Notes du cours d'Analyse Numérique Matricielle

Daniele A. Di Pietro

A.A. 2012–2013

Table des matières

Chapitre 1

Rappels et compléments d'algèbre linéaire

1.1 Quelques définitions

Soit m et n deux entiers positifs et soit K un corps commutatif. Par la suite seuls les cas $K = \mathbb{R}$ et $K = \mathbb{C}$ seront considérés, et les éléments de K seront de ce fait appelés *scalaires*. Une *matrice* à m lignes et n colonnes est un ensemble de $m\,n$ scalaires (dits *éléments de la matrice*) indexés par les éléments du produit cartésien $I \times J$ avec $I := [\![1, m]\!]$ et $J := [\![1, n]\!]$: $a_{ij} \in K$, $1 \leq i \leq m$, $1 \leq j \leq n$. Les indices i et j sont dits, respectivement *ligne* et *colonne* de a_{ij}. En effet, on peut interpréter le couple d'indices (i, j) comme les coordonnées de l'élément a_{ij} dans le tableau suivant :

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Par la suite on utilisera la notation $A = (a_{ij})$ pour indiquer que l'élément générique de la matrice A est noté a_{ij} (les intervalles I et J ne seront précisés si on peut les déduire du contexte). La notation $(A)_{ij}$ pourra également être utilisée pour indiquer l'élément d'indices (i, j). Pour tout $1 \leq i \leq m$ on définit le *vecteur ligne* i de A (noté $A_{i:}$) comme suit :

$$A_{i:} := \begin{pmatrix} a_{i1} & a_{i2} & \dots & a_{im} \end{pmatrix} \in \mathbb{R}^m.$$

De manière analogue, pour tout $1 \leq j \leq m$ on définit le *vecteur colonne* j de A (noté $A_{:j}$) par

$$A_{:j} := \begin{pmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{jn} \end{pmatrix} \in \mathbb{R}^n.$$

On peut identifier chaque ligne i (resp. colonne j) de A avec le vecteur ligne $A_{i:}$ (resp. vecteur colonne $A_{:j}$) correspondant, et on parlera alors tout simplement de lignes et de colonnes de A. Une matrice qui a autant de lignes que de colonnes est dite *carrée*. Les matrices non carrées sont dites *rectangulaires*.

Définition 1.1 (Matrice triangulaire, strictement triangulaire et diagonale)**.** *Une matrice* $A \in \mathbb{C}^{n,n}$ *est dite* triangulaire supérieure *(resp.* inférieure*) si* $\big(i > j \implies a_{ij} = 0\big)$ *(resp.* $\big(i < j \implies a_{ij} = 0\big)$*) pour tout* $1 \leq i, j \leq n$*. Si les inégalité strictes sont remplacées par des inégalités simples (à savoir, on admet des éléments non nuls sur la diagonale de A) on parle alors de matrice* strictement triangulaire supérieure *(resp.* inférieure*). Une matrice (strictement) triangulaire inférieure ou supérieure est dite (strictement) triangulaire. La matrice A est dite* diagonale *si* $i \neq j \implies a_{ij} = 0$*. pour tout* $1 \leq i, j \leq n$*.*

Exercice 1.2 (Matrice triangulaire, strictement triangulaire et diagonale)**.** *Dire si les matrices suivantes sont triangulaires, strictement triangulaires ou diagonales :*

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Définition 1.3 (Matrice de Hessenberg)**.** *Une matrice* $A \in \mathbb{C}^{n+1,n}$ *est dite* de Hessenberg supérieure *si* $i > j+1 \implies a_{ij} = 0$*.*

Par la suite nous omettrons d'indiquer les éléments nuls des matrices triangulaires, diagonales ou de Hessenberg.

Définition 1.4 (Transposée et transconjuguée d'une matrice)**.** *Soit* $A = (a_{ij}) \in \mathbb{C}^{m,n}$*. On définit la* transposée *et la* transconjuguée *(ou* matrice adjointe*) de A respectivement par*

$$A^T := (a_{ji}) \in \mathbb{C}^{n,m}, \qquad A^H := (\overline{a}_{ji}) \in \mathbb{C}^{n,m}.$$

Example 1.5 (Transposée et transconjuguée)**.** *On considère la matrice suivante :*

$$A = \begin{pmatrix} 2+3i & 3+4i \\ 1+5i & 3+7i \end{pmatrix}.$$

Sa transposée et transconjuguée sont données, respectivement, par

$$A^T = \begin{pmatrix} 2+3i & 1+5i \\ 3+4i & 3+7i \end{pmatrix}, \qquad A^H = \begin{pmatrix} 2-3i & 1-5i \\ 3-4i & 3-7i \end{pmatrix}.$$

Soient $l, m, n \in \mathbb{N}_*$, $A = (a_{ij}) \in \mathbb{C}^{n,m}$, $B = (b_{ij}) \in \mathbb{C}^{m,l}$. On définit le *produit matriciel* $AB \in \mathbb{C}^{n,l}$ comme la matrice telle que

$$(AB)_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj} \qquad \forall i \in [\![1, n]\!], \forall j \in [\![1, l]\!].$$

Le produit matriciel est illustré par la Figure 1.1.

Proposition 1.6 (Propriétés du produit matriciel)**.** *Le produit matriciel est (i)* associatif*, à savoir, pour tout A, B, et C pour lesquelles les produits ont un sens, on a* $(AB)C = A(BC) = ABC$ *; (ii)* distributif par rapport à addition*, à savoir, pour toutes matrices A, B, C pour lesquelles l'écriture* $A(B+C)$ *a un sens on a* $A(B+C) = AB + AC$*.*

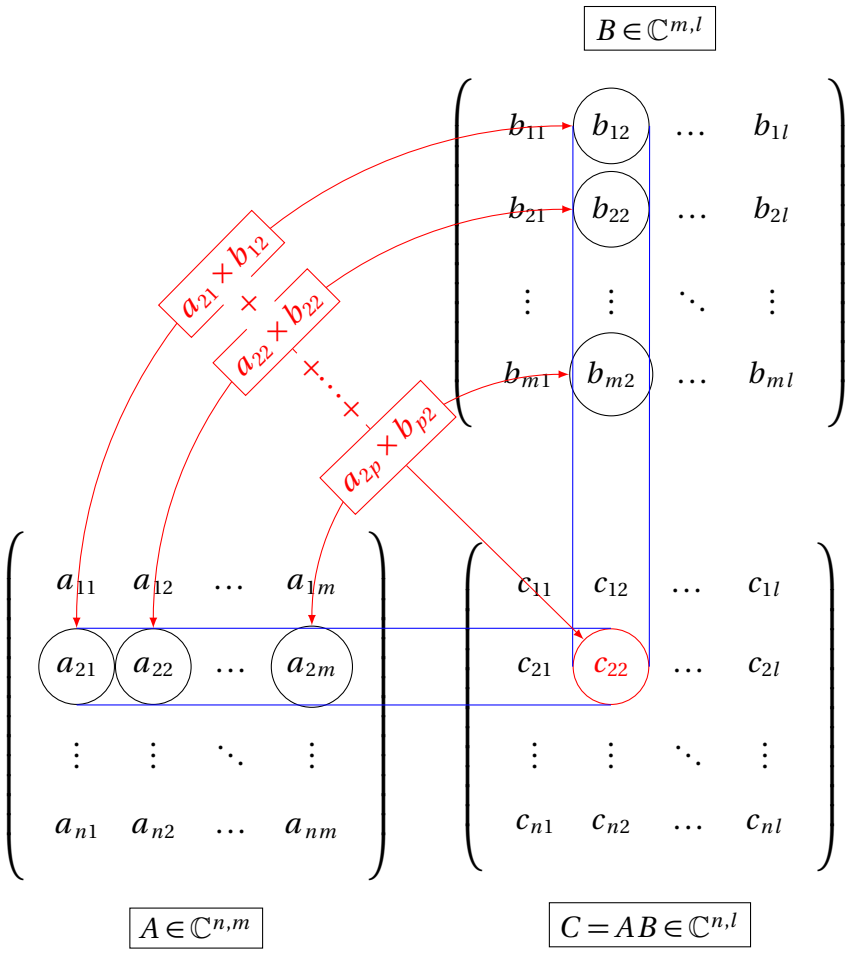


FIGURE 1.1 – Produit matriciel. Figure adaptée de Alain Matthes, `http://altermundus.com/pages/examples.html`

Il est important de retenir que le produit matriciel *n'est pas commutatif.*

Exercice 1.7 (Produit matriciel)**.** *Soient $A \in \mathbb{R}^{2,3}$ et $B \in \mathbb{R}^{3,3}$ définies comme suit :*

$$A := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \qquad B := \begin{pmatrix} 2 & 3 & 1 \\ 7 & 3 & 2 \\ 8 & 1 & 2 \end{pmatrix}.$$

On a

$$AB = \begin{pmatrix} 40 & 12 & 11 \\ 28 & 16 & 9 \end{pmatrix}.$$

Précisez si les produits BA et BA^T sont définis et, si c'est le cas, les calculer.

Proposition 1.8 (Produit de deux matrices triangulaires)**.** *Soient A, $B \in \mathbb{R}^{n,n}$ deux matrices triangulaires inférieures (resp. supérieures). Alors, AB est triangulaire inférieure (resp. supérieure).*

Démonstration. On détaille la preuve pour le cas triangulaire inférieure, l'autre cas pouvant se traiter de manière analogue. Par définition nous avons $(i < k \implies a_{ik} = 0)$ et $\big(k < j \implies b_{kj} = 0\big)$. Par conséquent,

$$i < j \implies (AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \sum_{k=1}^{i} a_{ik} \underbrace{b_{kj}}_{=0} + \sum_{k=i+1}^{n} \underbrace{a_{ik}}_{=0} b_{kj} = 0,$$

qui est le résultat cherché. □

Example 1.9 (Produit de deux matrices triangulaires)**.** *On considère les deux matrices triangulaires inférieures*

$$A = \begin{pmatrix} 1 & & \\ 2 & 3 & \\ 2 & 4 & 5 \end{pmatrix}, \qquad B = \begin{pmatrix} 8 & & \\ 2 & 4 & \\ 5 & 9 & 2 \end{pmatrix}.$$

On a

$$AB = \begin{pmatrix} 8 & & \\ 22 & 12 & \\ 49 & 61 & 10 \end{pmatrix}.$$

Proposition 1.10 (Transposé et transconjugué d'un produit)**.** *Pour tout $A \in \mathbb{C}^{n,m}$ et $B \in \mathbb{C}^{m,l}$ on a*

$$(AB)^H = B^H A^H, \qquad (AB)^T = B^T A^T.$$

Exercice 1.11 (Transposé et transconjugué d'un produit)**.** *Reprendre les matrices de l'Exercice 1.7 et vérifier que $(AB)^T = B^T A^T$.*

Définition 1.12 (Matrice hermitienne et symétrique)**.** *Une matrice complexe $A \in \mathbb{C}^{n,n}$ est dite* hermitienne *si $A^H = A$. Une matrice réelle $A \in \mathbb{R}^{n,n}$ est dite* symétrique *si $A^T = A$.*

Example 1.13 (Matrice hermitienne et symétrique)**.** *Les matrices suivantes sont, respectivement, hermitienne et symétrique.*

$$A = \begin{pmatrix} 1 & 1+i & 2+i \\ 1-i & 2 & 3+i \\ 2-i & 3-i & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 3 \end{pmatrix}.$$

Que peut-on dire des éléments diagonaux d'une matrice hermitienne ?

Définition 1.14 (Matrice normale, unitaire et orthogonale)**.** *Soit $A \in \mathbb{C}^{n,n}$. On dit que A est* normale *si $AA^H = A^H A$. Si, de plus, $AA^H = A^H A = I_n$ on dit que A est* unitaire. *Si $A \in \mathbb{R}^{n,n}$, on dit que A est* orthogonale *si $AA^T = A^T A = I_n$.*

Il est utile de rappeler que les produits scalaires canoniques sur $\mathbb{R}^n$ et $\mathbb{C}^n$ sont définis, respectivement, par

$$(x,y)_{\mathbb{R}^n} := y^T x, \quad \forall x, y \in \mathbb{R}^n, \qquad (x,y)_{\mathbb{C}^n} := y^H x, \quad \forall x, y \in \mathbb{C}^n.$$

Quand le contexte rende la notation non ambiguë, nous allons omettre l'indice pour le produit interne de $\mathbb{R}^n$.

Remarque 1.15 (Matrices unitaires et norme 2)**.** *La dénomination des matrices unitaires se justifie par la remarque suivante. Soit $\|\cdot\|_2$ la norme vectorielle sur $\mathbb{C}^n$ définie par $\|x\|_2^2 := (x,x)_{\mathbb{C}^n} = x^H x$. Alors, pour tout $x \in \mathbb{C}^n$,*

$$\|Ux\|_2^2 = (Ux, Ux)_{\mathbb{C}^n} = (Ux)^H Ux = x^H U^H U x = x^H x = \|x\|_2^2,$$

à savoir, la multiplication par une matrice unitaire ne modifie pas la norme d'un vecteur. *Plus généralement, pour tout $x, y \in \mathbb{C}^n$, on a*

$$(Ux, Uy)_{\mathbb{C}^n} = (Uy)^H Ux = y^H U^H U x = y^H x = (x,y)_{\mathbb{C}^n},$$

à savoir, les matrices unitaires préservent le produit scalaire.

Example 1.16 (Matrice élémentaire unitaire)**.** *Soit $w \in \mathbb{C}^n$ tel que $(w,w)_{\mathbb{C}^n} = w^H w = 1$. On définit*

$$A := I_n - 2ww^H. \tag{1.1}$$

Une matrice de la forme (1.1) *est dite* élémentaire. *On peut vérifier que A est unitaire. En effet*

$$\begin{aligned}
A^H A &= (I_n - 2ww^H)^H (I_n - 2ww^H) \\
&= (I_n - 2ww^H)(I_n - 2ww^H) && \textit{Proposition 1.10} \\
&= I_n - 4ww^H + 4(ww^H)(ww^H) && \textit{Proposition 1.6, distributivité} \\
&= I_n - 4ww^H + 4w\underbrace{(w^H w)}_{=1}w^H && \textit{Proposition 1.6, associativité} \\
&= I_n - 4ww^H + 4ww^H = I_n.
\end{aligned}$$

Définition 1.17 (Inverse d'une matrice)**.** *Soit $A \in \mathbb{C}^{n,n}$. On dit que A est* inversible *s'il existe $B \in \mathbb{C}^{n,n}$ telle que $AB = BA = I_n$. B est dit* inverse *de la matrice A et elle est notée A^{-1}.*

On vérifie aisément que l'inverse d'une matrice, si elle existe, est unique. Pour s'en convaincre, soit $A \in \mathbb{C}^{n,n}$ une matrice inversible et $B, C \in \mathbb{C}^{n,n}$ deux matrices telles que $AB = BA = AC = CA = I_n$. De par la Proposition 1.6 on a

$$AB = BA \implies CAB = CBA \implies (CA)B = C(BA) \implies I_n B = C I_n \implies B = C.$$

Par définition, en outre, toute matrice $A \in \mathbb{C}^{n,n}$ (resp. $A \in \mathbb{R}^{n,n}$) unitaire (resp. orthogonale) est inversible avec $A^{-1} = A^H$ (resp. $A^{-1} = A^T$).

Il est utile de considérer l'inverse de la transposée d'une matrice. On a, par définition,

$$I_n = (A^T)^{-1} A^T \iff I_n^T = \left[(A^T)^{-1} A^T\right]^T \iff I_n = A\left[(A^T)^{-1}\right]^T = AA^{-1},$$

à savoir $(A^{-1})^T = (A^T)^{-1}$. Cette remarque suggère la définition suivante.

Définition 1.18 (Inverse transposée)**.** *Soit $A \in \mathbb{C}^{n,n}$ inversible. On définit la matrice* inverse transposée *de A par $A^{-T} := (A^{-1})^T = (A^T)^{-1}$.*

Proposition 1.19 (Inverse d'un produit)**.** *Pour tout $A, B \in \mathbb{C}^{n,n}$ inversibles telles que AB est inversible on a*

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

1.2 Valeurs et vecteurs propres, rayon spectral déterminant

Définition 1.20 (Valeurs et vecteurs propres)**.** *Soit $A \in \mathbb{C}^{n,n}$. Un nombre $\lambda \in \mathbb{C}$ est une* valeur propre *de A s'il existe un vecteur $x \in \mathbb{C}^n$* non nul *tel que*

$$Ax = \lambda x.$$

On dit dans ce cas que $x \in \mathbb{C}^n$ est un vecteur propre *associé à la valeur propre λ. L'ensemble des valeurs propres d'une matrice A, noté $\lambda(A)$, est dit* spectre *de A.*

Les valeurs propres sont par définition les solutions de l'*équation caractéristique*

$$p_A(\lambda) := \det(A - \lambda I) = 0.$$

Comme p_A est un polynôme de degré n, il admet précisément n racines complexes (non nécessairement distinctes).

Proposition 1.21 (Valeurs propres de l'inverse d'une matrice)**.** *Pour toute matrice $A \in \mathbb{C}^{n,n}$ inversible on a*

$$\lambda \in \lambda(A) \implies \lambda^{-1} \in \lambda(A^{-1}). \tag{1.2}$$

Démonstration. On verra plus loin (Théorème 1.30) que l'inversibilité de A implique que toutes ses valeurs propres soient non nulles. Comme $\lambda \in \lambda(A)$, il existe $x \in \mathbb{C}^n$ non nul tel que

$$Ax = \lambda x \iff x = A^{-1}Ax = \lambda A^{-1}x \iff \lambda^{-1}x = A^{-1}x,$$

ce qui prouve que $\lambda^{-1} \in \lambda(A^{-1})$. □

Définition 1.22 (Matrice HDP et SDP)**.** *Une matrice $A \in \mathbb{C}^{n,n}$ est HDP si elle est* hermitienne *et* définie positive*, à savoir,*

$$(Ax, x)_{\mathbb{C}^n} > 0 \qquad \forall x \in \mathbb{R}^n.$$

Une matrice $A \in \mathbb{R}^{n,n}$ est SDP si elle est symétrique *et* définie positive.

Proposition 1.23 (Valeurs propres d'une matrice HDP/SDP)**.** *Soit $A \in \mathbb{C}^{n,n}$ (resp. $A \in \mathbb{R}^{n,n}$) une matrice HDP (resp. SDP). Alors, toutes les valeurs propres de A sont réelles et strictement positives.*

Démonstration. On détaille la preuve uniquement pour le cas HDP, l'autre étant similaire. Soit λ une valeur propre de A et $x \in \mathbb{C}^n$ un vecteur propre associé. Alors,

$$Ax = \lambda x \implies 0 < x^H Ax = \lambda x^H x = \lambda \|x\|_2^2 \implies \lambda = \frac{x^H Ax}{\|x\|_2^2} \in \mathbb{R}_*^+.$$

□

Nous admettrons le résultat suivant.

Proposition 1.24 (Valeurs propres d'une matrice triangulaire)**.** *Les valeurs propres d'une matrice triangulaire sont ses éléments diagonaux.*

Définition 1.25 (Rayon spectral)**.** *Soit $A \in \mathbb{C}^{n,n}$. On définit son* rayon spectral *comme la plus grande valeur propre en valeur absolue,*

$$\rho(A) := \max_{\lambda \in \lambda(A)} |\lambda|.$$

Le rayon spectral n'est pas une norme matricielle (voir Section 1.5), car on peut avoir $\rho(A) = 0$ sans que A soit nulle (il suffit de prendre une matrice triangulaire non nulle avec diagonale nulle pour s'en convaincre).

Définition 1.26 (Déterminant)**.** *Soit $A \in \mathbb{C}^{n,n}$. Le* déterminant *de A est donné par*

$$\det(A) := \prod_{\lambda \in \lambda(A)} \lambda.$$

Le déterminant d'une matrice permet de décider de son inversibilité. Plus précisément, une matrice est inversible si et seulement si son déterminant est non nul. De par la Proposition 1.21 on a

$$\det(A^{-1}) = \prod_{\lambda \in \lambda(A^{-1})} \lambda = \prod_{\tilde{\lambda} \in \lambda(A)} \frac{1}{\tilde{\lambda}} = \frac{1}{\det(A)}.$$

1.3 Noyau et image d'une matrice

Soit $A \in \mathbb{R}^{m,n}$ et $\mathscr{C}_A$ le sous-espace vectoriel de $\mathbb{R}^m$ engendré par ses colonnes,

$$\mathscr{C}_A := \text{span}(A_{:j})_{1 \le j \le n}.$$

L'*image* de A est le sous-espace vectoriel de $\mathbb{R}^m$ défini par

$$\text{range}(A) := \{y \in \mathbb{R}^m \mid y = Ax,\, x \in \mathbb{R}^n\}.$$

Le *noyau* de A est la préimage du vecteur nul de $\mathbb{R}^m$ par l'application linéaire associée à A,

$$\ker(A) := \{x \in \mathbb{R}^n \mid Ax = 0 \in \mathbb{R}^m\}.$$

Proposition 1.27 (Image d'une matrice)**.** *On a* $\mathscr{C}_A = \text{range}(A)$.

Démonstration. Il suffit de remarquer que le résultat de la multiplication (à droite) d'une matrice $A \in \mathbb{R}^{m,n}$ par un vecteur $x \in \mathbb{R}^n$ est la combinaison linéaire des colonnes de la matrice avec coefficients donnés par les composantes de x,

$$(Ax)_i = \sum_{j=1}^{n} A_{ij} x_j \iff Ax = \sum_{j=1}^{n} A_{:j} x_j.$$

□

Définition 1.28 (Rang d'une matrice)**.** *On définit le* rang *d'une matrice comme la dimension de son image,*

$$\text{rank}(A) := \dim(\text{range}(A)).$$

Proposition 1.29 (Propriétés du rang d'une matrice)**.** *Soit* $A \in \mathbb{R}^{m,n}$*. Nous avons* $\text{rank}(A) = \text{rank}(A^T)$ *et*

$$\text{rank}(A) + \dim(\ker(A)) = n.$$

Le résultat suivant, que l'on admettra, établit un lien entre le déterminant, le rang, et le noyau d'une matrice et son inversibilité.

Théorème 1.30 (Caractérisation des matrices inversibles)**.** *Soit* $A \in \mathbb{R}^{n,n}$*. Les propriétés suivantes sont équivalentes : (i)* A *est inversible ; (ii)* $\det(A) \neq 0$ *; (iii)* $\ker(A) = \{0 \in \mathbb{R}^n\}$ *; (iv)* $\text{rank}(A) = n$ *; (v) les colonnes et les lignes de* A *sont linéairement indépendantes.*

Les résultats de cette section s'étendent sans difficulté au cas de matrices complexes.

1.4 Décompositions d'une matrice

Il est souvent utile de décomposer une matrice en un produit de plusieurs matrices. Dans cette section nous allons rappeler quelques décompositions importantes dans les applications numériques.

1.4.1 Matrices diagonalisables

Définition 1.31 (Matrice diagonalisable)**.** *Une matrice $A \in \mathbb{C}^{n,n}$ est* diagonalisable *s'il existe une matrice $Q \in \mathbb{C}^{n,n}$ inversible telle que $A = Q^{-1}\Lambda Q$ avec $\Lambda \in \mathbb{C}^{n,n}$ matrice diagonale.*

Proposition 1.32 (Valeurs et vecteurs propres d'une matrice diagonalisable)**.** *Soit $A \in \mathbb{C}^{n,n}$ une matrice diagonalisable. Alors, en reprenant la notation de la Définition 1.31, Λ contient les valeurs propres de A et les colonnes de Q^{-1} sont des vecteurs propres associés.*

Démonstration. En multipliant à droite par Q^{-1} la relation $A = Q^{-1}\Lambda Q$ on obtient $AQ^{-1} = Q^{-1}\Lambda$, ou, de façon équivalente,

$$AQ^{-1}_{:j} = Q^{-1}\Lambda_{:j} = \lambda_j Q^{-1}_{:j} \qquad \forall 1 \le j \le n,$$

où nous avons noté $\lambda_j = (\Lambda)_{jj}$. □

La Proposition 1.32 suggère le résultat suivant, que nous admettrons.

Théorème 1.33 (Caractérisation d'une matrice diagonalisable)**.** *Une matrice $A \in \mathbb{C}^{n,n}$ (resp. $A \in \mathbb{R}^{n,n}$) est diagonalisable si et seulement si on peut construire une base de $\mathbb{C}^n$ (resp. $\mathbb{R}^n$) formée de vecteurs propres de A.*

Dans certains cas, on peut prouver des propriétés additionnelle pour la matrice Q qui apparaît dans la Définition 1.31. Un exemple à retenir est donné dans le lemme suivant, qui affirme que les matrices normales sont les seules matrices unitairement semblables à des matrices diagonales.

Lemme 1.34 (Diagonalisation d'une matrice normale)**.** *La matrice $A \in \mathbb{C}^{n,n}$ est normale si et seulement s'il existe une matrice unitaire $U \in \mathbb{C}^{n,n}$ telle que*

$$U^{-1}AU = U^H AU = \Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_n),$$

avec $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ valeurs propres de A.

1.4.2 Décomposition de Schur

Pour une matrice générique, on peut prouver uniquement qu'elle est unitairement semblable à une matrice triangulaire, comme l'affirme le résultat suivant.

Théorème 1.35 (Décomposition de Schur (DS))**.** *Pour toute $A \in \mathbb{C}^{n,n}$ il existe $U \in \mathbb{C}^{n,n}$ unitaire telle que*

$$U^{-1}AU = U^H AU = \begin{pmatrix} \lambda_1 & b_{12} & \cdots & b_{1n} \\ & \lambda_2 & & b_{2n} \\ & & \ddots & \vdots \\ & & & \lambda_n \end{pmatrix} := T,$$

où $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ sont les valeurs propres de A.

On pourra remarquer que, si T et une matrice carrée triangulaire supérieure, sa décomposition de Schur s'obtient en posant $U = I_n$. Ceci implique, en particulier, que les valeurs propres de T sont ses éléments diagonaux. Ce résultat s'applique également aux matrices triangulaires inférieures.

Le corollaire suivant montre une autre conséquence importante de la DS, à savoir, toute matrice hermitienne est diagonalisable.

Corollaire 1.36 (Valeurs propres et DS d'une matrice hermitienne)**.** *Si la matrice $A \in \mathbb{C}^{n,n}$ est hermitienne, T est diagonale, $\lambda_i \in \mathbb{R}$ pour tout $1 \leq i \leq n$, et les lignes de U (où, de manière équivalent, les colonnes de $U^H = U^{-1}$) sont des vecteurs propres de A. Si $A \in \mathbb{R}^{n,n}$, la matrice U est à valeurs réelles et on écrit U^T au lieu de U^H.*

Démonstration. Comme A est hermitienne on a

$$T^H = (U^H A U)^H = U^H A^H U = U^H A U = T,$$

ce qui implique $T = \text{diag}(\lambda_1, \ldots, \lambda_n)$ et, pour tout $1 \leq i \leq n$, $\lambda_i = \overline{\lambda_i} \iff \Im(\lambda_i) = 0 \iff \lambda_i \in \mathbb{R}$. Pour prouver que les lignes de U sont des vecteurs propres on procède comme dans la preuve de la Proposition 1.32 en observant que $U^{-1} = U^H$. □

1.4.3 Décomposition en valeurs singulières

Toute matrice $A \in \mathbb{C}^{m,n}$ peut être transformée en une matrice diagonale rectangulaire à l'aide de deux matrices unitaires.

Théorème 1.37 (Décomposition en valeurs singulières (DVS))**.** *Pour toute $A \in \mathbb{C}^{m,n}$ il existent deux matrices unitaires $U \in \mathbb{C}^{m,m}$ et $V \in \mathbb{C}^{n,n}$ telles que*

$$U^H A V = \Sigma = \text{diag}(\sigma_1, \ldots \sigma_p) \in \mathbb{R}^{m,n} \qquad p := \min(m, n).$$

Les réels $0 \leq \sigma_1 \leq \ldots \leq \sigma_p$ sont dits valeurs singulières *de A. Si, de plus, $A \in \mathbb{R}^{m,n}$, on a $U \in \mathbb{R}^{m,m}$ et $V \in \mathbb{R}^{n,n}$ et on peut écrire $U^T A V = \Sigma$.*

Proposition 1.38 (Caractérisation des valeurs singulières)**.** *Soit $A \in \mathbb{C}^{n,n}$. On a*

$$\sigma_i(A) = \sqrt{\lambda_i(A^H A)} \qquad \forall 1 \leq i \leq n.$$

Démonstration. On a

$$\Sigma^H \Sigma = (U^H A V)^H U^H A V = V^H A^H U^H U A V = V^H A^H A V = V^{-1} A^H A V$$

où nous avons utilisé le fait que U et V sont unitaires pour conclure $U^H U = I_n$ et $V^H = V^{-1}$. La matrice diagonale $\Sigma^H \Sigma$ contient les valeurs propres de $A^H A$ et V ses vecteurs propres. Pour s'en convaincre, il suffit de procéder comme dans la preuve de la Proposition 1.32. Comme $\Sigma^H \Sigma = \text{diag}(\sigma_i^2(A))_{1 \leq i \leq n}$ et $\sigma_i(A) \geq 0$ pour tout $1 \leq i \leq n$, on a donc

$$\lambda_i(A^H A) = \sigma_i(A)^2 \iff \sigma_i(A) = \sqrt{\lambda_i(A^H A)}.$$

□

Corollaire 1.39 (Valeurs singulières et rayon spectral d'une matrice hermitienne)**.** *Si $A \in \mathbb{C}^{n,n}$ est une matrice hermitienne, on a*

$$\sigma_i(A) = \sqrt{\lambda_i(A)^2} = |\lambda_i(A)|, \qquad \rho(A) = \max_{1\leq i\leq n} \sigma_i(A) = \|A\|_2.$$

Démonstration. Comme A est hermitienne on a $A^H A = A^2$. Or, soit λ un vecteur propre de A et $x \in \mathbb{C}^n$ un vecteur propre associé. Alors

$$Ax = \lambda x \implies AAx = \lambda Ax = \lambda^2 x,$$

à savoir, $\lambda_i(A^H A) = \lambda_i(A^2) = \lambda_i(A)^2 = \sigma_i(A)^2$ pour tout $1 \leq i \leq n$. De par le Corollaire **??** on a de plus

$$\max_{1\leq i\leq n} \sigma_i(A) = \max_{1\leq i\leq n} |\lambda_i(A)| = \rho(A).$$

La conclusion s'ensuit de la Proposition 1.44. □

1.5 Normes matricielles

Définition 1.40 (Norme matricielle)**.** *Soit K un corps commutatif. Une norme matricielle sur $K^{n,n}$ est une application $\|\cdot\|$ de $K^{n,n}$ dans $\mathbb{R}$ telle que, pour tout $A \in K^{n,n}$,*

(i) $\|A\| > 0$ *si* $A \neq 0$ *et* $\|A\| = 0$ *si et seulement si* $A = 0_{K^{n,n}}$ *;*

(ii) $\|\alpha A\| = |\alpha|\|A\|$ *pour tout* $\alpha \in K$ *;*

(iii) $\|A + B\| \leq \|A\| + \|B\|$ *pour tout* $B \in K^{n,n}$*.*

Dans le reste de cette section nous allons supposer $A \in \mathbb{C}^{n,n}$ sans nécessairement le préciser à chaque fois.

Définition 1.41 (Norme matricielle subordonnée à une norme vectorielle)**.** *Soit $\|\cdot\|$ une norme vectorielle sur $\mathbb{C}^n$. On définit la* norme matricielle subordonné à la norme vectorielle *$\|\cdot\|$ par*

$$\forall A \in \mathbb{C}^{n,n}, \qquad \|A\| := \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Des normes matricielles particulièrement importantes sont les normes p subordonnées aux normes vectorielles définies par

$$\forall x \in \mathbb{C}^n, \qquad \|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}},$$

avec $p \in \mathbb{N}_*$ et

$$\|x\|_\infty := \max_{1\leq i\leq n} |x_i|.$$

Remarque 1.42 (Définition alternatives d'une norme subordonnée)**.** *Soit $\|\cdot\|$ une norme vectorielle sur $\mathbb{C}^n$. On prouve aisément que la norme matricielle subordonnée à $\|\cdot\|$ vérifie pour tout $A \in \mathbb{C}^{n,n}$,*

$$\|A\| = \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\| = \sup_{x\in\mathbb{C}^n, \|x\|\leq 1} \|Ax\|.$$

Proposition 1.43 (Norme d'un produit matrice-vecteur)**.** *Soit $\|\cdot\|$ une norme matricielle subordonnée à la norme vectorielle $\|\cdot\|$. Alors pour toute matrice $A \in \mathbb{C}^{n,n}$ et tout vecteur $x \in \mathbb{C}^n$,*

$$\|Ax\| \leq \|A\|\|x\|.$$

Démonstration. Par définition on a

$$\|A\| := \sup_{y \in \mathbb{C}^n, y \neq 0} \frac{\|Ay\|}{\|y\|} \geq \frac{\|Ax\|}{\|x\|},$$

où la deuxième inégalité vient de la définition de supremum. □

Proposition 1.44 (Normes 1, 2 et ∞)**.** *On a*

$$\|A\|_2 = \|A^H\|_2 = \max_{1 \leq i \leq n} \sigma_i(A),$$

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|,$$

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Démonstration. On commence par remarquer que la matrice $A^H A$ est normale. Elle admet donc une DVS, à savoir, il existe une matrice unitaire $U \in \mathbb{C}^{n,n}$ telle que $U^H A^H A U = \Lambda = \text{diag}(\lambda_1(A^H A), \ldots \lambda_n(A^H A))$ où on a noté $\lambda_i(A^H A)$, $1 \leq i \leq n$, les valeurs propres de $A^H A$. Nous avons alors

$$\begin{aligned}
\|A\|_2^2 &= \sup_{x \in \mathbb{C}^n, \|x\|_2 = 1} \|Ax\|_2^2 = \sup_{x \in \mathbb{C}^n, \|x\|_2 = 1} x^H A^H A x \\
&= \sup_{x \in \mathbb{C}^n, \|x\|_2 = 1} x^H U^H \Lambda U x && A^H A = U^H \Lambda U \\
&= \sup_{y \in \mathbb{C}^n, \|U^H y\|_2 = 1} y^H \Lambda y && y := Ux \\
&= \sup_{y \in \mathbb{C}^n, \|y\|_2 = 1} y^H \Lambda y = \max_{1 \leq i \leq n} \lambda_i(A^H A) = \max_{1 \leq i \leq n} \sigma_i(A), && \text{Remarque 1.15}
\end{aligned}$$

ce qui prouve le premier point. [COMPLETER] □

Proposition 1.45 (Propriétés des normes subordonnées)**.** *Soit $\|\cdot\|$ une norme matricielle subordonnée sur $\mathbb{C}^{n,n}$. Alors, (i) pour toute matrice $A \in \mathbb{C}^{n,n}$ il existe $x_A \in \mathbb{C}^n \setminus \{0\}$ tel que $\|A\| = \|Ax_A\|/\|x_A\|$; (ii) on a $\|I_n\| = 1$; (iii) pour toutes matrices A, $B \in \mathbb{C}^{n,n}$, $\|AB\| \leq \|A\|\,\|B\|$.*

Démonstration. (i) La fonction $\|Ax\|$ définie sur $\mathbb{C}^n$ et à valeurs réels est par définition continue car $\|Ax\| \leq \|A\|\|x\|$. Par conséquent, elle atteint son maximum sur le compact $\|x\| = 1$. (ii) On a par définition $\|I_n\| = \sup_{x \in \mathbb{C}^n, \|x\| = 1} \|I_n x\| = 1$. (iii) Conséquence immédiate de la définition de norme matricielle. □

On amdettra le résultat suivant qui relie le rayon spectral avec les normes matricielles subordonnées.

Lemme 1.46 (Rayon spectral et normes subordonnées)**.** *Soit $\|\cdot\|$ une norme matricielle subordonnée. Alors, pour toute matrice $A \in \mathbb{C}^{n,n}$, $\rho(A) \leq \|A\|$. Reciproquement, pour toute matrice A et tout réel $\epsilon > 0$, il existe une norme subordonnée $\|\cdot\|$ telle que*

$$\|A\| \leq \rho(A) + \epsilon.$$

1.6 Exercices

Exercice 1.47 (Matrice SDP)**.** *Soit $A=(a_{ij})\in\mathbb{R}^{n,n}$ une matrice SDP. Prouver que $a_{kk}>0$ pour tout $1\leq k\leq n$.*

Comme A est SDP, on a pour tout $x\in\mathbb{R}^n$

$$x^TAx=(Ax,x)=(x,x)_A>0.$$

En prénant $x=e^k$ avec e^k k-ème vecteur de la base canonique de $\mathbb{R}^n$ tel que $e^k_l=\delta_{kl}$ pour tout $1\leq l\leq n$, nous avons

$$0<(Ae^k,e^k)=\sum_{i=1}^n e^k_i\sum_{j=1}^n a_{ij}e^k_j=\sum_{i=1}e^k_i a_{ik}=a_{kk},$$

qui est le résultat cherché.

Exercice 1.48 (Matrice antisymétrique)**.** *Soit $A\in\mathbb{R}^{n,n}$ une matrice* antisymétrique*, à savoir $A^T=-A$. On pose*

$$B_\pm:=I_n\pm A.$$

On supposera B_- inversible. (i) Vérifier que B_+ est inversible et identifier son inverse. (ii) Vérifier que B_- est normale. (iii) Montrer que la matrice $C:=B_+B_-^{-1}$ est orthogonale, à savoir, $B^{-1}=B^T$.

(i) Comme A est antisymétrique on a

$$B_+=B_-^T. \tag{1.3}$$

Par conséquent

$$I_n=B_-^{-1}B_-\implies I_n=I_n^T=(B_-^{-1}B_-)^T=B_-^TB_-^{-T}=B_+B_-^{-T},$$

à savoir, B_+ est inversible avec $B_+^{-1}=B_-^{-T}$.

(ii) Un calcul direct montre que

$$B_-^TB_-=B_+B_-=I_n-A^2,\qquad B_-B_-^T=B_-B_+=I_n-A^2,$$

ce qui prouve que B_- est normale car nous avons $B_-^TB_-=B_-B_-^T$.

(iii) Comme B_- est normale, d'après le Lemme 1.34 elle est unitairement semblable à une matrice diagonale, à savoir, il existe $U\in\mathbb{R}^{n,n}$ orthogonale et $\Lambda\in\mathbb{R}^{n,n}$ diagonale telles que

$$\begin{aligned}
B_-&=U^T\Lambda U, && \text{Lemme 1.34}\\
B_+&=B_-^T=U^T\Lambda U, && \text{eq. (1.3)}\\
B_-^{-1}&=U^{-1}\Lambda^{-1}U^{-T}=U^T\Lambda^{-1}U, && \text{Proposition 1.19, } U^T=U^{-1}\\
B_+^{-1}&=U^{-1}\Lambda^{-1}U^{-T}=U^T\Lambda^{-1}U. && \text{eq. (1.3), } U^T=U^{-1}
\end{aligned}$$

En utilisant les rélations ci-dessus et le fait que $B_+^{-1} = B_-^{-T}$ (voir point (i)) on trouve

$$C = B_+ B_-^{-1} = (U^T \Lambda U)(U^T \Lambda^{-1} U), \qquad C^T = B_-^{-T} B_+^T = B_+^{-1} B_- = (U^T \Lambda^{-1} U)(U^T \Lambda U),$$

et

$$C^T C = (U^T \Lambda^{-1} U)(U^T \Lambda U)(U^T \Lambda U)(U^T \Lambda^{-1} U) = I_n,$$

où nous avons utilisé à plusieures réprises le fait que $U^T = U^{-1}$ pour simplifier.

Exercice 1.49 (Matrice hermitienne et antihermitienne)**.** *Une matrice $A \in \mathbb{C}^{n,n}$ est dite* antihermitienne *si $A^H = -A$. Montrer que (i) les éléments diagonaux d'une matrice hermitienne sont des réels tandis que ceux d'une matrice antihermitienne sont des imaginaires purs ; (ii) montrer que, si une matrice triangulaire est hermitienne ou antihermitienne, elle est diagonale.*

(i) Si A est hermitienne on a $a_{ii} = \overline{a_{ii}}$ pour tout $1 \le i \le n$, à savoir, $\Im(a_{ii}) = 0 \implies a_{ii} \in \mathbb{R}$. Si A est antihermitienne, $a_{ii} = -\overline{a_{ii}} \implies \Re(a_{ii}) = 0$ pour tout $1 \le i \le n$. (ii) Supposons A triangulaire supérieure (resp. inférieure). Si A est hermitienne ou antihermitienne on a $a_{ij} = \pm 0 \implies a_{ij} = 0$ pour tout $j > i$ (resp. $i > j$), à savoir, A est diagonale.

Exercice 1.50 (Matrices de Hilbert)**.** *La* matrice de Hilbert $H(n) = (h_{ij}) \in \mathbb{R}^{n,n}$ *d'ordre $n \ge 1$ est la matrice carrée symétrique à valeurs réelles telle que*

$$h_{ij} = \frac{1}{i+j-1}.$$

Montrer que $H(n)$ est définie positive.

On remarque que

$$\frac{1}{i+j-1} = \int_0^1 t^{j-1} t^{i-1} \mathrm{d}t. \tag{1.4}$$

Soit maintenant $x = (x_i)_{1 \le i \le n} \in \mathbb{R}^n$ non nul. On a

$$\begin{aligned}
x^T H(n) x &= \sum_{i=1}^n \sum_{j=1}^n x_j h_{ij} x_i \\
&= \sum_{i=1}^n \sum_{j=1}^n \int_0^1 (x_j t^{j-1})(x_i t^{i-1}) \mathrm{d}t && (1.4) \\
&= \int_0^1 \left(\sum_{j=1}^n x_j t^{j-1} \right) \left(\sum_{i=1}^n x_i t^{i-1} \right) \mathrm{d}t \\
&= \int_0^1 w(t)^2 \mathrm{d}t > 0, && w(t) := \sum_{i=1}^n x_i t^{i-1}
\end{aligned}$$

où on a conclu grâce au fait que, par construction, w n'est pas la fonction identiquement nulle.

Exercice 1.51. *Soient $A, B \in \mathbb{R}^{n,n}$ deux matrices inversibles telles que $A+B$ est inversible. Montrer que $A^{-1}+B^{-1}$ est inversible et qu'on a*

$$\left(A^{-1}+B^{-1}\right)^{-1} = B(A+B)^{-1}A = A(A+B)^{-1}B.$$

On observe que

$$A^{-1}+B^{-1} = A^{-1}\left(I_n + AB^{-1}\right) = A^{-1}(B+A)B^{-1}.$$

Or, A, B et $A+B$ étant inversibles, nous avons que

$$B(A+B)^{-1}A = \left[A^{-1}(B+A)B^{-1}\right]^{-1} = (A^{-1}+B^{-1})^{-1},$$

ce qui prouve la première égalité. Pour prouver la deuxième on procède de manière analogue en factorisant A^{-1} à droite et B^{-1} à gauche.

Exercice 1.52 (Valeurs propres d'un polynôme à variable matricielle)**.** *Soit $A \in \mathbb{C}^{n,n}$. Montrer que, si $P(A) := \sum_{k=0}^{n} c_k A^k$ et $\lambda(A)$ est le spectre de A, on a $\lambda(P(A)) = P(\lambda(A))$.*

Soit $\lambda \in \lambda(A)$ une valeur propre de A et $x \in \mathbb{C}^{n,n}$ un vecteur propre associé, à savoir, $Ax = \lambda x$. En multipliant la relation précédent à gauche par A^{k-1} on obtient

$$Ax = \lambda x \implies A^{k-1}Ax = \lambda A^{k-1}x \implies A^k x = \lambda^k x,$$

car on prouve aisément de façon récursive que $A^{k-1}x = \lambda A^{k-2}x = \ldots = \lambda^{k-1}x$. L'égalité précédente permet de conclure

$$P(A)x = \left(\sum_{k=0}^{n} c_k A^k\right)x = \left(\sum_{k=0}^{n} c_k \lambda^k\right)x,$$

et, par définition, $\sum_{k=0}^{n} c_k \lambda^k \in \lambda(P(A))$. Comme $\mathrm{card}(\lambda(A)) = \mathrm{card}(\lambda(P(A))) = n$ nous avons trouvé toutes les n valeurs propres de $P(A)$, ce qui permet de conclure.

Exercice 1.53 (Valeurs singulières de l'inverse d'une matrice)**.** *Soit $A \in \mathbb{C}^{n,n}$ inversible. Exprimer les valeurs singulières de A^{-1} en fonction de celles de A.*

De par le Théorème 1.37 il existent deux matrices unitaires U, $V \in \mathbb{C}^{n,n}$ telles que $U^H AV = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ avec $(\sigma_i)_{1\le i\le n}$ valeurs singulières de A. On a donc,

$$(U^H AV)^{-1} = \Sigma^{-1} \iff V^{-1}A^{-1}(U^H)^{-1} = \Sigma^{-1} \iff V^H A^{-1}U = \Sigma^{-1},$$

qui est la DVS de A^{-1}. Par conséquent $\Sigma^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots \frac{1}{\sigma_n}\right)$ contient les valeurs singulières de A^{-1}. De par le Corollaire 1.36, il existe une matrice orthogonale U telle que

$$A = UDU^T = UDU^{-1}, \quad D = \mathrm{diag}(\lambda)_{\lambda\in\lambda(A)}.$$

Soit $t := \sqrt{\lambda_{\min}\lambda_{\max}}$. On a

$$\frac{1}{t}A + tA^{-1} = U\left(\frac{1}{t}D + tD^{-1}\right)U^{-1} := U\Delta U^{-1}, \tag{1.5}$$

avec $\Delta = \mathrm{diag}(\lambda/t + t/\lambda)_{\lambda\in\lambda(A)}$. La fonction $f(\xi) := \xi/t + t/\xi$ est convexe sur $\mathbb{R}_*^+$ et elle atteint son infimum en $\xi = t$, comme on le voit en résolvant

$$0 = f'(\xi) = \frac{1}{t} - \frac{t}{\xi^2} \iff \left(\xi^2 - t^2 = 0 \text{ et } t > 0\right) \iff \xi = t.$$

De plus, on a

$$\max_{\xi\in[\lambda_{\min},\lambda_{\max}]} f = f(\lambda_{\min}) = f(\lambda_{\max}) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}.$$

Par conséquent, pour tout $\xi \in [\lambda_{\min}, \lambda_{\max}]$,

$$f(\xi) \le \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}. \tag{1.6}$$

On a donc, pour tout $x \in \mathbb{R}^n$,

$$\begin{aligned}
\sqrt{(Ax,x)(A^{-1}x,x)} &\le \frac{1}{2}\left(\frac{1}{t}(Ax,x) + t(A^{-1}x,x)\right) && \sqrt{ab} \le \tfrac{a}{2\epsilon} + \tfrac{b\epsilon}{2} \quad \forall a,b \ge 0, \forall \epsilon > 0 \\
&= \frac{1}{2}\left(\left(\frac{1}{t}A + tA^{-1}\right)x, x\right) \\
&= \frac{1}{2}(U\Delta U^{-1}x, x) && (1.5) \\
&= \frac{1}{2}(\Delta y, y) && y := U^T x,\ U^{-1} = U^T \\
&\le \frac{1}{2}\left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\right)\|y\|_2^2 && (1.6) \\
&\le \frac{1}{2}\left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\right)\|x\|_2^2, && \text{Remarque (1.15)}
\end{aligned}$$

ce qui permet de conclure.

Exercice 1.54. *Soit A une matrice HDP et $\alpha \in \mathbb{R}_*^+$. Montrer que la matrice $I_n + \alpha A$ est inversible, que le rayon spectral de la matrice $B := (I_n - \alpha A)(I_n + \alpha A)^{-1}$ est < 1 et que la matrice B est hermitienne.*

La matrice A est hermitienne donc diagonalisable (voir Corollaire 1.36), à savoir, il existe une matrice unitaire $Q \in \mathbb{C}^{n,n}$ telle que $A = Q^{-1}\Lambda Q$, avec $\Lambda \in \mathbb{R}^{n,n}$ matrice diagonale contenant les valeurs propres de A (le fait que les valeurs propres de A sont des réels strictement positifs est une conséquence de la Proposition 1.23). Nous avons donc

$$I_n \pm \alpha A = Q^{-1}Q \pm \alpha Q^{-1}\Lambda Q = Q^{-1}(I_n \pm \alpha\Lambda)Q,$$

ce qui permet de conclure que la matrice $I_n \pm \alpha A$ est diagonalisable et ses valeurs propres sont de la forme $1 \pm \alpha\lambda$, $\lambda \in \lambda(A)$. De par le Théorème 1.30, la matrice $I_n + \alpha A$ est inversible si et seulement si toutes ses valeurs propres sont non nulles, à savoir $\alpha \neq -1/\lambda$ pour tout $\lambda \in \lambda(A)$, ce qui est toujours vérifié car $\alpha > 0$ par hypothèse. De plus,

$$(I_n + \alpha A)^{-1} = \left[Q^{-1}(I_n + \alpha\Lambda)Q\right]^{-1} = Q^{-1}(I_n + \alpha\Lambda)^{-1}Q,$$

et $(I_n+\alpha\Lambda)^{-1} = \operatorname{diag}\left(\frac{1}{1+\alpha\lambda}\right)_{\lambda\in\lambda(A)}$. De par les résultats précédents, on a

$$B = (I_n-\alpha A)(I_n+\alpha A)^{-1} = Q^{-1}(I_n-\alpha\Lambda)QQ^{-1}(I_n+\alpha\Lambda)^{-1}Q = Q^{-1}\tilde{\Lambda}Q, \tag{1.7}$$

avec $\tilde{\Lambda} = \operatorname{diag}(\tilde{\lambda})_{\tilde{\lambda}\in\lambda(B)} = \operatorname{diag}\left(\frac{1-\alpha\lambda}{1+\alpha\lambda}\right)_{\lambda\in\lambda(A)}$ matrice diagonale contenant les valeurs propres de B. Comme les valeurs propres de A sont >0, on vérifie aisément que $|\tilde{\lambda}|<1$ pour tout $\tilde{\lambda}\in\lambda(B)$ si $\alpha>0$ et, par conséquent, $\rho(B) := \max_{\tilde{\lambda}\in\lambda(B)}|\tilde{\lambda}|<1$. Pour prouver que B est hermitienne, on utilise la décomposition (1.7) pour écrire

$$B^H = (Q^H\tilde{\Lambda}Q)^H = Q^H\tilde{\Lambda}^HQ = Q^H\tilde{\Lambda}Q = B,$$

où nous avons utilisé le fait que $\tilde{\Lambda}$ est une matrice diagonale à valeurs réells pour conclure $\tilde{\Lambda}^H = \tilde{\Lambda}$.

Exercice 1.55 (Pséudo-inverse de Moore–Penrose)**.** *Soient m et n deux entiers avec $m\geq n$, et soit $A\in\mathbb{R}^{m,n}$ une matrice de rang n. On admettra par la suite que $A^TA\in\mathbb{R}^{n,n}$ est définie positive et donc inversible. On définit la* pseudo-inverse de Moore–Penrose *par*

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

Prouver les propriétés suivantes :

$$A^\dagger A = I_n, \qquad A^\dagger AA^\dagger = A^\dagger, \qquad \text{si } m=n,\ A^\dagger = A^{-1};$$

Soit $b\in\mathbb{R}^m$. On considère le problème dit aux moindres carrés

$$\min_{y\in\mathbb{R}^n}\left\{J(y) := \|Ay-b\|_2^2\right\}.$$

Ce problème admet une unique solution x caractérisée par la propriété $\nabla J(y)=0$. Montrer que $x=A^\dagger b$.

On a $A^\dagger A = (A^TA)^{-1}(A^TA) = I_n$, d'où $A^\dagger AA^\dagger = I_nA^\dagger = A^\dagger$. La dernière propriété est une conséquence de l'unicité de l'inverse d'une matrice inversible. Pour prouver le deuxième point on observe que, pour tout $y\in\mathbb{R}^n$,

$$\begin{aligned} J(y) &= (Ay-b)^T(Ay-b)\\ &= (Ay)^TAy-(Ay)^Tb-b^TAy+b^Tb\\ &= y^T(A^TA)y-y^TA^Tb-b^TAy+b^Tb.\\ &= y_j(A^TA)_{ij}y_i-y_i(A^Tb)_i-(b^TA)_iy_i+b_ib_i \end{aligned}$$

où par brévité nous avons utilisé la notation de Einstein qui sous-entend les sommes sur les indices repétés. Par conséquent, pour tout $1\leq k\leq n$,

$$\frac{\partial J(y)}{\partial y_k} = (A^TA)_{ik}y_i + y_j(A^TA)_{kj} - (A^Tb)_k - (b^TA)_k,$$

à savoir,

$$\nabla J(y) = 2(A^TA)y - 2A^Tb.$$

La condition d'optimalité donne

$$\nabla J(x) = 2(A^TA)x - 2A^Tb = 0 \iff x = (A^TA)^{-1}A^Tb = A^\dagger b.$$

Exercice 1.56 (Valeurs singulières d'une matrice normale)**.** *Soit $A \in \mathbb{C}^{n,n}$ une matrice normale. Montrer que les valeurs singulières de A sont les modules de ses valeurs propres.*

COMPLETER

Exercice 1.57 (Inégalité triangulaire)**.** *Soit $\|\cdot\|$ une norme matricielle subordonnée. Prouver que pour tout A, $B \in \mathbb{C}^{n,n}$,*

$$\|A+B\| \leq \|A\| + \|B\|.$$

On a

$$\begin{aligned}\|A+B\| &:= \sup_{x\in\mathbb{C}^n,\|x\|=1} \|(A+B)x\| = \sup_{x\in\mathbb{C}^n,\|x\|=1} \|Ax+Bx\| \\ &\leq \sup_{x\in\mathbb{C}^n,\|x\|=1} (\|Ax\|+\|Bx\|) \\ &\leq \sup_{x\in\mathbb{C}^n,\|x\|=1} \|Ax\| + \sup_{x\in\mathbb{C}^n,\|x\|=1} \|Bx\| := \|A\|+\|B\|,\end{aligned}$$

où nous avons utilisé l'inégalité triangulaire pour la norme vectorielle pour passer à la deuxième ligne.

Exercice 1.58 (Norme d'une matrice unitaire)**.** *Soit $U \in \mathbb{C}^{n,n}$ une matrice unitaire. Prouver que $\|U\|_2 = 1$ et que pour tout $A \in \mathbb{C}^{n,n}$ on a $\|AU\|_2 = \|UA\|_2 = \|A\|_2$.*

Soit $x \in \mathbb{C}^n$. Par définition on a $\|Ux\|_2^2 = (Ux)^H(Ux) = x^H U^H U x = x^H x = \|x\|_2^2$. Par conséquent,

$$\|U\|_2 = \sup_{x\in\mathbb{C}^n,\|x\|_2=1} \|Ux\|_2 = 1.$$

Un raisonnement similaire montre que $\|U^H\|_2 = 1$. D'autre part,

$$\|AU\|_2 = \sup_{x\in\mathbb{C}^n,\|x\|_2=1} \|AUx\|_2, = \sup_{y\in\mathbb{C}^n,\|U^{-1}y\|_2=1} \|Ay\|_2, = \sup_{y\in\mathbb{C}^n,\|y\|_2=1} \|Ay\|_2,$$

où nous avons utilisé le point précédent pour conclure $\|U^{-1}y\|_2 = \|U^H y\|_2 = \|y\|_2$.

Exercice 1.59 (Norme de Frobenius)**.** *Soit $A = (a_{ik}) \in \mathbb{R}^{n,n}$. La norme de Frobenius est définie par*

$$\|A\|_{\mathrm{F}} := \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2.$$

Cette norme n'est pas subordonnée à une norme vectorielle. Prouver que

$$\frac{1}{n}\|A\|_1 \leq \|A\|_{\mathrm{F}} \leq \sqrt{n}\|A\|_1, \qquad \frac{1}{n}\|A\|_\infty \leq \|A\|_{\mathrm{F}} \leq \sqrt{n}\|A\|_\infty.$$

Calculer $\|I_n\|_{\mathrm{F}}$.

Il est utile de rappeler que

$$\|A\|_{\mathrm{F}}^2 := \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2, \quad \|A\|_1 := \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \quad \|A\|_\infty := \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Or,

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \leq \sum_{j=1}^{n}\sum_{i=1}^{n} 1\times|a_{ij}| \leq \left(\sum_{j=1}^{n}\sum_{i=1}^{n} 1^2\right)^{\frac{1}{2}} \times \left(\sum_{j=1}^{n}\sum_{i=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}} = n\|A\|_{\mathrm{F}},$$

ou nous avons utilisé l'inégalité de Cauchy–Schwarz dans le deuxième passage. La preuve que $\|A\|_\infty \leq n\|A\|_F$ est similaire. Pour prouver la borne supérieure, on observe que

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2 \leq n \max_{1\leq j\leq n}\sum_{i=1}^{n}|a_{ij}|^2 \leq n \max_{1\leq j\leq n}\left(\sum_{i=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}} = n\|A\|_1^2,$$

d'où $\|A\|_F \leq \sqrt{n}\|A\|_1$. La preuve de $\|A\|_F \leq \sqrt{n}\|A\|_\infty$ suit un raisonnement analogue. On a $\|I_n\|_F = \sqrt{n}$.

Exercice 1.60 (Rayon spectral)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP. Montrer que $\rho(A) < 1$ si et seulement s'il existe une matrice $Q \in \mathbb{R}^{n,n}$ SDP telle que $B := Q - A^TQA$ est SDP.*

(i) $\boxed{\rho(A)<1 \implies (\exists Q \in \mathbb{R}^{n,n},\ B \textbf{ SDP})}$ Il suffit de prendre $Q = I_n$. En effet, avec ce choix on a pour tout $x \in \mathbb{R}^n \setminus \{0 \in \mathbb{R}^n\}$,

$$\begin{aligned} x^TBx &= x^T\left(I_n - A^TI_nA\right)x = \|x\|_2^2 - \|Ax\|_2^2 \\ &\geq \|x\|_2^2 - \|A\|_2^2\|x\|_2^2 && \text{Proposition 1.43} \\ &= \left(1-\rho(A)^2\right)\|x\|_2^2 \geq 0, && \text{Corollaire 1.36} \end{aligned}$$

ce qui montre le caractère définie positif de B.

(i) $\boxed{(\exists Q \in \mathbb{R}^{n,n},\ B \textbf{ SDP}) \implies \rho(A)<1}$ Soit $x \in \mathbb{R}^n$ et $y := Ax$. On a par hypothèse

$$x^TBx > 0 \implies y^TQy < x^TQx. \tag{1.8}$$

Soit $\overline{\lambda}$ la plus grande valeur propre (en valeur absolue) de A, $\overline{x} \in \mathbb{R}^n$ un vecteur propre associé avec $\|\overline{x}\|_2 = 1$, et $\overline{y} := A\overline{x} = \overline{\lambda}\overline{x}$. De par la rélation (1.8) on a

$$\overline{x}^TQ\overline{x} > \overline{y}^TQ\overline{y} = \overline{\lambda}^2\overline{x}^TQ\overline{x},$$

ce qui prouve $\overline{\lambda} < 1$ et, par conséquent, $\rho(A) < 1$.

Chapitre 2

Méthodes directes

Dans ce chapitre on étudie quelques exemples de méthodes *directes* pour la résolution du système

$$Ax = b. \tag{2.1}$$

Aux erreurs d'arrondi près, ces méthodes permettent de résoudre le système de manière exacte. Par simplicité nous allons nous restreindre au cas réel, et on supposera dans toute la section $A \in \mathbb{R}^{n,n}$, $x \in \mathbb{R}^n$ et $b \in \mathbb{R}^n$ non nul.

2.1 Solution numérique des systèmes linéaires

On s'intéresse à l'effet des erreurs d'arrondi sur la précision de la solution obtenue par la méthode de Gauss.

2.1.1 Conditionnement d'une matrice

Définition 2.1 (Conditionnement d'une matrice)**.** *Soit $A \in \mathbb{C}^{n,n}$ inversible et $\|\cdot\|$ une norme matricielle subordonnée. Le* conditionnement *de A est le réel*

$$\mathrm{cond}(A) := \|A\| \, \|A^{-1}\|.$$

Comme $\|\cdot\|$ est une norme subordonnée, nous avons

$$1 = \|I_n\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| = \mathrm{cond}(A),$$

à savoir, le conditionnement d'une matrice est toujours ≥ 1. Pour $p \in \mathbb{N}$ on définit $\mathrm{cond}_p(A) := \|A\|_p \|A^{-1}\|_p$ le conditionnement associé à la norme p. Un cas particulièrement important est $p = 2$, pour lequel on a d'après la Proposition 1.44,

$$\mathrm{cond}_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)},$$

où $\sigma_1(A)$ et $\sigma_n(A)$ dénotent, respectivement, la plus grande et la plus petite valeur singulière de A (le fait que $1/\sigma_n(A)$ est la plus grande valeur singulière de A^{-1} est justifié dans l'Exercice 1.53). Si, de plus, A est hermitienne, de par le Corollaire 1.39 on a

$$\mathrm{cond}_2(A) = \frac{\max_{\lambda \in \lambda(A)} |\lambda|}{\min_{\lambda \in \lambda(A)} |\lambda|} = \frac{\rho(A)}{\rho(A^{-1})} = \|A\|_2 \|A^{-1}\|_2, \tag{2.2}$$

ce qui justifie l'appellation de *nombre de conditionnement spectral.*

2.1.2 Analyse a priori

Proposition 2.2. *Soit $A \in \mathbb{C}^{n,n}$ telle que $\rho(A) < 1$. Alors, la matrice $I_n - A$ est inversible et on a pour toute norme matricielle subordonnée telle que $\|A\| < 1$,*

$$\frac{1}{1+\|A\|} \leq \|(I_n - A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Démonstration. La preuve directe de l'inversibilité de $I_n - A$ est laissée en exercice. Puisque $\|\cdot\|$ est une norme subordonnée, de par la Proposition 1.45 nous avons

$$\begin{aligned} 1 = \|I_n\| &= \|(I_n - A)(I_n - A)^{-1}\| \leq \|I_n - A\| \|(I_n - A)^{-1}\| \\ &\leq (\|I_n\| + \|A\|) \|(I_n - A)^{-1}\| \\ &= (1 + \|A\|) \|(I_n - A)^{-1}\| \iff \frac{1}{1+\|A\|} \leq \|(I_n - A)^{-1}\|. \end{aligned}$$

Pour prouver la deuxième inégalité il suffit d'observer que

$$(I_n - A)^{-1} = (I_n - A)^{-1} I_n = (I_n - A)^{-1}(I_n - A + A) = I_n + (I_n - A)^{-1} A,$$

d'où, en utilisant l'inégalité triangulaire,

$$\begin{aligned} \|(I_n - A)^{-1}\| &\leq \|I_n\| + \|(I_n - A)^{-1} A\| \\ &\leq 1 + \|(I_n - A)^{-1}\| \, \|A\| \iff \|(I_n - A)^{-1}\| \leq \frac{1}{1-\|A\|}. \quad \square \end{aligned}$$

Théorème 2.3 (Estimation d'erreur)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice inversible et $\delta A \in \mathbb{R}^{n,n}$ une perturbation telle que pour une norme matricielle subordonnée $\|\cdot\|$,*

$$\|A^{-1}\| \, \|\delta A\| < 1. \tag{2.3}$$

On note $x \in \mathbb{R}^n$ la solution du système linéaire $Ax = b$ avec $b \in \mathbb{R}^n \setminus \{0\}$ et, pour une perturbation du seconde membre $\delta b \in \mathbb{R}^n$, soit $\delta x \in \mathbb{R}^n$ tel que

$$(A + \delta A)(x + \delta x) = b + \delta b.$$

Alors,

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\|\delta A\|/\|A\|} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right). \tag{2.4}$$

Remarque 2.4 (Hypothèse (2.3))**.** *L'hypothèse* (2.3) *a une interprétation naturelle dans la mesure où elle implique $\|A^{-1}\delta A\| < 1$, à savoir, les erreurs d'arrondi sur la matrice A sont supposées petites par rapport aux valeurs des entrées de A.*

Démonstration. Comme $b \neq 0$ et A est inversible, $x \neq 0 \implies \|x\| \neq 0$. De plus, la matrice $I + A^{-1}\delta A$ est inversible car, par hypothèse, $\rho(A^{-1}\delta A) \leq \|A^{-1}\delta A\| \leq \|A^{-1}\| \|\delta A\| < 1$ (la première inégalité est une conséquence du Lemme 1.46). On peut donc appliquer le résultat de la

Proposition 2.2 à la matrice $-A^{-1}\delta A$, obtenant ainsi

$$
\begin{aligned}
& (A+\delta A)(x+\delta x) = b+\delta b && \text{définition de } \delta x \\
\Longrightarrow\ & \cancel{Ax} + \delta Ax + A(I_n + A^{-1}\delta A)\delta x = \cancel{b} + \delta b && Ax=b,\ A+\delta A = A(I_n+A^{-1}\delta A) \\
\Longrightarrow\ & \frac{\delta x}{\|x\|} = (I+A^{-1}\delta A)^{-1}A^{-1}\left(\frac{\delta b}{\|x\|} - \frac{\delta Ax}{\|x\|}\right) && \text{mult. à gauche par } \frac{(I_n+A^{-1}\delta A)^{-1}A^{-1}}{\|x\|} \\
\Longrightarrow\ & \frac{\delta x}{\|x\|} \le \|(I+A^{-1}\delta A)^{-1}\|\,\|A^{-1}\|\left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\right) && \|\delta Ax\| \le \|\delta A\|\,\|x\| \\
\Longrightarrow\ & \frac{\delta x}{\|x\|} \le \|(I+A^{-1}\delta A)^{-1}\|\,\mathrm{cond}(A)\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right) && \|A\|\,\|x\| \ge \|Ax\| = \|b\| \\
\Longrightarrow\ & \frac{\delta x}{\|x\|} \le \frac{\mathrm{cond}(A)}{1-\|A^{-1}\delta A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right) && \text{Proposition 2.2} \\
\Longrightarrow\ & \frac{\delta x}{\|x\|} \le \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right). && (2.3)
\end{aligned}
$$

□

Remarque 2.5. *Si les perturbations δb et δA sont telles que $\delta b = \gamma\|b\|$ et $\delta A = \gamma\|A\|$ avec $\delta \in \mathbb{R}_*^+$ et $\gamma\,\mathrm{cond}(A) < 1$, l'estimation* (2.4) *se reduit à*

$$\frac{\|\delta x\|}{\|x\|} \le \frac{2\,\mathrm{cond}(A)}{1-\gamma\,\mathrm{cond}(A)}.$$

En pratique cette hypothèse est raisonnable, et le paramètre γ depend de l'epsilon machine ϵ, à savoir, le plus petit nombre tel que $1+\epsilon > 1$ en arithmétique flottante.

2.2 Opérations élémentaires

Pour fixer les idées, soit $A \in \mathbb{R}^{n,n}$ une matrice carré d'ordre n. La généralisation des résultats de cette section au cas de matrices rectangulaires est laissée au lecteur. Les méthodes directes que nous allons examiner reposent sur deux opérations élémentaires, à savoir, (i) l'échange de deux lignes de la matrice A, que l'on notera $A_{i:} \longleftrightarrow A_{j:}$, $1 \le i, j \le n$. Lorsqu'on souhaite nommer différemment la matrice obtenue en échangeant les lignes i et j de A on notera

$$B \leftarrow (A_{i:} \longleftrightarrow A_{j:});$$

(ii) la combinaison linéaire d'une ligne $A_{i:}$ avec une ligne $A_{j:}$, $1 \le i \ne j \le n$, que l'on notera $A_{i:} \leftarrow \lambda A_{i:} + \mu A_{j:}$, $\lambda, \mu \in \mathbb{R}_*$. Lorsqu'on souhaite nommer différemment la matrice obtenue en combinant linéairement les lignes i et j de A on notera

$$B_{i:} \leftarrow \lambda A_{i:} + \mu A_{j:},$$

en sous-entendant $B_{l:} = A_{l:}$ pour tout $l \in [\![1, n]\!] \setminus \{i\}$.

Remarque 2.6 (Interprétation alternative de l'échange de deux lignes)**.** *Si on relâche l'hypothèse $\lambda, \mu \ne 0$, l'échange de deux lignes de la matrice A peut être interprété comme le résultat des deux combinaisons linéaires*

$$B_{i:} \leftarrow 0A_{i:} + 1A_{j:}, \qquad B_{j:} \leftarrow 1A_{i:} + 0A_{j:}.$$

Cependant, sans l'hypothèse $\lambda, \mu \neq 0$ on ne pourra plus garantir des propriétés d'équivalence entre les deux matrices.

Nous allons par la suite montrer qu'on peut interpréter ces deux opérations élémentaires comme des multiplications à gauche par des matrices opportunes. On commencera par remarquer que, étant donné une matrice $P = (p_{ij}) \in \mathbb{R}^{n,n}$, nous avons pour tout $1 \leq i, j \leq n$,

$$(PA)_{i:} = p_{i1}A_{1:} + p_{i2}A_{2:} + \ldots + p_{in}A_{n:} = \sum_{k=1}^{n} p_{ik}A_{k:}, \tag{2.5}$$

à savoir, l'i-ème ligne de PA est obtenue par combinaison linéaire des lignes de A avec coefficients $p_{i1}, \ldots, p_{in}$. Soient maintenant $1 \leq i \neq j \leq n$ deux indices. La matrice *de permutation* P qui réalise l'échange $A_{i:} \longleftrightarrow A_{j:}$ doit satisfaire les conditions suivantes :

$$(PA)_{i:} = A_{j:}, \quad (PA)_{j:} = A_{i:}, \quad (PA)_{l:} = A_{l:} \quad \forall l \in [\![1, n]\!] \setminus \{i, j\}.$$

Par conséquent, les seuls éléments non nuls de P sont les suivants :

$$p_{ij} = 1, \quad p_{ji} = 1, \quad p_{ll} = 1 \quad \forall l \in [\![1, n]\!] \setminus \{i, j\}. \tag{2.6}$$

Exercice 2.7 (Échange de deux lignes)**.** *On considère la matrice*

$$A = \begin{pmatrix} 2 & 3 & 4 & 5 \\ 5 & 8 & 9 & 8 \\ 0 & 3 & 6 & 1 \\ 2 & 3 & 4 & 9 \end{pmatrix}.$$

Écrire la matrice de permutation P qui permet d'effectuer l'échange $A_{1:} \longleftrightarrow A_{3:}$ et vérifier le résultat.

Proposition 2.8 (Matrice de permutation de lignes)**.** *La matrice de permutation de lignes P définie par* (2.6) *est inversible et on a*

$$P^{-1} = P^T = P.$$

Démonstration. Soit $\tilde{A} = PA$ la matrice obtenue à partir de A en effectuant l'échange $A_{i:} \longleftrightarrow A_{j:}$ pour $1 \leq i \neq j \leq n$. On vérifie aisément qu'on peut revenir à A à partir de $\tilde{A}$ en effectuant l'échange de lignes $\tilde{A}_{i:} \longleftrightarrow \tilde{A}_{j:}$, qui correspond à la matrice de permutation $P = P^T$. Nous avons donc

$$A = P\tilde{A} = P^T PA,$$

ce qui montre que $P^T P = I_n$, à savoir, P est inversible et $P^{-1} = P^T = P$. □

On peut appliquer un raisonnement similaire à la combinaison linéaire de deux lignes $A_{i:} \leftarrow \lambda A_{i:} + \mu A_{j:}$, $1 \leq i \neq j \leq n$, $\lambda, \mu \in \mathbb{R}_*$. D'après (2.5), les seuls coefficients non nuls de la matrice P sont dans ce cas

$$p_{ii} = \lambda, \quad p_{ij} = \mu, \quad p_{ll} = 1 \quad \forall l \in [\![1, n]\!] \setminus \{i, j\}.$$

Exercice 2.9 (Combinaison linéaire de deux lignes)**.** *Écrire la matrice P qui correspond à la combinaison linéaire $A_{1:} \leftarrow 2A_{1:} + 3A_{4:}$ pour la matrice de l'Exercise 2.7 et vérifier le résultat.*

Exercice 2.10 (Échange de deux colonnes)**.** *Soient* $1 \leq i \neq j \leq n$*. Montrer que l'échange de colonnes*

$$A_{:i} \longleftrightarrow A_{:j}$$

peut s'interpréter comme la multiplication à droite par la matrice de permutation $P = (p_{ij})$ *dont les seuls éléments non nuls sont*

$$p_{ij} = 1, \quad p_{ji} = 1, \quad p_{ll} = 1 \quad \forall l \in [\![1, n]\!] \setminus \{i, j\}.$$

Suggestion : remarquer d'abord que, pour tout $1 \leq l \leq n$*,*

$$(AP)_{:l} = A_{:1} p_{1l} + A_{:2} p_{2l} + \ldots + A_{:n} p_{nl}.$$

2.3 Méthode de Gauss

2.3.1 Factorisation $A = LU$

Les résultats de la section précédent permettent d'interpréter la méthode du pivot de Gauss comme un factorisation de la matrice A. La stratégie de la méthode de Gauss consiste, à chaque itération $1 \leq k \leq n-1$, à éliminer la variable d'indice k des équations $k+1 \ldots n$ de la matrice. On se situe à l'itération k et on cherche à interpréter les opérations élémentaires correspondantes en termes de multiplication à gauche par des matrices opportunes. La matrice $A^{(k)}$ prend la forme suivante :

$$A^{(k)} = \begin{pmatrix} U^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{pmatrix} \in \mathbb{R}^{k+(n-k),k+(n-k)}, \tag{2.7}$$

avec, en particulier,

$$U^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} \\ & \ddots & \vdots \\ & & a_{kk}^{(k)} \end{pmatrix}, \quad A_{21}^{(k)} = \begin{pmatrix} 0 & \cdots & 0 & a_{k+1k}^{(k)} \\ 0 & \cdots & 0 & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} \end{pmatrix}$$

On supposera pour l'instant que $u_{kk} = a_{kk}^{(k)}$ (le *pivot*) est non nul. Notre objectif consiste à annuler les coefficients de la variable k dans les lignes $k+1, \ldots n$ de A, à savoir, annuler l'unique colonne non nulle de $A_{21}^{(k)}$. Cela revient à effectuer les combinaisons linéaires suivantes :

$$A_{i:}^{(k+1)} \leftarrow A_{i:}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} A_{k:}^{(k)} \qquad \forall i \in [\![k+1, n]\!]. \tag{2.8}$$

D'après (2.5), les seuls éléments non nuls de la matrice $E^{(k)} = (e_{ij}) \in \mathbb{R}^{n,n}$ telle que $A^{(k+1)} = E^{(k)} A^{(k)}$ sont

$$\begin{aligned} e_{ii}^{(k)} &= 1 && \forall i \in [\![1, n]\!] \\ e_{ik}^{(k)} &= -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} && \forall i \in [\![k+1, n]\!]. \end{aligned} \tag{2.9}$$

La matrice $E^{(k)}$ est donc de la forme

$$E^{(k)} = \begin{pmatrix} 1 & & & & & \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \\ \vdots & & -\frac{a_{k+1k}^{(k)}}{a_{kk}^{(k)}} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & & -\frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} & & & 1 \end{pmatrix}.$$

La nouvelle matrice $A^{(k+1)} = E^{(k)}A^{(k)}$ est par construction telle que $a_{ij}^{(k+1)} = 0$ pour tout $j \in [\![1, k+1]\!]$ et $1 \leq i < j$. Un point important à noter est que la matrice $E^{(k)}$ est triangulaire inférieure car, pour tout $1 \leq i < j \leq n$ on a $e_{ij}^{(k)} = 0$. De plus, on a $\mathrm{diag}(E^{(k)}) = (1, \ldots, 1)$. En supposant les pivots non nuls, à savoir, $a_{kk}^{(k)} \neq 0$ pour tout $k \in [\![1, n-1]\!]$, la méthode de Gauss s'écrit synthétiquement

$$E^{(n-1)} \ldots E^{(1)} A = U \iff A = LU,$$

où nous avons posé $L := (E^{(n-1)} \ldots E^{(1)})^{-1}$ et $U := A^{(n)} \in \mathbb{R}^{n,n}$ est une matrice triangulaire supérieure. La proposition suivante montre que la matrice L est obtenue au cours des itérations sans coût supplémentaire.

Proposition 2.11 (Calcul du facteur L)**.** *Le facteur $L \in \mathbb{R}^{n,n}$ est triangulaire inférieure et on a*

$$L = L^{(1)} \ldots L^{(n-1)},$$

avec, pour tout $k \in [\![1, n-1]\!]$, $L^{(k)} = (l_{ij}^{(k)}) := (E^{(k)})^{-1}$ et

$$\begin{aligned} l_{ii}^{(k)} &= 1 && \forall i \in [\![1, n]\!] \\ l_{ik}^{(k)} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} && \forall i \in [\![k+1, n]\!]. \end{aligned}$$

La matrice $L^{(k)}$ prend donc la forme

$$L^{(k)} = \begin{pmatrix} 1 & & & & & \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \\ \vdots & & \frac{a_{k+1k}^{(k)}}{a_{kk}^{(k)}} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & & \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} & & & 1 \end{pmatrix} = 2I_n - E^{(k)}.$$

Démonstration. On commence par remarquer que, pour tout $k \in [\![1, n-1]\!]$ la matrice $A^{(k)}$ peut s'obtenir à partir de la matrice $A^{(k+1)}$ en effectuant les combinaisons linéaires suivantes :

$$A_{i:}^{(k)} \leftarrow A_{i:}^{(k+1)} + \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} A_{kk}^{(k+1)} \qquad \forall i \in [\![k+1, n]\!].$$

où nous avons utilisé le fait que $A^{(k+1)}_{kk} = A^{(k)}_{kk}$. A partir de la formule (2.5) on vérifie aisément que la matrice correspondante est précisément $L^{(k)}$. Nous avons donc

$$A^{(k)} = L^{(k)}A^{(k+1)} = L^{(k)}E^{(k)}A^{(k)},$$

et encore

$$A^{(k+1)} = E^{(k)}A^{(k)} = E^{(k)}L^{(k)}A^{(k+1)},$$

ce qui montre que $L^{(k)}E^{(k)} = E^{(k)}L^{(k)} = I_n$, à savoir, $E^{(k)}$ est inversible et $(E^{(k)})^{-1} = L^{(k)}$. Enfin, de par la Proposition 1.8, L est triangulaire inférieure car produit de matrices triangulaires inférieures. □

Proposition 2.12 (Expression pour L)**.** *Pour tout $k \in [\![1, n-1]\!]$ on a*

$$\tilde{L}^{(k)} := L^{(1)} \dots L^{(k)} = \begin{pmatrix} 1 & & & & & \\ \frac{a^{(1)}_{21}}{a^{(1)}_{11}} & \ddots & & & & \\ \vdots & \ddots & 1 & & & \\ \vdots & & \frac{a^{(k)}_{k+1k}}{a^{(k)}_{kk}} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ \frac{a^{(1)}_{n1}}{a^{(1)}_{11}} & \cdots & \frac{a^{(k)}_{nk}}{a^{(k)}_{kk}} & & & 1 \end{pmatrix}.$$

Exercice 2.13 (Factorisation LU)**.** *Calculer la factorisation LU de la matrice*

$$A = \begin{pmatrix} 2 & 4 & 3 \\ 8 & 11 & 3 \\ 2 & 1 & 4 \end{pmatrix},$$

en précisant à chaque étape la matrice $L^{(k)}$. Vérifier que $L = L^{(n-1)} \dots L^{(1)}$ est triangulaire inférieure et que $LA^{(n)} = LU = A$.

Exercice 2.14 (Compte d'opérations)**.** *Prouver que le nombre d'opérations élémentaires effectué pour accomplir la factorisation LU est de l'ordre de $2n^3/3$ (on négligera les termes d'ordre inférieur).*

Proposition 2.15 (Calcul du déterminant de A)**.** *Soit A une matrice admettant une factorisation LU. Alors,*

$$\det(A) = \det(U) = \prod_{i=1}^{n} u_{ii}.$$

Démonstration. On a $\det(A) = \det(L)\det(U)$. Puisque L et U sont triangulaires, leurs déterminants sont donnés par les produits des éléments diagonaux, à savoir $\det(L) = 1$ et $\det(U) = \prod_{i=1}^{n} u_{ii}$, ce qui conclut la preuve. □

2.3.2 Existence et unicité de la factorisation $A = LU$

Dans la section précédent nous avions supposé que les pivots $a_{kk}^{(k)}$, $k \in [\![1, n-1]\!]$ soient non nuls. Le théorème suivant fournit une condition suffisante pour que cette hypothèse soit vérifiée.

Théorème 2.16 (Existence et unicité de la factorisation LU)**.** *Soit $A = (a_{ij}) \in \mathbb{R}^{n,n}$ et posons pour tout $1 \leq k \leq n$,*

$$\Delta^{(k)} := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}.$$

Alors si toutes les matrices $\Delta^{(k)}$, $1 \leq k \leq n-1$, sont inversibles il existe une unique factorisation $A = LU$ avec U triangulaire supérieure et L triangulaire inférieure inversible ayant une diagonale de 1. Si, de plus, $\Delta^{(n)} = A$ est inversible, la matrice U l'est aussi.

Remarque 2.17 (Matrices de rang $n-1$)**.** *D'après le théorème précédent, on peut obtenir une factorisation LU pour toute matrice de rang $n-1$.*

Démonstration. La preuve se fait par récurrence sur l'indice d'itération k. A la première étape de l'algorithme de Gauss nous pouvons éliminer la variable x_1 des lignes $2 \dots n$ de A car $a_{11} = \Delta^{(1)} \neq 0$ par hypothèse. Prouvons maintenant que, si $\Delta^{(k)}$ est non nul, on peut avancer de l'étape k à l'étape $k+1$. Soit $\tilde{L}^{(k)} := L^{(k)} \dots L^{(1)}$. On partitionne $\tilde{L}^{(k)}$ et $A^{(k)}$ selon (2.7),

$$\tilde{L}^{(k)} = \begin{pmatrix} \tilde{L}_{11}^{(k)} & \tilde{L}_{12}^{(k)} \\ \tilde{L}_{21}^{(k)} & \tilde{L}_{22}^{(k)} \end{pmatrix} \in \mathbb{R}^{k+(n-k),k+(n-k)}, \quad A^{(k)} = \begin{pmatrix} U^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{pmatrix} \in \mathbb{R}^{k+(n-k),k+(n-k)}.$$

A partir de l'égalité

$$\tilde{L}^{(k)} A^{(k)} = A = \begin{pmatrix} \Delta^{(k)} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

on déduit $\tilde{L}^{(k)} U^{(k)} = \Delta^{(k)} \implies U^{(k)} = (\tilde{L}^{(k)})^{-1} \Delta^{(k)}$, à savoir, $U^{(k)}$ est inversible en tant que produit de deux matrices inversibles. Nous avons donc $\det(U^{(k)}) = \prod_{1 \leq i \leq k} a_{ii}^{(k)} \neq 0$, ce qui implique, en particulier, $a_{kk}^{(k)} \neq 0$. On peut donc avancer à l'étape $k+1$. Venons maintenant à la question de l'unicité. Le raisonnement ci-dessus reste vrai jusqu'à l'étape $(n-1)$ de la factorisation. A ce point il ne reste qu'une ligne dans la matrice A, et deux cas peuvent se produire : soit la matrice A est de rang $(n-1)$, et alors cette ligne est nulle et on a $u_{nn} = a_{nn}^{(n-1)} = 0 \implies \det(U) = 0$, soit la matrice A est de rang plein (ce qui équivaut à supposer $\Delta^{(n)} = A$ inversible), et dans ce cas $u_{nn} = a_{nn}^{(n-1)} \neq 0 \implies \det(U) \neq 0$ et U est inversible □

2.3.3 *Pivoting* partiel et factorisation $PA = LU$

La méthode présentée à la section précédente échoue lorsqu'un pivot nul est rencontré. Il est toutefois possible d'introduire une modification qui permet d'aboutir dans un plus grand nombre de cas. L'idée consiste, à chaque étape k, d'effectuer un échange de lignes pour assurer un pivot non nul. Soit $A^{(k)}$ de la forme (2.7). Le *pivoting* par colonnes consiste à chercher l'indice l tel que

$$a_{lk}^{(k)} = \max_{k \leq i \leq n} |a_{ik}^{(k)}|, \tag{2.10}$$

à savoir, on choisit le plus grand élément en valeur absolue dans la colonne k de A. On admettra par la suite qu'il existe un indice l tel que $a_{lk}^{(k)} \neq 0$. Comme on le verra plus loin, choisir le plus grand pivot en valeur absolue permet de réduire les erreurs d'arrondi associées à la méthode de Gauss. La nouvelle itération k consiste maintenant à effectuer deux opérations élémentaires, à savoir, l'échange de la ligne k avec la ligne l puis les combinaisons linéaires (2.8) :

$$A_{k:}^{(k)} \longleftrightarrow A_{l:}^{(k)}, \qquad A_{i:}^{(k+1)} \leftarrow A_{i:}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} A_{k:}^{(k)} \qquad \forall i \in [\![k+1, n]\!].$$

Il faut bien noter que le pivot $a_{kk}^{(k)}$ est non nul On vérifie aisement à l'aide de (2.5) que ces deux opérations s'écrivent de manière synthétique comme suit :

$$A^{(k+1)} = E^{(k)} P^{(k)} A^{(k)},$$

avec $E^{(k)}$ définie par (2.9) et les seuls éléments non nuls de la matrice de permutation $P^{(k)} = (p_{ij}^{(k)}) \in \mathbb{R}^{n,n}$ sont

$$p_{kl}^{(k)} = 1, \qquad p_{lk}^{(k)} = 1, \qquad p_{ii}^{(k)} = 1 \quad \forall i \in [\![1, n]\!] \setminus \{l, k\}. \tag{2.11}$$

Nous avons alors

$$U = A^{(n)} = E^{(n-1)} P^{(n-1)} \dots E^{(1)} P^{(1)} A.$$

En posant

$$P := P^{(n-1)} \dots P^{(1)}, \quad E := (E^{(n-1)} P^{(n-1)} \dots E^{(1)} P^{(1)}) P^{-1}, \quad L := E^{-1}$$

nous avons

$$U = EPA \iff LU = PA.$$

Une amélioration ultérieure (mais coûteuse) du *pivoting* consiste à autoriser des permutations de colonnes, et à chercher le pivot parmi *tous* les éléments de la matrice $A_{22}^{(k)}$ de (2.7). Voir l'Exercice 2.27 pour plus de détails.

2.3.4 Résolution de systèmes linéaires

La factorisation $A = LU$ (ou $PA = LU$) est utilisée notamment pour la résolution de systèmes linéaires. Une fois calculée la factorisation LU de la matrice A, on réécrit le system (2.1) comme deux systèmes triangulaires moyennant l'introduction de la variable auxiliaire $y \in \mathbb{R}^n$:

$$\begin{cases} Ly &= b, \\ Ux &= y. \end{cases} \tag{2.12}$$

La résolution d'un système triangulaire est particulièrement simple, car elle peut s'effectuer par substitutions successives. On parle alors de *remontée* si le système est triangulaire supérieur et de *descente* s'il est triangulaire inférieur. Pour fixer, soit $A \in \mathbb{R}^{n,n}$ une matrice triangulaire supérieure inversible, à savoir

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix},$$

où nous avons omis d'indiquer les éléments nuls. Puisque A est inversible et donc, $\det(A) = \prod_{i=1}^{n} a_{ii} \neq 0$, on a $a_{ii} \neq 0$ pour tout $1 \leq i \leq n$. Alors,

$$x_n = \frac{b_n}{a_{nn}}, \qquad x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{k=i+1}^{n} a_{ik} x_k \right) \quad \forall i = n-1, \ldots, 1. \tag{2.13}$$

Il est utile de compter le nombre d'opérations élémentaires nécessaires pour la résolution complète du système. On a besoin d'une division pour calculer x_n, de $n-i$ multiplications, de $n-i$ sommes, et d'une division pour calculer x_i. Le nombre d'opérations élémentaires est donc

$$\begin{aligned} 1 + \underbrace{\sum_{i=1}^{n-1}(n-i)}_{\text{multiplications}} + \underbrace{\sum_{i=1}^{n-1}(n-i)}_{\text{sommes}} + \underbrace{\sum_{i=1}^{n-1} 1}_{\text{divisions}} &= 1 + 2\sum_{i=1}^{n-1} - 2\sum_{i=1}^{n-1} i + (n-1) \\ &= 1 + 2n(n-1) - n(n-1) + (n-1) = n^2. \end{aligned}$$

Comme la factorisation LU demande $\approx 2n^3/3$ opérations élémentaires, on voit bien que le coût de résolution des deux systèmes (2.12) est négligeable.

2.4 Autres factorisations

2.4.1 Matrices SDP : La factorisation de Cholesky

On commence par remarquer que, s'il existe une factorisation LU de la matrice $A \in \mathbb{R}^{n,n}$ (que l'on supposera inversible) on peut poser

$$A = LU = LD(D^{-1}U) := LDM^T,$$

où $D = \mathrm{diag}((u_{ii})_{1 \leq i \leq n})$ contient la diagonal de U et M^T est une matrice triangulaire supérieure. L'existence de D^{-1} est garantie par l'inversibilité de A, qui assure $\det(A) = \det(U) \neq 0$ (voir la Proposition 2.15). Cette décomposition est unique sous les hypothèses du Théorème 2.16. Si A est symétrique on peut conclure

$$LDM^T = A = A^T = MD^T L^T = MDL^T,$$

et, par l'unicité de la factorisation, $M = L$. Si, de plus, A est SDP, d'après la Proposition 1.23, elle admet n valeur propres réelles strictement positives (non nécessairement toutes distinctes), et on peut définir

$$D^{\frac{1}{2}} := \mathrm{diag}(\sqrt{d_{ii}}).$$

En posant $H := L^T D^{\frac{1}{2}}$ (H est triangulaire supérieure) nous avons donc

$$A = H^T H. \tag{2.14}$$

Cette factorisation est appelée *de Cholesky*.

Lemme 2.18 (Calcul de la factorisation de Cholesky)**.** *Soit $A \in \mathbb{R}^{n,n}$ SDP. Alors, il existe une unique factorisation* (2.14) *de A avec H triangulaire supérieure avec diagonale positive. Les éléments de H^T peuvent être calculés à partir des formules suivantes : $h_{11} = \sqrt{a_{11}}$ et, pour tout, $i = 2, \dots, n$,*

$$h_{ij} = \frac{1}{h_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} h_{ik} h_{jk} \right) \quad j = 1, \dots, i-1, \tag{2.15a}$$

$$h_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} h_{ik}^2 \right)^{\frac{1}{2}}. \tag{2.15b}$$

Démonstration. On raisonne par récurrence sur la taille i de la matrice. Pour $i = 1$ le résultat est clairement vrai. Supposons-le vrai pour $i-1$ et prouvons-le pour i. D'après l'hypothèse de récurrence il existe une matrice H_{i-1} triangulaire supérieure telle que $A_{i-1} = H_{i-1}^T H_{i-1}$. Alors on cherche $\beta \in \mathbb{R}$ tel que l'égalité par blocs suivante soit vérifiée :

$$A_i = \begin{pmatrix} A_{i-1} & v \\ v^T & a_{ii} \end{pmatrix} = \begin{pmatrix} H_{i-1}^T & 0 \\ h^T & \beta \end{pmatrix} \begin{pmatrix} H_{i-1} & h \\ 0 & \beta \end{pmatrix} = H_i^T H_i,$$

où v, $h \in \mathbb{R}^{i-1}$ et a_{ii} est un réel *strictement positif* (ceci est nécessaire pour que A_i soit définie positive). En imposant l'équivalence bloc par bloc on a les équations

$$H_{i-1}^T h = v, \tag{2.16a}$$

$$h^T h + \beta^2 = a_{ii}, \tag{2.16b}$$

où on remarquera que le vecteur v est unique car H_{i-1} est non singulière. De plus, comme A_i est SDP on a

$$0 < \det(A_i) = \det(H_i^T) \det(H_i) = \beta^2 \det(H_{i-1})^2,$$

et on conclut que β est un réel et $\beta = \sqrt{a_{ii} - h^T h}$. La formule (2.15a) n'est rien d'autre que la descente pour le système triangulaire supérieur (2.16a) (voir l'Exercice 2.28 pour plus de détails), tandis que (2.16b) n'est rien d'autre que la formulation vectorielle de (2.15b). □

2.4.2 Matrices rectangulaires : La factorisation $A = QR$

2.5 Matrices creuses

2.5.1 Matrices tridiagonales : La méthode de Thomas

Pour certaines matrices dotées d'une structure particulière la factorisation LU peut être obtenue avec un nombre d'opérations élémentaires significativement inférieur à celui de la méthode de Gauss. C'est le cas notamment des matrices tridiagonales de la forme

$$A = \text{tridiag}(c, a, b), \tag{2.17}$$

où $a \in \mathbb{R}^n$ et c, $b \in \mathbb{R}^{n-1}$. On supposera dans le reste de cette section que A est définie positive, ce qui assure la bonne définition de tous les termes qui apparaissent dans l'algorithme.

Remarque 2.19 (Matrice définie positive non symétrique)**.** *On peut se poser la question s'il existe des matrices définies positives et non symétriques (tel est le cas pour la matrice* (2.17) *si $b \neq c$). Soit $A \in \mathbb{R}^{n,n}$. On commence par remarquer que l'on peut décomposer A en la somme d'une partie symétrique A_s et une anti-symétrique A_{ss},*

$$A = \frac{1}{2}\left(A + A^T\right) + \frac{1}{2}\left(A - A^T\right) := A_s + A_{ss}.$$

Par définition de matrice définie positive on a pour tout $x \in \mathbb{R}^n$, $x \neq 0$,

$$0 < (Ax, x) = (A_s x, x) + (A_{ss} x, x). \tag{2.18}$$

Comme A_{ss} est anti-symétrique, on a $A_{ss} = -A_{ss}^T$ donc

$$(A_{ss}x, x) = -(A_{ss}^T x, x) = -x^T A_{ss}^T x = -\sum_{1 \leq i,j \leq n} a_{ji} x_i x_j = -x^T A_{ss} x,$$

à savoir, $(A_{ss}x, x) = 0$, ce qui montre que la partie anti-symétrique d'une matrice ne contribue pas à son caractère défini positif. De plus, on peut déduire de (2.18) *qu'une matrice est positive définie si et seulement si sa partie symétrique l'est.*

Pour toute matrice tridiagonale définie positive de la forme (2.17) on a

$$A = LU, \quad L = \text{tridiag}(\gamma, 1, 0), \quad U = \text{tridiag}(0, \alpha, b),$$

avec $\alpha \in \mathbb{R}^n$ et $\gamma \in \mathbb{R}^{n-1}$. On vérifie aisément que

$$LU = \begin{pmatrix} \alpha_1 & b_1 & \\ \gamma_i \alpha_{i-1} & \gamma_i b_{i-1} + \alpha_i & b_i \\ & \gamma_n \alpha_{n-1} & \gamma_n b_{n-1} + \alpha_n \end{pmatrix}$$

Les composantes des vecteurs α et γ peuvent être calculées par comparaison de manière récursive selon le schéma suivant :

$$\begin{pmatrix} \alpha_1 & b_1 & \\ \downarrow & & \\ \gamma_i \alpha_{i-1} \rightarrow & \gamma_i b_{i-1} + \alpha_i & b_i \\ & \downarrow & \\ & \gamma_n \alpha_{n-1} \longrightarrow & \gamma_n b_{n-1} + \alpha_n \end{pmatrix}$$

Plus précisément on a $\alpha_1 = a_1$ et, pour $2 \leq i \leq n$,

$$\gamma_i = \frac{c_i}{\alpha_{i-1}}, \qquad \alpha_i = a_i - \gamma_i b_{i-1}.$$

On voit facilement que le nombre d'opérations élémentaires nécessaires pour conclure le calcul des facteurs L et U est de l'ordre de $3n$. L'exemple suivant montre que l'intérêt de la méthode de Thomas n'est pas uniquement théorique.

Example 2.20 (Discrétisation différences finies du problème de Poisson)**.** *Soit* $\Omega = (0,1)$ *et* $f \in C^0(\Omega, \mathbb{R})$*. On considère le problème de Poisson*

$$-u'' = f \quad \textit{dans}\,\Omega, \qquad u(0) = u(1) = 0, \tag{2.19}$$

Ce problème peut être approché numériquement par la méthode des différences finie*, consistante à remplacer les dérivées par des approximations construites à partir de la formule de Taylor. Soit* N *un entier* ≥ 1*. On introduit la famille de points* $(x_i)_{0\leq i\leq N+1}$ *tels que* $x_i = ih$*,* $h := 1/(N+1)$ *et, pour toute fonction* φ *suffisamment régulière, on pose* $\varphi_i := \varphi(x_i)$*. Pour tout point intérieur* x_i*,* $1 \leq i \leq N$*, et une fonction la formule de Taylor donne*

$$\phi_{i-1} = \phi_i - \phi_i' h + \phi_i'' \frac{h^2}{2} + \mathrm{o}(h^2), \qquad \phi_{i+1} = \phi_i + \phi_i' h + \phi_i'' \frac{h^2}{2} + \mathrm{o}(h^2),$$

où $\phi \in C^2(\Omega, \mathbb{R})$*. En sommant les deux égalités ci-dessus on trouve*

$$\phi_{i-1} + \phi_{i+1} = 2\phi_i + \phi_i'' h^2 + \mathrm{o}(h^2) \iff \phi_i'' = \frac{\phi_{i-1} - 2\phi_i + \phi_{i+1}}{h^2} + \frac{\mathrm{o}(h^2)}{h^2}.$$

En négligeant le reste dans l'expression de ϕ_i'' *on obtient une approximation consistante de la dérivée seconde de* ϕ *au sens où, pour tout* $1 \leq i \leq N$*,*

$$\lim_{h\to 0^+} \phi_i'' - \frac{\phi_{i-1} - 2\phi_i + \phi_{i+1}}{h^2} = \lim_{h\to 0^+} \frac{\mathrm{o}(h^2)}{h^2} = 0.$$

On peut donc approcher le problème (2.19) *comme suit :*

$$-\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = f_i \quad \forall 1 \leq i \leq N, \qquad u_0 = u_{N+1} = 0, \tag{2.20}$$

où les inconnues $(u_i)_{0\leq i\leq N+1}$ *représentent des valeurs nodales. La résolution du problème* (2.20) *demande l'inversion d'une matrice tridiagonale pour déterminer les inconnues* u_i*,* $1 \leq i \leq N$*, associées aux nœuds intérieurs. Cette matrice prend la forme*

$$A = \frac{1}{h^2}\,\mathrm{tridiag}(-1, 2, -1).$$

2.5.2 Matrices creuses non structurées

Nous allons enfin traiter le cas des matrices creuses qui contiennent un nombre d'éléments non nuls de l'ordre d'une de leurs dimensions linéaires (nombre de lignes ou nombre de colonnes), mais qui ne possèdent pas une structure particulière.

Un exemple important

Un contexte très important où l'on retrouve des matrices creuses non structurées est celui de la discrétisation des équations aux dérivées partielles, comme le montre l'exemple abordé dans cette section. Soit $\Omega \in \mathbb{R}^2$ un ouvert polygonal borné. On considère le problème de Poisson

$$-\triangle u = f \quad \text{dans}\,\Omega, \qquad u = 0 \quad \text{sur}\,\partial\Omega. \tag{2.21}$$

On supposera par la suite que le terme source f est à carré intégrable, à savoir, $f \in L^2(\Omega)$. Par brévité de notation on pose $V := H_0^1(\Omega)$, où $H_0^1(\Omega)$ est l'*espace de Sobolev* qui contient les fonctions à carré intégrable, dont les dérivées (au sens faible) sont à carré intégrable, et qui s'annulent sur $\partial\Omega$. La formulation faible de (2.21) consiste à : Trouver $u \in V$ tel que

$$a(u,v) = \int_\Omega fv \qquad \forall v \in V, \tag{2.22}$$

où $a(u,v) := \int_\Omega \nabla u \cdot \nabla v$ est une forme bilinéaire de $V \times V$ dans $\mathbb{R}$. On peut facilement se convaincre que V est un espace vectoriel de dimension infinie. La méthode des éléments finis consiste à approcher la solution de (2.22) en remplaçant V par un sous-espace $V_h \subset V$ de dimension finie. Ceci revient à considérer le problème suivant : Trouver $u_h \in V_h$ tel que

$$a(u_h, v_h) = \int_\Omega f v_h \qquad \forall v_h \in V_h. \tag{2.23}$$

On peut montrer qu'un choix judicieux de l'espace V_h s'obtient comme suit : (i) on définit une triangulation $\mathcal{T}_h = \{T\}$ de Ω (dite *maillage*) telle que l'intersection de deux triangles est soit un nœud, soit une arête [AJOUTER FIGURE] ; (ii) on pose

$$V_h := \{v_h \in H_0^1(\Omega) \mid v_{h|T} \in \mathbb{P}^1(T) \quad \forall T \in \mathcal{T}_h\},$$

où $\mathbb{P}^1(T)$ désigne la restriction à T des fonctions affines de deux variables. Soit N le nombre de nœuds de la triangulation $\mathcal{T}_h$ qui sont des points intérieur à Ω. Pour tout $1 \le i \le N$ on notera $x_i \in \mathbb{R}^2$ le vecteur des coordonnées du nœud i. On montre aisement que l'espace vectoriel V_h est de dimension N, et une base de V_h est donnée par la famille des *fonctions chapeaux* $(\varphi_i)_{1\le i\le N}$ telles que [AJOUTER FIGURE]

$$\varphi_i(x_j) = \delta_{ij}, \qquad \varphi_{i|T} \in \mathbb{P}^1(T) \quad \forall T \in \mathcal{T}_h.$$

Le support des fonctions chapeau est compact, et, pour chaque φ_i, $1 \le i \le N$, il coïncide avec les éléments qui partagent le nœud i. En décomposant la solution discrète dans la base, le problème (2.23) consiste à : Trouver $u_h = \sum_{j=1}^N U_j \varphi_j$ tel que

$$a(u_h, \varphi_i) = \int_\Omega f\varphi_i \qquad \forall 1 \le i \le n.$$

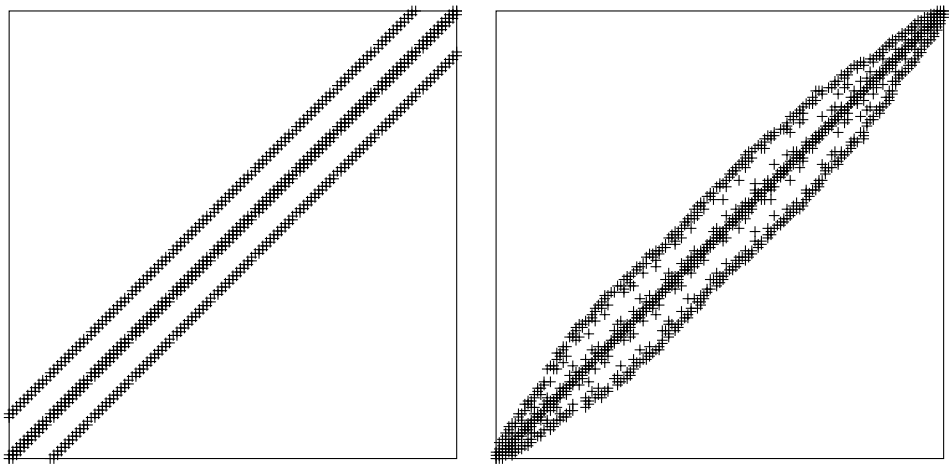
En utilisant la linéarité de a et en posant

$$A = (a_{ij})_{1\le i,j\le N} := \big(a(\varphi_j,\varphi_i)\big)_{1\le i\, j\le N} \in \mathbb{R}^{n,n}, \qquad U := (U_j)_{1\le j\le N}, \qquad b := \left(\int_\Omega f\varphi_i\right)_{1\le i\le N},$$

on obtient

$$a(u_h,\varphi_i) = a\left(\sum_{j=1}^N U_j\varphi_j, \varphi_i\right) = \sum_{j=1}^N a(\varphi_j,\varphi_i)U_j = b(f,\varphi_i) \iff AU = b. \tag{2.24}$$

Ceci montre que le problème (2.23) est équivalent au système linéaire $AU = b$. Examinons plus de près la structure de la matrice A. Par définition on a $a_{ij} = a(\varphi_j,\varphi_i) \neq 0$ si et seulement si les supports des fonctions chapeaux ont une intersection d'aire non nulle, ce qui vrai est s'il existe un triangle $T \in \mathcal{T}_h$ dont les nœuds i et j sont deux sommets. De ce fait la disposition des éléments non nuls de A dépend fortement de la triangulation $\mathcal{T}_h$, comme le montrent la Figure 2.1, où l'on compare le remplissage de la matrice A pour un maillage structuré et un maillage non structuré du domaine $\Omega = (0,1)^2$.



(a) Maillage structuré (b) Maillage non structuré

FIGURE 2.1 – Remplissage de la matrice A de (2.24)

Factorisation LU et remplissage

On peut se demander quel est l'effet de la factorisation LU sur le nombre d'éléments non nuls d'une matrice. Cette question est importante dans le cas des matrices creuses car elle a un impact sur leur stockage. On retiendra que, en général, le nombre total d'éléments non nuls dans les facteurs L et U est supérieur à celui de la matrice de départ. Par la suite on va considérer par simplicité une matrice SPD car, dans ce cas, la factorisation LU peut être obtenue sans *pivoting*. Pour énoncer un résultat plus précis il faut d'abord introduire la notion d'enveloppe convexe d'une matrice.

Définition 2.21 (Enveloppe convexe d'une matrice)**.** *Soit $A \in \mathbb{R}^{n,n}$ SPD et, pour tout $1 \leq i \leq N$,*

$$m_i(A) := i - \min\{j < i \mid a_{ij} \neq 0\}.$$

*On définit l'*enveloppe convexe *de A comme*

$$\mathcal{E}(A) := \left\{(i,j) \mid 0 < i - j \leq m_i(A)\right\}.$$

On a le résultat remarquable suivant, qui affirme que, au cours de la factorisation LU, d'éventuels nouveaux éléments non nuls ne peuvent apparaitre qu'à l'intérieure de l'enveloppe convexe.

Lemme 2.22 (Factorisation LU et remplissage)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SPD et $A = H^T H$ sa factorisation de Cholesky. Alors,*

$$\mathcal{E}(H + H^T) = \mathcal{E}(A).$$

Une conséquence importante du Lemme 2.22 est que la factorisation LU peut être stockée à la place de la matrice A sans besoin de mémoire supplémentaire si on mémorise tous les éléments de $\mathcal{E}(A)$ (y inclus les éléments non nuls). En effet, pour le facteur L il suffit de stocker le triangle inférieur, et poser implicitement les éléments diagonaux $= 1$. Cependant, lorsque des éléments très éloignés de la diagonale sont présents, cette solution n'est pas efficace.

Stockage des matrices creuses non structurées

On a déjà remarqué que, dans le stockage d'une matrice creuse, on peut tirer profit de la présence d'un nombre important d'éléments nuls. Il existe, en effet, différentes format de stockage qui permettent de réduire significativement l'occupation mémoire. Le but de cette section est de rappeller les plus importants.

Le format COO

Le format *skyline*

Le format CSR

Le format MSR

2.6 Exercices

Exercice 2.23 (Conditionnement). *Soit $A \in \mathbb{C}^{n,n}$. Prouver que*

(i) si $\|\cdot\|$ est une norme matricielle subordonnée, alors $\mathrm{cond}(A) = \mathrm{cond}(A^{-1}) \geq 1$, $\mathrm{cond}(\alpha A) = \alpha\,\mathrm{cond}(A)$ *for all* $\alpha \in \mathbb{R}^*$;

(ii) $\mathrm{cond}_2(A) = \sigma_n(A)/\sigma_1(A)$ *où* $\sigma_1(A)$ *et* $\sigma_n(A)$ *sont respectivement la plus petite et la plus grande valeur singulière de A ;*

(iii) si A est normale, $\mathrm{cond}_2(A) = |\lambda_n(A)|/|\lambda_1(A)|$ *où* $\lambda_1(A)$ *et* $\lambda_n(A)$ *sont respectivement la plus petite et la plus grande valeur propre en module de A ;*

(iv) pour toute matrice unitaire U, $\mathrm{cond}_2(U) = 1$ *et* $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(A)$.

(i) Puisque $\|\cdot\|$ est subordonnée nous avons

$$1 = \mathrm{cond}(I_n) = \mathrm{cond}(A^{-1}A) = \|A^{-1}A\|^2 \leq \|A^{-1}\|^2\|A\|^2 = \mathrm{cond}(A)^2,$$

à savoir, $\mathrm{cond}(A) \geq 1$. D'autre part, comme

$$\|\alpha A\| = \sup_{x\in\mathbb{C}^n,\|x\|=1} \|\alpha A x\| = \alpha \sup_{x\in\mathbb{C}^n,\|x\|=1} \|Ax\| = \alpha\|A\|,$$

et que $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$, nous avons d'après l'Exercice 1.58,

$$\mathrm{cond}(\alpha A) = \|\alpha A\|\|(\alpha A)^{-1}\| = \alpha\|A\|\frac{1}{\alpha}\|A^{-1}\| = \mathrm{cond}(A).$$

(ii) On peut diagonaliser A à l'aide de deux matrices unitaires U, $V \in \mathbb{C}^{n.n}$ (décomposition en valeurs singulières) comme suit :

$$U^H A V = \Sigma := \mathrm{diag}(\sigma_1, \ldots, \sigma_n).$$

Comme U et V sont unitaires nous avons que $\|U^H A V\|_2 = \|A\|_2$ et $\|(U^H A V)^{-1}\|_2 = \|V^H A^{-1} U\|_2 = \|A^{-1}\|_2$. Par conséquent,

$$\mathrm{cond}_2(\Sigma) = \mathrm{cond}_2(U^H A V) = \|U^H A V\|_2\|(U^H A V)^{-1}\|_2 = \|A\|_2\|A^{-1}\|_2 = \mathrm{cond}_2(A).$$

D'autre part, puisque Σ est diagonal, $\mathrm{cond}_2(\Sigma) = \sigma_n(A)/\sigma_1(A)$, ce qui conclut la preuve.

(iii) Puisque A est normale, elle admet une décomposition de la forme

$$W^H A W = \Lambda = \mathrm{diag}(\lambda_1(A), \ldots, \lambda_n(A)),$$

avec W matrice unitaire. Par conséquent, $A = W\Lambda W^H$ et

$$AA^H = A^H A = (W\Lambda W^H)^H(W\Lambda W^H) = W\Lambda^H W^H W\Lambda W^H = W\Lambda^H\Lambda W^H,$$

à savoir, $\lambda_i(A^H A) = \overline{\lambda}_i(A)\lambda_i(A) = |\lambda_i(A)|^2$. Pour s'en convaincre, il suffit de multiplier à droite l'inégalité ci-dessus par W et utiliser le fait que $W^H W = I_n$, ce qui donne

$$AA^H W = W\,\mathrm{diag}(|\lambda_i(A)|^2, \ldots, |\lambda_n(A)|^2).$$

On remarquera aussi que les colonnes de W sont des vecteurs propres de AA^H. D'après le point précédent on a aussi $V^H A^H A V = \Sigma^2$, à savoir, $\lambda_i(A^H A) = \sigma_i(A)^2$. Enfin,

$$|\lambda_i(A)|^2 = \lambda_i(A^H A) = \sigma_i(A)^2,$$

et la conclusion suit.

(iv) Conséquence de l'Exercice 1.58.

Exercice 2.24 (Conditionnement de la matrice de l'Exemple 2.20)**.** *L'objectif de cet exercice est d'étudier le conditionnement de la matrice définie dans l'Exemple 2.20 en fonction de sa taille. Vérifier que les valeurs propres de la matrice sont*

$$\lambda_k = 4h^{-1}\sin^2\left(\frac{k\pi}{2(n+1)}\right) \qquad 1 \le k \le n,$$

avec vecteurs propres

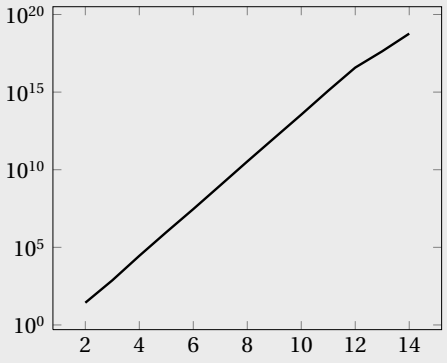
$$u^k = (u_i^k)_{1\le i\le n} = \left(\frac{ik\pi}{n+1}\right)_{1\le i\le n}.$$

On a pour tout $1 \le j \le n$,

$$\begin{aligned}(Au^k)_i &= \frac{1}{h}\left(-u_{i-1}^k + 2u_i^k - u_{i+1}^k\right)\\ &= \frac{1}{h}\left(\sin\left({}^{(i-1)k\pi}/_{(n+1)}\right) + 2\sin\left({}^{ik\pi}/_{(n+1)}\right) - \sin\left({}^{(i+1)k\pi}/_{(n+1)}\right)\right)\end{aligned}$$

COMPLETER

Exercice 2.25 (Factorisation $PA = LU$)**.** *Donner l'expression de $A := H(4)$ avec $H(4)$ matrice de Hilbert d'ordre 4 (voir l'Exercice 1.50) et calculer sa décomposition $PA = LU$ en effectuant un* pivoting *partiel. Estimer le conditionnement* $\mathrm{cond}_2(A)$*. Les matrices de Hilbert sont un exemple classique de matrices mal conditionnées, comme le montre la figure suivante où on trace une estimation du conditionnement de $H(n)$ en fonction de n :*



On a

$$A = H(4) = \begin{pmatrix} 1 & {}^1/_2 & {}^1/_3 & {}^1/_4 \\ {}^1/_2 & {}^1/_3 & {}^1/_4 & {}^1/_5 \\ {}^1/_3 & {}^1/_4 & {}^1/_5 & {}^1/_6 \\ {}^1/_4 & {}^1/_5 & {}^1/_6 & {}^1/_7 \end{pmatrix}.$$

Les étapes de la factorisation de Gauss sont données ci-dessous. Par brévité seuls les éléments non nuls des matrices $A^{(k)}$, $E^{(k)}$ et $P^{(k)}$ sont affichés. Le pivot satisfaisant la condi-

tion (2.10) est contourné par un rectangle.

$$A^{(1)} = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ & \boxed{1/12} & 1/12 & 3/40 \\ & 1/12 & 4/45 & 1/12 \\ & 3/40 & 1/12 & 9/112 \end{pmatrix}, \qquad L^{(1)} = \begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 1/3 & & 1 & \\ 1/4 & & & 1 \end{pmatrix}, \qquad P^{(1)} = I_n,$$

$$A^{(2)} = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ & 1/12 & 1/12 & 3/40 \\ & & 1/180 & 1/120 \\ & & \boxed{1/120} & 9/700 \end{pmatrix}, \qquad L^{(2)} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & 1 & 1 & \\ & 9/10 & & 1 \end{pmatrix}, \qquad P^{(2)} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{pmatrix},$$

$$A^{(3)} = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ & 1/12 & 1/12 & 3/40 \\ & & 1/120 & 9/700 \\ & & & -1/4200 \end{pmatrix}, \qquad L^{(3)} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & \frac{2}{3} & 1 \end{pmatrix}, \qquad P^{(3)} = I_n.$$

On a donc $PA = LU$ avec

$$P = P^{(2)}, \quad L = PL^{(1)}P^{(1)}L^{(2)}P^{(2)}L^{(3)}P^{(3)}, \quad U = A^{(3)}.$$

En utilisant les résultats énoncés dans les Propositions 2.8 et 2.11 pour le calcul de $(P^{(k)})^{-1}$ et $(L^{(k)})^{-1}$ on obtient

$$L = \begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 1/4 & 9/10 & 1 & \\ 1/3 & 1 & 2/3 & 1 \end{pmatrix}.$$

Comme A est symétrique, la formule (2.2) donne $\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$ et

$$\frac{1}{n}\|A\|_\infty\|A^{-1}\|_\infty \leq \|A\|_2\|A^{-1}\|_2 \leq n\|A\|_\infty\|A^{-1}\|_\infty,$$

où nous avons utilisé le fait que $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\|A\|_\infty$ dans la dernière estimation. L'inverse A^{-1} peut être calculée en utilisant la factorisation $PA = LU$ pour résoudre le système suivant

$$AA^{-1} = I_n \iff PAA^{-1} = P \iff LUA^{-1} = P \iff LY = P \text{ et } UA^{-1} = Y.$$

On obtient numériquement $\|A\|_\infty \approx 2.08{\cdot}10^0$ et $\|A^{-1}\|_\infty \approx 1.36{\cdot}10^4$, à savoir, $1.42{\cdot}10^4 \leq \mathrm{cond}_2(A) \leq 5.67{\cdot}10^4$.

Exercice 2.26 (Factorisation de Cholesky)**.** *Calculer la factorisation de Cholesky de la matrice suivante :*

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 8 & 12 & 16 \\ 3 & 12 & 27 & 30 \\ 4 & 16 & 30 & 37 \end{pmatrix}.$$

Proposer un critère pour arrêter la méthode de Cholesky si la matrice n'est pas définie positive.

La factorisation cherchée est de la forme

$$A = H^T H, \qquad H^T = \begin{pmatrix} h_{11} & 0 & & \\ h_{12} & h_{22} & & \\ h_{13} & h_{23} & h_{33} & \\ h_{14} & h_{24} & h_{34} & h_{44} \end{pmatrix}.$$

Les étapes du calcul des éléments de H^T sont détaillées ci-dessous en utilisant la notation de la formule (2.15).

$$\left(\begin{array}{l} \sqrt{a_{11}} = 1 \\ \quad \downarrow \\ \frac{a_{21}}{h_{11}} = 2 \longrightarrow \sqrt{a_{22} - h_{21}^2} = 2 \\ \quad \downarrow \\ \frac{a_{31}}{h_{11}} = 3 \longrightarrow \frac{a_{32} - h_{31} h_{21}}{h_{22}} = 3 \longrightarrow \sqrt{a_{33} - h_{31}^2 - h_{32}^2} = 3 \\ \quad \downarrow \\ \frac{a_{41}}{h_{11}} = 4 \longrightarrow \frac{a_{42} - h_{41} h_{21}}{h_{22}} = 4 \longrightarrow \frac{a_{43} - h_{41} h_{31} - h_{42} h_{32}}{h_{33}} = 2 \longrightarrow \sqrt{a_{44} - h_{41}^2 - h_{42}^2 - h_{43}^2} = 1 \end{array} \right)$$

On peut arrêter la factorisation de Cholesky si on trouve la racine carré d'un nombre négatif pendant le calcul d'un élément diagonal.

Exercice 2.27 (*Pivoting* total)**.** *On considère une version alternative de la procédure de* pivoting *décrite dans la Section 2.3.3 où on étend la recherche du meilleur pivot à toute la sous-matrice d'indices $k \leq i, j \leq n$, à savoir*

$$a_{lm}^{(k)} = \max_{k \leq i, j \leq n} |a_{ik}^{(k)}|.$$

Montrer que cette stratégie, dite de pivoting *total, équivaut à trouver une factorisation de la forme $PAQ = LU$, P et Q étant deux matrices de permutation.*

Exercice 2.28 (Descente pour un système triangulaire inférieure)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice triangulaire inférieure inversible, à savoir*

$$A = \begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Prouver des formules analogues à (2.13) *pour la résolution du système $Ax = b$, $b \in \mathbb{R}^n$.*

On trouve

$$x_1 = \frac{b_1}{x_{11}} \qquad x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{k=1}^{i-1} a_{ik} x_k \right) \quad \forall i = 2, \ldots, n-1. \tag{2.25}$$

Exercice 2.29 (Méthode de Gram–Schmidt)**.** *Soit $n \geq 1$ un entier et $X := (x_i)_{1\leq i\leq n}$ une famille de vecteurs libre de $\mathbb{R}^m$, $m \geq 1$. On considère la famille de vecteurs $Y := (y_i)_{1\leq i\leq n}$ obtenue à partir de X comme suit :*

$$y_1 = x_1, \qquad y_{i+1} = x_{i+1} - \sum_{j=1}^{i} \frac{(x_{i+1}, y_j)}{\|y_j\|_2^2} y_j \qquad \forall 1 \leq i \leq n-1, \tag{2.26}$$

où par brévité nous avons noté le produit interne canonique de $\mathbb{R}^m$ $(\cdot,\cdot)$ au lieu de $(\cdot,\cdot)_{\mathbb{R}^m}$. Montrer que Y est une famille libre et orthogonale*, à savoir*

$$(y_i, y_j) = \begin{cases} \|y_i\|_2^2 \neq 0 & \text{si } j = i, \\ 0 & \text{sinon.} \end{cases}$$

Proposer une modification de l'algorithme permettant d'obtenir une famille orthonormale*, à savoir telle que $\|y_i\|_2 = 1$ pour tout $1 \leq i \leq n$.*

On prouve le résultat par récurrence sur la taille n de la famille. Si $n = 1$ la famille Y est orthogonale car la famille X est libre (ce qui implique, en particulier $\|y_1\|_2 \neq 0$). Supposons la propriété vérifiée pour $n-1$ et prouvons-là pour n. Soit $\tilde{X} = (x_i)_{1\leq i\leq n-1}$ et $\tilde{Y} = (y_i)_{1\leq i\leq n-1}$ obtenue en appliquant l'algorithme (2.26). La famille $\tilde{Y}$ est orthogonale par l'hypothèse de récurrence. On a

$$y_n = x_n - \sum_{j=1}^{n-1} \frac{(x_n, y_j)}{\|x_n\|_2^2} y_j.$$

En multipliant scalairement l'égalité précédente par y_i, $1 \leq i \leq n-1$ on obtient

$$(y_n, y_i) = (x_n, y_i) - \sum_{j=1}^{n-1} \frac{(y_j, x_n)}{\|x_n\|_2^2} (y_j, y_i) = (x_n, y_i) - \frac{(y_i, x_n)}{\cancel{\|y_i\|_2^2}} \cancel{(y_i, y_i)} = 0,$$

ce qui montre que y_n est orthogonale aux vecteurs de $\tilde{Y}$. De plus, la famille Y est libre, car sinon on pourrait exprimer x_n comme combinaison linéaire des vecteurs de $\tilde{X}$ (ce qui est absurde car X est libre par hypothèse). Pour obtenir une famille orthonormale il suffit de diviser chaque vecteur de Y par sa norme euclidienne. On obtient ainsi l'algorithme suivant :

$$y_1 = x_1, \qquad \tilde{y}_{i+1} = x_{i+1} - \sum_{j=1}^{i} \frac{(x_{i+1}, y_j)}{\|y_j\|_2^2} y_j, \quad y_{i+1} = \frac{\tilde{y}}{\|\tilde{y}\|_2} \qquad \forall 1 \leq i \leq n-1,$$

Chapitre 3

Méthodes itératives

Dans ce chapitre on étudie quelques exemples de méthode *itératives* pour la résolution du système

$$Ax = b. \tag{3.1}$$

Ces méthodes ne fournissent en général pas la solution exacte du système, mais elle peuvent en principe l'approcher à une précision arbitraire. Comme dans le chapitre précédent, nous allons nous restreindre au cas où $A \in \mathbb{R}^{n,n}$, $x \in \mathbb{R}^n$, et $b \in \mathbb{R}^n$ non nul.

3.1 Généralités

A COMPLETER

3.2 Méthodes de point fixe

La première famille de méthodes que nous allons considérer repose sur la caractérisation de la solution du système linéaire (3.1) comme point fixe d'une fonction obtenue à partir de la matrice A et du seconde membre b.

3.2.1 Formulation abstraite basée sur une décomposition régulière

Soit $A \in \mathbb{C}^{n,n}$ une matrice inversible, $b \in \mathbb{C}^n$ non nul, et $P, N \in \mathbb{C}^{n,n}$ deux matrices telles que

$$A = P - N \text{ et } P \text{ est (facilement) inversible.} \tag{3.2}$$

Une décomposition de la forme (3.2) est dite *régulière*. On vérifie aisément que la solution du système est point fixe de la fonction $\Phi : y \to P^{-1}Ny - b$. En effet,

$$\begin{aligned} &Ax = (P-N)x = b \\ \iff &x = P^{-1}Nx + P^{-1}b = \Phi(x). \end{aligned} \tag{3.3}$$

Cette remarque nous conduit à considérer la méthode itérative suivante : Pour tout $k \geq 0$,

$$x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b = \Phi(x^{(k)}), \tag{3.4}$$

avec $x^{(0)} \in \mathbb{C}^n$ estimation initiale. A chaque itération de la méthode (3.4) on résout un système linéaire de matrice P, ce qui justifie l'hypothèse que P soit facile à inverser. Par exemple, si on choisit P diagonale ou triangulaire le coût lié à l'inversion sera de l'ordre de n ou n^2 opérations respectivement. Le résultat suivant est une conséquence naturelle de l'interprétation de la méthode (3.4) comme recherche d'un point fixe.

Remarque 3.1 (Hypothèse sur P)**.** *Nous avons supposé dans* (3.2) *que P soit* facilement *inversible. En effet, le choix $A = P$ (qui est toujours possible car A est inversible) donne une méthode itérative qui converge en une itération, mais dont la complexité équivaut à celle de la résolution du système linéaire $Ax = b$. Le choix de P doit donc représenter un compromis qui garantisse que le coût pour l'inversion de P à chaque itération reste raissonnable tout en évitant que le nombre d'itérations pour arriver à convergence explose.*

Lemme 3.2 (Condition nécessaire et suffisante pour la convergence de la méthode (3.4))**.** *La méthode* (3.4) *converge vers l'unique solution du système linéaire* (3.1) *si et seulement si* $\rho(P^{-1}N) < 1$.

Démonstration. Pour tout $k \geq 0$, soit $e^{(k)} := x^{(k)} - x$ l'erreur associée à l'estimation $x^{(k)}$. Comme $\rho(P^{-1}N) < 1$, il existe une norme matricielle subordonnée $\|\cdot\|$ telle que $\|P^{-1}N\| < 1$ (il s'agit d'une conséquence du Lemme 1.46). De par (3.3) et (3.4) nous avons pour $k \geq 1$,

$$e^{(k)} = x^{(k)} - x = \Phi(x^{(k-1)}) - \Phi(x) = P^{-1}N(x^{(k-1)} - x) = P^{-1}Ne^{(k-1)} = (P^{-1}N)^k e^{(0)},$$

donc

$$\|e^{(k)}\| = \|(P^{-1}N)^k e^{(0)}\| \leq \|P^{-1}N\|^k \|e^{(0)}\|,$$

et le seconde membre tend vers 0 lorsque $k \to +\infty$ si et seulement si $\|P^{-1}N\| < 1$. □

Remarque 3.3 (Choix de la norme dans la preuve du Lemme 3.2)**.** *Dans la preuve du Lemme 3.2 nous avons utilisé une norme $\|\cdot\|$ subordonnée satisfaisant la condition $\|P^{-1}N\| < 1$. Cependant, il s'ensuit du fait que $e^{(k)} \to 0 \in \mathbb{R}^n$ pour $k \to +\infty$ que, pour toute norme $\|\cdot\|_*$ sur $\mathbb{R}^n$, nous avons $\|e^{(k)}\|_* \to 0$ lorsque $k \to +\infty$.*

La condition du Lemme 3.2 est optimale mais difficile à vérifier en pratique. Une condition suffisante souvent plus simple à prouver est identifiée dans le lemme suivant.

Lemme 3.4 (Condition suffisante pour la convergence de la méthode (3.4))**.** *Soit $A \in \mathbb{C}^{n,n}$ hermitienne définie positive. On considère une décomposition régulière de la forme* (3.2) *avec $P \neq A$ et $P^H + N$ définie positive. Alors la méthode* (3.4) *converge vers la solution de* (3.1).

Démonstration. On commence par remarquer que

$$P^H + N = P^H + P - A. \tag{3.5}$$

Considérons maintenant la norme matricielle $\|\cdot\|_A$ subordonnée à la norme vectorielle définie par le produit scalaire $(x, y)_A := (Ax, y)_{\mathbb{C}^n}$ (que l'on notera toujours $\|\cdot\|_A$). De par la Proposition 1.45, il existe $y \in \mathbb{C}^n$ tel que $\|y\|_A = (Ay, y)_{\mathbb{C}^n} = 1$ et $\|P^{-1}N\|_A = \|P^{-1}Ny\|_A$. Nous avons

alors

$$\begin{aligned}
\|P^{-1}N\|_A^2 &= \|P^{-1}Ny\|_A^2 \\
&= (AP^{-1}Ny, P^{-1}Ny)_{\mathbb{C}^n} \\
&= (AP^{-1}(P-A)y, P^{-1}(P-A)y)_{\mathbb{C}^n} && (N = P-A) \\
&= ((A-AP^{-1}A)y, (I_n - P^{-1}A)y)_{\mathbb{C}^n} \\
&= (Ay,y)_{\mathbb{C}^n} - (AP^{-1}Ay,y)_{\mathbb{C}^n} - (Ay,P^{-1}Ay)_{\mathbb{C}^n} + (AP^{-1}Ay,P^{-1}Ay)_{\mathbb{C}^n} \\
&= 1 - (Az, A^{-1}Pz)_{\mathbb{C}^n} - (Pz,z)_{\mathbb{C}^n} + (Az,z)_{\mathbb{C}^n} && (z := P^{-1}Ay) \\
&= 1 - (P^H z,z)_{\mathbb{C}^n} - (Pz,z)_{\mathbb{C}^n} + (Az,z)_{\mathbb{C}^n} && (A^H = A) \\
&= 1 - ((P^H + P - A)z,z)_{\mathbb{C}^n} = 1 - ((P^H+N)z,z)_{\mathbb{C}^n} < 1, && (3.5)
\end{aligned}$$

où nous avons utilisé le fait que $P \neq A \implies y \neq 0$ pour conclure que $z \neq 0$ et le caractére défini positif de $P^H + N$ pour en déduire $((P^H+N)z,z)_{\mathbb{C}^n} > 0$. □

On conclut cette section en observant qu'une généralisation de (3.4) consiste à considérer des méthodes de la forme

$$x^{(k+1)} = Bx^{(k)} + f,$$

où $B \in \mathbb{C}^{n,n}$ est dite *matrice d'itération* et f est obtenu à partir du second membre b. Dans ce cas, la consistance de la méthode est garantie si en posant $x^{(k)} = x$ on trouve $x^{(k+1)} = x$, à savoir,

$$x = Bx + f, \tag{3.6}$$

et la solution exacte est un point fixe de $\Phi(y) = By + f$. Pour une méthode consistante on a

$$e^{(k)} = x^{(k)} - x = (Bx^{(k-1)} - f) - (Bx - f) = B^k e^{(0)}.$$

La condition nécessaire et suffisante pour la convergence de la méthode est donc

$$\rho(B) < 1. \tag{3.7}$$

3.2.2 Les méthodes de Jacobi et Gauss–Seidel

Les méthodes de Jacobi et Gauss–Seidel sont obtenues à partir de la décomposition suivante de la matrice A (*décomposition DEL*) :

$$A = D - E - L, \tag{3.8}$$

où D est la diagonale de A, E sa portion strictement triangulaire supérieure, L sa portion strictement triangulaire inférieure.

Example 3.5 (Décomposition DEL)**.** *La décomposition DEL de la matrice suivante :*

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 7 & 1 & 2 \\ 8 & 1 & 2 \end{pmatrix}$$

est

$$D = \begin{pmatrix} 2 & & \\ & 1 & \\ & & 2 \end{pmatrix}, \qquad E = \begin{pmatrix} & -3 & -1 \\ & & -2 \\ & & \end{pmatrix}, \qquad L = \begin{pmatrix} & & \\ -7 & & \\ -8 & -1 & \end{pmatrix}.$$

La *méthode de Jacobi* est obtenue avec $P = D$ et $N = E + L$. A l'itération $k+1$ la mise à jour est effectuée en posant

$$x^{(k+1)} = D^{-1}(E+L)x^{(k)} + D^{-1}b. \tag{3.9}$$

La matrice d'itération correspondante est

$$B_{\text{J}} = D^{-1}(E+L) = I - D^{-1}A. \tag{3.10}$$

Une condition suffisante pour la convergence de la méthode de Jacobi simple à vérifier est énoncée dans l'Exercice 3.17. Une variante de la méthode de Jacobi consiste à définir $x^{(k+1)}$ comme une moyenne pondérée de $x^{(k)}$ et de la valeur donnée par l'expression (3.9). Plus précisément, pour $\omega \in (0,1)$ on pose

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega\left[D^{-1}(E+L)x^{(k)} + D^{-1}b\right]. \tag{3.11}$$

Ce choix correspond à la méthode de surrélaxation JOR (*Jacobi Over-Relaxation*). La matrice d'itération correspondante est

$$B_{\text{JOR}} = (1-\omega)I_n + \omega B_{\text{J}}. \tag{3.12}$$

Lemme 3.6 (Convergence de la méthode JOR)**.** *Si $A = (a_{ij}) \in \mathbb{R}^{n,n}$ est SDP la méthode JOR converge pour tout $0 < \omega < {}^{2}/_{\rho(D^{-1}A)}$.*

Démonstration. On a

$$B_{\text{JOR}} = I_n - \omega D^{-1}D + \omega D^{-1}(E+L) = I_n - \omega D^{-1}A.$$

Comme A est SDP, $a_{kk} > 0$ pour tout $1 \le k \le n$ et $(D^{-1})_{kk} = 1/a_{kk} > 0$. De plus, les valeurs propres de A sont toutes strictement positives. Par conséquent, si $\omega > 0$,

$$\rho(I_n - \omega D^{-1}A) = \max_{\lambda \in \lambda(D^{-1}A)} (\omega|\lambda| - 1) = \rho(D^{-1}A) - 1.$$

La condition (3.7) est donc vérifiée si

$$\omega\rho(D^{-1}A) - 1 < 1 \iff \omega < \frac{2}{\rho(D^{-1}A)},$$

qui est la deuxième condition de l'énoncé. □

La *méthode de Gauss–Seidel* consiste à choisir $P = D - L$ et $N = E$. A l'itération $k+1$ la mise à jour est effectuée en posant

$$x^{(k+1)} = (D-L)^{-1}Ex^{(k)} + (D-L)^{-1}b. \tag{3.13}$$

La matrice d'itération correspondante est

$$B_{\text{GS}} = (D-L)^{-1}E.$$

Dans ce cas aussi on peut pondérer la valeur (3.13) par $x^{(k)}$, obtenant ainsi la méthode SOR (*Successive Over-Relaxation*). La mise à jour à l'itération $k+1$ revient donc à poser pour $\omega \in (0,1)$,

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega\left[(D-L)^{-1}Ex^{(k)} + (D-L)^{-1}b\right]. \tag{3.14}$$

La matrice d'itération correspondante est

$$B_{\text{SOR}} = (1-\omega)I_n + \omega B_{\text{GS}}.$$

3.2.3 La méthode du gradient

Le dernier algorithme de point fixe que nous allons examiner admet une interprétation qui permet de dépasser le cadre de la résolution de systèmes linéaires, et qui le rende applicable à des problèmes plus généraux d'optimisation. Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP. On définit la fonctionnelle quadratique $J : \mathbb{R}^n \to \mathbb{R}$ telle que

$$J(y) := \frac{1}{2} y^T A y - b^T y, \tag{3.15}$$

en on considère le problème de minimisation libre

$$\min_{y \in \mathbb{R}^n} J(y).$$

On peut prouver que ce problème admet une et une seule solution $x \in \mathbb{R}^n$ (dite *minimiseur global* de J sur $\mathbb{R}^n$) telle que

$$J(x) \leq J(y) \qquad \forall y \in \mathbb{R}^n,$$

et caractérisée par

$$\nabla J(x) = Ax - b = 0 \in \mathbb{R}^n.$$

Ce résultat est précisé dans le lemme suivant.

Lemme 3.7 (Minimisation d'une fonctionnelle quadratique et systèmes linéaires)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP. Alors, pour tout $b \in \mathbb{R}^n$,*

$$Ax = b \iff J(x) = \min_{y \in \mathbb{R}^n} J(y).$$

Démonstration. Pour tout $x, y \in \mathbb{R}^n$ et tout $t \in \mathbb{R}$ on a

$$\begin{aligned} J(z+ty) &= \frac{1}{2}(z+ty)^T A(z+ty) - b^T(z+ty) \\ &= \underbrace{\frac{1}{2} z^T A z - b^T z}_{=J(z)} + \frac{t}{2}\left(z^T A y + y^T A z\right) + \frac{t^2}{2} y^T A y - t\, b^T y \qquad \text{Proposition 1.6} \\ &= J(z) + t\, y^T (Az - b) + \frac{t^2}{2} y^T A y. \qquad \text{Symétrie de } A \end{aligned} \tag{3.16}$$

(i) *Implication* $J(x) = \min_{y \in \mathbb{R}^n} J(y) \implies Ax = b$. Puisque x est minimiseur global de J, l'identité (3.16) avec $z = x$ implique, pour tout $t \in \mathbb{R}$ et tout $y \in \mathbb{R}^n$,

$$J(x+ty) \geq J(x) \iff t\, y^T (Ax - b) + \frac{t^2}{2} y^T A y \geq 0.$$

En prenant $t > 0$, en divisant par t, et en faisant tendre $t \to 0^+$ il vient $y^T(Ax - b) \geq 0$, d'où $Ax - b = 0 \in \mathbb{R}^n$ car $y \in \mathbb{R}^n$ est générique.
(ii) *Implication* $Ax = b \implies J(x) = \min_{y \in \mathbb{R}^n} J(y)$. Pour tout $w \in \mathbb{R}^n$, en utilisant encore l'identité (3.16) avec $z = x$, $t = 1$, et $y = w - x$ il vient

$$J(w) = J(x) + t(w-x)^T \underbrace{(Ax - b)}_{=0 \in \mathbb{R}^n} + \frac{1}{2}(w-x)^T A (w-x).$$

La matrice A étant SDP, le dernier terme du membre de droite est positif. Par suite, x est minimiseur global de J sur V. □

L'idée de la *méthode du gradient* se base sur la remarque suivante : pour tout $y \in \mathbb{R}^n$ avec $\nabla J(y) \neq 0$, on peut réduire *au moins localement* la valeur de $J(y)$ en se déplaçant dans la direction $-\nabla J(y)$. En effet, $\nabla J(y)$ correspond à la direction de plus grande pente (positive) de J en y. Fixons une estimation initiale $x_0 \in \mathbb{R}^n$. Pour $k \geq 1$ et jusqu'à convergence on pose

$$r^{(k)} := -\nabla J(x^{(k)}) = b - Ax^{(k)}, \qquad x^{(k+1)} = x^{(k)} + \alpha r^{(k)},$$

où le réel $\alpha > 0$ est lié au module du déplacement dans la direction de la plus profonde descente. La direction de descente $r^{(k)}$ est le *résidu* du système linéaire correspondant à l'approximation $x^{(k)}$. On peut ensuite se poser la question de combien faut-il avancer dans la direction $r^{(k)}$. Pour répondre à cette question on cherche à reformuler la méthode du gradient sous la forme (3.4). On a

$$x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)}) = (I_n - \alpha A)x^{(k)} + \alpha b := B_G x^{(k)} + f_G.$$

Afin de maximiser la vitesse de convergence, le paramètre α doit être choisi de façon à minimiser le rayon spectral de la matrice d'itération B_G.

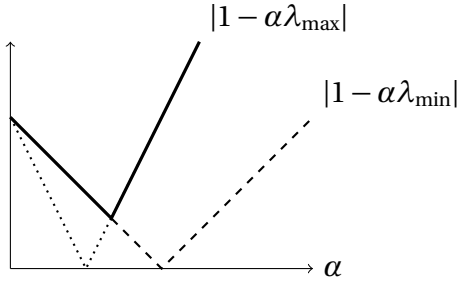
Lemme 3.8 (Vitesse de convergence de la méthode du gradient)**.** *Soit $A \in \mathbb{R}^{n,n}$ SDP et on note respectivement $\lambda_{\max}$ et $\lambda_{\min}$ la plus grande et la plus petite valeur propre de A. Alors la méthode du gradient converge pour toute $x^{(0)} \in \mathbb{R}^n$ et, en choisissant $\alpha = \frac{2}{\lambda_{\max}+\lambda_{\min}}$, on a pour $k \geq 1$,*

$$\|e^{(k+1)}\|_2 \leq \frac{\mathrm{cond}_2(A)-1}{\mathrm{cond}_2(A)+1}\|e^{(k)}\|_2.$$

Démonstration. On a par définition

$$\lambda(B_G) = 1 - \alpha\lambda(A) \Longrightarrow \rho(B_G) = \max_{\lambda\in\lambda(A)} |1-\alpha\lambda|.$$

Le graphe de la fonction $\varphi(\alpha) := \max_{\lambda\in\lambda(A)} |1-\alpha\lambda|$ est représenté en trait plein dans la figure suivante.



Le minimum de $\varphi(\alpha)$ est donc atteint pour α tel que $|1-\alpha\lambda_{\min}| = |1-\alpha\lambda_{\max}|$, à savoir, $\alpha = \frac{2}{\lambda_{\max}+\lambda_{\min}}$. De plus, on a

$$\varphi\left(\frac{2}{\lambda_{\min}+\lambda_{\max}}\right) = \rho(B_G) = \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}} = \frac{\cancel{\lambda_{\min}}(\mathrm{cond}_2(A)-1)}{\cancel{\lambda_{\min}}(\mathrm{cond}_2(A)+1)},$$

où nous avons utilisé l'équation (2.2) pour conclure $\mathrm{cond}_2(A) = \lambda_{\max}/\lambda_{\min}$. □

Example 3.9 (Convergence lente de la méthode du gradient)**.** *Le Lemme 3.8 montre que convergence de la méthode du gradient peut devenir lente lorsque la matrice A est mal conditionnée. [EXEMPLE NUMERIQUE]*

Le choix du paramètre d'accélération α suggéré par le Lemme 3.8 est souvent remplacé en pratique par d'autres expressions plus simples à calculer. Un choix qui est *localement optimal* est obtenu en posant, à chaque itération k, $\alpha = \alpha^{(k)}$ avec $\alpha^{(k)}$ unique solution du problème de minimisation suivant :

$$\min_{\alpha\in\mathbb{R}} \left\{ \|e^{(k+1)}\|_A = \|x^{(k+1)} - x\|_A = \|x^{(k)} - \alpha r^{(k)} - x\|_A \right\}. \tag{3.17}$$

On peut montrer (voir Exercice 3.21) que la solution de ce problème est

$$\alpha^{(k)} = \frac{\|r^{(k)}\|_2^2}{\|r^{(k)}\|_A^2}. \tag{3.18}$$

3.3 Méthode du gradient conjugué

Comme nous l'avons vu dans la section précédente, la méthode du gradient devient assez inefficace lorsque la matrice est mal conditionnée. En effet, le problème vient du fait que la direction de descente est choisie en utilisant uniquement des informations locales. L'idée de la méthode des méthodes conjuguées consiste à se déplacer dans une direction obtenue en tenant compte des directions déjà explorées au cours des itérations précédentes.

3.3.1 Vecteurs A-conjuguées

Définition 3.10 (Vecteurs A-conjugués)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP. Deux vecteurs* non nuls *$x, y \in \mathbb{R}^n$ sont dits A-conjugués si $y^T Ax = x^T Ay = 0$. Une famille $(x_i)_{1\le i\le p}$ de vecteurs de $\mathbb{R}^n$* non nuls *est dite A-conjuguée si*

$$\forall 1 \le i, j \le n \qquad \left(i \neq j \implies x_i^T A x_j = x_j^T A x_i = 0 \right).$$

Lemme 3.11 (Famille de vecteurs A-conjuguée)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP et $\mathcal{F} := (x_i)_{1\le i\le p}$, une famille de vecteurs A-conjuguée. Alors $\mathcal{F}$ est libre et $p \le n$. Si $p = n$, alors $\mathcal{F}$ est une base de $\mathbb{R}^n$.*

Démonstration. On cherche les combinaisons linéaires des vecteurs de $\mathcal{F}$ telles que

$$\lambda_1 x_i + \ldots + \lambda_p x_p = 0, \qquad \lambda_i \in \mathbb{R} \quad \forall 1 \le i \le p.$$

Soit $1 \le i \le p$. En multipliant à gauche par $x_i^T A$ l'égalité ci-dessus on obtient

$$x_i^T \sum_{i=1}^{p} \lambda_j x_j = \sum_{j=1}^{p} \lambda_j x_i^T A x_j = \lambda_i \|x_i\|_A^2 = 0.$$

Comme $x_i \neq 0$ par définition, ceci implique $\lambda_i = 0$, à savoir, la famille $\mathcal{F}$ est libre et, par conséquent, $p \le n$. Si $p = n$ on a une famille libre de cardinalité égale à la dimension de l'espace $\mathbb{R}^n$, et il s'agit donc d'une base. □

3.3.2 La méthode du gradient conjugué

Une méthode qui utilise des directions conjuguées est dite conjuguée. Nous allons montrer dans cette section comment construire une méthode conjuguée pour la solution du système linéaire (3.1) avec $A \in \mathbb{R}^{n,n}$ SDP. Par brévité dans ce qui suit on notera $(\cdot,\cdot)$ le produit scalaire canonique de $\mathbb{R}^n$ et $(\cdot,\cdot)_A := (A\cdot,\cdot)$. Soit $x^{(0)} \in \mathbb{R}^n$ une estimation initial de la solution et définissons le résidu $r^{(0)} := b - Ax^{(0)}$. Les directions de descente $w^{(k)}$, $k = 0,\ldots$, sont obtenues comme suit : $w^{(0)} = r^{(0)}$ et

$$w^{(k)} = r^{(k)} - \beta^{(k)} w^{(k-1)} \qquad k = 1,\ldots, \tag{3.19}$$

Une fois identifiée une direction $w^{(k)}$ appropriée, la mise à jour de la solution se fait en posant

$$x^{(k+1)} = x^{(k)} + \alpha^{(k+1)} w^{(k)}. \tag{3.20}$$

Ceci implique, en particulier,

$$r^{(k+1)} = b - Ax^{(k+1)} = b - Ax^{(k)} - \alpha^{(k+1)} Aw^{(k)} = r^{(k)} - \alpha^{(k+1)} Aw^{(k)}. \tag{3.21}$$

Pour déterminer les scalaires $\alpha^{(k)}$ et $\beta^{(k)}$ nous demandons pour tout $0 \le j \le k-1$,

$$(w^{(k)}, w^{(j)})_A = 0, \tag{CG1}$$

$$(r^{(k)}, w^{(j)}) = 0. \tag{CG2}$$

Les conditions (CG1) et (CG2) correspondent à la *méthode du gradient conjugué* (CG) et expriment le fait que la nouvelle direction $w^{(k)}$ et le résidu $r^{(k)}$ sont respectivement A-orthogonale et orthogonal aux directions précédentes. Ces deux conditions montrent une caractéristique importante de la méthode CG : la mise à jour n'est pas uniquement basée sur une information locale (direction de plus grande pente), mais elle tient compte également des itérations précédentes. Comme on le verra plus loin, on peut assurer les deux conditions (CG1)–(CG2) *sans mémoriser* les directions $w^{(j)}$, $1 \le j < k-1$. Ce point est très important en pratique, car il implique que la mise en œuvre de la méthode CG ne demande pas plus de mémoire que les autres méthodes itératives que l'on a étudié jusqu'à là.

Le lemme suivant montre de façon constructive l'existence d'une famille de directions de descente $(w^{(k)})_{k=0,\ldots}$ qui remplissent les conditions (CG1)–(CG2) et permet de préciser l'expression des paramètres $\alpha^{(k)}$ et $\beta^{(k)}$.

Lemme 3.12 (Existence des directions $(w^{(k)})_{k=0,\ldots}$)**.** *Pour tout $x^{(0)} \in \mathbb{R}^n$ il existe des valeurs des paramètres $\alpha^{(k)}$ et $\beta^{(k)}$, $k = 1,\ldots$, et une famille de vecteurs $(w^{(k)})_{k=0,\ldots}$ telle que les conditions* (CG1)–(CG2) *sont satisfaites pour l'itération définie par* (3.19) *et* (3.20)*.*

Démonstration. La preuve procède par induction Posons $w^{(0)} = r^{(0)}$. Pour $k = 1$ on a

$$w^{(1)} = r^{(1)} - \beta^{(1)} w^{(0)}.$$

On choisit $\beta^{(1)}$ tel que (CG1) soit vérifiée :

$$0 = (w^{(0)}, w^{(1)})_A = (w^{(0)}, r^{(1)} - \beta^{(1)} w^{(0)})_A \iff \beta^{(1)} = \frac{(w^{(0)}, w^{(1)})_A}{\|w^{(0)}\|_A^2}.$$

On identifie ensuite la valeur de $\alpha^{(1)}$ qui assure la condition (CG2). De par (3.21) on a

$$0=(w^{(0)},r^{(1)})=(r^{(0)},r^{(0)}-\alpha^{(1)}Aw^{(0)})=\|r^{(0)}\|_2^2-\alpha^{(1)}\|w^{(0)}\|_A^2 \iff \alpha^{(1)}=\frac{(w^{(0)},r^{(0)})}{\|w^{(0)}\|_A^2},$$

ce qui prouve l'existence de $w^{(1)}$. Supposons maintenant (CG1)–(CG2) vérifiées pour $k\geq 1$ et prouvons l'existence de $\alpha^{(k+1)}$, $\beta^{(k+1)}$ et $w^{(k+1)}$ tels les deux conditions restent vraies. En utilisant l'expression (3.21) pour $r^{(k+1)}$ on a

$$(w^{(k)},r^{(k+1)})=(w^{(k)},r^{(k)}-\alpha^{(k+1)}Aw^{(k)})=(w^{(k)},r^{(k)})-\alpha^{(k+1)}\|w^{(k)}\|_A^2,$$

à savoir, (CG2) pour $j=k$ est vérifiée pour

$$\boxed{\alpha^{(k+1)}=\frac{(w^{(k)},r^{(k)})}{\|w^{(k)}\|_A^2}.} \tag{3.22}$$

Prouvons maintenant (CG2) pour tout $0\leq j\leq n-1$. Il suffit d'observer que, pour tout $0\leq j\leq n-1$,

$$\begin{aligned}(w^{(j)},r^{(k+1)})&=(w^{(j)},r^{(k)}-\alpha^{(k+1)}Aw^{(k)}) && (3.21)\\ &=(w^{(j)},r^{(k)})-\alpha^{(k+1)}(w^{(j)},w^{(k)})_A=0. && \text{(recurrence (CG1)–(CG2))}\end{aligned}$$

Venons maintenant à (CG1). Nous avons

$$\begin{aligned}(w^{(k)},w^{(k+1)})_A&=(w^{(k)},r^{(k+1)}-\beta^{(k+1)}w^{(k)})_A && (3.19)\\ &=(w^{(k)},r^{(k+1)})_A-\beta^{(k+1)}\|w^{(k)}\|_A^2,\end{aligned}$$

et (CG1) pour $j=k$ est donc vérifiée pour

$$\boxed{\beta^{(k+1)}=\frac{(w^{(k)},r^{(k+1)})_A}{\|w^{(k)}\|_A^2}.}$$

Il ne reste plus qu'à prouver (CG1) pour $0\leq j\leq k-1$. Puisque $w^{(0)}=r^{(0)}$ et chaque nouvelle direction $w^{(k)}$ est obtenue à partir de $r^{(k)}$ et des directions $w^{(j)}$, $0\leq j\leq k-1$, on a

$$V_{k+1}:=\operatorname{span}(w^{(0)},\dots,w^{(k)})=\operatorname{span}(r^{(0)},\dots,r^{(k)}).$$

La condition (CG2) pour $0\leq j\leq k$ implique $r^{(k+1)}\in V_{k+1}^{\perp}$. D'autre part, pour tout $0\leq j\leq k-1$, on a

$$Aw^{(j)}=\frac{1}{\alpha^{(j+1)}}\left(r^{(j)}-r^{(j+1)}\right)\in V_{k+1}\Longrightarrow (w^{(j)},r^{(k+1)})_A=0. \tag{3.23}$$

En utilisant les remarques précédentes on trouve

$$\begin{aligned}(w^{(j)},w^{(k+1)})_A&=(w^{(j)},r^{(k+1)}-\beta^{(k+1)}w^{(k)})_A && (3.19)\\ &=(Aw^{(j)},r^{(k+1)})-\beta^{(k+1)}(w^{(j)},w^{(k)})_A=0. && \text{(3.23), récurrence (CG1)}\end{aligned}$$

Ceci conclut la preuve. □

Compte tenu du lemme précédent, la méthode CG est définie comme suit : Pour une estimation initiale $x^{(0)} \in \mathbb{R}^n$, poser $r^{(0)} = b - Ax^{(0)}$, $w^{(0)} = r^{(0)}$ et, pour $k = 0, \ldots,$

$$\alpha^{(k+1)} = \frac{(w^{(k)}, r^{(k)})}{\|w^{(k)}\|_A^2}, \tag{3.24a}$$

$$x^{(k+1)} = x^{(k)} + \alpha^{(k+1)} w^{(k)}, \tag{3.24b}$$

$$r^{(k+1)} = r^{(k)} - \alpha^{(k+1)} A w^{(k)}, \tag{3.24c}$$

$$\beta^{(k+1)} = \frac{(w^{(k)}, r^{(k+1)})_A}{\|w^{(k)}\|_A^2}, \tag{3.24d}$$

$$w^{(k+1)} = r^{(k+1)} - \beta^{(k+1)} w^{(k)}. \tag{3.24e}$$

Théorème 3.13 (Convergence de la méthode CG)**.** *Soit $A \in \mathbb{R}^{n,n}$ SDP et $b \in \mathbb{R}^n$. La méthode CG pour la solution du système linéaire $Ax = b$ converge en au plus n itérations.*

Démonstration. La famille $\mathscr{F} := (w^{(0)}, \ldots, w^{(n-1)})$ de cardinalité n est libre d'après le Lemme 3.11, et elle est donc une base de l'espace $\mathbb{R}^n$. Comme $r^{(n)}$ est orthogonal à tout vecteur de $\mathscr{F}$ on doit avoir $r^{(n)} = 0$, à savoir, $x^{(n)} = x$. □

Le Théorème 3.13 montre que la méthode du gradient conjugué n'est pas une méthode itérative au sens strict, car elle converge après un nombre fini d'itérations. Cependant, son intérêt est lié au fait que, dans de nombreux cas pratiques, le résidu décroît rapidement, et le critère $\|r^{(k)}\| \leq \epsilon$ est vérifié après un petit nombre d'itérations. En effet, le résultat suivant montre que la valeur de la fonctionnelle J définie par (3.15) décroît à chaque itération de la méthode (de par le Lemme 3.7, on sait que résoudre le système linéaire $Ax = b$ avec A SDP équivaut à minimiser J). Dans le jargon de l'optimisation on dit que $w^{(k)}$ est une *direction de descente stricte* de J en $x^{(k-1)}$.

Proposition 3.14 (Monotonicité des itérations CG)**.** *Soit $A \in \mathbb{R}^{n,n}$ une matrice SDP, $b \in \mathbb{R}^n$ et J définie par* (3.15)*. On considère l'itération $(k+1)$ de la méthode du gradient définie par* (3.24)*. Alors, si $w^{(k)} \neq 0$ et $\alpha^{(k+1)} \neq 0$,*

$$J(x^{(k+1)}) < J(x^{(k)}).$$

Si $\alpha^{(k+1)} = 0$, $x^{(k)}$ est le minimiseur de J et $Ax^{(k)} = b$.

Démonstration. On commence par remarquer que

$$\begin{aligned}(w^{(k)}, r^{(k)}) &= (w^{(k)}, r^{(k+1)} + \alpha^{(k+1)} A w^{(k)}) && (3.21)\\ &= \underbrace{(w^{(k)}, r^{(k+1)})}_{=0} + \alpha^{(k+1)} \|w^{(k)}\|_A^2 > 0, && (\text{CG2})\end{aligned}$$

où l'inégalité est stricte car $w^{(k)} \neq 0$. Nous avons alors

$$\begin{aligned}J(x^{(k+1)}) &= \frac{1}{2}(x^{(k)} + \alpha^{(k+1)} w^{(k)}, x^{(k)} + \alpha^{(k+1)} w^{(k)})_A - (b, x^{(k)} + \alpha^{(k+1)} w^{(k)}) && (3.15) \text{ et } (3.20)\\ &= \underbrace{\frac{1}{2}(x^{(k)}, x^{(k)})_A - (b, x^{(k)})}_{=J(x^{(k)})} + \alpha^{(k+1)} \underbrace{(Ax^{(k)} - b, w^{(k)})}_{=-(r^{(k)}, w^{(k)})} + \underbrace{\frac{(\alpha^{(k+1)})^2}{2} \|w^{(k)}\|_A^2}_{=\frac{\alpha^{(k+1)}}{2}(w^{(k)}, r^{(k)})} && (3.22)\\ &= J(x^{(k)}) - \frac{\alpha^{(k+1)}}{2}(w^{(k)}, r^{(k)}) < J(x^{(k)}),\end{aligned}$$

où nous avons conclu grâce au fait que $(w^{(k)}, r^{(k)}) > 0$ prouvé précédemment. □

Il est important de retenir que la méthode CG accompagnée d'un bon préconditionneur est la méthode de choix pour la résolution de systèmes linéaires caractérisés par une matrice SDP.

3.4 Méthodes basées sur les espaces de Krylov

Les méthodes de cette section sont caractérisées par le fait que l'itération k consiste à chercher une solution de la forme $x^{(k)} = x^{(0)} + y$ avec y appartenant à l'espace de Krylov de dimension k et $x^{(k)}$ satisfaisant un critère de distance minimal de x.

3.4.1 Espaces de Krylov

Définition 3.15 (Espace de Krylov $K_k(A, v)$)**.** *Soit $A \in \mathbb{R}^{n,n}$ et $v \in \mathbb{R}^n$ non nul. Pour tout $k \in \mathbb{N}$ on appelle* espace de Krylov associé au vecteur v *et on note $K_k(A, v)$ le sous-espace vectoriel de $\mathbb{R}^n$*

$$K_k(A, v) := \text{span}(v, Av, \dots, A^{k-1}v).$$

Nous avons de manière générale $K_k(A, v) \subset K_{k+1}(A, v)$ et, comme $K_k(A, v) \subset \mathbb{R}^n$ pour tout $k \geq 0$, il existe un indice k_0 à partir duquel l'inclusion devient une égalité, à savoir

$$\begin{cases} \dim(K_k(A, v)) = k & \text{si } 0 \leq k \leq k_0, \\ \dim(K_k(A, v)) = k_0 & \text{si } k \geq k_0. \end{cases}$$

La suite $(K_k(A, v))_{k \in \mathbb{N}}$ devient donc stationnaire pour $k \geq k_0$.

3.4.2 Retour sur la méthode du gradient conjugué

On commence par un rappel. Soit V un espace vectoriel, K un sous-espace vectoriel de V. Pour tout $v \in V$ on appelle *projection orthogonale de v sur K* l'élément de $\Pi_K v \in K$ qui minimise la distance de v,

$$\|\Pi_K v - v\|_V = \min_{w \in K} \|w - v\|_V.$$

Cette élément est unique et il est caractérisé par la relation suivante :

$$(\Pi_K v - v, w)_V = 0 \qquad \forall w \in K. \tag{3.25}$$

Proposition 3.16. *Soient $w^{(0)}, \dots, w^{(k)}$ le directions de descente engendrées par la méthode CG jusqu'à l'itération k. On a*

$$K_{k+1}(A, r^{(0)}) = V_{k+1} := \text{span}(w^{(0)}, \dots, w^{(k)}).$$

Si, de plus, $k \leq k_0$, la famille $(w^{(0)}, \dots, w^{(k)})$ est une base de $K_{k+1}(A, r^{(0)})$.

Démonstration. On procède par induction. Puisque $r^{(0)} = w^{(0)}$, $r^{(0)} \in V_1$. Supposons maintenant le résultat vrai pour k et prouvons-le pour $(k+1)$:

$$\left(K_k(A, r^{(0)}) = V_k\right) \Longrightarrow \left(K_{k+1}(A, r^{(0)}) = V_{k+1}\right).$$

Par l'hypothèse de récurrence, $A^i r^{(0)} \in V_k \subset V_{k+1}$ pour tout $0 \leq i \leq k-1$. Si le résultat n'était pas vrai pour $(k+1)$ on devrait donc avoir $A^k r^{(0)} \in V_{k+1}^{\perp}$, à savoir,

$$(A^k r^{(0)}, w^{(j)}) = (A^{k-1} r^{(0)}, w^{(j)})_A = 0 \qquad \forall 0 \leq j \leq k. \tag{3.26}$$

D'autre part, comme $A^{k-1} r^{(0)} \in V_k$, il existe des réels $\lambda_0, \ldots, \lambda_{k-1}$ non tous nuls (car, comme A est SDP, $\ker(A) = \{0 \in \mathbb{R}^n\}$ et $r^{(0)} \neq 0 \in \mathbb{R}^n \Longrightarrow A^{k-1} r^{(0)} \neq 0 \in \mathbb{R}^n$) tels que

$$A^{k-1} r^{(0)} = \lambda_0 w^{(0)} + \ldots + \lambda_{k-1} w^{(k-1)}. \tag{3.27}$$

En remplaçant (3.27) dans (3.26) et en utilisant (CG1) on obtient

$$\lambda_i = 0 \qquad \forall 0 \leq j \leq k-1,$$

ce qui est absurde. □

Grâce au résultat prouvé dans la proposition précédente, on peut reformuler la condition d'orthogonalité des résidus (CG2) comme suit :

$$r^{(k)} \in K_k(A, r^{(0)})^{\perp}. \tag{3.28}$$

De plus, on peut conclure que $x^{(k)} = x^{(0)} + y^{(k)}$ avec $y^{(k)} \in V_{k+1} = K_{k+1}(A, r^{(0)})$, et, par conséquent, $r^{(k)} = b - Ax^{(k)} = r^{(0)} - Ay^{(k)}$. La relation (3.28) devient alors

$$(r^{(0)} - Ay^{(k)}, y) = (A^{-1} r^{(0)} - y^{(k)}, y)_A = 0 \qquad \forall y \in K_k(A, r^{(0)}), \tag{3.29}$$

et, compte tenu de (3.25), ceci équivaut à

$$y^{(k)} = \Pi_{K_k(A, r^{(0)})}(A^{-1} r^{(0)}).$$

Cette remarque permet une nouvelle interprétation de la méthode CG : il s'agit en effet de la méthode qui obtient à l'itération k l'estimation $x^{(k)}$ en projetant le résidu modifié $A^{-1} r^{(0)}$ sur l'espace de Krylov $K_{k-1}(A, r^{(0)})$. Ainsi, la convergence a lieu à l'itération k_0 car la suite $(K_k(A, r^{(0)}))_{k \in \mathbb{N}}$ devient stationnaire au delà de ce point.

3.4.3 L'algorithme de Gram–Schmidt–Arnoldi

Lorsque $A \in \mathbb{R}^{n,n}$ est SDP, la méthode CG construit des directions de descente qui forment une base pour l'espace de Krylov $K_k(A, r^{(0)})$ (pour $k \leq k_0$). Dans le cas d'une matrice générale, il est possible de construire une base en utilisant l'*algorithme de Gram–Schmidt–Arnoldi* (GSA) basé sur la procédure d'orthonormalisation de Gram–Schmidt (voir l'Exercice 2.29 pour plus de détails). Une différence majeure par rapport à la méthode CG est que, dans ce cas, il est nécessaire de mémoriser la base, ce qui limite en pratique la dimension de

l'espace de Krylov que l'on peut considérer. Soit $v \in \mathbb{R}^n$, $v \neq 0$. L'algorithme GSA consiste à poser $v_1 := v/\|v\|_2$ puis à calculer pour $k = 1, \ldots$

$$\begin{aligned} h_{ik} &= v_i^T A v_k \qquad \forall 1 \leq i \leq k \\ w_k &= A v_k - \sum_{i=1}^{k} h_{ik} v_i, \qquad h_{k+1,k} = \|w\|_2. \end{aligned} \tag{3.30}$$

Si $w_k = 0$ l'algorithme s'arrête et $k = k_0$, sinon on pose $v_{k+1} = w_k/\|w_k\|_2$, on incrément k d'une unité, et on reprend. Les vecteurs $v_1, \ldots, v_{k_0}$ ainsi obtenus forment une base pour l'espace de Krylov $K_{k_0}(A, v)$. En pratique, on peut arrêter l'algorithme à toute itération $1 \leq m \leq k_0$. La mémorisation de la base demande alors de stocker une matrice de Hessenberg supérieure $H \in \mathbb{R}^{m+1,m}$ contenant les coefficients h_{ij} donnés par (3.30).

3.4.4 Principe des méthodes de Arnoldi et GMRes

Une fois construite une base pour l'espace $K_k(A, r^{(0)})$, on peut obtenir des méthodes pour la résolution du système linéaire $Ax = b$ (avec A inversible mais en général non SDP) en cherchent $x^{(k)}$ de la forme $x^{(k)} = x^{(0)} + y^{(k)}$ avec $y^{(k)}$ choisi de façon à satisfaire un critère de distance minimale de x. Nous avons, alors,

$$r^{(k)} = b - Ax^{(k)} = b - A(x^{(0)} + y^{(k)}) = r^{(0)} - Ay^{(k)}.$$

Une première idée consiste à s'inspirer de la méthode du gradient et imposer cette fois-ci que le *résidu* $r^{(k)}$ (et non pas le résidu modifié $A^{-1}r^{(k)}$) soit orthogonal à tout vecteur de $K_k(A, r^{(0)})$, à savoir on cherche $y^{(k)} \in \mathbb{R}^n$ tel que

$$(r^{(0)} - Ay^{(k)}, w) = 0 \qquad \forall w \in K_k(A, r^{(0)}).$$

Ce choix correspond à la *méthode de Arnoldi* ou FOM (*Full Orthogonalization Method*). Un deuxième choix possible consiste à minimiser la norme éuclidienne du résidu en cherchant $y^{(k)} \in \mathbb{R}^n$ solution de

$$\min_{y \in K_k(A, r^{(0)})} \|r^{(0)} - Ay\|_2.$$

On obtient ainsi la méthode GMRes (*Generalized Minimum Residual*). Comme c'était le cas pour la méthode CG (voir Théorème 3.13), les méthodes de Arnoldi et GMRes sont en effet des méthodes directes, car, si on néglige les erreurs d'arrondi, elles convergent en au plus n itérations. Cependant, d'un point de vue pratique, aucune de ces méthodes n'est utilisée comme méthode directe, d'un part parce que n peut devenir très grand, d'autre part parce que, dans le cas des méthodes de Arnoldi et GMRes, il est nécessaire de mémoriser la base de l'espace de Krylov $K_k(A, r^{(0)})$. La stratégie pour limiter le coût computationnel consiste à fixer la dimension maximale de l'espace de Krylov considéré et à utiliser des préconditionneurs pour accélérer la convergence. Une variation très populaire est le *restart* (réinitialisation), qui consiste à effectuer un maximum de m itérations successives et, si la convergence n'a pas eu lieu, poser $x^{(0)} = x^{(m)}$ et recommencer. Les méthodes correspondantes sont alors notées FOM(m) et GMRes(m) respectivement.

3.5 Exercices

Exercice 3.17 (Matrice à diagonale dominante)**.** *Soit $A=(a_{ij})\in\mathbb{R}^{n,n}$ une matrice à diagonale dominante par lignes, à savoir*

$$|a_{ii}|>\sum_{j=1,\,j\neq i}^{n}|a_{ij}| \qquad \forall 1\leq i\leq n.$$

Soit, de plus, $b\in\mathbb{R}^n$, $b\neq 0$. Montrer que la méthode de Jacobi pour la résolution du système linéaire $Ax=b$ est convergente.

La matrice d'itération de la méthode de Jacobi est

$$B:=D^{-1}(D-A),$$

où nous avons noté $D\in\mathbb{R}^{n,n}$ la matrice diagonale telle que $D=\mathrm{diag}(A)$. Pour tout $1\leq i\leq n$ nous avons donc

$$b_{ii}=0 \quad \forall 1\leq i\leq n, \qquad b_{ij}=\frac{a_{ij}}{a_{ii}} \quad \forall 1\leq j\leq n,\, j\neq i,$$

et, par conséquent,

$$\|B\|_\infty=\max_{1\leq i\leq n}\frac{1}{|a_{ii}|}\sum_{j=1,\,j\neq i}^{n}|a_{ij}|<1.$$

Comme $\rho(B)\leq\|B\|_\infty<1$ ($\|\cdot\|_\infty$ étant une norme matricielle subordonnée), la méthode est convergente.

Exercice 3.18 (Méthode des directions alternées)**.** *Soit $A\in\mathbb{R}^{n,n}$ une matrice SDP décomposée en $A=A_1+A_2$ avec A_1 et A_2 matrices SDP et $b\in\mathbb{R}^n$, $b\neq 0$. Soient $\alpha_1,\alpha_2\in\mathbb{R}^+_*$. On considère la méthode itérative suivante pour résoudre le système linéaire $Ax=b$:*

$$(I_n+\alpha_1A_1)x^{(k+1/2)}=(I_n-\alpha_1A_2)x^{(k)}+\alpha_1b, \tag{3.31a}$$
$$(I_n+\alpha_2A_2)x^{(k+1)}=(I_n-\alpha_2A_1)x^{(k+1/2)}+\alpha_2b. \tag{3.31b}$$

On souhaite analyser la convergence de cette méthode. A ce propos,

(i) prouver que, pour toute matrice SDP $M\in\mathbb{R}^{n,n}$ et tout $\xi\in\mathbb{R}^+_*$, la matrice $I_n+\xi M$ est inversible. En déduire que les matrices $I_n+\alpha_iA_i$, $i\in\{1,2\}$, sont définies positives ;

(ii) montrer la méthode est consistante, à savoir, si $x^{(k)}=x$, alors $x^{(k+1/2)}=x^{(k+1)}=x$;

(iii) interpreter la méthode comme une méthode de point fixe de la forme

$$x^{(k+1)}=Bx^{(k)}+f;$$

(iv) montrer la convergence de la méthode en évaluant le rayon spectral de B.

(i) Il suffit de prouver que $I_n+\xi M$ est à son tour SDP. La symétrie est évidente. Soit $x\in\mathbb{R}^n$, $x\neq 0$. Par définition on a

$$x^T(I_n+\xi M)x=x^Tx+\xi x^TMx>0,$$

ce qui prouve que $I_n+\xi M$ est définie positive. Le fait que les matrices $I_n+\alpha_iA_i$, $i\in\{1,2\}$, sont définies positives est une conséquence immédiate.

(ii) Comme $I_n + \alpha A_1$ est SDP, elle est aussi inversible. Pour une valeur fixée de $x^{(k)}$ il existe alors un unique $x^{(k+1/2)}$ satisfaisant (3.31a). En remplaçant dans (3.31a) on trouve

$$\begin{aligned}(I_n+\alpha_1 A_1)x^{(k+1/2)} &= x-\alpha_1 A_2 x+\alpha_1 b \\ &= x-\alpha_1(A-A_1)x+\alpha_1 b \qquad (A_2=A-A_1)\\ &= x+\alpha_1 A_1 x=(I_n+\alpha_1 A_1)x, \qquad (Ax=b)\end{aligned}$$

à savoir, $x^{(k+1/2)} = (I_n + \alpha_1 A_1)^{-1}(I_n + \alpha - 1A_1)x = x$. Pour prouver que $x^{(k+1)} = x$ on peut procéder de façon similaire en remplaçant $x^{(k+1/2)}$ par x dans (3.31b) :

$$\begin{aligned}(I_n+\alpha_2 A_2)x^{(k+1)} &= (I_n-\alpha_2 A_1)x+\alpha_2 b \\ &= x-\alpha_2(A-A_2)x+\alpha_2 b \qquad (A_2=A-A_2)\\ &= x+\alpha_2 A_2 x=(I_n+\alpha_2 A_2)x, \qquad (Ax=b)\end{aligned}$$

et, par conséquent, $x^{(k+1)} = (I_n + \alpha_2 A_2)^{-1}(I_n + \alpha_2 A_2)x = x$, ce qui prouve le résultat souhaité.

(iii) De par (3.31a) on a $x^{(k+1/2)} = (I_n + \alpha_1 A_1)^{-1}(I_n - \alpha_1 A_2)x^{(k)} + \alpha_1(I_n + \alpha_1 A_1)^{-1}b$. En remplaçant dans (3.31b) on obtient

$$x^{(k+1)} = Bx^{(k)} + f,$$

avec

$$\begin{aligned}B &:= (I_n+\alpha_2 A_2)^{-1}(I_n-\alpha_2 A_1)(I_n+\alpha_1 A_1)^{-1}(I_n-\alpha_1 A_2)\\ f &:= \alpha_1(I_n+\alpha_2 A_2)^{-1}(I_n-\alpha_2 A_1)(I_n+\alpha_1 A_1)^{-1}b+\alpha_2(I_n+\alpha_2 A_2)^{-1}b.\end{aligned}$$

La consistance de la méthode prouvée au point précédent se traduit par la rélation

$$x = Bx + f.$$

(iv) En utilisant l'expression trouvée au point précédent on peut estimer l'erreur $e^{(k)} := x^{(k)} - x$ par récurrence :

$$e^{(k)} = x^{(k)} - x = (Bx^{(k-1)} + f) - (Bx + f) = Be^{(k-1)} = B^k e^{(0)}.$$

La méthode est donc convergente si et seulement si $\rho(B) < 1$. On peut estimer (voir aussi l'Exercice 1.54)

$$\rho(B) \le \max_{\lambda\in\lambda(A_1)}\left|\frac{1-\alpha_2\lambda}{1+\alpha_1\lambda}\right| \times \max_{\lambda\in\lambda(A_2)}\left|\frac{1-\alpha_1\lambda}{1+\alpha_2\lambda}\right|. \tag{3.32}$$

Or, comme les valeurs propres de A_1 et A_2 sont > 0, il suffit de prendre $\alpha_1 = \alpha_2 = \alpha > 0$ pour que les deux facteurs soient < 1 et, donc, $\rho(B) < 1$.

Exercice 3.19 (Analyse d'une méthode itérative)**.** *Pour la résolution du système $Ax = b$ avec*

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} 3 \\ 5 \end{pmatrix},$$

on considère la méthode itérative suivante : Pour $k \geq 0$,

$$x^{(k+1)} = B(\theta)x^{(k)} + f(\theta),$$

avec $x^{(0)} \in \mathbb{R}^n$ estimation initiale, $\theta \in \mathbb{R}$, et

$$B = \frac{1}{4}\begin{pmatrix} 2\theta^2+2\theta+1 & -2\theta^2+2\theta+1 \\ -2\theta^2+2\theta+1 & 2\theta^2+2\theta+1 \end{pmatrix}, \qquad f(\theta) = \frac{1}{2}\begin{pmatrix} 1-2\theta \\ 1-2\theta \end{pmatrix}.$$

Vérifier que la méthode est consistante pour tout $\theta \in \mathbb{R}$, à savoir, $x = B(\theta)x + f(\theta)$. Préciser ensuite pour quelles valeurs de θ (i) les valeurs propres de B sont réelles et la méthode est convergente ; (ii) la convergence de la méthode est plus rapide.

La solution du système est $x = (1,1)^T$ et on vérifie aisément pour tout $\theta \in \mathbb{R}$,

$$B(\theta)x + f(\theta) = \frac{1}{4}\begin{pmatrix} \cancel{2\theta^2}+2\theta+1\cancel{-2\theta^2}+2\theta+1 \\ \cancel{-2\theta^2}+2\theta+1+\cancel{2\theta^2}+2\theta+1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1-2\theta \\ 1-2\theta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

ce qui prouve la consistance de la méthode. Condition nécessaire et suffisante pour que la méthode soit convergente est que $\rho(B(\theta)) < 1$. Calculons donc les valeurs propres de B. Il s'agit de résoudre

$$\begin{aligned} 0 = \det(B(\theta) - \lambda I_n) &= \frac{1}{4}\begin{vmatrix} 2\theta^2+2\theta+1-4\lambda & -2\theta^2+2\theta+1 \\ -2\theta^2+2\theta+1 & 2\theta^2+2\theta+1-4\lambda \end{vmatrix} \\ &= \frac{1}{4}\left\{\left(2\theta^2+2\theta+1-4\lambda\right)^2 - \left(-2\theta^2+2\theta+1\right)^2\right\} \\ &= \frac{1}{4}\left\{16\lambda^2 - 8\lambda\left(2\theta^2+2\theta+1\right) + \left(2\theta^2+2\theta+1\right)^2 - \left(-2\theta^2+2\theta+1\right)^2\right\} \\ &= 4\lambda^2 - 2\lambda\left(2\theta^2+2\theta+1\right) + 2\theta^2(2\theta+1), \end{aligned}$$

qui est un polynôme de seconde degré en λ, dont le déterminant vaut

$$\Delta = 4\left(2\theta^2+2\theta+1\right)^2 - 32\theta^2(2\theta+1) = \left[4\theta^2 - 2(2\theta+1)\right]^2.$$

Les valeurs propres sont donc

$$\lambda_{1,2} = \frac{2(2\theta^2+2\theta+1) \pm \left|4\theta^2 - 2(2\theta+1)\right|}{8} = \begin{cases} \theta + \frac{1}{2} \\ \theta^2 \end{cases}.$$

La condition $\rho(B(\theta)) < 1$ équivaut donc à imposer

$$\left|\theta + \frac{1}{2}\right| < 1 \text{ et } |\theta^2| < 1 \iff -1 < \theta + \frac{1}{2} < 1 \text{ et } -1 \leq \theta \leq 1 \iff -1 < \theta < \frac{1}{2}.$$

La valeur optimale du paramètre θ s'obtient en correspondance du minimum de la fonction $\rho(B(\theta)) = \max(\theta^2, |\theta + 1/2|)$, à savoir pour $\theta < 0$ tel que $\theta^2 = \theta + \frac{1}{2}$. Il vient $\theta = (1-\sqrt{3})/2$.

Exercice 3.20 (Forte convexité d'une fonctionnelle quadratique)**.** *Montrer que la fonctionnelle J définie par* (3.15) *est* fortement convexe, *à savoir, il existe $\eta \in \mathbb{R}_*^+$ tel que, pour tout $x, y \in \mathbb{R}^n$ et tout $\vartheta \in [0,1]$,*

$$J(\vartheta x + (1-\vartheta)y) \leq \vartheta J(x) + (1-\vartheta)J(y) - \eta \frac{\vartheta(1-\vartheta)}{2}\|x-y\|^2.$$

Préciser la valeur de η.

On a

$$\begin{aligned} J(\vartheta x + (1-\vartheta)y) &= \frac{1}{2}(\vartheta x + (1-\vartheta)y)^T A(\vartheta x + (1-\vartheta)y) - b^T(\vartheta x + (1-\vartheta)y) \\ &= \frac{\vartheta^2}{2}x^T Ax\cdot + \frac{(1-\vartheta)^2}{2}y^T Ax + \vartheta(1-\vartheta)y^T Ax - \vartheta b^T x - (1-\vartheta)b^T y, \end{aligned}$$

où nous avons utilisé la symétrie de A pour conclure $x^T Ay = y^T Ax$. Pour le troisième terme au seconde membre on observera que

$$2y^T Ax = \frac{1}{2}x^T Ax + \frac{1}{2}y^T Ay - (y-x)^T A(y-x).$$

En remplaçant on obtient

$$J(\vartheta x + (1-\vartheta)y) = \vartheta J(x) + (1-\vartheta)J(y) - \frac{\vartheta(1-\vartheta)}{2}(x-y)^T A(x-y),$$

Or, en notant $\lambda_{\min}(A) > 0$ la plus petite valeur propre de A, nous avons pour tout $z \in \mathbb{R}^n$,

$$z^T Az \geq \lambda_{\min}(A)\|z\|^2 \iff -z^T Az \leq -\lambda_{\min}(A)\|z\|^2.$$

L'inégalité ci-dessus pour $z = y - x$ donne finalement

$$J(\vartheta x + (1-\vartheta)y) \leq \vartheta J(x) + (1-\vartheta)J(y) - \lambda_{\min}(A)\frac{(1-\vartheta)\vartheta}{2}\|x-y\|^2,$$

qui est l'inégalité cherchée avec $\eta = \lambda_{\min}(A)$.

Exercice 3.21 (Paramètre d'accéleration pour la méthode du gradient)**.** *Le but de cet exercice est de prouver la formule* (3.18)*. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ la fonction définie par $\varphi(\alpha) := \|x^{(k)} - \alpha r^{(k)}\|_A^2$. Le problème* (3.17) *équivaut à*

$$\min_{\alpha \in \mathbb{R}} \varphi(\alpha).$$

Prouver que φ est fortement convexe, à savoir, il existe $\eta \in \mathbb{R}_*^+$ tel que pour tout $\alpha,, \beta \in \mathbb{R}$ et tout $\vartheta \in [0,1]$,

$$\varphi(\vartheta\alpha + (1-\vartheta)\beta) \leq \vartheta\varphi(\alpha) + (1-\vartheta)\varphi(\beta) - \eta\frac{\vartheta(1-\vartheta)}{2}|\alpha-\beta|^2.$$

La forte convexité de φ assure que le problème (3.17) *admet unique solution $\alpha^{(k)}$ caractérisée par la propriété suivante :*

$$\varphi'(\alpha^{(k)}) = 0.$$

Montrer que cette propriété équivaut à (3.18)*.*

Par brévité on note $(\cdot,\cdot)$ le produit interne canonique de $\mathbb{R}^n$ et $(\cdot,\cdot)_A = (A\cdot,\cdot)$. On commence par prouver la forte convexité de φ. On a

$$\begin{aligned}
\varphi(\vartheta\alpha+(1-\vartheta)\beta) &= J(x^{(k)}+(\vartheta\alpha+(1-\vartheta)\beta)r^{(k)}) \\
&= J(\vartheta w_\alpha+(1-\vartheta)w_\beta) && (w_\xi := x^{(k)}+\xi r^{(k)}) \\
&= \vartheta J(w_\alpha)+(1-\vartheta)J(w_\beta)-\frac{\vartheta(1-\vartheta)}{2}\|w_\alpha-w_\beta\|_A^2 \\
&= \vartheta\varphi(\alpha)+(1-\vartheta)\varphi(\beta)-\|r^{(k)}\|_A^2\frac{\vartheta(1-\vartheta)}{2}|\alpha-\beta|^2, && \text{(Exercice 3.20)}
\end{aligned}$$

et φ est donc fortement convexe de paramètre $\eta = \|r^{(k)}\|_A^2$. Calculons maintenant la dérivée de φ. Nous avons

$$\begin{aligned}
\varphi'(\alpha) &= \left\{\frac{1}{2}(x^{(k)}+\alpha r^{(k)}, x^{(k)}+\alpha r^{(k)})_A-(b, x^{(k)}+\alpha r^{(k)})\right\}' \\
&= (r^{(k)}, x^{(k)}+\alpha r^{(k)})_A-(b, r^{(k)}) \\
&= (Ax^{(k)}, r^{(k)})+\alpha\|r^{(k)}\|_A^2-(b, r^{(k)}) \\
&= -(r^{(k)}, r^{(k)})+\alpha\|r^{(k)}\|_A^2. && (r^{(k)} = b-Ax^{(k)})
\end{aligned}$$

Par conséquent $\varphi'(\alpha)=0 \iff \alpha = \frac{\|r^{(k)}\|_2^2}{\|r^{(k)}\|_A^2}$, ce qui prouve (3.18).

Exercice 3.22 (Méthode du gradient conjugué)**.** *Utiliser la méthode CG avec donnée initiale $x^{(0)}=(0,0,0)^T$ pour résoudre le système linéaire $Ax=b$ avec*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}.$$

Par brévité on note $(\cdot,\cdot)$ le produit interne canonique de $\mathbb{R}^n$ et $(\cdot,\cdot)_A = (A\cdot,\cdot)$. On a $r^{(0)} = w^{(0)} = (-1,2,-1)^T$ et

$$\alpha^{(1)} = \frac{3}{10}, \qquad x^{(1)} = \frac{1}{10}\begin{pmatrix} -3 \\ 6 \\ -3 \end{pmatrix}, \qquad r^{(1)} = \frac{1}{5}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \beta^{(1)} = -\frac{1}{50}, \qquad w^{(1)} = \frac{1}{50}\begin{pmatrix} 9 \\ 12 \\ 9 \end{pmatrix},$$

$$\alpha^{(2)} = \frac{5}{3}, \qquad x^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad r^{(2)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

La méthode a donc convergé à la deuxième itération.

Exercice 3.23 (Paramètre d'accélération pour la méthode CG)**.** *Avec les techniques de l'Exercice 3.21 montrer que le choix* (3.22) *minimise la fonctionnelle* (3.15) *pour des arguments de la forme $x^{(k)}+\alpha w^{(k)}$ (avec $x^{(k)}$ et $w^{(k)}$ fixés).*

Exercice 3.24 (Mise en œuvre efficace de la méthode CG)**.** *On se place à l'itération k de la méthode CG et on suppose $\alpha^{(k+1)} \neq 0$). Prouver la formule récursive suivante, dont l'intérêt est d'éviter le produit matrice-vecteur $Aw^{(k)}$ pour le calcul de $r^{(k+1)}$ (voir la formule* (3.21)*) :*

$$r^{(k+1)} = -\alpha^{(k+1)} A r^{(k)} + \alpha^{(k+1)} \left(\frac{1}{\alpha^{(k+1)}} - \frac{\beta^{(k)}}{\alpha^{(k)}} \right) r^{(k)} + \alpha^{(k+1)} \frac{\beta^{(k)}}{\alpha^{(k)}} r^{(k-1)}.$$

On pourra remarquer que le résultat du produit matrice-vecteur $Ar^{(k)}$ peut être obtenu sans coût additionnel en mettant en œuvre de manière opportune l'étape (3.24d).

On a

$$\begin{aligned}
r^{(k+1)} &= r^{(k)} - \alpha^{(k+1)} A w^{(k)} && \text{eq. (3.21)} \\
&= r^{(k)} - \alpha^{(k+1)} A \left(r^{(k)} - \beta^{(k)} w^{(k-1)} \right) && \text{eq. (3.24e)} \\
&= r^{(k)} - \alpha^{(k+1)} A r^{(k)} + \alpha^{(k+1)} \beta^{(k)} A w^{(k-1)} \\
&= r^{(k)} - \alpha^{(k+1)} A r^{(k)} + \alpha^{(k+1)} \frac{\beta^{(k)}}{\alpha^{(k)}} \left(r^{(k-1)} - r^{(k)} \right) && \text{eq. (3.24c)} \\
&= -\alpha^{(k+1)} A r^{(k)} + \alpha^{(k+1)} \left(\frac{1}{\alpha^{(k+1)}} - \frac{\beta^{(k)}}{\alpha^{(k)}} \right) r^{(k)} + \alpha^{(k+1)} \frac{\beta^{(k)}}{\alpha^{(k)}} r^{(k-1)}.
\end{aligned}$$

Bibliographie

[1] G. Allaire. *Analyse numérique et optimisation.* Éditions de l'École Polytechnique, Palaiseau, 2009.

[2] P. G. Ciarlet, B. Miara, and J. M. Thomas. *Exercices d'analyse numérique matricielle et d'optimisation.* Masson, Paris, 2e edition, 1986.

[3] A. Ern. Calcul scientifique. Lecture notes, 2010.

[4] L. Formaggia, F. Saleri, and A. Veneziani. *Solving numerical PDEs : Problems, applications, exercises.* Springer, Milan, 2012.

[5] A. S. Householder. *The theory of matrices in numerical analysis.* Dover, 2006. Originally published in 1964 by the Blaisdell Publishing Company, New York.

[6] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics.* Texts in Applied Mathematics. Springer, New York, 2000.