 4 : Localisation des valeurs propres d'une matrice

Soit $A \in \mathcal{M}_d(\mathbb{C})$, $A = (a_{i,j})_{1\leq i,j\leq d}$.

1) Soit λ une valeur propre de A et $v \in \mathbb{C}^d$, $v \neq 0$, un vecteur propre associé : $Av = \lambda v$. Soit i_0 tel que $|v_{i_0}| = \|v\|_\infty$.
Montrer que

$$|\lambda - a_{i_0,i_0}| \leq \sum_{j=1, j\neq i_0}^{d} |a_{i_0,j}|.$$

2) En déduire que le spectre de A (i.e. l'ensemble des valeurs propres de A) est contenu dans l'ensemble

$$\bigcup_{i=1}^{d}\left\{ z \in \mathbb{C}, |z - a_{i,i}| \leq \sum_{j=1, j\neq i}^{d} |a_{i,j}| \right\}.$$

Quelle est la nature géométrique de cet ensemble ?

3)
a) En utilisant la définition du polynôme caractéristique et le fait que pour toute matrice $B \in \mathcal{M}_d(\mathbb{C})$, $\det(B^*) = \overline{\det(B)}$, montrer que λ est valeur propre de A si et seulement si $\bar{\lambda}$ est valeur propre de A^*.
b) En déduire que le spectre de A est contenu dans l'ensemble

$$\bigcup_{i=1}^{d}\left\{ z \in \mathbb{C}, |z - a_{i,i}| \leq \sum_{j=1, j\neq i}^{d} |a_{j,i}| \right\}.$$

9.5 Examen du 30 Janvier 2003.

Exercice 1

Soit B la matrice

$$B = \begin{pmatrix} \frac{1}{\sqrt{2}} & 1 & 1 \\ -\frac{1}{\sqrt{2}} & 1 & 1 \\ 0 & -\frac{3}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{pmatrix}$$

Calculer une décomposition en valeurs singulières de la matrice B.
c) Donner $\|B\|_1$,$\|B\|_2$, $\|B\|_\infty$ et $\|B\|_F$.

Exercice 2

On veut résoudre au sens des moindres carrés le système d'équations linéaires

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sqrt{3} \\ 0 \\ 1 \end{pmatrix} \qquad \text{où} \qquad A = \begin{pmatrix} 0 & \frac{\sqrt{3}}{2} \\ \frac{3}{5} & -1 \\ \frac{4}{5} & -\frac{1}{2} \end{pmatrix} \tag{9.15}$$

a) Donner la factorisation de Cholesky de $A^T A$.
b) Écrire et résoudre l'équation normale associée au problème aux moindres carrés.

Exercice 3

Soit $\lambda \in \mathbb{R}$ et soit A la matrice

$$A = \begin{pmatrix} 2 & \lambda & 0 \\ \lambda & 2 & \lambda \\ 0 & \lambda & 2 \end{pmatrix} \tag{9.16}$$

a) Écrire les matrices d'itérations $\mathcal{L}_J$ et $\mathcal{L}_{GS}$ des méthodes de Jacobi et de Gauss-Seidel pour résoudre un système $Ax = b$.
b) Discuter la convergence de ces deux méthodes en fonction de λ.

Exercice 4

Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$ à diagonale strictement dominante et dont les coefficients extra-diagonaux sont négatifs ou nuls. On dit que A est une M-matrice.
Soit D la matrice diagonale construite à partir de la diagonale de A. On appelle $-L$ la matrice triangulaire inférieure stricte obtenue à partir de A en annulant les coefficients de A au dessus de la diagonale (diagonale comprise), et U la matrice triangulaire supérieure telle que $A = D - L - U$.
Soient deux matrices L_1 et L_2 triangulaires inférieures strictes à coefficients positifs ou nuls telles que $L = L_1 + L_2$. De même, soient deux matrices U_1 et U_2 triangulaires supérieures strictes à coefficients positifs ou nuls telles que $U = U_1 + U_2$.
On pourra appeler $l_{1,i,j}$ les coefficients de L_1 et $l_{2,i,j}$ les coefficients de L_2. De même, on pourra appeler $u_{1,i,j}$ les coefficients de U_1 et $u_{2,i,j}$ les coefficients de U_2. On a $a_{i,j} = -l_{1,i,j} - l_{2,i,j}$ si

$i > j$ et $a_{i,j} = -u_{1,i,j} - u_{2,i,j}$ si $i < j$.
a) Montrer que

$$\|(D - L_1 - U_1)^{-1}(L_2 + U_2)\|_\infty < 1$$

Pour montrer l'inégalité ci-dessus, on montrera qu'il existe une constante $c \in \mathbb{R}$, $0 < c < 1$ telle que pour tout $x \in \mathbb{R}^n$, $x \neq 0$, on a $\|y\|_\infty \leq c\|x\|_\infty$ où y est défini par

$$(D - L_1 - U_1)y = (L_2 + U_2)x,$$

en écrivant les équations de ce système linéaire.
On a de la même manière

$$\|(D - L_2 - U_2)^{-1}(L_1 + U_1)\|_\infty < 1$$

b) En déduire que pour résoudre le système $AX = F$, la méthode itérative

$$\begin{aligned}(D - L_1 - U_1)X^{n+\frac{1}{2}} &= F + (L_2 + U_2)X^n \\ (D - L_2 - U_2)X^{n+1} &= F + (L_1 + U_1)X^{n+\frac{1}{2}}\end{aligned}$$

converge.

Exercice 5

On considère une matrice $A \in \mathcal{M}_n(\mathbb{R})$, symétrique définie positive. Soient $\lambda_1 \leq \cdots \leq \lambda_n$ les valeurs propres de A. Pour $x \in \mathbb{R}^n$, on note $r(x)$ le quotient de Rayleigh

$$r(x) = \frac{x^T Ax}{x^T x}. \tag{9.17}$$

1) Montrer que

$$\lambda_1 = \min_{x \in \mathbb{R}^n, x \neq 0} r(x), \quad \text{et} \quad \lambda_n = \max_{x \in \mathbb{R}^n, x \neq 0} r(x). \tag{9.18}$$

2) Soit $x \neq 0$. Montrer que $r(x+w) - r(x) = 2\frac{w^T Ax}{x^T x} - 2r(x)\frac{w^T x}{x^T x} + o(w)$, où $\frac{\|o(w)\|_2}{\|w\|_2} \to 0$ quand $w \to 0$. En déduire que

$$\text{grad } r(x) = \frac{2}{x^T x}(Ax - r(x)x). \tag{9.19}$$

Remarque 9.1 *On remarque que* grad $r(x) \in \text{Vect}(x, Ax)$.

3) Soient $(q_1, \ldots, q_j)$ une famille orthonormée de vecteurs de $\mathbb{R}^n$ et $H_j = \text{Vect}(q_1, \ldots, q_j)$. On note M_j et m_j les réels

$$M_j = \max_{y \in H_j, y \neq 0} r(y) \quad \text{et} \quad m_j = \min_{y \in H_j, y \neq 0} r(y) \tag{9.20}$$

Expliquer pourquoi on peut parler de max et de min dans (9.20) et montrer que

$$M_j \leq \lambda_n \quad \text{et que} \quad m_j \geq \lambda_1. \tag{9.21}$$

On notera u_j et v_j des vecteurs $u_j, v_j \in H_j$ pour lesquels le maximum et le minimum dans (9.20) sont respectivement atteints : $r(u_j) = M_j$, $r(v_j) = m_j$.

4) Montrer que si grad $r(u_j) \neq 0$ et si q_{j+1} est tel que

$$\text{grad } r(u_j) \in H_{j+1} = \text{Vect}(q_1, \ldots, q_j, q_{j+1}), \tag{9.22}$$

alors on a l'inégalité stricte

$$M_{j+1} > M_j. \tag{9.23}$$

De même, on peut vérifier que si grad $r(v_j) \neq 0$ *et si* q_{j+1} *est tel que*

$$\text{grad } r(v_j) \in H_{j+1}, \tag{9.24}$$

alors

$$m_{j+1} < m_j. \tag{9.25}$$

On va donc chercher à construire le vecteur q_{j+1} *tel que la famille* $(q_1, \ldots, q_{j+1})$ *soit orthonormée et tel que les conditions (9.22) et (9.24) soient vérifiées : si on y arrive, on a*

$$\lambda_1 \leq m_{j+1} < m_j < \ldots m_1 = M_1 < \cdots < M_j < M_{j+1} \leq \lambda_n$$

Il est alors raisonnable de penser que la suite M_j *(respectivement* m_j*) converge vers* λ_n *(respectivement* λ_1*).*
A priori, les conditions (9.22) et (9.24) semblent trop fortes pour définir le vecteur q_{j+1}*, mais on va voir qu'il n'en est rien.*

5) On choisit un vecteur normé q_1 et on appelle K_j l'espace $K_j = \text{Vect}(q_1, Aq_1, \ldots, A^{j-1}q_1)$ pour $1 \leq j \leq n$.
a) On va commencer par construire q_2 en fonction de q_1 : en supposant que Aq_1 n'est pas colinéaire à q_1, montrer que pour que les conditions (9.22) et (9.24) soient vérifiées pour $j = 1$, il faut que $q_2 \in K_2$. En déduire les choix possibles pour q_2 si $\dim(K_2) = 2$ et montrer qu'on a alors $H_2 = K_2$.
b) On suppose que pour tout $j \leq k$, on a $H_j = K_j$ et que $\dim(H_j) = j$. En supposant que $A^k q_1 \notin H_k$, montrer que pour que les conditions (9.22) et (9.24) soient vérifiées pour $j = k$, il faut que

$$H_{k+1} = K_{k+1}. \tag{9.26}$$

6) On suppose que pour $1 \leq j \leq n$, les sous-espaces K_j vérifient $\dim(K_j) = j$.
a) Montrer par récurrence que l'on peut choisir deux familles $(\alpha_1, \ldots, \alpha_{n-1})$ et $(\beta_1, \ldots, \beta_{n-1})$, les β_i étant pris strictement positifs, telles que les vecteurs définis par

$$Aq_j = \beta_{j-1}q_{j-1} + \alpha_j q_j + \beta_j q_{j+1}, \quad j < n \tag{9.27}$$

(en posant $q_0 = 0$ et $\beta_0 = 0$) forment une famille orthonormale. Il est clair qu'on obtient ainsi $q_{j+1} \in K_{j+1}$ pour que

$$Aq_n = \beta_{n-1}q_{n-1} + \alpha_n q_n, \tag{9.28}$$

c) Montrer que si $Q \in \mathcal{M}_n(\mathbb{R})$ est la matrice unitaire

$$Q = (q_1, q_2, \ldots, q_n)$$

et si $T \in \mathcal{M}_n(\mathbb{R})$ est la matrice tridiagonale

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & \dots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & \dots & 0 & \beta_{n-1} & \alpha_n \end{pmatrix}$$

alors $T = Q^T AQ$. En déduire que A et T ont les mêmes valeurs propres.
L'intérêt de l'algorithme ci-dessus, appelé algorithme de Lanczos, est qu'on se ramène à trouver les valeurs propres d'une matrice tridiagonale symétrique définie positive, ce qui peut être fait facilement à l'aide d'une méthode QR.

9.6 Examen du 28 Janvier 2004.

Exercice 1

a) Montrer que la matrice

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 10 & 7 \\ 2 & 7 & 21 \end{pmatrix}$$

est symétrique définie positive.
b) Calculer la factorisation de Cholesky de A.
c) Résoudre $Ax = b$ avec $b = (0, 0, 96)^T$ en utilisant b).

Exercice 2

Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$ inversible. On utilise la notation $A = D - E - F$ où D est la matrice diagonale construite à partir de la diagonale de A, $-E$ est la matrice triangulaire inférieure stricte obtenue à partir de A en annulant les coefficients de A au dessus de la diagonale (diagonale comprise), et $-F$ la matrice triangulaire supérieure obtenue à partir de A en annulant les coefficients de A au dessous de la diagonale (diagonale comprise).
Soit ω un nombre réel non nul.
Pour résoudre le système $Ax = b$, on considère la méthode itérative

$$\frac{1}{\omega} D x^{(k+1)} = b + (\frac{1-\omega}{\omega} D + E + F) x^{(k)}. \tag{9.29}$$

1) En quoi s'agit-il d'une généralisation de la méthode de Jacobi ? Quelle condition doit vérifier A pour que cette méthode puisse être appliquée ?

2) Établir une relation entre la matrice d'itération $\mathcal{L}_\omega$ de cette méthode et la matrice d'itération de la méthode de Jacobi $\mathcal{L}_J$. En déduire une relation entre les valeurs propres de $\mathcal{L}_\omega$ et celles de $\mathcal{L}_J$.

3) Soit μ un réel non nul et $M(\mu)$ la matrice tridiagonale de $\mathcal{M}_n(\mathbb{R})$

$$M(\mu) = \begin{pmatrix} b_1 & \mu^{-1}c_1 & 0 & \dots & \dots & 0 \\ \mu a_2 & b_2 & \mu^{-1}c_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \mu a_{n-1} & b_{n-1} & \mu^{-1}c_{n-1} \\ 0 & \dots & \dots & 0 & \mu a_n & b_n \end{pmatrix}$$

3.a) Montrer qu'il existe une matrice diagonale inversible $Q(\mu)$ telle que $M(\mu)Q(\mu) = Q(\mu)M(1)$.
3.b) En déduire que le déterminant de $M(\mu)$ ne dépend pas de μ.

4) Dans toute la suite de l'exercice, la matrice A est tridiagonale et symétrique définie positive. Montrer en utilisant la question précédente que $2D - A$ est symétrique définie positive. **Indication** On pourra montrer que toutes les valeurs propres de $2D - A$ sont positives. En déduire que la méthode de Jacobi converge, et donc que $\rho(\mathcal{L}_J) < 1$.

5) Déduire aussi de la question 3) que si λ est valeur propre de $\mathcal{L}_J$ alors $-\lambda$ l'est aussi.

6) Montrer alors que la méthode (9.29) converge si et seulement si ω appartient à l'intervalle $]0, \frac{2}{1+\rho(\mathcal{L}_J)}[$.

Exercice 3 : Méthode de résidu minimum généralisé

On considère une matrice inversible $A \in \mathcal{M}_n(\mathbb{R})$ (sans symétrie particulière), et on veut étudier une méthode itérative appelée *méthode de résidu minimum généralisé* (ou GMRES dans la littérature anglo-saxonne) pour résoudre le système linéaire $Ax = b$, où $b \in \mathbb{R}^n$. On approche donc la solution x par une suite $(x^{(k)})_{k\in\mathbb{N}}$. L'utilisateur choisit le vecteur $x^{(0)}$.

1) Principe de la méthode
On appelle *résidu* à la $k^{\text{ème}}$ itération, le vecteur

$$r^{(k)} = b - Ax^{(k)}.$$

On note $K^{(k)}$ l'espace de Krylov

$$K^{(k)} = \text{Vect}(r^{(0)}, \ldots, A^{k-1}r^{(0)}). \tag{9.30}$$

On cherche le vecteur $x^{(k)}$ sous la forme $x^{(k)} = x^{(0)} + z^{(k)}$ où $z^{(k)}$ résout le problème de moindres carrés

$$z^{(k)} \in K^{(k)} \text{ réalise le minimum sur } K^{(k)} \text{ de } z \mapsto \|r^{(0)} - Az\|_2^2. \tag{9.31}$$

1.a) Montrer que si $z^{(k)}$ vérifie $Az^{(k)} = r^{(0)}$, alors $x^{(k)} = x$.
1.b) Si $dim(K^{(k)}) = k$, soit $(v^{(1)}, \ldots, v^{(k)})$ une base de $K^{(k)}$. On peut chercher $z^{(k)}$ sous la forme $z^{(k)} = \sum_{j=1}^k y_j v^{(j)}$. Montrer qu'une condition nécessaire et suffisante pour que $z^{(k)}$ soit solution de (9.31) est que

$$\sum_{j=1}^{k} y_j (v^{(i)})^T A^T A v^{(j)} = (v^{(i)})^T A^T r^{(0)}, \quad \forall i = 1, \ldots, k \tag{9.32}$$

1.c) Écrire (9.32) sous la forme d'un système (équations normales)

$$MY = F, \tag{9.33}$$

où $Y = (y_1, \ldots, y_k)^T$, en précisant la matrice M et le second membre F.
1.d) Montrer que la matrice M est inversible.

2. Méthode d'Arnoldi Si la base $(v^{(j)})_{1\le j\le k}$ de $K^{(K)}$ ci-dessus est quelconque, la matrice M dans (9.33) n'a pas de structure particulière, et la complexité algorithmique de la résolution du système (9.33) rend la méthode inefficace. L'idée est de construire une base orthonormée $(v^{(j)})_{1\le j\le k}$ de $K^{(k)}$ par une méthode de Gram-Schmidt.
Si $r^{(0)} \neq 0$, on pose $v^{(1)} = \frac{r^{(0)}}{\|r^{(0)}\|_2}$, et on utilise la relation de récurrence

$$v^{(i+1)} = \frac{Av^{(i)} - \sum_{j=1}^{i} \left((v^{(i)})^T A^T v^{(j)}\right) v^{(j)}}{\|Av^{(i)} - \sum_{j=1}^{i} \left((v^{(i)})^T A^T v^{(j)}\right) v^{(j)}\|_2}, \tag{9.34}$$

quand la division est possible. Nous allons montrer qu'une division par zéro dans (9.34) signifie que la solution x de $Ax = b$ appartient à $K^{(i)}$, et qu'on a donc pas à calculer $v^{(i+1)}$.
2.a) Soit $k \geq 1$ un entier tel que $v^{(1}, \ldots, v^{(k)}$ sont définis. Montrer par récurrence que $(v^{(1}, \ldots, v^{(j)})$ est une base orthonormée de $K^{(j)}$, pour tout j, $1 \leq j \leq k$. En déduire qu'il existe un (unique) entier $p \leq n$ tel que $v^{(1}, \ldots, v^{(p)}$ sont définis et $Av^{(p)} - \sum_{j=1}^{p} \left((v^{(p)})^T A^T v^{(j)}\right) v^{(j)} = 0$.
2.b) Montrer que

$$AK^{(p)} \subset K^{(p)}. \tag{9.35}$$

2.c) Pour $j = 1, \ldots, p$, soit $V^{(j)} \in \mathcal{M}_{n,j}(\mathbb{R})$ la matrice dont les colonnes sont les vecteurs $v^{(1)}, \ldots, v^{(j)}$. Déduire de 2.a) et de (9.35) qu'il existe une matrice $H \in \mathcal{M}_p(\mathbb{R})$ telle que

$$AV^{(p)} = V^{(p)} H. \tag{9.36}$$

2.d) Montrer que H est inversible, en utilisant l'inversibilité de A.
2.e) Montrer que pour $1 \leq j \leq p$, $V^{(j)}$ vérifie

$$(V^{(j)})^T V^{(j)} = I_j. \tag{9.37}$$

2.f) Déduire des questions précédentes et de (9.31) que $r^{(p)} = 0$. On a donc montré que $Av^{(p)} - \sum_{j=1}^{p} \left((v^{(p)})^T A^T v^{(j)}\right) v^{(j)} = 0$ implique $r^{(p)} = 0$, et il est alors inutile de construire $K^{(p+1)}$.
2.g) Conclure que la méthode itérative donne la solution x en au plus n itérations.

3. Méthode des moindres carrés dans $K^{(k)}$
3.a) Soit $k < p$, où p est défini dans 2.a). Montrer que

$$AV^{(k)} = V^{(k+1)} H^{(k)}, \tag{9.38}$$

en donnant la matrice $H^{(k)} \in \mathcal{M}_{k+1,k}(\mathbb{R})$. On notera qu'on a en particulier

$$H^{(k)}_{l,m} = 0 \text{ si } l > m + 1. \tag{9.39}$$

3.b) Soit $\beta \in \mathbb{R}$ tel que $r^{(0)} = \beta v^{(1)} = \beta V^{(k+1)} e^1$, où $(e^j)_{j=1,\ldots,k+1}$ est la base canonique de $\mathbb{R}^{k+1}$. Montrer en utilisant (9.37) que le problème aux moindres carrés (9.31) est équivalent à trouver un vecteur $Y \in \mathbb{R}^k$ tel que

$$(H^k)^T H^{(k)} Y = \beta (H^{(k)})^T e^1. \tag{9.40}$$

4. Conclusion (cette partie ne contient pas de question)
À partir de l'observation (9.39), on peut très facilement factoriser $H^{(k)}$ sous la forme

$$H^{(k)} = Q^{(k)} R^{(k)}$$

où

- $Q^{(k)} \in \mathcal{M}_{k+1}(\mathbb{R})$ est une matrice unitaire (produit de $k+1$ rotations planes)
- $R^{(k)} \in \mathcal{M}_{k+1,k}(\mathbb{R})$ est une matrice triangulaire supérieure.

On peut alors résoudre le système (9.40) avec une complexité algorithmique de l'ordre de k^2.

9.7 Examen du 3 Septembre 2003.

Exercice 1

Soit B la matrice

$$B = \begin{pmatrix} \frac{3}{5} & -\frac{32}{25} & -\frac{24}{25} \\ \frac{4}{5} & \frac{24}{25} & \frac{18}{25} \\ 0 & -\frac{3}{5} & \frac{4}{5} \end{pmatrix}$$

1) Calculer les valeurs singulières de la matrice B. **Indication** On admettra que le noyau de $B^TB - I$ a pour base $((1,0,0)^T; (0,3,-4)^T)$.

2) Donner $\|B\|_1$, $\|B\|_2$, $\|B\|_\infty$ et $\|B\|_F$.

Exercice 2

1) Montrer comment obtenir une factorisation

$$A = UL$$

d'une matrice $A \in \mathcal{M}_N(\mathbb{R})$ où U est triangulaire supérieure avec des 1 sur la diagonale et L est triangulaire inférieure. On pourra procéder en $N-1$ étapes en annulant les coefficients de A colonne par colonne en commençant par la dernière colonne.

2) Donner une condition nécessaire et suffisante pour qu'une matrice $A \in \mathcal{M}_N(\mathbb{R})$ inversible possède une factorisation UL avec U est triangulaire supérieure avec des 1 sur la diagonale et L est triangulaire inférieure inversible.

3) Calculer la factorisation UL de la matrice

$$\begin{pmatrix} 4 & 4 & 3 \\ 4 & 2 & 4 \\ 1 & 0 & 1 \end{pmatrix}.$$

Exercice 3

Étant donnée une matrice $A \in \mathcal{M}_N(\mathbb{R})$ **symétrique définie positive** décomposée sous la forme $A = D - L - U$, (D diagonale, L strictement triangulaire inférieure, et $U = L^T$), on étudie la méthode itérative de résolution du système linéaire $Ax = b$: étant donné $x^{(0)}$ arbitraire, on définit la suite $x^{(k)}$ par :

$$\begin{aligned} (D-L)x^{(k+\frac{1}{2})} &= Ux^{(k)} + b, \\ (D-U)x^{(k+1)} &= Lx^{(k+\frac{1}{2})} + b, \end{aligned}$$

1) Écrire le vecteur $x^{(k+1)}$ sous la forme

$$x^{(k+1)} = Bx^{(k)} + c,$$

en donnant la matrice B et le vecteur c.

2) Soit λ une valeur propre de B : il existe $p \in \mathbb{C}^n$, $p \neq 0$ tel que $Bp = \lambda p$.
a) Montrer que

$$L(D-L)^{-1}Up = \lambda(A+L)p \tag{9.41}$$

b) En déduire que

$$LD^{-1}Up + LD^{-1}L(D-L)^{-1}Up = \lambda(A+L)p, \tag{9.42}$$

c) De (9.41) et (9.42), déduire que

$$LD^{-1}Up + \lambda LD^{-1}(A+L)p = \lambda(A+L)p, \tag{9.43}$$

puis que

$$(1-\lambda)LD^{-1}Up = \lambda Ap, \tag{9.44}$$

3)
a) En multipliant (9.44) à gauche par p^*, montrer que λ est un réel et que $0 < \lambda < 1$.
b) En déduire la convergence de la méthode. Cette méthode est une version symétrisée de la méthode de Gauss-Seidel.

Exercice 4

Soit β un réel strictement positif.
On considère la suite de vecteurs de $\mathbb{R}^N$ construite par la récurrence

$$E^{(m)} = (I - \beta A)E^{(m-1)}, \quad m > 0,$$

où $E^{(0)}$ est donné et A est la matrice tridiagonale dont les coefficients sont

$$\begin{array}{rcll} A_{j,j} & = & 2a + c, & j = 1, \ldots, N, \\ A_{j,j+1} & = & -a - b, & j = 1, \ldots, N-1, \\ A_{j,j-1} & = & -a + b, & j = 2, \ldots, N, \end{array}$$

où $a > 0$, $c \geq 0$ et b est un réel tel que $a^2 > b^2$.

1) Vérifier que pour $l = 1, \ldots, N$, si W_l est le vecteur de $\mathbb{R}^N$:

$$W_l = \left(\left(\frac{a-b}{a+b}\right)^{\frac{1}{2}} \sin(\frac{l\pi}{N+1}), \ldots, \left(\frac{a-b}{a+b}\right)^{\frac{k}{2}} \sin(\frac{lk\pi}{N+1}), \ldots, \left(\frac{a-b}{a+b}\right)^{\frac{N}{2}} \sin(\frac{Nl\pi}{N+1}) \right)^T$$

alors on a

$$AW_l \quad = \quad \lambda_l W_l, \quad \lambda_l = c + 2a - 2\sqrt{a^2 - b^2}\cos(\tfrac{l\pi}{N+1}). \tag{9.45}$$

2) En déduire que si $\beta > 2(c + 2a + 2\sqrt{a^2 - b^2})^{-1}$, alors pour N assez grand, la suite $E^{(m)}$ diverge en général.

3) a) Montrer que pour tout $U \in \mathbb{R}^N$, $U^T AU \geq 0$.
b) Montrer que pour tout réel positif β, $I + \beta A$ est inversible.
c) Montrer qu'il existe $\lambda > 0$ tel que pour tout $U \in \mathbb{R}^N$, $U^T AU \geq \lambda \|U\|_2^2$.
d) En déduire que que pour tout réel positif β,

$$\|(I + \beta A)^{-1}\|_2 < 1$$

e) Que peut on dire de la suite de vecteurs de $\mathbb{R}^N$ construite par la récurrence

$$F^{(m)} = (I + \beta A)^{-1} F^{(m-1)}, \quad m > 0,$$

où $F^{(0)}$ est donné ?

9.8 Examen Janvier 2005

Exercice 1

On considère le système d'équations linéaire :

$$\begin{aligned} 3x - \tfrac{1}{3}y &= 6, \\ \tfrac{4\sqrt{5}}{3}y &= 3\sqrt{5}, \\ 4x + 4y &= \tfrac{11}{2} \end{aligned}$$

Résoudre ce système au sens des moindres carrés. On devra utiliser la méthode de Choleski pour résoudre l'équation normale.

Exercice 2

Soit n, m deux entiers tels que $n \geq m > 0$. Soient
- $C_1 \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice symétrique définie positive,
- $C_2 \in \mathcal{M}_{m,m}(\mathbb{R})$ une matrice symétrique semi-définie positive,
- $D \in \mathcal{M}_{n,m}(\mathbb{R})$ une matrice de rang m.

Montrer qu'il existe
- une unique matrice $L_{11} \in \mathcal{M}_{n,n}(\mathbb{R})$ triangulaire inférieure avec des coefficients strictement positifs sur sa diagonale,
- une unique matrice $L_{21} \in \mathcal{M}_{m,n}(\mathbb{R})$
- une unique matrice $L_{22} \in \mathcal{M}_{m,m}(\mathbb{R})$ triangulaire inférieure avec des coefficients strictement positifs sur sa diagonale,

telles que

$$\begin{pmatrix} L_{11} & 0 \\ L_{21} & -L_{22} \end{pmatrix} \begin{pmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix} = \begin{pmatrix} C_1 & D \\ D^T & -C_2 \end{pmatrix}.$$

L'unicité devra être soigneusement démontrée.

Exercice 3

On considère la matrice inversible $A \in \mathcal{M}_d(\mathbb{R})$ et on veut résoudre le système $Au = b$ par la méthode semi-itérative décrite ci-dessous.
On considère la décomposition $A = M - N$ où M est inversible. Pour résoudre le système $Au = b$, on part de la méthode itérative

$$Mu^{(n+1)} = Nu^{(n)} + b, \tag{9.46}$$

qui est équivalente, en posant $L = M^{-1}N$, à

$$u^{(n+1)} = Lu^{(n)} + (I_d - L)u. \tag{9.47}$$

On se donne des paramètres réels $\alpha_{n,j}$, $n \in \mathbb{N}$, $j = 0, \ldots, n$, et on construit la suite $(y^{(n)})_{n\in\mathbb{N}}$ par

$$y^{(n)} = \sum_{j=0}^{n} \alpha_{n,j} u^{(j)}. \tag{9.48}$$

On veut choisir les coefficients réels $\alpha_{n,j}$, $n \in \mathbb{N}$, $j = 0, \ldots, n$ pour que la suite $(y^{(n)})_{n\in\mathbb{N}}$ converge vers la solution u du système si possible plus vite que la suite $(u^{(n)})_{n\in\mathbb{N}}$.

1) On dit que la méthode est consistante si pour tous vecteurs u, b tels que $Au = b$, pour la suite $u^{(n)}$ construite par (9.47) et la suite $y^{(n)}$ construite par (9.48), on a, pour tout $n \in \mathbb{N}$,

$$\Big(\forall i, 0 \leq i \leq n, \quad u^{(i)} = u\Big) \Rightarrow y^{(n)} = u.$$

Montrer que la méthode est consistante si et seulement si

$$\forall n \in \mathbb{N}, \quad \sum_{i=0}^{n} \alpha_{n,i} = 1. \tag{9.49}$$

2) a) On supposera désormais que la condition (9.49) est vérifiée. Montrer que la méthode semi-itérative (9.48) s'écrit

$$y^{(n)} = p_n(L)u^{(0)} + (I_d - p_n(L))u, \tag{9.50}$$

où $(p_n)_{n\in\mathbb{N}}$ est une suite de polynômes tels que

$$\text{degré}(p_n) \leq n, \quad \text{et} \quad p_n(1) = 1. \tag{9.51}$$

On donnera p_n en fonction des coefficients $\alpha_{n,j}$, $0 \leq j \leq n$.
b) Exprimer l'erreur $e^{(n)} = u - y^{(n)}$ en fonction de p_n, de L et de $e^{(0)}$.
c) En déduire qu'une condition nécessaire et suffisante pour que la méthode semi-itérative (9.48) (supposée consistante) converge vers la solution u quelle que soit l'initialisation $u^{(0)}$ est que

$$\lim_{n\to\infty} p_n(L) = 0.$$

d) Soit $\|.\|$ une norme vectorielle sur $\mathbb{R}^d$. On note aussi $\|.\|$ la norme matricielle subordonnée. Que dire de la convergence de $y^{(n)}$ vers u si

$$\lim_{n\to\infty} \|p_n(L)\|^{\frac{1}{n}} = r < 1?$$

e) On appelle $\rho(p_n(L))$ le rayon spectral de $p_n(L)$. Montrer que si L est normale, alors $p_n(L)$ l'est aussi. Montrer alors que si la suite $(\rho(p_n(L)))^{\frac{1}{n}}$ admet une limite quand n tend vers $+\infty$, alors

$$\lim_{n\to\infty} \|p_n(L)\|^{\frac{1}{n}} = \lim_{n\to\infty} (\rho(p_n(L)))^{\frac{1}{n}}. \tag{9.52}$$

Indication : on pourra utiliser une norme particulière et utiliser la relation d'équivalence avec $\|.\|$.
f) Supposons que 1 n'est pas une valeur propre de L. Montrer que pour $n \geq d$, le polynôme

$$p_n(x) = \frac{\det(xI_d - L)}{\det(I_d - L)}$$

vérifie (9.51) et $p_n(L) = 0$.
On ne peut cependant pas utiliser ce polynôme car son calcul est plus coûteux que la résolution approchée du problème de départ et car on souhaite utiliser moins de d itérations de (9.48).

4) a) Montrer qu'à toute suite de réels non nuls $(\mu_n)_{n\geq 1}$ et à toute suite de réels $(\nu_n)_{n\geq 1}$, on peut associer une suite de polynômes p_n par la récurrence

$$p_0(x)=1, \quad p_1(x)=(1-\frac{1}{2}\nu_1)x+\frac{1}{2}\nu_1, \quad p_n(x)=(\mu_n x+\nu_n)p_{n-1}(x)+(1-\mu_n-\nu_n)p_{n-2}(x),\ n\geq 2$$

qui vérifie (9.51).
b) Montrer que la suite $y^{(n)}$ correspondante vérifie la relation de récurrence

$$\begin{aligned} y^{(0)} &= u^{(0)}, \qquad y^{(1)}=(1-\frac{1}{2}\nu_1)(Lu^{(0)}+M^{-1}b)+\frac{1}{2}\nu_1 u^{(0)}, \\ y^{(n)} &= \mu_n(Ly^{(n-1)}+M^{-1}b-y^{(n-2)})+\nu_n(y^{(n-1)}-y^{(n-2)})+y^{(n-2)}. \end{aligned} \tag{9.53}$$

c) Expliquer pourquoi (9.53) est plus utilisable en pratique que (9.48).

5) Jusqu'à la fin du problème, on suppose que L est normale et que les valeurs propres de L sont réelles. Soient $a\leq b$ deux réels tels que, pour toute valeur propre λ de L, on a $a\leq\lambda\leq b$. On rappelle que la méthode itérative (9.46) converge si et seulement si chaque valeur propre λ vérifie $-1<\lambda<1$. Jusqu'à la fin du problème, on supposera donc que $-1<a\leq b<1$.
a) Démontrer que les fonctions $t_n:\mathbb{R}\mapsto\mathbb{R}$ définies par

$$\begin{array}{lcll} t_n(z) & = & \cos(n\arccos(z)) & \text{si } |z|\leq 1, \\ t_n(z) & = & \cosh(n\,\mathrm{arg\,cosh}(z)) & \text{si } z\geq 1, \\ t_n(z) & = & (-1)^n t_n(|z|) & \text{si } z<-1, \end{array} \tag{9.54}$$

sont telles que $\forall z\in\mathbb{R}$,

$$t_0(z)=1, \quad t_1(z)=z, \quad \forall n>1,\ t_n(z)+t_{n-2}(z)=2zt_{n-1}(z). \tag{9.55}$$

En déduire que t_n est un polynôme de degré n, dont le coefficient de degré n est 2^{n-1} si $n\geq 1$. Ces polynômes sont appelés polynômes de Chebyshev.
b) On rappelle que $\mathrm{arg\,cosh}(z)=\ln(z+\sqrt{z^2-1})$ si $z>1$. Montrer que si $z>1$,

$$t_n(z)=\frac{1}{2}\left((z+\sqrt{z^2-1})^n+(z+\sqrt{z^2-1})^{-n}\right). \tag{9.56}$$

c) On définit le polynôme

$$p_n(x)=\frac{1}{t_n\left(\frac{2-(a+b)}{b-a}\right)}t_n\left(\frac{2x-(a+b)}{b-a}\right). \tag{9.57}$$

Montrer que p_n vérifie (9.51) et que

$$\max_{x\in[a,b]}|p_n(x)|=\frac{1}{t_n\left(\frac{2-(a+b)}{b-a}\right)}=\frac{2}{\left(\frac{\sqrt{1-a}+\sqrt{1-b}}{\sqrt{b-a}}\right)^{2n}+\left(\frac{\sqrt{1-a}+\sqrt{1-b}}{\sqrt{b-a}}\right)^{-2n}}. \tag{9.58}$$

6) La méthode semi-itérative (9.48) ou de manière équivalente (9.50) où p_n est donné par (9.57) est appelée méthode semi-itérative de Chebyshev. On va étudier ses propriétés :
a) Calculer en fonction de b

$$\lim_{n\to\infty}\|p_n(L)\|^{\frac{1}{n}} \tag{9.59}$$

dans le cas où $a=0$, $b=\rho(L)<1$.
b) Dans le cas précédent avec l'hypothèse supplémentaire $b=\rho(L)=1-h$ avec $h\ll 1$, comparer les convergences de la méthode semi-itérative de Chebyshev et de la méthode (9.46).

Bibliographie

[1] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A multigrid tutorial*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2000.

[2] P.G. Ciarlet. *Introduction a l'analyse numerique matricielle et a l'optimisation*. Masson, 1988.

[3] G. Golub and C. Van Loan. *Matrix computations (second edition)*. John Hopkins, 1989.

[4] D. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison Wesley, 1973.

[5] G. Meurant. *Computer solution of large linear systems*, volume 28 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1999.

[6] Alfio Quarteroni. *Méthodes Numériques pour le Calcul Scientifique*. collection IRIS. Springer-Verlag France, 2000.

[7] Jacques Rappaz and Marco Picasso. *Introduction à l'analyse numérique*. Presses Polytechniques et Universitaires Romandes, Lausanne, 1998.

[8] Saad Y. *Iterative Methods for Sparse Linear Systems*. PWS Publishing company, 1996.