ANALYSE NUMÉRIQUE MATRICIELLE

- Cours
- Exercices
- Corrigés détaillés

Luca Amodei
Jean-Pierre Dedieu

DUNOD

Série « Mathématiques pour le Master/SMAI »

La série « Mathématiques pour le Master/SMAI » propose une nouvelle génération de livres adaptés aux étudiants de Master niveau M1 et aux élèves ingénieurs. Leur adéquation au cursus LMD et aux outils de calcul modernes sont au service de la qualité scientifique.

La SMAI (Société de Mathématiques Appliquées et Industrielles) assure la direction éditoriale grâce à un comité renouvelé périodiquement et largement représentatif des différents thèmes des mathématiques appliquées et de leur évolution : analyse numérique, probabilités appliquées, statistique, optimisation, systèmes dynamiques et commande, traitement d'images et du signal, finance, recherche opérationnelle, etc. Son ambition est de constituer un ensemble d'ouvrages de référence.

Illustration de couverture : © *Digitalvision*

ISBN 978-2-10-052085-5

Avant-propos

Ce livre est issu des cours d'analyse numérique matricielle que nous avons enseignés pendant plusieurs années en licence de mathématique et en licence d'ingénierie mathématique. Il s'adresse aux étudiants de licence, à ceux de mastère ou préparant l'agrégation, aux élèves ingénieurs et aux chercheurs confirmés.

Cet ouvrage s'articule autour de quatre thèmes principaux qui sont :

- Les décompositions matricielles,
- La résolution des systèmes d'équations linéaires,
- Le calcul des valeurs propres,
- Le problème des erreurs en algèbre linéaire.

À vrai dire, cette division est informelle et plusieurs thèmes peuvent être abordés au sein d'un même chapitre.

En rédigeant ce manuscrit, nous avons adopté le point de vue d'un numéricien : pour chaque problème étudié nous décrivons :

- Les résultats théoriques qui y sont associés,
- Les problèmes de robustesse et de sensibilité,
- L'algorithmique et les problèmes de complexité,
- La stabilité des algorithmes.

Par *robustesse et sensibilité* nous entendons l'étude locale de la fonction *problème - solution* c'est-à-dire l'étude des variations de la solution d'un problème en fonction des variations des données. Elle conduit au concept de *conditionnement* d'un problème qui est un des concepts clé de l'analyse numérique. Le conditionnement est une mesure de la difficulté intrinsèque d'un problème.

Le problème de la *stabilité* est, quant à lui, lié à l'utilisation d'une arithmétique de précision finie au lieu de l'arithmétique des nombres réels. Les erreurs que l'on commet pour calculer la solution d'un problème dépendent alors de l'algorithme choisi pour mener ce calcul : des algorithmes différents peuvent donner des solutions (approchées) différentes. La recherche d'algorithmes stables est un souci majeur de l'analyse numérique.

Nous définissons la *complexité* d'un algorithme par le nombre d'opérations arithmétiques sur le corps des nombres réels (ou des nombres complexes) que requiert l'algorithme considéré. Un tel modèle continu (modèle Blum-Shub-Smale par exemple) est cohérent avec l'usage de l'arithmétique *virgule flottante*.

Nous nous écartons du calcul formel sur ces deux derniers points. En effet, l'usage de l'arithmétique (exacte) des nombres entiers rend inutile l'étude de la stabilité des algorithmes, quant aux problèmes de complexité sur des modèles discrets ils font aussi intervenir la taille des entiers considérés alors que, dans notre modèle, chaque opération sur les nombres réels compte pour une unité quelle que soit la taille des nombres ou la nature de l'opération.

Ce livre débute par un chapitre de rappels. Il sert à fixer les notations utilisées et il contient l'énoncé de théorèmes fondamentaux de l'algèbre linéaire. Les quatre chapitres suivants (2 à 5) sont consacrés aux normes matricielles, à l'arithmétique *virgule flottante*, au conditionnement et au problème des erreurs. On passe ensuite aux décompositions matricielles : LU, QR, Cholesky, SVD, à leur application à la résolution des systèmes et au problème des moindres carrés (chapitres 6 à 9). Les deux chapitres suivants étudient les méthodes itératives pour la résolution des systèmes. Elles sont fondées soit sur un schéma de type *approximations successives* (chapitre 10) soit sur des méthodes de projections sur des *espaces de Krylov* (chapitre 11). Les chapitres 12 à 15 sont consacrés aux problèmes de sensibilité des problèmes de valeurs propres, à leur calcul et à celui des sous-espaces invariants. Les chapitres 16 et 17 présentent des exemples de matrices et de problèmes d'algèbre linéaire : matrices classiques, systèmes obtenus via l'approximation d'équations aux dérivées partielles, problèmes industriels et assimilation des données.

Chaque chapitre se termine par un paragraphe d'exercices. Certains sont de simples applications numériques, d'autres de véritables prolongements du cours. Ces exercices sont corrigés en fin d'ouvrage.

Luca Amodei, Jean-Pierre Dedieu, Toulouse, juillet 2007.

Table des matières

AVANT-PROPOS v

CHAPITRE 1 • **RAPPELS D'ALGÈBRE LINÉAIRE** 1

1.1 Notations 1

1.2 Rang et noyau d'une matrice 2

1.3 Déterminant 3

1.4 Valeurs propres et vecteurs propres 3

1.5 Sous-espaces caractéristiques et théorème de Cayley-Hamilton 5

1.6 Décomposition de Jordan 6

1.7 Trace 6

1.8 Produit hermitien 7

1.9 Produit scalaire 8

1.10 Matrices unitaires 9

1.11 Matrices orthogonales 10

1.12 Matrices hermitiennes, symétriques, normales 10

1.13 Projections orthogonales 11

1.14 Matrices par blocs 12

1.15 Décomposition de Schur 15

1.16 Notes et références 17

EXERCICES 19

CHAPITRE 2 • **L'ARITHMÉTIQUE « VIRGULE FLOTTANTE »** 25
2.1 Les nombres flottants 25
2.2 Arrondis 26
2.3 L'arithmétique flottante 28
2.4 Exemple : le calcul du produit scalaire 28
2.5 Notes et références 31
EXERCICES 32
CHAPITRE 3 • **NORMES SUR LES ESPACES DE MATRICES** 33
3.1 Norme d'opérateur 33
3.2 Rayon spectral 35
3.3 La norme spectrale 36
3.4 La norme de Frobenius 37
3.5 Le théorème de perturbation de Neumann 39
3.6 Notes et références 40
EXERCICES 41
CHAPITRE 4 • **LA DÉCOMPOSITION EN VALEURS SINGULIÈRES** 47
4.1 Définition 47
4.2 Calcul des valeurs singulières 49
4.3 Notes et références 49
EXERCICES 50
CHAPITRE 5 • **LE PROBLÈME DES ERREURS** 53
5.1 Introduction 53
5.2 Concepts généraux 55
5.3 Le théorème des fonctions implicites 57
5.4 Le cas des systèmes linéaires : conditionnement d'une matrice. 61
5.5 Le cas des systèmes linéaires : erreurs inverses 64

5.6 Préconditionnement d'un système linéaire 64
5.7 Notes et références 65
EXERCICES 67

CHAPITRE 6 • **PIVOT DE GAUSS ET DÉCOMPOSITION LU** 69
6.1 Résolution des systèmes triangulaires 69
6.2 L'élimination de Gauss 70
6.3 Décomposition LU 71
6.4 Pivot partiel, pivot total 74
6.5 Complexité 80
6.6 Conditionnement de la décomposition LU 81
6.7 Notes et références 83
EXERCICES 84

CHAPITRE 7 • **MATRICES DÉFINIES POSITIVES ET DÉCOMPOSITION DE CHOLESKY** 87
7.1 Matrices définies positives 87
7.2 Quadriques et optimisation 91
7.3 Racine carrée d'une matrice, décomposition polaire 95
7.4 La décomposition de Cholesky 96
7.5 Complexité de la décomposition 97
7.6 Conditionnement de la décomposition de Cholesky 98
7.7 Notes et références 100
EXERCICES 101

CHAPITRE 8 • **LA DÉCOMPOSITION QR** 107
8.1 Matrices de Stiefel 107
8.2 Décomposition QR 108
8.3 L'orthonormalisation de Gram-Schmidt 109
8.4 Rotations de Givens 113

8.5 La méthode de Householder 116
8.6 Réduction à la forme de Hessenberg 121
8.7 Tridiagonalisation d'une matrice hermitienne 124
8.8 L'algorithme d'Arnoldi 125
8.9 L'algorithme de Lanczos 129
8.10 Conditionnement de la décomposition QR 132
8.11 Notes et références 137
EXERCICES 138

CHAPITRE 9 • **INVERSES GÉNÉRALISÉS ET MOINDRES CARRÉS** 141
9.1 Inverses généralisés 141
9.2 Moindres carrés 146
9.3 Problèmes surdéterminés 151
9.4 Etude d'un exemple : l'équation $AX = B$ 154
9.5 Notes et références 155
EXERCICES 156

CHAPITRE 10 • **MÉTHODES ITÉRATIVES** 161
10.1 Résultats généraux 162
10.2 Choix d'un test d'arrêt 164
10.3 Exemples de méthodes itératives 166
10.4 Convergence des méthodes itératives 168
10.5 Exemples 173
10.6 Méthodes itératives et préconditionnement 174
10.7 Notes et références 175
EXERCICES 176

CHAPITRE 11 • **MÉTHODES DE PROJECTION SUR DES SOUS-ESPACES DE KRYLOV** 181
11.1 Structure générale d'une méthode de projection 182

11.2 Espaces de Krylov et réduction de Hessenberg 182
11.3 La méthode GMRES 185
11.4 La méthode du gradient conjugué 188
11.5 Analyse d'erreur 195
11.6 Notes et références 199
EXERCICES 201

CHAPITRE 12 • **VALEURS PROPRES : SENSIBILITÉ** 203
12.1 Le théorème de Gershgorin 203
12.2 Le théorème d'Elsner 204
12.3 Sensibilité via le théorème des fonctions implicites 206
12.4 Notes et références 210
EXERCICES 211

CHAPITRE 13 • **SOUS-ESPACES INVARIANTS** 213
13.1 Sous-espaces invariants, simples, complémentaires 213
13.2 Forme réduite 216
13.3 Équation de Sylvester 217
13.4 Diagonalisation par blocs d'une matrice 220
EXERCICES 222

CHAPITRE 14 • **LE CALCUL DES VALEURS PROPRES** 225
14.1 La méthode de la puissance 225
14.2 Itération de sous-espaces 230
14.3 La méthode QR 236
14.4 Le cas des matrices réelles 241
14.5 L'utilisation de la forme Hessenberg 242
14.6 La stratégie du décalage 243
14.7 Remarques finales 246
14.8 Notes et références 247

EXERCICES 249

CHAPITRE 15 • **MÉTHODES DE PROJECTION POUR LE PROBLÈME DES VALEURS PROPRES** 251

15.1 Principe d'une méthode de projection pour le problème des valeurs propres 251

15.2 Méthode de projection sur des sous-espaces de Krylov 253

15.3 Notes et références 256

EXERCICES 257

CHAPITRE 16 • **EXEMPLES DE SYSTÈMES LINÉAIRES** 259

16.1 Le problème de Poisson discrétisé par différences finies 259

16.2 Le problème de Poisson sur un carré discrétisé par différences finies 261

16.3 Le problème de Poisson discrétisé par éléments finis 262

16.4 La matrice de Vandermonde 264

16.5 La matrice de Fourier 265

16.6 Système linéaire associé à la spline cubique d'interpolation 267

16.7 Notes et références 268

EXERCICES 270

CHAPITRE 17 • **GAUSS-NEWTON ET L'ASSIMILATION DES DONNÉES** 271

17.1 La méthode de Newton 271

17.2 Gauss-Newton et moindres carrés 272

17.3 Le problème de l'assimilation des données 273

CORRIGÉS DES EXERCICES 281

BIBLIOGRAPHIE 309

INDEX 313

Chapitre 1

Rappels d'algèbre linéaire

1.1 NOTATIONS

Ce paragraphe a pour but de fixer les notations qui sont utilisées tout au long de ce livre.

- L'espace des matrices complexes (resp. réelles) à m lignes et n colonnes est noté $\mathbb{C}^{m\times n}$ (resp. $\mathbb{R}^{m\times n}$). Pour une matrice $A = (a_{ij}) \in \mathbb{C}^{m\times n}$, i est l'indice de la ligne et j celui de la colonne.

- $0_{mn} \in \mathbb{C}^{m\times n}$ (aussi notée 0) est la matrice nulle et $I_n \in \mathbb{C}^{n\times n}$ est la matrice identité.

- Les vecteurs $x \in \mathbb{C}^n$ (resp. $x \in \mathbb{R}^n$) sont identifiés à des matrices $n \times 1$ donc à des vecteurs-colonne [1]. Avec cette convention, $A \in \mathbb{C}^{m\times n}$ s'identifie à l'application linéaire $A : \mathbb{C}^n \to \mathbb{C}^m$ qui à $x \in \mathbb{C}^n$ associe le produit matrice-vecteur $Ax \in \mathbb{C}^m$.

- Lorsqu'une matrice A a pour colonnes a_i, $1 \leqslant i \leqslant n$, on la note $A = (a_1 \dots a_n)$.

- Soient $A \in \mathbb{C}^{m\times n}$ et les indices $1 \leqslant i_1 < i_2 < \dots < i_p \leqslant m$ et $1 \leqslant j_1 < j_2 < \dots < j_q \leqslant n$. Alors, la matrice de taille $p \times q$ constituée par les éléments aux intersections des lignes $1 \leqslant i_1 < i_2 < \dots < i_p \leqslant m$ et des colonnes

1. Suivant un usage déjà ancien les mots composés *vecteur-colonne*, *vecteur-ligne*, *matrice-colonne*, *matrice-ligne* sont unis par un trait d'union ; ils font leur pluriel en *vecteurs-colonne* ... sur le modèle de *timbres-poste*.

$1 \leqslant j_1 < j_2 < \ldots < j_q \leqslant n$ de A est appelée sous-matrice de A. Autrement dit, une sous-matrice de A est obtenue en supprimant dans A un certain nombre de lignes et de colonnes.

- Pour toute matrice $A \in \mathbb{C}^{m\times n}$ et les entiers p, q, r, s tels que $1 \leqslant p \leqslant q \leqslant m$, $1 \leqslant r \leqslant s \leqslant n$ on note $A(p : q, r : s)$ la sous-matrice de A de terme général a_{ij}, $p \leqslant i \leqslant q$, $r \leqslant j \leqslant s$. De telles sous-matrices sont parfois qualifiées de « contigües ».

- Pour tout vecteur $a \in \mathbb{C}^m$, $a(p : q)$ est le sous-vecteur de a de coordonnées a_i, $p \leqslant i \leqslant q$.

- Une matrice (pas nécessairement carrée) est triangulaire supérieure si $a_{ij} = 0$ pour $i > j$, triangulaire inférieure si $a_{ij} = 0$ pour $i < j$ et diagonale si $a_{ij} = 0$ pour $i \neq j$.

- Une matrice diagonale D est notée $D = \mathrm{diag}(d_i)$ où les d_i sont les entrées diagonales.

- Soit $A = (a_{ij}) \in \mathbb{C}^{m\times n}$. On note $A^T \in \mathbb{C}^{n\times m}$ la transposée de A et $A^* = \overline{A}^T \in \mathbb{C}^{n\times m}$ son adjointe : $A^T = (a_{ji})$ et $A^* = (\overline{a}_{ji})$ (conjuguée et transposée).

- $\mathbb{GL}_n(\mathbb{C})$ (resp. $\mathbb{GL}_n(\mathbb{R})$) ou plus simplement $\mathbb{GL}_n$ est l'ensemble des matrices $A \in \mathbb{C}^{n\times n}$ (resp. $A \in \mathbb{R}^{n\times n}$) qui sont inversibles. C'est un groupe pour la multiplication des matrices appelé *groupe linéaire*. Les notations A^{-T} et A^{-*} (inverse de la transposée et inverse de l'adjointe) ne sont pas ambigües parce que $(A^T)^{-1} = (A^{-1})^T$, de même pour A^*.

1.2 RANG ET NOYAU D'UNE MATRICE

Étant donné une matrice $A \in \mathbb{C}^{m\times n}$, l'image par A d'un vecteur $x \in \mathbb{C}^n$ est le vecteur $Ax = \sum_{i=1}^n x_i a_i \in \mathbb{C}^m$ où les a_i sont les colonnes de A. L'*image* de A est définie par

$$\mathrm{Im}\, A = \{Ax \ : \ x \in \mathbb{C}^n\} = \left\{ \sum_{i=1}^n x_i a_i \ : \ x \in \mathbb{C}^n \right\}.$$

C'est un sous-espace vectoriel de $\mathbb{C}^m$ engendré par les vecteurs-colonne de A. Sa dimension est le *rang* de A. Le rang de A est donc le nombre maximum de vecteurs-colonne indépendants de A.

Une caractérisation utile du rang est la suivante : rang $A = r$ si et seulement s'il existe dans A une sous-matrice carrée $r \times r$ de déterminant non nul et si toute sous-matrice carrée $s \times s$ avec $s > r$ a un déterminant égal à 0.

Cette caractérisation montre que

$$\text{rang } A = \text{rang } A^T = \text{rang } A^*.$$

Le *noyau* de A est le sous-espace vectoriel

$$\text{Ker } A = \{x \in \mathbb{C}^n \ : \ Ax = 0\}.$$

Le rang et la dimension du noyau de A sont reliés par la formule célèbre :

$$\text{rang } A + \dim \text{Ker } A = n.$$

1.3 DÉTERMINANT

- Le *déterminant* d'une matrice $A \in \mathbb{C}^{n\times n}$ est une forme multilinéaire alternée des colonnes de A, on le note $\det A$.
- $\det I_n = 1$,
- Le déterminant d'un produit de matrices est le produit de ses déterminants : $\det(AB) = \det A \det B$,
- $\det A = \det A^T$,
- A est inversible si et seulement si $\det A \neq 0$, dans ce cas $\det(A^{-1}) = 1/\det A$,
- Si deux matrices sont semblables elles ont même déterminant.

1.4 VALEURS PROPRES ET VECTEURS PROPRES

- On appelle *valeur propre* d'une matrice $A \in \mathbb{C}^{n\times n}$ toute racine du *polynôme caractéristique*

$$P_A(\lambda) = \det(A - \lambda I_n) = 0.$$

La *multiplicité algébrique* d'une valeur propre est sa multiplicité en tant que racine de l'équation caractéristique. Lorsque l'on parle de multiplicité d'une valeur propre c'est de la multiplicité algébrique dont il s'agit.

- L'ensemble des valeurs propres de A est appelé le *spectre* de A et se note spec A.

- On appelle *vecteur propre* de A associé à la valeur propre λ tout vecteur non nul $x \in \mathbb{C}^n$ vérifiant $Ax = \lambda x$. La réunion du vecteur 0 et des vecteurs propres associés à la valeur propre λ est un sous-espace vectoriel E_λ de $\mathbb{C}^n$ appelé *sous-espace propre* associé à λ. Sa dimension $\dim E_\lambda$ est la *multiplicité géométrique* de λ. Elle est toujours inférieure ou égale à la multiplicité algébrique.

- Deux matrices $A, B \in \mathbb{C}^{n\times n}$ sont *semblables* lorsqu'il existe une matrice $P \in \mathbb{GL}_n(\mathbb{C})$ telle que $A = PBP^{-1}$.

- Une matrice $A \in \mathbb{C}^{n\times n}$ est *diagonalisable* lorsqu'elle est semblable à une matrice diagonale, c'est-à-dire s'il existe des matrices $D \in \mathbb{C}^{n\times n}$ diagonale et $P \in \mathbb{GL}_n(\mathbb{C})$ telles que
$$A = PDP^{-1}.$$
Dans ce cas, écrivons $D = \mathrm{diag}(\lambda_i)$ où les λ_i sont les valeurs propres de A ; on peut prendre pour P une matrice dont les colonnes $p_1, \ldots, p_n \in \mathbb{C}^n$ sont indépendantes et où p_i est un vecteur propre associé à λ_i.

- Une matrice $A \in \mathbb{C}^{n\times n}$ est diagonalisable si et seulement si $\mathbb{C}^n$ possède une base de vecteurs propres de A ou encore lorsque, pour toute valeur propre λ de A, les multiplicités algébrique et géométrique de λ sont les mêmes.

- Lorsque $A \in \mathbb{R}^{n\times n}$, que les valeurs propres de A sont réelles et que A est diagonalisable, il existe une base de $\mathbb{R}^n$ faite de vecteurs propres et l'on peut prendre $P \in \mathbb{GL}_n(\mathbb{R})$.

Proposition 1.1 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$, on a les propriétés suivantes :*

1. *Les valeurs propres de $A - \mu I_n$ sont les scalaires $\lambda - \mu$, $\lambda \in \mathrm{spec}\, A$,*
2. *Les valeurs propres de A^k, $k \geqslant 0$, sont les λ^k, $\lambda \in \mathrm{spec}\, A$, et, plus généralement, pour tout polynôme $p(x)$, les scalaires $p(\lambda)$, $\lambda \in \mathrm{spec}\, A$, sont les valeurs propres de $p(A)$ (définition paragraphe 1.5),*
3. *Lorsque A est inversible, les valeurs propres de A^{-1} sont les scalaires λ^{-1}, $\lambda \in \mathrm{spec}\, A$,*
4. *Le déterminant de A est le produit des valeurs propres de A, chacune comptée autant de fois que sa multiplicité algébrique :*
$$\det A = \Pi_{\lambda_i \in \mathrm{spec}\, A}\, \lambda_i^{m_i},$$
où m_i est la multiplicité algébrique de λ_i.

- Une matrice $A \in \mathbb{C}^{n\times n}$ est *triangularisable* lorsqu'elle est semblable à une matrice triangulaire supérieure, c'est-à-dire s'il existe une matrice $P \in \mathbb{C}^{n\times n}$ inversible et une matrice $T \in \mathbb{C}^{n\times n}$ triangulaire supérieure telles que $A = PTP^{-1}$.

- Nous verrons ci-dessous (décomposition de Jordan, décomposition de Schur) que toute matrice $A \in \mathbb{C}^{n\times n}$ est triangularisable.

- Lorsque $A \in \mathbb{C}^{n\times n}$ s'écrit $A = PTP^{-1}$ avec T triangulaire supérieure, la diagonale de T contient les valeurs propres de A.

1.5 SOUS-ESPACES CARACTÉRISTIQUES ET THÉORÈME DE CAYLEY-HAMILTON

À tout polynôme complexe $P(z) = a_0 + a_1 z + \ldots + a_d z^d$ et à toute matrice $A \in \mathbb{C}^n$ on associe le *polynôme matriciel*

$$P(A) = a_0 I_n + a_1 A + \ldots + a_d A^d.$$

On dit qu'un polynôme $P(z)$ est un *polynôme annulateur* de $A \in \mathbb{C}^{n\times n}$ lorsque $P(A) = 0$.

Théorème 1.2 *(Cayley-Hamilton) Le polynôme caractéristique de A est un polynôme annulateur de A : $P_A(A) = 0$.*

Définition 1.3 *(Sous-espaces caractéristiques) Soit $A \in \mathbb{C}^{n\times n}$. Écrivons son polynôme caractéristique*

$$P_A(\lambda) = \det(A - \lambda I_n) = \prod_{i=1}^{q} (\lambda_i - \lambda)^{m_i}$$

où les valeurs propres λ_i, $1 \leqslant i \leqslant q$, sont deux à deux distinctes, de multiplicité algébrique m_i avec $m_1 + \ldots + m_q = n$. Les sous-espaces caractéristiques de A sont les ensembles

$$E_i = \operatorname{Ker}(\lambda_i I_n - A)^{m_i}.$$

Théorème 1.4 *(Décomposition en sous-espaces caractéristiques) Les sous-espaces caractéristiques de A vérifient les propriétés suivantes :*

1. *E_i est un sous-espace de $\mathbb{C}^n$ de dimension m_i,*
2. *$AE_i \subset E_i$,*
3. *$\mathbb{C}^n = E_1 \oplus \ldots \oplus E_q$.*

1.6 DÉCOMPOSITION DE JORDAN

Théorème 1.5 *(Décomposition de Jordan) Pour toute matrice $A \in \mathbb{C}^{n\times n}$, il existe une matrice $P \in \mathbb{GL}_n(\mathbb{C})$ et une matrice $J \in \mathbb{C}^{n\times n}$ telles que $A = PJP^{-1}$ et où J a la structure diagonale par blocs suivante :*

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{pmatrix}.$$

Chaque bloc diagonal $J_k \in \mathbb{C}^{n_k\times n_k}$ $(n_1 + \ldots + n_p = n)$ est soit du type $J_k = \lambda_k I_{n_k}$, c'est-à-dire un multiple de l'identité, soit du type $J_k = \lambda_k I_{n_k} + N_{n_k}$, où $N_{n_k} \in \mathbb{C}^{n_k\times n_k}$ est la matrice nilpotente

$$N_{n_k} = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{pmatrix}.$$

Les scalaires $\lambda_1, \ldots, \lambda_p$ (qui ne sont pas nécessairement distincts) sont les valeurs propres de A :

$$\text{spec } A = \{\lambda_1 \ldots, \lambda_p\}.$$

1.7 TRACE

- La *trace* d'une matrice carrée $A \in \mathbb{C}^{n\times n}$ est la somme de ses entrées diagonales :
$$\text{trace } A = \sum_{i=1}^{n} a_{ii}.$$
- Pour deux matrices $M \in \mathbb{C}^{m\times n}$ et $N \in \mathbb{C}^{n\times m}$ on a
$$\text{trace } (MN) = \text{trace } (NM)$$
de sorte que, pour toute matrice $A \in \mathbb{C}^{n\times n}$ et $P \in \mathbb{GL}_n$,
$$\text{trace } (P^{-1}AP) = \text{trace } A.$$
- La trace de A est égale à la somme des valeurs propres de A comptées avec leur multiplicité (cela se prouve en écrivant $A = PTP^{-1}$ avec T triangulaire).

1.8 PRODUIT HERMITIEN

- Un *produit hermitien* sur un espace vectoriel complexe E est une application $\langle .,. \rangle : E \times E \to \mathbb{C}$ qui vérifie les propriétés suivantes :
  1. Pour tout $y \in E$, l'application $x \in E \to \langle x, y \rangle \in \mathbb{C}$ est linéaire,
  2. Pour tout $x, y \in E$, $\langle x, y \rangle = \overline{\langle y, x \rangle}$,
  3. Pour tout $x \in E$, $\langle x, x \rangle \geqslant 0$,
  4. Pour tout $x \in E$, $\langle x, x \rangle = 0$ si et seulement si $x = 0$.

  Un espace vectoriel complexe E muni d'un produit hermitien est appelé *espace préhilbertien complexe* ; si de plus E est de dimension finie on dit que c'est un *espace hermitien*.

- Un exemple fondamental d'espace hermitien est donné par

$$E = \mathbb{C}^n, \ \langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y}_i.$$

  Avec les notation matricielles, x et y sont des vecteurs-colonne et

$$\langle x, y \rangle = y^* x$$

  en identifiant la matrice $y^* x \in \mathbb{C}^{1 \times 1}$ au scalaire correspondant. Attention, xy^* est une matrice $n \times n$!

- Lorsque $\langle x, y \rangle$ est un produit hermitien sur E, $\|x\| = \langle x, x \rangle^{\frac{1}{2}}$ est une norme sur E et

$$|\langle x, y \rangle| \leqslant \|x\| \|y\|$$

  avec égalité si et seulement si x et y sont colinéaires (*inégalité de Cauchy-Schwarz*).

- Lorsque $E = \mathbb{C}^n$ est muni du produit hermitien canonique, la norme associée est notée

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

- Soient E et F deux espaces hermitiens. L'*adjoint* d'un opérateur linéaire $L : E \to F$ est l'unique opérateur linéaire $L^* : F \to E$ tel que

$$\langle Lx, y \rangle_F = \langle x, L^* y \rangle_E$$

  pour tout $x \in E$ et $y \in F$.

- Lorsque $E = \mathbb{C}^n$ et $F = \mathbb{C}^m$ sont munis de leur structure hermitienne canonique, l'*adjoint* de l'opérateur défini par une matrice $A \in \mathbb{C}^{m\times n}$ est l'opérateur défini par la matrice adjointe $A^* = \overline{A^T}$.

- Lorsque L est un endomorphisme de E, on dit que L est *hermitien* lorsque $L^* = L$ c'est-à-dire si
$$\langle Lx, y\rangle_E = \langle x, Ly\rangle_E$$
pour tout $x, y \in E$.

- Lorsque $E = \mathbb{C}^n$, l'opérateur défini par une matrice $A \in \mathbb{C}^{n\times n}$ est hermitien lorsque la matrice A est *hermitienne* c'est-à-dire lorsque $A^* = A$.

1.9 PRODUIT SCALAIRE

- Un *produit scalaire* sur un espace vectoriel réel E est une application $\langle .,. \rangle : E \times E \to \mathbb{R}$ qui vérifie les propriétés suivantes :
  1. Pour tout $y \in E$, l'application $x \in E \to \langle x, y\rangle \in \mathbb{R}$ est linéaire,
  2. Pour tout $x, y \in E$, $\langle x, y\rangle = \langle y, x\rangle$,
  3. Pour tout $x \in E$, $\langle x, x\rangle \geqslant 0$,
  4. Pour tout $x \in E$, $\langle x, x\rangle = 0$ si et seulement si $x = 0$.

  Un espace vectoriel réel E muni d'un produit scalaire est appelé *espace préhilbertien*; si de plus E est de dimension finie on dit que c'est un *espace euclidien.*

- Le produit scalaire canonique sur $E = \mathbb{R}^n$ est donné par
$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i.$$
Avec les notation matricielles, x et y sont des vecteurs-colonne et
$$\langle x, y\rangle = y^T x$$
en identifiant la matrice $y^T x \in \mathbb{R}^{1\times 1}$ au scalaire correspondant.

- Lorsque $\langle x, y\rangle$ est un produit scalaire sur E, $\|x\| = \langle x, x\rangle^{\frac{1}{2}}$ est une norme sur E et
$$|\langle x, y\rangle| \leqslant \|x\|\|y\|$$
avec égalité si et seulement si x et y sont colinéaires (*inégalité de Cauchy-Schwarz*).

- Lorsque $E = \mathbb{R}^n$ est muni du produit scalaire canonique, la norme associée est notée, comme dans le cas complexe,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

- Soient E et F deux espaces euclidiens. L'*adjoint* d'un opérateur linéaire $L : E \to F$ est l'unique opérateur linéaire $L^* : F \to E$ tel que

$$\langle Lx, y\rangle_F = \langle x, L^* y\rangle_E$$

pour tout $x \in E$ et $y \in F$.

- Lorsque $E = \mathbb{R}^n$ et $F = \mathbb{R}^m$ sont munis de leur structure euclidienne canonique, l'adjoint de l'opérateur défini par une matrice $A \in \mathbb{R}^{m\times n}$ est l'opérateur défini par la matrice transposée A^T.

- Lorsque L est un endomorphisme de E, on dit que L est *symétrique* lorsque $L^* = L$ c'est-à-dire si

$$\langle Lx, y\rangle_E = \langle x, Ly\rangle_E$$

pour tout $x, y \in E$.

- Lorsque $E = \mathbb{R}^n$, l'opérateur défini par une matrice $A \in \mathbb{R}^{n\times n}$ est symétrique lorsque la matrice A est *symétrique* c'est-à-dire si $A^T = A$.

1.10 MATRICES UNITAIRES

- Une matrice $U \in \mathbb{C}^{n\times n}$ est *unitaire* lorsqu'elle conserve le produit hermitien de $\mathbb{C}^n$:

$$\langle Ux, Uy\rangle = \langle x, y\rangle$$

pour tout $x, y \in \mathbb{C}^n$.

- Une matrice $U \in \mathbb{C}^{n\times n}$ est unitaire si et seulement si $U^*U = UU^* = I_n$. L'ensemble de ces matrices est un sous-groupe (pour la multiplication) du groupe linéaire $\mathbb{GL}_n(\mathbb{C})$: c'est le *groupe unitaire*, il est noté $\mathbb{U}_n$.

- Une matrice U est unitaire si et seulement si les vecteurs-colonne de U constituent une base orthonormée de $\mathbb{C}^n$ pour le produit hermitien canonique.

- Les valeurs propres d'une matrice unitaire sont des nombres complexes de module 1. Le déterminant d'une telle matrice est aussi un nombre complexe de module 1. Les matrices unitaires dont le déterminant est égal à 1 constituent un sous-groupe de $\mathbb{U}_n$ appelé *groupe spécial unitaire*. Il est noté $\mathbb{SU}_n$.

1.11 MATRICES ORTHOGONALES

- Une matrice $U \in \mathbb{R}^{n\times n}$ est *orthogonale* lorsqu'elle conserve le produit scalaire de $\mathbb{R}^n$:

$$\langle Ux.Uy\rangle = \langle x, y\rangle$$

pour tout $x, y \in \mathbb{R}^n$.

- $U \in \mathbb{R}^{n\times n}$ est orthogonale lorsque $U^TU = UU^T = I_n$. Une matrice est orthogonale si elle est à la fois réelle et unitaire. Ces matrices constituent un sous-groupe (pour la multiplication) du groupe linéaire $\mathbb{GL}_n(\mathbb{R})$ appelé *groupe orthogonal* et noté $\mathbb{O}_n$.

- Une matrice $U \in \mathbb{R}^{n\times n}$ est orthogonale si et seulement si ses vecteurs-colonne constituent une base orthonormée de $\mathbb{R}^n$ pour le produit scalaire canonique.

- Les valeurs propres d'une matrice orthogonale sont des nombres complexes de module 1. Le déterminant d'une telle matrice est égal à 1 ou -1. Les matrices orthogonales dont le déterminant est égal à 1 constituent un sous-groupe de $\mathbb{O}_n$ appelé *groupe spécial orthogonal* ou *groupe des rotations*. Il est noté $\mathbb{SO}_n$.

1.12 MATRICES HERMITIENNES, SYMÉTRIQUES, NORMALES

- Les matrices hermitiennes ou symétriques réelles possèdent la propriété fondamentale suivante :

Théorème 1.6 *(Théorème spectral) Si $A \in \mathbb{C}^{n\times n}$ est hermitienne (resp. symétrique réelle) alors*

1. *Les valeurs propres de A sont réelles,*
2. *$\mathbb{C}^n$ (resp. $\mathbb{R}^n$) possède une base orthonormée constituée de vecteurs propres de A.*

   Ces deux propriétés sont équivalentes à la suivante :
3. *Il existe une matrice diagonale réelle $D \in \mathbb{R}^{n\times n}$ et une matrice unitaire (resp. une matrice orthogonale) U telles que $A = UDU^*$.*

- Une matrice $A \in \mathbb{C}^{n\times n}$ est *normale* lorsque $AA^* = A^*A$. Pour une matrice $A \in \mathbb{R}^{n\times n}$ cette condition devient $AA^T = A^TA$. Ces matrices possèdent la caractérisation suivante :

Théorème 1.7 *$A \in \mathbb{C}^{n\times n}$ est normale si et seulement si $\mathbb{C}^n$ possède une base orthonormée constituée de vecteurs propres de A, c'est-à-dire s'il existe une matrice diagonale $D \in \mathbb{C}^{n\times n}$ et une matrice unitaire U telles que $A = UDU^*$.*

Les matrices hermitiennes, les matrices unitaires, les matrices réelles et *antisymétriques* ($A^T = -A$) sont des matrices normales.

1.13 PROJECTIONS ORTHOGONALES

- Soit E un espace vectoriel réel ou complexe. On appelle *projecteur* un endomorphisme p de E qui vérifie $p \circ p = p$. L'espace E se décompose alors en somme directe

$$E = F \oplus G \text{ avec } F = \text{Im } p,\ G = \text{Ker } p.$$

De plus

$$\begin{aligned} p(y) &= y \quad \text{pour tout } y \in F, \\ p(y) &= 0 \quad \text{pour tout } y \in G. \end{aligned}$$

On dit aussi que p est la *projection sur F parallèlement à G*.

- Lorsque E est un espace hermitien ou euclidien et que p est un projecteur hermitien, c'est-à-dire lorsque

$$p \circ p = p \text{ et } p^* = p,$$

on a Ker $p = (\text{Im } p)^\perp$ de sorte que

$$E = F \oplus G \text{ avec } F = \text{Im } p \text{ et } G = F^\perp.$$

p est appelé la *projection orthogonale* de E sur F et noté $p = \Pi_F$

- Pour tout $x \in E$ la projection orthogonale de x sur F est l'unique vecteur $y \in F$ qui rende minimum la distance de x à F :

$$y \in F \text{ et } \|x - y\| = \min_{z \in F} \|x - z\|.$$

- Lorsque $E = \mathbb{C}^n$ et que F est le sous-espace vectoriel engendré par r vecteurs indépendants y_i, $1 \leqslant i \leqslant r$, la matrice de Π_F est égale à $Y(Y^*Y)^{-1}Y^*$ avec $Y = (y_1 \ldots y_p)$ (l'inversibilité de la matrice Y^*Y est prouvée au théorème 7.2, voir aussi la remarque 7.1). Si les vecteurs y_i sont orthonormés, alors $\Pi_F = YY^*$.

1.14 MATRICES PAR BLOCS

1.14.1 Définition

Une *matrice par blocs* $M = (M_{ij})$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$, est une matrice dont les entrées M_{ij} sont des matrices au lieu d'être des scalaires. On doit toutefois respecter les deux règles suivantes :

- Toutes les matrices d'une même ligne (M_{ij} avec $1 \leqslant j \leqslant n$) ont le même nombre de lignes,
- Toutes les matrices d'une même colonne (M_{ij} avec $1 \leqslant i \leqslant m$) ont le même nombre de colonnes.

Ainsi, il existe des nombres entiers m_i et n_j tels que $M_{ij} \in \mathbb{C}^{m_i \times n_j}$.

On étend aux matrices par blocs les concepts de matrice diagonale, de matrice triangulaire supérieure ou de matrice triangulaire inférieure :

- $M = (M_{ij})$ est *triangulaire supérieure par blocs* si $M_{ij} = 0$ pour $i > j$,
- M est *triangulaire inférieure par blocs* si $M_{ij} = 0$ pour $i < j$,
- M est *diagonale par blocs* si $M_{ij} = 0$ pour $i \neq j$.

1.14.2 Produit par blocs

Donnons-nous deux matrices par blocs :

- $M = (M_{ij})$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$ où $M_{ij} \in \mathbb{C}^{m_i \times n_j}$,
- $N = (N_{kl})$, $1 \leqslant k \leqslant n$, $1 \leqslant l \leqslant p$ où $N_{kl} \in \mathbb{C}^{n_k \times p_l}$.

Tous les produits $M_{ik}N_{kl}$ sont bien définis et $M_{ik}N_{kl} \in \mathbb{C}^{m_i \times p_l}$. Le *produit par blocs* des matrices M et N est défini par :

$$MN = ((MN)_{il}),\ 1 \leqslant i \leqslant m,\ 1 \leqslant l \leqslant p \text{ avec } (MN)_{il} = \sum_{k=1}^{n} M_{ik}N_{kl}.$$

La propriété essentielle de ce produit est qu'il coïncide avec le produit usuel ; c'est la raison pour laquelle on les note tous deux de la même manière. Toutefois, il faut prendre garde à la non-commutativité du produit $M_{ik}N_{kl}$ et respecter l'ordre de ces facteurs.

Etudions un exemple : prenons

- $M = \begin{pmatrix} \alpha & \beta & 0 \\ \gamma & \delta & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}$ où $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathbb{R}^{2\times 2}$ et où l'on note par 0 le scalaire $0 \in \mathbb{R}$ dans M, la matrice $\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{2\times 1}$ et la matrice $\begin{pmatrix} 0 & 0 \end{pmatrix} \in \mathbb{R}^{1\times 2}$ dans la description de M par blocs,
- $N = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 3 \end{pmatrix} = \begin{pmatrix} I_2 \\ a \end{pmatrix}$ où $I_2 \in \mathbb{R}^{2\times 2}$ est la matrice identité et $B = \begin{pmatrix} 3 & 3 \end{pmatrix} \in \mathbb{R}^{1\times 2}$.

Le calcul du produit par blocs s'écrit :

$$MN = \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I_2 \\ B \end{pmatrix} = \begin{pmatrix} AI_2 + 0B \\ 0I_2 + 1B \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \\ 3 & 3 \end{pmatrix}.$$

On vérifie facilement qu'il s'agit bien du produit usuel :

$$MN = \begin{pmatrix} \alpha & \beta & 0 \\ \gamma & \delta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 3 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \\ 3 & 3 \end{pmatrix}.$$

1.14.3 Matrices triangulaires par blocs

Théorème 1.8 *Étant donné une matrice triangulaire par blocs*

$$M = \begin{pmatrix} M_{11} & 0 & \dots & 0 \\ M_{21} & M_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \dots & M_{nn} \end{pmatrix}$$

on a :

1. *Le déterminant de M est le produit des déterminants des matrices M_{ii} :*

$$\det M = \prod_{1\leqslant i\leqslant n} \det M_{ii},$$

2. *Le polynôme caractéristique de M est le produit des polynômes caractéristiques des matrices M_{ii} :*

$$P_M(\lambda) = \prod_{1\leqslant i\leqslant n} P_{M_{ii}}(\lambda),$$

3. *Le spectre de M est la réunion des spectres des matrices M_{ii} :*

$$\text{spec } M = \cup_{1 \leqslant i \leqslant n} \text{spec } M_{ii}$$

et la multiplicité algébrique d'une valeur propre de M est la somme de ses multiplicités en tant que valeur propre de matrices M_{ii}.

Démonstration. La propriété sur les spectres découle de celle sur les polynômes caractéristiques qui est elle-même une conséquence de la formule donnant le déterminant. Pour prouver cette dernière il suffit de l'établir pour $n = 2$ puis de raisonner par récurrence. Traitons donc le cas

$$M = \begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{pmatrix} \in \mathbb{C}^{(n_1+n_2)\times(n_1+n_2)}.$$

Le théorème de Jordan (théorème 1.5) appliqué à la transposée de la matrice M_{11} montre que l'on peut écrire $M_{11} = VTV^{-1}$ où T est triangulaire inférieure. On obtient

$$M = \begin{pmatrix} V & 0 \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} T & 0 \\ M_{21}V & M_{22} \end{pmatrix} \begin{pmatrix} V^{-1} & 0 \\ 0 & I_{n_2} \end{pmatrix}.$$

Comme les matrices $\begin{pmatrix} V & 0 \\ 0 & I_{n_2} \end{pmatrix}$ et $\begin{pmatrix} V^{-1} & 0 \\ 0 & I_{n_2} \end{pmatrix}$ sont inverses l'une de l'autre on a

$$\det M = \det \begin{pmatrix} T & 0 \\ M_{21}V & M_{22} \end{pmatrix}.$$

En développant ce déterminant par rapport à la première ligne et après n_1 telles opérations cela donne

$$\det M = t_{11} \ldots t_{n_1 n_1} \det M_{22} = \det T \det M_{22} = \det M_{11} \det M_{22}.$$

Il est bien évident qu'un énoncé similaire au théorème 1.8 à lieu pour des matrices triangulaires supérieures par blocs.

1.14.4 Le complément de Schur

Proposition 1.9 *Considérons la matrice par blocs*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

où les dimensions des blocs A, B, C et D sont $n \times n$, $n \times m$, $m \times n$ et $m \times m$. Supposons que A soit inversible. On a alors :

$$M = \begin{pmatrix} I_n & 0 \\ CA^{-1} & I_m \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I_n & A^{-1}B \\ 0 & I_m \end{pmatrix}.$$

De plus

$$\det(M) = \det(A)\det(D - CA^{-1}B).$$

Démonstration. La décomposition est immédiate et le calcul du déterminant est une conséquence du théorème 1.8.

Définition 1.10 *La matrice $D - CA^{-1}B$ s'appelle le complément de Schur de la matrice A dans M.*

1.14.5 La formule de Sherman-Morrison-Woodbury

Le complément de Schur est à la base de la formule de Sherman-Morrison-Woodbury de mise à jour de l'inverse d'une matrice :

Proposition 1.11 *Soient A, B, C, D des matrices de dimensions $n \times n$, $n \times m$, $m \times n$ et $m \times m$. Supposons que A et $D - CA^{-1}B$ soient inversibles. Alors $A - BD^{-1}C$ est inversible et*

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Cette proposition est prouvée à l'exercice 1.9.

1.15 DÉCOMPOSITION DE SCHUR

Théorème 1.12 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ il existe une matrice unitaire $U \in \mathbb{U}_n$ et une matrice triangulaire supérieure R telles que*

$$A = URU^*.$$

Cette décomposition est appelée décomposition de Schur de A. Pour une matrice réelle, la décomposition de Schur est réelle si et seulement si les valeurs propres de A sont réelles.

Démonstration. Par récurrence sur n. Le cas $n = 1$ est immédiat. Supposons que le théorème soit vrai pour des matrices de taille $n - 1 \times n - 1$. Soit λ une valeur propre de A et soit $x \in \mathbb{C}^n$ un vecteur propre associé avec

$\|x\|_2 = 1$. Soit $V = (x\ Z)$ une matrice unitaire : $V \in \mathbb{U}_n$ et sa première colonne est le vecteur x. On a :

$$V^*AV = \begin{pmatrix} x^* \\ Z^* \end{pmatrix} A \begin{pmatrix} x & Z \end{pmatrix} = \begin{pmatrix} x^*Ax & x^*AZ \\ Z^*Ax & Z^*AZ \end{pmatrix}.$$

Puisque $Ax = \lambda x$ et que V est unitaire, on a $Z^*x = 0$ et donc $Z^*Ax = \lambda Z^*x = 0$. Ceci prouve que

$$V^*AV = \begin{pmatrix} \lambda & x^*AZ \\ 0 & Z^*AZ \end{pmatrix}.$$

Appliquons l'hypothèse de récurrence à la matrice $B = Z^*AZ \in \mathbb{C}^{n-1\times n-1}$. Il existe des matrices W et $T \in \mathbb{C}^{n-1\times n-1}$, W unitaire et T triangulaire supérieure, telles que $B = WTW^*$. Ainsi

$$V^*AV = \begin{pmatrix} \lambda & x^*AZ \\ 0 & WTW^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix} \begin{pmatrix} \lambda & x^*AZW \\ 0 & T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & W^* \end{pmatrix}$$

On obtient la décomposition de Schur $A = URU^*$ en prenant

$$R = \begin{pmatrix} \lambda & x^*AZW \\ 0 & T \end{pmatrix} \text{ et } U = V \begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix}.$$

Remarque 1.1. La démonstration précédente montre que l'on peut choisir la matrice unitaire U telle que les valeurs propres de A qui apparaissent sur la diagonale de R aient un ordre spécifique. On dit dans ce cas que l'on a une *décomposition de Schur ordonnée.*

Lorsque A est une matrice réelle, la décomposition de Schur $A = URU^*$ ne fait pas nécessairement intervenir des matrices réelles. On peut toutefois introduire une *décomposition de Schur réelle* à condition de prendre R triangulaire supérieure par blocs :

Théorème 1.13 *Pour toute matrice $A \in \mathbb{R}^{n\times n}$ il existe des matrices $Q \in \mathbb{O}_n$ orthogonale et $R \in \mathbb{R}^{n\times n}$ triangulaire supérieure par blocs telles que $A = QRQ^T$. De plus*

$$R = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ & R_{22} & \dots & R_{2p} \\ & & \ddots & \vdots \\ & & & R_{pp} \end{pmatrix},$$

chaque bloc diagonal R_{ii} est soit de taille 1×1 soit de taille 2×2 avec un spectre constitué de deux valeurs propres complexes conjuguées.

Démonstration. On démontre qu'il existe une matrice orthogonale $Q \in \mathbb{O}_n$ telle que

$$Q^T A Q = \begin{pmatrix} R_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}$$

et où R_{11} a les propriétés requises puis on raisonne par récurrence. Montrons comment construire ces matrices. Soit λ une valeur propre de A.

Si elle est réelle on procède comme au théorème 1.12 : on prend $q_1 \in \mathbb{R}^n$ de norme 1 tel que $Aq_1 = \lambda q_1$ que l'on complète par $n-1$ autres vecteurs pour en faire une base orthonormée (q_i) de $\mathbb{R}^n$. Soit Q la matrice orthogonale dont les colonnes sont les q_i. On a :

$$Q^T A Q = \begin{pmatrix} \lambda & B_{12} \\ 0 & B_{22} \end{pmatrix}.$$

Si $\lambda = \alpha + i\beta$ avec $\beta \neq 0$, soient $x, y \in \mathbb{R}^n$ tels que

$$A(x + iy) = (\alpha + i\beta)(x + iy),$$

c'est-à-dire

$$A(x\ y) = (x\ y)\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}.$$

Notons que $x+iy$ et $x-iy$ sont linéairement indépendants parce qu'associés aux valeurs propres distinctes $\alpha \pm i\beta$. On en déduit que x et y sont aussi linéairement indépendants. Soit (q_1, q_2) une base orthonormée de l'espace engendré par x et y et soit $B \in \mathbb{R}^{2\times 2}$ telles que $(q_1\ q_2)B = (x\ y)$. On a

$$A(q_1\ q_2) = (q_1\ q_2)B\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}B^{-1} = (q_1\ q_2)R_{11}.$$

On construit Q en prenant pour colonnes q_1 et q_2 complétés par $n-2$ autres vecteurs pour en faire une base orthonormée (q_i) de $\mathbb{R}^n$. On obtient $Q^T A Q = \begin{pmatrix} R_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}$ avec les propriétés souhaitées.

1.16 NOTES ET RÉFÉRENCES

Le terme *matrice* qui est au cœur du sujet de ce livre est utilisé pour la première fois par le mathématicien anglais James-Joseph Sylvester (1814-1897) en 1850 dans un

texte intitulé *Sur une nouvelle classe de théorèmes*. Ce mot provient de la racine indo-européenne *m—a* qui désigne la mère et qui a donné les mots latins *mater* (mère) et *matrix* : femelle reproductrice, puis l'organe qui sert de réceptacle au fœtus (l'utérus) et, par extension de sens, au moule (fonderie, sculpture), au contenant, à un registre (la matrice des impôts).

La notion de matrice est définie de manière générale par Arthur Cayley (1821-1895) dans son traité *Mémoire sur la théorie des matrices* (1858).

Il est difficile de recommander un ouvrage d'algèbre linéaire générale tant ce sujet a fait l'objet de publications. Osons toutefois le « Cours d'algèbre » de R. Godement [12] et « Algèbre linéaire » de notre collègue J. Grifone [16].

EXERCICES

Exercice 1.1 Matrices triangulaires inférieures

Notons $\mathcal{L}_n$ l'ensemble des matrices $A \in \mathbb{C}^{n\times n}$ qui sont triangulaires inférieures et $\mathcal{GL}_n$ le sous-ensemble de celles qui sont inversibles. Montrer que :

1. $\mathcal{L}_n$ est stable pour le produit des matrices,
2. Pour tout $A \in \mathcal{L}_n$, $\det A = a_{11} \dots a_{nn}$,
3. Les valeurs propres de $A \in \mathcal{L}_n$ sont les entrées diagonales a_{ii},
4. $A \in \mathcal{L}_n$ est inversible si et seulement si $a_{ii} \neq 0$ pour tout i,
5. Pour tout $A \in \mathcal{GL}_n$, la diagonale de A^{-1} est donnée par a_{ii}^{-1},
6. L'inverse d'une matrice $A \in \mathcal{GL}_n$ est lui aussi dans $\mathcal{GL}_n$ (ceci fait de cet ensemble un groupe multiplicatif).
7. Calculer l'inverse de la matrice $n \times n$

$$\begin{pmatrix} 1 & & & & & \\ a_2 & 1 & & & & \\ a_3 & 0 & 1 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ a_{n-1} & 0 & 0 & \dots & 1 & \\ a_n & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Exercice 1.2 Matrices de rang 1

Soient u et $v \in \mathbb{C}^n$ non nuls. On note R la matrice $uv^* \in \mathbb{C}^{n\times n}$.

1. Montrer que R est une matrice de rang 1 et que toute matrice de rang 1 est de ce type.
2. Montrer que les valeurs propres de R sont 0 et $\langle u, v\rangle$, déterminer les sous-espaces propres correspondants.
3. Montrer que R est diagonalisable si et seulement si $\langle u, v\rangle \neq 0$.

Exercice 1.3

Soient a et b deux nombres réels. Calculer une matrice unitaire $U \in \mathbb{U}_2$ et une matrice diagonale $D \in \mathbb{C}^{2\times 2}$ telles que

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = UDU^*.$$

Exercice 1.4

Soient α et β deux nombres réels. Calculer une matrice orthogonale $O \in \mathbb{O}_2$ et une matrice diagonale $D \in \mathbb{R}^2$ telles que

$$\begin{pmatrix} 1+\alpha^2 & \alpha\beta \\ \alpha\beta & 1+\beta^2 \end{pmatrix} = ODO^T.$$

Exercice 1.5

Soient a et $b \in \mathbb{R}^n$ non nuls. Calculer les valeurs propres et les sous-espaces propres de la matrice $2n \times 2n$ suivante :

$$\begin{pmatrix} (1+\|a\|_2^2)I_n & ba^T \\ ab^T & (1+\|b\|_2^2)I_n \end{pmatrix}.$$

Exercice 1.6

Soit $A \in \mathbb{C}^{m\times n}$ et soient $x \in \mathbb{C}^n$ et $y \in \mathbb{C}^m$ tels que $y^*Ax \neq 0$. Posons

$$B = A - \frac{Axy^*A}{y^*Ax}.$$

Montrer que :

1. $\operatorname{Im} B \subset \operatorname{Im} A \subset \operatorname{Im} B + \mathbb{C}\,Ax$,
2. $\operatorname{Ker} A \subset \operatorname{Ker} B$,
3. $x \in \operatorname{Ker} B$ et $x \notin \operatorname{Ker} A$.
4. En déduire que $\operatorname{rang} B = \operatorname{rang} A - 1$.

Exercice 1.7 Matrice compagnon

Étant donnés n nombres complexes $a_0, \ldots, a_{n-1}$, la matrice

$$A = \begin{pmatrix} 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & \ldots & 0 & -a_1 \\ 0 & 1 & \ldots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -a_{n-1} \end{pmatrix}$$

est appelée *matrice compagnon* du polynôme

$$P(z) = a_0 + a_1 z + \ldots + a_{n-1}z^{n-1} + z^n.$$

Montrer que le polynôme caractéristique de A est égal à $(-1)^n P(z)$.

Exercice 1.8

Soit la matrice par blocs :

$$M = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix}$$

avec $A \in \mathbb{C}^{n\times n}$, $D \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times m}$ et $C \in \mathbb{C}^{m\times n}$. Montrer que M est inversible si et seulement si A et D sont inversibles. Calculer alors l'inverse de M à l'aide de A, B et D. Donner l'inverse de la matrice

$$M = \begin{pmatrix} I_n & B \\ 0 & I_m \end{pmatrix}.$$

Exercice 1.9

Soit $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ avec $A \in \mathbb{C}^{n\times n}$, $D \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times m}$ et $C \in \mathbb{C}^{m\times n}$.

1. On suppose que A est inversible. Démontrer l'égalité

$$M = \begin{pmatrix} I_n & 0 \\ CA^{-1} & I_m \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I_n & A^{-1}B \\ 0 & I_m \end{pmatrix}.$$

   En déduire que $\det(M) = \det(A)\det(D - CA^{-1}B)$. Montrer que si de plus $n = m$ et $AC = CA$, alors $\det(M) = \det(AD - CB)$.

2. On suppose que D est inversible. Démontrer de la même façon l'égalité

$$M = \begin{pmatrix} I_n & BD^{-1} \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I_n & 0 \\ D^{-1}C & I_m \end{pmatrix}.$$

   En déduire que $\det(M) = \det(D)\det(A - BD^{-1}C)$. Montrer que si de plus $n = m$ et $BD = DB$, alors $\det(M) = \det(DA - BC)$.

3. On suppose que A et $D - CA^{-1}B$ sont inversibles. À l'aide de la question 1 donner une expression de M^{-1} utilisant A^{-1} et $(D - CA^{-1}B)^{-1}$.

4. On suppose que D et $A - BD^{-1}C$ sont inversibles. Calculer de même M^{-1} en utilisant D^{-1} et $(A - BD^{-1}C)^{-1}$.

5. Sous l'hypothèse que A et $D - CA^{-1}B$ sont inversibles, montrer grâce aux questions 3 et 4 que $A - BD^{-1}C$ est inversible et que

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$$

(formule de Sherman-Morrison-Woodbury de mise à jour de l'inverse d'une matrice).

Application : soient x et y deux vecteurs-colonne de $\mathbb{C}^n$. On suppose que A est inversible. Montrer que $A + xy^*$ est inversible si et seulement si $y^*A^{-1}x \neq -1$ et que $(A + xy^*)^{-1} = A^{-1} - A^{-1}xy^*A^{-1}/\left(1 + y^*A^{-1}x\right)$.

Exercice 1.10

Soient A et $B \in \mathbb{C}^{n\times n}$. Montrer que

$$\det\begin{pmatrix} A & B \\ B & A \end{pmatrix} = \det(A - B)\ \det(A + B).$$

Exercice 1.11

Soit la matrice par blocs :

$$M = \frac{1}{2}\begin{pmatrix} I_n & iI_n \\ iI_n & I_n \end{pmatrix}.$$

Montrer que

$$M^{-1} = \begin{pmatrix} I_n & -iI_n \\ -iI_n & I_n \end{pmatrix}.$$

Soient A et B dans $\mathbb{R}^{n\times n}$. Montrer que

$$M^{-1}\begin{pmatrix} A & B \\ -B & A \end{pmatrix} M = \begin{pmatrix} A+iB & 0 \\ 0 & A-iB \end{pmatrix}.$$

En déduire que

$$\det\begin{pmatrix} A & B \\ -B & A \end{pmatrix} = \det(A + iB)\det(A - iB) = |\det(A + iB)|^2.$$

Si de plus $AB = BA$ on a

$$\det\begin{pmatrix} A & B \\ -B & A \end{pmatrix} = \det(A^2 + B^2).$$

Exercice 1.12

Utiliser le complément de Schur pour déterminer l'inverse de la matrice 5×5 suivante :

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 & 2 \\ 4 & 1 & 0 & 1 & 2 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Exercice 1.13

Soit $A \in \mathbb{C}^{n\times n}$ tridiagonale

$$A(a,b,c) = \begin{pmatrix} a & b & 0 & \cdots & 0 \\ c & a & b & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c & a & b \\ 0 & \cdots & 0 & c & a \end{pmatrix}$$

où $a, b, c \in \mathbb{C}$ avec $bc \neq 0$. Nous allons calculer les valeurs propres et les vecteurs propres de $A(a,b,c)$.

1. On commence par traiter le cas des matrices $A(0,b,c)$. Montrer que pour tout $p = 1 \ldots n$, le vecteur $v^{(p)} = (v_1^{(p)}, \ldots, v_n^{(p)})^T$ où $v_k^{(p)} = \left(\sqrt{\frac{c}{b}}\right)^k \sin \frac{kp\pi}{n+1}$ est un vecteur propre de $A(0,b,c)$ relatif à une valeur propre λ_p que l'on précisera. Montrer que ces valeurs propres sont distinctes et que les vecteurs $v^{(p)}$ sont indépendants.
2. Déterminer les valeurs propres et les vecteurs propres de $A(a,b,c)$.

Exercice 1.14

Soit $A \in \mathbb{C}^{n\times n}$ antihermitienne, c'est-à-dire telle que $A^* = -A$.

1. Montrer que les valeurs propres de A sont des nombres complexes imaginaires purs.
2. Montrer que $I_n - A$ est inversible.
3. Montrer que $Q = (I_n - A)^{-1}(I_n + A)$ (connue sous le nom de transformation de Cayley) est unitaire et que $-1 \notin \text{spec } Q$.

Exercice 1.15

Soient x et y deux vecteurs linéairement indépendants de $\mathbb{C}^n$. On considère la matrice $A = xy^* + yx^* \in \mathbb{C}^{n\times n}$.

1. Montrer que A est hermitienne, de rang 2 et déterminer Im A.
2. Déterminer les valeurs propres de A ainsi que les sous-espaces propres associés (utiliser la décomposition $\mathbb{C}^n = \text{Im } A \oplus (\text{Im } A)^{\perp}$). Préciser le cas $x, y \in \mathbb{R}^n$.
3. Étudier par la même méthode le cas de la matrice $B = xy^* - yx^* \in \mathbb{C}^{n\times n}$. Montrer qu'elle est antihermitienne, c'est-à-dire telle que $B^* = -B$, de rang 2, déterminer ses valeurs propres, ses vecteurs propres et préciser le cas $x, y \in \mathbb{R}^n$.

Chapitre **2**

L'arithmétique « virgule flottante »

Dans ce chapitre nous allons étudier l'arithmétique « virgule flottante » puisque c'est elle (ou l'une de ses versions) que le calcul scientifique utilise en machine au lieu de celle du corps $\mathbb{R}$ des nombres réels. Il faut bien avoir conscience du fait que cela introduit des erreurs d'arrondi et qu'à force d'empiler de telles erreurs on peut aboutir à des résultats sans signification. Nous allons définir les nombres flottants, le concept d'arrondi, les opérations sur les flottants et nous étudierons le problème du calcul d'un produit scalaire.

2.1 LES NOMBRES FLOTTANTS

Définition 2.1 *Donnons-nous quatre nombres entiers :* $\beta > 1$, $t \geqslant 1$, $e_{\min} \in -\mathbb{N}$ *et* $e_{\max} \in \mathbb{N}$. *Les nombres flottants qui leur sont associés sont les nombres réels suivants*

$$y = \pm m \times \beta^{e-t}$$

β *est la base, en général* $\beta = 2, 10$ *ou* 16,

t *est la précision,*

e *est l'exposant, c'est un entier qui vérifie* $e_{\min} \leqslant e \leqslant e_{\max}$ *;* $e_{\min}$ *est l'exposant minimum et* $e_{\max}$ *est l'exposant maximum,*

m *est la mantisse, c'est un entier qui vérifie* $0 \leqslant m \leqslant \beta^t - 1$.

La mantisse m d'un nombre flottant peut s'écrire

$$m = d_1\beta^{t-1} + \ldots + d_{t-1}\beta + d_t$$

avec $0 \leqslant d_i \leqslant \beta - 1$ ($d_1 \ldots d_{t-1}d_t$ est donc l'écriture en base β de l'entier m). On peut donc écrire un flottant

$$y = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right)\beta^e$$

avec $e_{\min} \leqslant e \leqslant e_{\max}$ et $0 \leqslant d_i \leqslant \beta - 1$. On utilise plutôt la notation suivante :

$$y = \pm .d_1d_2 \ldots d_t \times \beta^e.$$

Définition 2.2 *On appelle nombre flottant normalisé les nombres flottants*

$$y = \pm .d_1d_2 \ldots d_t \times \beta^e$$

avec $d_1 \neq 0$. *Leur ensemble est noté* $\mathbb{F}$.

Le choix des nombres flottants normalisés privilégie l'écriture $.1234\ 10^{-6}$ plutôt que $.01234\ 10^{-5}$.

2.2 ARRONDIS

Définition 2.3 *On appelle fonction d'arrondi toute fonction*

$$fl : \mathbb{R} \to \mathbb{F}$$

qui associe à un nombre réel x *le flottant* $fl(x)$ *le plus proche de* x.

Cette définition n'est pas complètement déterministe. Il y a plusieurs stratégies possibles pour définir l'arrondi de x lorsque celui-ci est équidistant de deux flottants. Par exemple on peut prendre le flottant le plus éloigné de 0.

Définition 2.4 *Les concepts d'overflow et d'underflow pour un nombre réel* x *sont définis par les inégalités* $|x| > \max_{y\in\mathbb{F}} |y|$ *et* $0 < |x| < \min_{y\in\mathbb{F},\ y\neq 0} |y|$.

Définition 2.5 *L'unité d'arrondi est*

$$u = \frac{\beta^{1-t}}{2}.$$

L'énoncé suivant prouve que, dans le changement d'un nombre réel par son arrondi, on commet une erreur relative constante : c'est la propriété fondamentale des nombres flottants.

Théorème 2.6 *Pour tout nombre réel x contenu dans l'intervalle*

$$\left[\min_{y\in\mathbb{F},\, y>0} y,\ \max_{y\in\mathbb{F},\, y>0} y\right],$$

il existe un autre nombre réel δ, $|\delta| < u$, tel que

$$fl(x) = x(1+\delta)$$

ou bien

$$\left|\frac{fl(x)-x}{x}\right| \leqslant u.$$

Démonstration. Prenons $x \in [\beta^e, \beta^{e+1}]$. Lorsque $x = \beta^e$ le résultat est évident. Lorsque $x > \beta^e$, les nombres flottants contenus dans l'intervalle $[\beta^e, \beta^{e+1}]$ sont du type $.d_1 d_2 \ldots d_t \times \beta^{e+1}$ avec $d_1 \neq 0$, $0 \leqslant d_i \leqslant \beta - 1$. La distance entre deux tels nombres consécutifs est constante et égale à $\beta^{e+1-t} = 2\beta^e u$. La distance entre x et $fl(x)$ est au plus la moitié de cette distance c'est à dire

$$|fl(x) - x| \leqslant 2\beta^e u/2 < |x|u.$$

Remarque 2.1. La norme IEEE utilise la base $\beta = 2$. Pour cette norme, les *nombres flottants en double précision* ont une précision $t = 53$ et l'unité d'arrondi vaut donc $u = 2^{-53} \approx 1.11 \times 10^{-16}$. Pour un nombre écrit en base 10, on a ainsi 16 chiffres significatifs après la virgule.

Remarque 2.2. Certains auteurs considèrent également l'*epsilon machine* ε_M qui est la distance entre 1 et le plus petit élément de $\mathbb{F}$ strictement plus grand. Pour la norme IEEE, on a $\varepsilon_M = 2u$ et donc, pour la double précision, $\varepsilon_M = 2^{-52} \approx 2.22 \times 10^{-16}$.

Remarque 2.3. Le fait que les erreurs d'arrondi soient relativement constantes n'est pas sans conséquences lorsque ce sont les erreurs absolues qui importent. Voici un calcul effectué à l'aide de Maple ($\beta = 10, t = 10$) :

```
> Digits := 10;
                              Digits := 10
> evalf(13 + 2000 * Pi);
                              6296.185308
> sin(6296.185308);
                              .4201677813
> sin(13.);
                              .4201670368
```

2.3 L'ARITHMÉTIQUE FLOTTANTE

Définition 2.7 *Notons* $\circ$ *l'une des opérations arithmétiques suivantes :* $+, -, \times$ *et* $/$ *. On définit l'opération flottante correspondante par*

$$x \tilde{\circ} y = fl(x \circ y)$$

pour tous x *et* y *réels.*

Notons, en vertu du théorème 2.6, qu'il existe δ, $|\delta| < u$, tel que

$$x \tilde{\circ} y = (x \circ y)(1 + \delta).$$

En général, x et y sont eux-mêmes flottants mais il n'y a aucune raison pour que l'opération $x \circ y$ fournisse un résultat dans $\mathbb{F}$. Reprenons l'exemple des flottants de l'exercice 2.1.

- $.99 \tilde{+} .099 = fl(1.089) = 1.1$
- $.02 \tilde{\times} 9.9 = fl(.198) = .20$,
- $9.9 \tilde{/} .02 = fl(495) = overflow$.

2.4 EXEMPLE : LE CALCUL DU PRODUIT SCALAIRE

Le calcul d'un produit scalaire est une opération essentielle que l'on retrouve lors d'un produit matrice-vecteur ou matrice-matrice. Nous allons analyser ce qui se passe lorsqu'un tel produit est calculé en arithmétique flottante.

2.4.1 Calcul en série

Soient $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ deux vecteurs-ligne. On souhaite calculer $S_n = x_1y_1 + \ldots + x_ny_n$. Bien sûr, on doit spécifier quel est l'algorithme de calcul suivi. Nous prenons :

$$P_1 = x_1y_1,\ S_1 = P_1.\ P_{k+1} = x_{k+1}y_{k+1},\ S_{k+1} = S_k + P_{k+1},$$

ce qui conduit à un algorithme flottant de même conception mais en y remplaçant les opérations usuelles par leurs contreparties flottantes. On obtient donc, à la place des S_k, des quantités $\tilde{S}_k$ qui vérifient

- $\tilde{P}_1 = fl(x_1y_1) = x_1y_1(1 + \delta_1)$,
- $\tilde{S}_1 = \tilde{P}_1 = x_1y_1(1 + \delta_1)$,
- $\tilde{P}_2 = fl(x_2y_2) = x_2y_2(1 + \delta_2)$,
- $\tilde{S}_2 = fl(\tilde{S}_1 + \tilde{P}_2) = (x_1y_1(1 + \delta_1) + x_2y_2(1 + \delta_2))(1 + \delta_3)$,
- $\tilde{S}_n = x_1y_1(1 + \delta)^n + x_2y_2(1 + \delta)^n + x_3y_3(1 + \delta)^{n-1} + \ldots + x_ny_n(1 + \delta)^2$,

avec les notations suivantes : tous les δ vérifient $|\delta| < u$ et une expression du type $(1 + \delta)^n$ est écrite pour un produit $(1 + \delta_1)\ldots(1 + \delta_n)$. Ils ont la propriété suivante :

Proposition 2.8 *Considérons des nombres $|\delta_i| < u$, $1 \leqslant i \leqslant n$, avec $nu < 1$. Notons*

$$\gamma_n = \frac{nu}{1 - nu}.$$

Alors

$$\Pi_{i=1}^{n}(1 + \delta_i) = 1 + \theta_n$$

pour un réel θ_n qui vérifie l'inégalité $|\theta_n| < \gamma_n$.

Dans le cas du produit scalaire nous avons obtenu :

- $S_n = x_1y_1 + \ldots + x_ny_n$
- $\tilde{S}_n = x_1y_1(1 + \theta_n) + x_2y_2(1 + \theta'_n) + x_3y_3(1 + \theta_{n-1}) + \ldots + x_ny_n(1 + \theta_2)$,

À ce stade de notre analyse nous avons deux interprétations possibles de ce résultat. La première est de constater que l'on a obtenu une estimation de l'erreur absolue commise dans ce calcul :

$$\tilde{S}_n - S_n = x_1y_1\theta_n + x_2y_2\theta'_n + x_3y_3\theta_{n-1} + \ldots + x_ny_n\theta_2.$$

Remarquons que cette erreur dépend de l'ordre dans lequel on effectue les calculs, un ordre pour lequel

$$|x_1 y_1| \leqslant |x_2 y_2| \leqslant \ldots \leqslant |x_n y_n|$$

semblant préférable. Notons que

$$|\tilde{S}_n - S_n| \leqslant \gamma_n \sum_{i=1}^{n} |x_i y_i|.$$

La seconde interprétation découle des identités précédentes et montre que

$$\tilde{S}_n = x_1(y_1(1+\theta_n)) + x_2(y_2(1+\theta'_n)) + x_3(y_3(1+\theta_{n-1})) + \ldots + x_n(y_n(1+\theta_2))$$

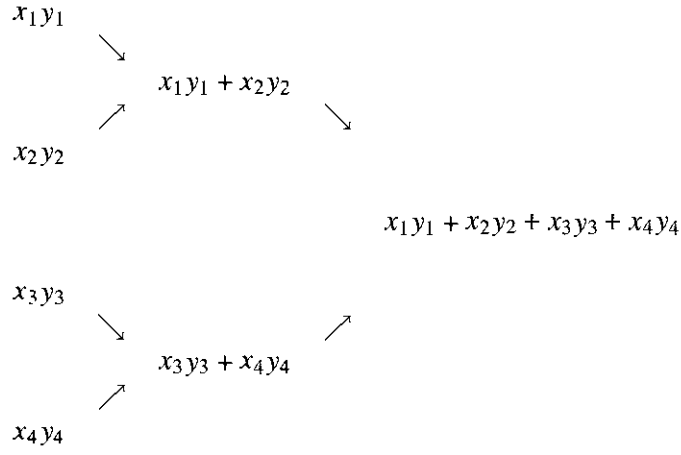
autrement dit $\tilde{S}_n$ peut être vu comme le produit scalaire exact de deux vecteurs proches des vecteurs de données : $x' = (x_1, \ldots, x_n)$ et $y' = (y_1(1+\theta_n), \ldots, y_n(1+\theta_2))$. Le choix fait de x' et y' n'est bien sûr pas unique, toute mixture $x' = (\alpha_1 x_1, \ldots, \alpha_n x_n)$ et $y' = (\beta_1 y_1, \ldots, \beta_n y_n)$ avec $\alpha_1 \beta_1 = (1+\theta_n)$ et cetera est acceptable.

Cette interprétation, introduite par Givens et Wilkinson, est connue sous le nom de « backward error analysis » ou « analyse rétrograde des erreurs ». Sa signification est importante pour un numéricien. Si l'on suppose que les vecteurs x et y sont donnés de façon approchée (soit résultant de calculs approchés, soit donnés expérimentalement), le calcul flottant du produit scalaire $\langle x, y \rangle$ qui nous a conduit à $\tilde{S}_n = \langle x', y' \rangle$ sera tout à fait acceptable si les vecteurs x' et y' sont dans les tolérances du problème.

Insistons sur le fait que ces deux points de vue sur l'analyse des erreurs peuvent être transposés à tout calcul approché : une analyse directe fournit une estimation de la taille de l'erreur commise, une analyse rétrograde donne au calcul sa signification.

2.4.2 Calcul en éventail

La stratégie du calcul en éventail consiste à effectuer tous les produits $x_i y_i$ puis, pour les additionner, à les séparer en deux sous-ensembles, à sommer chacun d'eux et enfin à additionner ces deux sommes. Cette procédure s'applique aussi aux sous-ensembles en question puis aux sous-sous-ensembles et ainsi de suite ... ce qui conduit au schéma :



Le calcul en flottant du produit scalaire $S_n = x_1y_1 + \ldots + x_ny_n$ selon cet algorithme conduit au résultat suivant

$$\hat{S}_n = x_1y_1(1+\theta^{(1)}) + x_2y_2(1+\theta^{(2)}) + \ldots + x_ny_n(1+\theta^{(n)})$$

où chacun des nombres $\theta^{(i)}$ vérifie

$$\left|\theta^{(i)}\right| \leqslant \gamma_N = \frac{Nu}{1-Nu},\ N = 1 + \lceil \log_2 n \rceil .$$

Ainsi

$$|\hat{S}_n - S_n| \leqslant \gamma_N \sum_{i=1}^{n} |x_i y_i|$$

et la borne obtenue pour cette erreur absolue est bien moindre que pour le calcul en série.

2.5 NOTES ET RÉFÉRENCES

Le lecteur voulant en savoir plus peut consulter le livre de J.-M. Muller « Arithmétique des ordinateurs » [25] qui est téléchargeable gratuitement via le site

http://prunel.ccsd.cnrs.fr/ensl-00086707

ou bien l'ouvrage plus récent de J.-C. Bajard et J.-M. Muller « Calcul et arithmétique des ordinateurs » [4].

EXERCICES

Exercice 2.1

Quels sont les nombres flottants normalisés correspondant aux paramètres $\beta = 10$, $t = 2$, $e_{\min} = -1$, $e_{\max} = 1$.

Exercice 2.2

Montrer que le plus grand des nombres flottants normalisés positif est $\beta^{e_{\max}}(1-\beta^{-t})$ et le plus petit $\beta^{e_{\min}-1}$.

Exercice 2.3

Dans le système des nombres flottants de l'exercice 2.1 calculer l'expression $3(4/3-1)-1$ en suivant le schéma

$$4/3 \rightarrow 4/3-1 \rightarrow 3(4/3-1) \rightarrow 3(4/3-1)-1.$$

Effectuer ce même calcul sur votre calculette ou bien à l'aide de Maple en prenant « Digits := 20 ».

Exercice 2.4

Prouver la proposition 2.8.

Chapitre 3

Normes sur les espaces de matrices

Un espace de matrices tel que $\mathbb{R}^{m\times n}$ ou $\mathbb{C}^{m\times n}$ étant un espace vectoriel de dimension finie on peut lui associer toutes sortes de normes qui, rappelons le, sont toutes équivalentes, c'est-à-dire définissent la même topologie. Mais toutes ces normes possibles n'ont pas forcément un bon comportement vis-à-vis de la multiplication des matrices ou des produits matrice-vecteur, à la différence des normes matricielles que nous introduisons ici.

3.1 NORME D'OPÉRATEUR

Définition 3.1 *Une norme sur $\mathbb{C}^{n\times n}$ est multiplicative si $\|AB\| \leqslant \|A\|\,\|B\|$ quelles que soient les matrices A et $B \in \mathbb{C}^{n\times n}$.*

Définition 3.2 *Une norme sur $\mathbb{C}^{m\times n}$ est consistante avec des normes $\|.\|_m$ sur $\mathbb{C}^m$ et $\|.\|_n$ sur $\mathbb{C}^n$ si $\|Ax\|_m \leqslant \|A\|\,\|x\|_n$ pour toute matrice $A \in \mathbb{C}^{m\times n}$ et tout vecteur $x \in \mathbb{C}^n$.*

Étant donné deux espaces vectoriels normés de dimension finie E et F, nous notons par $\mathcal{L}(E, F)$ l'espace des applications linéaires $L : E \to F$.

Définition 3.3 *La norme d'opérateur sur l'espace $\mathcal{L}(E, F)$ est définie par*

$$\|L\| = \sup_{x\in E,\, x\neq 0} \frac{\|L(x)\|}{\|x\|}.$$

Remarquons que dans cette définition nous notons de la même manière la norme de E, celle de F et celle de $\mathcal{L}(E, F)$. Le contexte permet de s'y retrouver.

Par définition du supremum (le plus petit des majorants) $\|L\|$ est la plus petite des constantes $C \in \mathbb{R}$ qui vérifient

$$\|L(x)\| \leqslant C \|x\|$$

pour tout $x \in E$. D'autres caractérisations des normes d'endomorphisme sont données à l'exercice 3.1.

Proposition 3.4 *La norme d'opérateur possède les propriétés suivantes :*

1. *C'est une norme, c'est-à-dire que*
   a) $\|L\| \geqslant 0$,
   b) $\|L\| = 0$ *si et seulement si* $L = 0$,
   c) $\|\lambda L\| = |\lambda| \, \|L\|$ *pour tout scalaire* λ,
   d) $\|L + M\| \leqslant \|L\| + \|M\|$
2. *Elle est consistante : pour tout* $x \in E$

$$\|L(x)\| \leqslant \|L\| \, \|x\| ,$$

3. *Pour tout* $L \in \mathcal{L}(E, F)$ *et* $N \in \mathcal{L}(F, G)$, *où* G *est un troisième espace normé,*

$$\|N \circ L\| \leqslant \|N\| \, \|L\| .$$

Nous ne démontrerons pas cette proposition.

Exemple 3.1 :

1. La norme d'opérateur d'une matrice $A \in \mathbb{C}^{m \times n}$ associée à la norme

$$\|x\|_1 = \sum_{j=1}^{n} |x_j|$$

sur $\mathbb{C}^n$ et

$$\|y\|_1 = \sum_{i=1}^{m} |y_i|$$

sur $\mathbb{C}^m$ est donnée par

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} |a_{ij}| .$$

2. La norme d'opérateur d'une matrice $A \in \mathbb{C}^{m\times n}$ associée à la norme

$$\|x\|_\infty = \max_{1\leqslant j\leqslant n} |x_j|$$

sur $\mathbb{C}^n$ et

$$\|y\|_\infty = \max_{1\leqslant i\leqslant m} |y_i|$$

sur $\mathbb{C}^m$ est donnée par

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

3.2 RAYON SPECTRAL

Définition 3.5 *Le rayon spectral d'une matrice $A \in \mathbb{C}^{n\times n}$ est le nombre*

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

Le rayon spectral d'une matrice joue un rôle central dans l'analyse de nombreux phénomènes et il est important de pouvoir le calculer. Voici deux résultats en ce sens :

Proposition 3.6 *$\rho(A) \leqslant \|A\|$ pour toute norme consistante.*

Démonstration. Si x est un vecteur propre unitaire associé à la valeur propre λ de A on a :

$$|\lambda| = |\lambda|\,\|x\| = \|\lambda x\| = \|Ax\| \leqslant \|A\|\,\|x\| = \|A\| .$$

Théorème 3.7 (Théorème de Gelfand) *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ et pour toute norme sur $\mathbb{C}^{n\times n}$*

$$\rho(A) = \lim_{p\to\infty} \|A^p\|^{1/p} .$$

Démonstration. Montrons tout d'abord que la valeur de la limite est indépendante de la norme choisie. Notons $\|.\|_a$ une norme sur $\mathbb{C}^{n\times n}$ pour laquelle

$$\rho(A) = \lim_{p\to\infty} \|A^p\|_a^{1/p}$$

et établissons le résultat pour une seconde norme $\|.\|_b$. Elle est équivalente à la première c'est-à-dire qu'il existe deux constantes positives α et β telles que

$$\alpha\,\|B\|_a \leqslant \|B\|_b \leqslant \beta\,\|B\|_a$$

pour toute matrice $B \in \mathbb{C}^{n\times n}$. Prenons $B = A^p$ et passons aux racines p-ièmes, on obtient

$$\alpha^{1/p} \|A^p\|_a^{1/p} \leqslant \|A^p\|_b^{1/p} \leqslant \beta^{1/p} \|A^p\|_a^{1/p}.$$

Il suffit alors de passer à la limite lorsque $p \to \infty$ pour obtenir le résultat.

Nous allons prouver le théorème dans un cas simplifié : celui où A est diagonalisable. On peut alors écrire $A = PDP^{-1}$ avec $D = \mathrm{diag}(\lambda_i)$ où les λ_i sont les valeurs propres de A. Prenons pour norme sur $\mathbb{C}^{n\times n}$

$$\|B\| = \max_{1\leqslant i,j\leqslant n} \left|(P^{-1}BP)_{ij}\right|.$$

Puisque $A^p = PD^pP^{-1}$ nous obtenons :

$$\|A^p\|^{1/p} = \left(\max_{1\leqslant i\leqslant n} |\lambda_i^p|\right)^{1/p} = \max_{1\leqslant i\leqslant n} |\lambda_i^p|^{1/p} = \rho(A).$$

3.3 LA NORME SPECTRALE

Définition 3.8 *La norme spectrale d'une matrice $A \in \mathbb{C}^{m\times n}$ est la norme d'opérateur associée aux structures hermitiennes canoniques de $\mathbb{C}^n$ et $\mathbb{C}^m$:*

$$\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2$$

(voir l'exercice 3.1) avec

$$\|x\|_2^{\,2} = \sum_{i=1}^{n} |x_i|^2.$$

Théorème 3.9 *La norme spectrale d'une matrice $A \in \mathbb{C}^{m\times n}$ est égale à la racine carrée du rayon spectral de A^*A :*

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

et, lorsque A est hermitienne,

$$\|A\|_2 = \rho(A).$$

Démonstration. Remarquons que A^*A est une matrice hermitienne et que ses valeurs propres sont $\geqslant 0$. En effet $(A^*A)^* = A^*A^{**} = A^*A$ et si $A^*Ax = \lambda x$ avec $x \neq 0$ on obtient, en multipliant à gauche par x^*,

$$\|Ax\|_2^2 = x^*A^*Ax = \lambda x^*x = \lambda \|x\|_2^2$$

de sorte que $\lambda \geqslant 0$. En vertu du théorème spectral (théorème 1.6) on peut décomposer $A^*A = UDU^*$ avec U unitaire et $D = \mathrm{diag}(\lambda_i)$ où les λ_i sont les valeurs propres ($\geqslant 0$) de A^*A. Revenons à la définition de la norme spectrale. On a : $\|A\|_2^2 =$

$$\max_{\|x\|_2=1} \|Ax\|_2^2 = \max_{\|x\|_2=1} x^*A^*Ax = \max_{\|x\|_2=1} x^*UDU^*x = \max_{\|x\|_2=1} y^*Dy$$

avec $y = U^*x$. Mais puisque U est unitaire, lorsque x décrit la sphère unité dans $\mathbb{C}^n$ il en est de même pour y de sorte que

$$\|A\|_2^2 = \max_{\|y\|_2=1} y^*Dy.$$

Le maximum de $y^*Dy = \sum_{i=1}^n \lambda_i |y_i|^2$ sur la sphère unité est égal à $\lambda_{\bar{i}} = \max_i \lambda_i = \rho(A^*A)$, il est atteint lorsque $y = e_{\bar{i}}$ le vecteur dont les coordonnées sont nulles sauf celle d'indice $\bar{i}$ égale à 1. Ceci établit la première identité. Pour la seconde on note que

$$\rho(A^*A) = \rho(A^2) = \rho(A)^2$$

lorsque A est hermitienne.

Remarque 3.1. La norme spectrale d'une matrice-colonne $a \in \mathbb{C}^{m\times 1}$ est égale à sa norme en tant que vecteur de $\mathbb{C}^m$: $\|a\|_2 = \sqrt{\sum_{i=1}^m |a_i|^2}$. La notation $\|a\|_2$ n'est donc pas ambiguë. Il en est de même pour les vecteurs-ligne.

La norme spectrale est unitairement invariante :

Proposition 3.10 *Quelles que soient la matrice $A \in \mathbb{C}^{m\times n}$ et les matrices unitaires $U \in \mathbb{U}_m$ et $V \in \mathbb{U}_n$, on a*

$$\|UAV\|_2 = \|A\|_2 .$$

Démonstration. $\|UAV\|_2^2 = \rho(V^*A^*U^*UAV) = \rho(V^*A^*AV) = \rho(A^*A) = \|A\|_2^2$. La première égalité vient du théorème 3.9, la seconde a lieu parce que U est unitaire et la troisième parce que V^*A^*AV et A^*A sont semblables.

3.4 LA NORME DE FROBENIUS

Définition 3.11 *Étant donné deux matrices $A, B \in \mathbb{C}^{m\times n}$ posons*

$$\langle A, B\rangle_F = \mathrm{trace}\,(B^*A).$$

C'est un produit scalaire hermitien sur $\mathbb{C}^{m\times n}$ *et l'on a*

$$\langle A, B\rangle_F = \sum_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} a_{ij}\overline{b_{ij}}.$$

La norme associée à ce produit scalaire s'appelle la norme de Frobenius :

$$\|A\|_F^2 = \text{trace}\,(A^*A) = \sum_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} |a_{ij}|^2.$$

Voici quatre propriétés de la norme de Frobenius :

Proposition 3.12

1. *Elle est multiplicative : pour tout* $A\in\mathbb{C}^{m\times n}$ *et* $B\in\mathbb{C}^{n\times p}$

$$\|AB\|_F \leqslant \|A\|_F\,\|B\|_F,$$

2. *C'est une norme consistante avec* $\|.\|_2$ *: pour tout* $A\in\mathbb{C}^{m\times n}$ *et* $x\in\mathbb{C}^n$

$$\|Ax\|_2 \leqslant \|A\|_F\,\|x\|_2,$$

3. *Constantes d'équivalence : pour tout* $A\in\mathbb{C}^{m\times n}$

$$\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\,\|A\|_2,$$

4. *Pour tous* $A\in\mathbb{C}^{m\times n}$ *et* $B\in\mathbb{C}^{n\times p}$

$$\|AB\|_F \leqslant \|A\|_2\,\|B\|_F.$$

Démonstration. 1 se déduit de 3 et 4. 2 provient de 1 en prenant pour matrice B la matrice-colonne x. 3 se prouve ainsi : notons $\lambda_1\geqslant\lambda_2\geqslant\ldots\geqslant\lambda_n\geqslant 0$ les valeurs propres de A^*A (voir le théorème 3.9 et sa preuve). On a $\|A\|_2^2=\rho(A^*A)=\lambda_1$ et $\|A\|_F^2=\text{trace}\,(A^*A)=\lambda_1+\lambda_2+\ldots+\lambda_n\leqslant n\lambda_1$ ce qui prouve 3. Reste à prouver 4. Notons b_j la $j-$ième colonne de B. On a :

$$\|AB\|_F^2=\sum_{j=1}^{p}\|Ab_j\|_2^2\leqslant\sum_{j=1}^{p}\|A\|_2^2\,\|b_j\|_2^2=\|A\|_2^2\,\|B\|_F^2.$$

Une dernière propriété importante de la norme de Frobenius est son invariance unitaire :

Proposition 3.13 *Quelles que soient les matrices* $A\in\mathbb{C}^{m\times n}$ *et les matrices unitaires* $U\in\mathbb{U}_m$ *et* $V\in\mathbb{U}_n$ *on a*

$$\|UAV\|_F=\|A\|_F.$$

Démonstration. $\|UAV\|_F^2 = \text{trace}\,(V^*A^*U^*UAV) = \text{trace}\,(V^*A^*AV) = \text{trace}\,(A^*A) = \|A\|_F^2$: la première égalité vient de la définition de la norme, la seconde de $U^*U = I_m$ et la troisième d'une propriété classique de la trace : $\text{trace}\,(P^{-1}A^*AP) = \text{trace}\,(A^*A)$ pour toute matrice inversible P (voir le paragraphe 1.7).

3.5 LE THÉORÈME DE PERTURBATION DE NEUMANN

En perturbant une matrice carrée inversible on récupère une matrice qui est encore inversible ou, en termes plus savants, le groupe linéaire $\mathbb{GL}_n$ est ouvert dans $\mathbb{C}^{n\times n}$. Voici une version quantitative de ce résultat :

Proposition 3.14 *Notons $\|.\|$ une norme multiplicative sur $\mathbb{C}^{n\times n}$. Si $\|A\| < 1$ alors $I_n - A$ est inversible et son inverse est la somme de la série absolument convergente*

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

De plus

$$\left\|(I_n - A)^{-1}\right\| \leqslant \frac{1}{1 - \|A\|}.$$

Démonstration. La série est absolument convergente parce que $\left\|A^k\right\| \leqslant \|A\|^k$ qui est le terme général d'une série convergente. Pour calculer sa somme on passe à la limite dans l'identité

$$I_n - A^{p+1} = (I_n - A)\sum_{k=0}^{p} A^k$$

en remarquant que $A^{p+1} \to 0$ puisque c'est le terme général d'une série convergente. On a enfin

$$\left\|(I_n - A)^{-1}\right\| = \left\|\sum_{k=0}^{\infty} A^k\right\| \leqslant \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|}.$$

Corollaire 3.15 *Notons $\|.\|$ une norme multiplicative sur $\mathbb{C}^{n\times n}$ et soit $B \in \mathbb{GL}_n$. Si $\|A\| < \left\|B^{-1}\right\|^{-1}$ alors $B - A$ est inversible.*

Démonstration. On écrit $B - A = B(I_n - B^{-1}A)$ et on note que $\left\|B^{-1}A\right\| \leqslant \left\|B^{-1}\right\| \|A\| < 1$. On applique alors la proposition précédente.

Ce corollaire signifie que la boule ouverte de centre B et de rayon $\left\|B^{-1}\right\|^{-1}$ est contenue dans $\mathbb{GL}_n$. Ceci prouve que cet ensemble est ouvert.

3.6 NOTES ET RÉFÉRENCES

L'axiomatisation de ce que l'on appelle aujourd'hui « espace de Banach » a été formalisée par S. Banach dans sa dissertation (1920) bien que d'autres auteurs tels Wiener et Minkowski aient eu leur contribution. L'inégalité de Cauchy est due à Cauchy (1821), on lui associe souvent les noms de Schwarz et de Bunyakovski. La norme spectrale (notée ici $\|.\|_2$) a été introduite par Peano (1888) et la norme de Frobenius (notée $\|.\|_F$) par lui-même (1911). Nous avons aussi rencontré Israil Gelfand (1913-) et Carl Neumann (1832-1925) sur la série des puissances d'un opérateur ... à ne pas confondre avec d'autres Neumann !

EXERCICES

Exercice 3.1

Montrer que les quantités suivantes sont égales

1. $\|L\| = \sup_{x \in E,\, x \neq 0} \frac{\|Lx\|}{\|x\|}$,
2. $\sup_{\|x\| \leqslant 1,\, x \neq 0} \frac{\|Lx\|}{\|x\|}$,
3. $\sup_{\|x\|=1} \|Lx\|$,
4. $\sup_{\|x\| \leqslant 1} \|Lx\|$,
5. $\sup_{\|x\|<1,\, x \neq 0} \frac{\|Lx\|}{\|x\|}$,
6. $\sup_{\|x\|<1} \|Lx\|$.
7. Montrer que les cinq premiers supremums sont des maximums[1].

Exercice 3.2

Démontrer les affirmations contenues dans l'exemple 3.1.

Exercice 3.3

Montrer que les valeurs propres de $A \in \mathbb{C}^{n \times n}$ sont contenues dans le disque de centre 0 et de rayon $\max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|$.

Exercice 3.4

Soit $A \in \mathbb{C}^{n \times n}$. Montrer que si $\rho(A) < 1$ alors $I_n - A$ est inversible.

Exercice 3.5

Prouver le cas général du théorème 3.7. Raisonner de façon similaire mais au lieu de la forme diagonalisée $A = PDP^{-1}$ utiliser la décomposition de Jordan (théorème 1.5) $A = PJP^{-1}$. Remarquer que J s'écrit $J = D + N$ avec D diagonale, N nilpotente ($N^n = 0$) et $ND = DN$ pour calculer J^p.

1. *Maximum, minimum, supremum* et *infimum* font leur pluriel soit en *ums* comme dans *minimums* soit en *ma* comme dans *minima*.

Exercice 3.6

Soient a, b, c, d quatre nombres réels. Calculer les normes spectrale et de Frobenius de la matrice $\begin{pmatrix} a & b+ic \\ b-ic & d \end{pmatrix}$.

Exercice 3.7

Calculer la norme de I_n pour les normes $\|.\|_1$, $\|.\|_2$, $\|.\|_\infty$, $\|.\|_F$.

Exercice 3.8

Montrer que $\|U\|_2 = 1$ et $\|U\|_F = \sqrt{n}$ pour toute matrice unitaire $U \in \mathbb{U}_n$.

Exercice 3.9

Calculer les normes spectrale et de Frobenius de la matrice $A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix}$.

Exercice 3.10

Soit $A \in \mathbb{R}^{n\times n}$. Montrer que A est diagonale si et seulement si A est symétrique et a ses valeurs propres sur la diagonale (utiliser la norme de Frobenius).

Exercice 3.11

Soient $A \in \mathbb{C}^{n\times n}$ et $\varepsilon > 0$. Montrer qu'il existe une norme matricielle multiplicative N telle que :

$$N(A) \leqslant \rho(A) + \varepsilon.$$

On procède de la façon suivante : notons $\alpha = \rho(A) + \varepsilon$. D'après le théorème de Gelfand il existe un entier $p > 0$ tel que

$$\|A^p\|_2^{1/p} < \alpha.$$

On pose alors

$$N(x) = \sum_{i=0}^{p-1} \alpha^{p-i-1} \left\|A^i x\right\|_2$$

pour tout $x \in \mathbb{C}^n$.

1. Montrer que c'est une norme sur $\mathbb{C}^n$,
2. Montrer que la norme d'endomorphisme qui est associée à N (que l'on note aussi N) vérifie $N(A) \leqslant \alpha$,
3. Conclure.

Exercice 3.12

Soient $x, y \in \mathbb{C}^n$. Montrer que $\|xy^*\|_2 = \|xy^*\|_F = \|x\|_2 \|y\|_2$, $\|xy^*\|_1 = \|x\|_1 \|y\|_\infty$ et que $\|xy^*\|_\infty = \|x\|_\infty \|y\|_1$.

Exercice 3.13

Soient $A \in \mathbb{GL}_n$ et $H \in \mathbb{C}^{n\times n}$. Montrer, en utilisant le théorème de perturbation de Neumann, que

$$\lim_{H\to 0} \frac{(A+H)^{-1} - A^{-1} + A^{-1}HA^{-1}}{\|H\|} = 0.$$

Cette identité montre que l'application $\mathcal{I}nv : A \in \mathbb{GL}_n \to A^{-1} \in \mathbb{GL}_n$ est différentiable et en donne la différentielle en A dans la direction H : $D\mathcal{I}nv(A)H = -A^{-1}HA^{-1}$.

Exercice 3.14 Exponentielle de matrice

L'exponentielle d'une matrice $A \in \mathbb{C}^{n\times n}$ est définie par

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

Le but de cet exercice est d'établir quelques propriétés de l'exponentielle. Montrer que :

1. Cette série est absolument convergente et que, pour toute norme multiplicative, $\|\exp(A)\| \leqslant \exp\left(\|A\|\right)$,
2. $\exp(0) = I_n$,
3. Si $AB = BA$ alors $\exp(A+B) = \exp(A)\exp(B)$,
4. $\exp(A)$ est inversible, calculer son inverse,
5. Si $A = PDP^{-1}$ avec $P \in \mathbb{GL}_n$ alors $\exp(A) = P\exp(D)P^{-1}$,
6. En déduire que les valeurs propres de $\exp(A)$ sont les exponentielles des valeurs propres de A,
7. $\det\exp(A) = \exp(\text{trace } A)$,
8. $t \in \mathbb{R} \to \exp(tA) \in \mathbb{C}^{n\times n}$ est différentiable et
$$\frac{d}{dt}\exp(tA) = A\exp(tA).$$
9. Soient $x, y \in \mathbb{C}^n$ distincts de zéro. Calculer $\exp(xy^*)$ (distinguer les cas $x^*y \neq 0$ et $x^*y = 0$).

10. Soit A hermitienne et $A = U\Lambda U^*$ sa diagonalisation avec U unitaire et $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ diagonale. Écrire A sous forme d'une somme de matrices de rang un. En déduire l'expression de $\exp(A)$ sous forme d'un produit de matrices qui commutent entre elles.

Exercice 3.15

Soit $A \in \mathbb{C}^{m\times n}$. Montrer que

$$\left\| \begin{pmatrix} I_m & A \\ 0 & I_n \end{pmatrix} \right\|_2 = \sqrt{\frac{\|A\|_2 + 2 + \|A\|_2 \sqrt{\|A\|_2^2 + 4}}{2}}.$$

Exercice 3.16

Montrer que pour toute matrice $A \in \mathbb{C}^{n\times n}$ on a $\|A\|_1 = \|A^*\|_\infty$. En déduire que $\|A\|_2 \leqslant \sqrt{\|A\|_1 \, \|A\|_\infty}$.

Exercice 3.17

Un carré magique est une matrice $C \in \mathbb{R}^{n\times n}$ dont les entrées sont les entiers de 1 à n^2 rangés de telle sorte que la somme des termes d'une même ligne ou d'une même colonne soit la même. Par exemple

$$\begin{pmatrix} 6 & 7 & 2 \\ 1 & 5 & 9 \\ 8 & 3 & 4 \end{pmatrix}$$

est un carré magique 3×3. Montrer que dans un carré magique C d'ordre n la somme des lignes vaut $n(n^2 + 1)/2$ et que ce nombre est aussi égal à $\|C\|_2$.

Exercice 3.18 Matrices à diagonale strictement dominante

Une matrice $A \in \mathbb{C}^{n\times n}$ est à *diagonale strictement dominante* lorsque

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|$$

pour tout i. Montrer qu'une telle matrice est inversible (utiliser le corollaire 3.15 et la norme $\|.\|_\infty$ de l'exercice 3.2).

Exercice 3.19

Calculer la norme spectrale de la matrice $n+1 \times n+1$ suivante :

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

Exercice 3.20

Montrer que pour toute matrice $M \in \mathbb{C}^{m\times n}$ on a :

$$\det(I_n + M^* M) \leqslant \left(1 + \frac{\|M\|_F^2}{n}\right)^n.$$

Exercice 3.21

Montrer que $\mathbb{GL}_n$ est dense dans $\mathbb{C}^{n\times n}$: toute matrice $A \in \mathbb{C}^{n\times n}$ est limite d'une suite de matrices inversibles (utiliser la décomposition de Jordan de A ou bien la décomposition de Schur).

Chapitre **4**

La décomposition en valeurs singulières

4.1 DÉFINITION

Nous avons vu, au chapitre précédent, que la norme spectrale d'une matrice $A \in \mathbb{C}^{m\times n}$ est égale à la racine carrée de la plus grande valeur propre de A^*A. Plus généralement, posons :

Définition 4.1 *Les valeurs singulières de $A \in \mathbb{C}^{m\times n}$ sont les racines carrées des valeurs propres positives (> 0) de A^*A.*

Remarque 4.1.

1. Nous avons démontré au cours de la preuve du théorème 3.9 que les valeurs propres de A^*A sont positives ou nulles. Il est donc loisible d'en considérer les racines carrées.
2. On trouve ça et là une définition des valeurs singulières qui accepte 0 : ce sont alors les racines carrées des valeurs propres de A^*A. Nous ne trouvons aucun avantage à cette définition.
3. Les valeurs propres positives de A^*A et AA^* sont les mêmes. Il n'y a donc pas des valeurs singulières « à gauche » et des valeurs singulières « à droite » (voir l'exercice 4.1).
4. Si l'on note $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ les valeurs singulières de A alors

$$\|A\|_2^2 = \sigma_1^2 = \rho(A^*A)$$

et

$$\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2 = \text{trace}\,(A^*A).$$

Théorème 4.2 (Décomposition en valeurs singulières) *Pour toute matrice $A \in \mathbb{C}^{m\times n}$ de rang r, il existe des matrices unitaires $U \in \mathbb{U}_n$, $V \in \mathbb{U}_m$ et une matrice $\Sigma \in \mathbb{R}^{m\times n}$ telles que :*

$$A = V\Sigma U^*,\ \Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},\ D \in \mathbb{R}^{r\times r},\ D = \text{diag}(\sigma_1, \ldots, \sigma_r)$$

où $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ sont les valeurs singulières de A. Cette décomposition s'appelle la décomposition en valeurs singulières de A (singular value decomposition ou SVD en anglais).

Démonstration. Puisque A^*A est hermitienne, par le théorème spectral (théorème 1.6), on peut écrire que

$$U^*A^*AU = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

pour une matrice unitaire $U \in \mathbb{U}_n$. Notons $u_1, \ldots, u_n$ les colonnes de U. L'écriture précédente prouve que les vecteurs Au_i, $1 \leqslant i \leqslant n$, sont deux à deux orthogonaux, que $\|Au_i\|_2 = \sigma_i$, $1 \leqslant i \leqslant r$, et que $Au_i = 0$, $r+1 \leqslant i \leqslant n$. Posons

$$v_i = Au_i/\sigma_i,\ 1 \leqslant i \leqslant r.$$

Ces r vecteurs de $\mathbb{C}^m$ sont orthonormés. Complétons-les pour en faire une base orthonormée de $\mathbb{C}^m$. On obtient une matrice unitaire $V \in \mathbb{U}_m$ dont les colonnes sont $v_1, \ldots, v_m$ et, par construction, $AU = V\Sigma$. Cette identité prouve aussi que le rang de A est égal au nombre de valeurs singulières.

Remarque 4.2.

1. Il n'y a pas unicité de la décomposition en valeurs singulières. Par exemple $I_n = UI_nU^*$ pour toute matrice unitaire $U \in \mathbb{U}_n$. C'est donc un abus que d'utiliser l'article défini « la » !

2. Notons U_r et V_r les matrices obtenues à partir de U et V en ne conservant que les r premières colonnes. On a aussi
$$A = V_r D U_r^*,$$
c'est la *décomposition en valeurs singulières réduite.*
3. Lorsque A est une matrice réelle, on peut prendre pour U et V des matrices orthogonales.
4. La démonstration du théorème 4.2 montre que l'image par A de la sphère unité dans l'orthogonal du noyau de A (ensemble des vecteurs $x \in (\mathrm{Ker}\, A)^\perp$ de norme 1) est l'ellipsoïde dans le sous-espace $\mathrm{Im}\, A \subset \mathbb{C}^m$ dont les axes sont portés par les vecteurs v_i, $1 \leqslant i \leqslant r$, et la longueur des demi-axes σ_i : si $x = \sum_{i=1}^r x_i u_i$ avec $\sum_{i=1}^r |x_i|^2 = 1$ alors $Ax = \sum_{i=1}^r \sigma_i x_i v_i$ et
$$\sum_{i=1}^r \frac{|\sigma_i x_i|^2}{\sigma_i^2} = 1.$$

4.2 CALCUL DES VALEURS SINGULIÈRES

Les valeurs singulières de A sont les racines carrées des valeurs propres positives de A^*A ou de AA^* : il vaut mieux choisir celle de ces deux matrices qui a la plus petite taille ! Ceci ramène le problème du calcul des valeurs singulières à un problème de valeurs propres pour une matrice hermitienne.

Une autre approche possible est basée sur la proposition suivante :

Proposition 4.3 *Soit $A \in \mathbb{C}^{m\times n}$ dont les valeurs singulières sont σ_i, $1 \leqslant i \leqslant r$. Les valeurs propres non nulles de la matrice* $\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \in \mathbb{C}^{(m+n)\times(m+n)}$ *sont $\pm\sigma_i$, $1 \leqslant i \leqslant r$.*

4.3 NOTES ET RÉFÉRENCES

La décomposition en valeurs singulières a été introduite par Eugenio Beltrami (1873) et indépendamment par Camille Jordan (1874) à propos de leurs études sur les formes quadratiques. On doit à Jordan la forme normale du même nom (théorème 1.5).

La décomposition en valeurs singulières est un outil important de l'algèbre linéaire. Elle joue un rôle essentiel en statistique (analyse en composantes principales), compression des données (approximation d'une matrice par une matrice de rang donné), traitement du signal, reconnaissance des formes, linguistique (analyse sémantique latente) et cetera.

EXERCICES

Exercice 4.1

Montrer que si $A \in \mathbb{C}^{m\times n}$ et si $B \in \mathbb{C}^{n\times m}$ alors les valeurs propres non nulles de AB et de BA sont les mêmes. Montrer que lorsque $m = n$ les valeurs propres de AB et de BA sont les mêmes.

Exercice 4.2

Calculer une décomposition en valeurs singulières de la matrice $A = \begin{pmatrix} 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{pmatrix}$.

Exercice 4.3

Calculer les valeurs singulières ainsi que toutes les décompositions en valeurs singulières d'une matrice colonne.

Exercice 4.4

Déterminer la décomposition en valeurs singulières d'une matrice hermitienne en fonction de ses éléments propres.

Exercice 4.5

Démontrer la proposition 4.3. On procèdera de la façon suivante :

1. Soit λ une valeur propre non nulle de la matrice augmentée :

$$\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

avec x ou y non nul. Alors x et y sont tous deux non nuls et $A^*Ay = \lambda^2 y$.

2. Réciproquement, si $A^*Ay = \lambda^2 y$ avec $y \neq 0$, montrer que $\pm\lambda$ sont valeurs propres de la matrice augmentée.

Exercice 4.6

Montrer que le rayon spectral d'une matrice est plus petit que la plus grande de ses valeurs singulières.

Exercice 4.7

Montrer que, pour tout $A \in \mathbb{GL}_n$, les valeurs propres de A^{-1} sont les inverses des valeurs propres de A et que les valeurs singulières de A^{-1} sont les inverses des valeurs singulières de A.

Exercice 4.8

Étant donnés n nombres complexes $z_1, \ldots, z_n$ calculer le polynôme caractéristique et les valeurs singulières de la matrice

$$Z = \begin{pmatrix} 1 & & & \\ z_1 & 1 & & \\ \vdots & & \ddots & \\ z_n & & & 1 \end{pmatrix}$$

(les entrées de Z sont nulles hors de la diagonale et de la première colonne).

Exercice 4.9

Donner la décomposition en valeurs singulières de la matrice

$$Z = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \\ \beta & 0 \\ 0 & \beta \end{pmatrix}.$$

Chapitre **5**

Le problème des erreurs

5.1 INTRODUCTION

Nous allons analyser à partir de quelques exemples modèles le problème des erreurs en analyse numérique. Mais au juste, pourquoi fait-on des erreurs ? Trois causes principales peuvent être envisagées :

5.1.1 Les erreurs de modélisation

Nous entendons par là le fait de remplacer un modèle du problème initial par un modèle simplifié. Un exemple classique est celui du pendule simple : les oscillations d'un tel pendule sont données, en l'absence d'amortissement, par l'équation différentielle du second ordre

$$\theta'' = -\frac{g}{l}\sin\theta$$

où g est l'accélération de la pesanteur, l la longueur du pendule et θ l'angle que fait le pendule avec la verticale. Sous l'hypothèse des « petites oscillations », on estime que $\sin\theta \approx \theta$ et l'équation devient

$$\theta'' = -\frac{g}{l}\theta$$

qui est une équation linéaire à coefficients constants.

De telles simplifications du modèle sont le pain quotidien du physicien et du mathématicien appliqué : nos moyens d'investigation ne permettent que rarement de considérer les problèmes naturels dans toute leur complexité.

5.1.2 Les erreurs de données

Il arrive que les paramètres du problème soient des données expérimentales obtenues avec une marge d'erreur ou bien des données issues d'un calcul approché. Nous ne traitons donc pas le « vrai » problème mais un problème voisin et la question se pose de savoir comment une telle erreur sur les données se répercute sur la solution : nous devons estimer la distance de la solution $\mathcal{S}(a)$ associée au paramètre a à la solution $\mathcal{S}(a')$ associée à un paramètre a' proche de a, c'est le problème de la *sensitivité* aux erreurs.

Il y a deux approches possibles à ce type d'étude : une *approche directe* qui est très utilisée en algèbre linéaire et une *analyse au premier ordre* fondée sur le calcul différentiel. Par exemple, l'erreur commise dans le calcul de la racine carrée d'un nombre réel positif a est, pour tout $h > 0$,

$$\sqrt{a+h} - \sqrt{a} = \frac{h}{\sqrt{a+h} + \sqrt{a}} \leqslant \frac{h}{2\sqrt{a}}.$$

C'est un exemple d'approche directe. Par le calcul différentiel on obtient :

$$\sqrt{a+h} - \sqrt{a} = \frac{h}{2\sqrt{a}} + O(h^2).$$

Dans ces deux cas, l'expression $h/2\sqrt{a}$ fait intervenir l'erreur h d'une part et un facteur multiplicateur indépendant de cette erreur : $1/2\sqrt{a}$. Ce facteur ne dépend que du problème (ici le calcul des racines carrées) et d'une instance de ce problème (le nombre a). Il conduit au concept de *conditionnement du problème*. Dans notre exemple on pose

$$\mathrm{cond}(\sqrt{\cdot}, a) = \frac{1}{2\sqrt{a}}$$

qui est un nombre indépendant de la perturbation h, de sorte que, au premier ordre

$$\left|\sqrt{a+h} - \sqrt{a}\right| \leqslant \mathrm{cond}(\sqrt{\cdot}, a)\,|h|\,.$$

5.1.3 Les erreurs de calcul

Une fois le problème posé, vient le moment d'introduire un algorithme pour en calculer la solution. Cet algorithme va être une source d'erreurs pour trois raisons principales :

1. Les processus limites sont arrêtés après un nombre fini d'étapes,
2. Les nombres irrationnels, les fonctions transcendantes sont remplacés par des approximations,
3. L'utilisation d'une arithmétique de précision finie (virgule flottante par exemple).

L'effet de ces erreurs est de remplacer la solution $\mathcal{S}(a)$ par une *solution approchée* $\tilde{\mathcal{S}}(a)$ qui dépend du problème, de l'instance a de ce problème et de l'algorithme utilisé.

Pour analyser ce type d'erreur, il est d'usage de procéder en deux étapes :

1. L'estimation de l'erreur $\tilde{\mathcal{S}}(a) - \mathcal{S}(a)$ proprement dite,
2. L'analyse rétrograde de cette erreur.

Qu'est-ce que cela signifie ? Il s'agit de déterminer le (ou un) paramètre $\tilde{a}$, le plus proche possible de a, pour lequel $\mathcal{S}(\tilde{a}) = \tilde{\mathcal{S}}(a)$, autrement dit, étudier le problème de minimisation

$$\inf_{\mathcal{S}(\tilde{a})=\tilde{\mathcal{S}}(a)} \|\tilde{a} - a\| .$$

La valeur de ce minimum est l'*erreur rétrograde* ou *erreur inverse* du problème : elle permet de valider l'algorithme choisi dans la mesure où cette erreur rétrograde est du même ordre que la précision des données.

Dans le cas des racines carrées, supposons que l'on ait calculé $\sqrt{a} + e$ au lieu de $\sqrt{a}$. Si e est suffisamment petit pour que $\sqrt{a} + e > 0$ on a

$$\sqrt{a} + e = \sqrt{\tilde{a}}$$

avec

$$\tilde{a} = a + e^2 + 2e\sqrt{a}.$$

L'erreur inverse est donc

$$\tilde{a} - a = e^2 + 2e\sqrt{a} \approx 2e\sqrt{a}$$

au premier ordre. Si la précision avec laquelle a est donné est du même ordre que $2e\sqrt{a}$ nous pouvons estimer que le calcul a été effectué avec une précision suffisante.

Nous voyons, sur cet exemple, que l'erreur inverse est le produit de l'erreur e par le coefficient multiplicateur $2\sqrt{a}$ indépendant de cette erreur que nous appelons *conditionnement inverse* du problème.

5.2 CONCEPTS GÉNÉRAUX

Considérons un problème que nous modélisons par une application

$$\mathcal{S} : \mathbb{E} \to \mathbb{F}.$$

$\mathbb{E}$ est l'espace des instances du problème, $\mathbb{F}$ est l'espace des solutions du problème et $\mathcal{S}$ est l'application « *solution* ». Voici quelques exemples de telles situations :

1. *Systèmes d'équations linéaires.* Soit $A \in \mathbb{C}^{n\times n}$ une matrice inversible. Il s'agit de résoudre l'équation $Ax = b$. Ce problème est décrit par la donnée de $b \in \mathbb{E} = \mathbb{C}^n$ et sa solution est $\mathcal{S}(b) = A^{-1}b \in \mathbb{F} = \mathbb{C}^n$.
2. *Racines carrées.* $\mathbb{E} =]0, \infty[$ l'ensemble des réels positifs, $\mathbb{F} =]0, \infty[$ et il faut calculer la racine carrée d'un nombre $a > 0$. L'application solution est $\mathcal{S}(a) = \sqrt{a}$.
3. *Le problème symétrique des valeurs propres.* $\mathbb{E} = \mathcal{S}_n(\mathbb{R})$ est l'espace des matrices $n \times n$ réelles et symétriques, $\mathbb{F} = \mathbb{R}^n \times \mathbb{R}$. Etant donnée $A \in \mathcal{S}_n(\mathbb{R})$ le problème considéré consiste à rechercher un couple $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ tel que $Ax = \lambda x$ et $\|x\|^2 = 1$. La définition de $\mathcal{S}$ n'est pas explicite : la solution (x, λ) est décrite par un système d'équations algébriques. On va donc utiliser le théorème des fonctions implicites.
4. *Equations polynomiales.* $\mathbb{E} = \mathcal{P}_d(\mathbb{C})$ est l'espace des polynômes complexes de degré $\leqslant d$ et $\mathbb{F} = \mathbb{C}$. Le problème posé est le calcul des racines d'un polynôme $f \in \mathcal{P}_d(\mathbb{C})$. La définition de l'application solution fait elle aussi appel au théorème des fonctions implicites.

Le premier problème à envisager est celui de la sensitivité : on veut savoir comment varie la solution $\mathcal{S}(a) \in \mathbb{F}$ lorsque l'on fait varier $a \in \mathbb{E}$. On suppose ici que $\mathcal{S}$ est de classe C^1. Dans ce contexte, pour deux entrées voisines a et $a' \in \mathbb{E}$, on a

$$\mathcal{S}(a') - \mathcal{S}(a) = D\mathcal{S}(a)(a' - a) + o(\|a' - a\|)$$

d'où, au premier ordre (c'est le sens du 1 en indice),

$$\|\mathcal{S}(a') - \mathcal{S}(a)\| \leqslant_1 \|D\mathcal{S}(a)\| \, \|a' - a\| .$$

Le nombre $\|D\mathcal{S}(a)\|$ (norme de l'opérateur linéaire $D\mathcal{S}(a)$) est appelé le *conditionnement du problème*. Il dépend du problème $(\mathcal{S})$, de l'instance considérée (a) mais pas de l'erreur sur les données $(a' - a)$.

Le second problème à envisager est celui du calcul approché de $\mathcal{S}(a)$. Nous supposons que a est connu de façon exacte mais que l'on calcule une quantité $\tilde{\mathcal{S}}(a)$ proche de $\mathcal{S}(a)$. L'analyse rétrograde des erreurs consiste à considérer la quantité calculée $\tilde{\mathcal{S}}(a)$ comme la solution exacte $\mathcal{S}(\tilde{a})$ d'un problème associé à une instance $\tilde{a}$ voisine de a. On cherche alors à estimer $\|\tilde{a} - a\|$ ce qui permet de savoir si la réponse $\tilde{\mathcal{S}}(a)$ est plausible compte tenu de la précision avec laquelle on connait a. Cette approche conduit à l'étude du problème d'optimisation

$$\min_{\mathcal{S}(\tilde{a})=\tilde{\mathcal{S}}(a)} \|\tilde{a} - a\|$$

dont la valeur est appelée *erreur inverse* ou *erreur rétrograde*.

Supposons, pour simplifier l'exposé, que la solution de ce problème de minimisation soit donnée par une application de classe C^1

$$\mathcal{R} : \mathbb{F} \to \mathbb{E}.$$

Par construction nous avons

$$\mathcal{R}(\mathcal{S}(a)) = a$$

pour tout a. L'erreur inverse du problème est donnée au premier ordre par

$$\min_{\mathcal{S}(\tilde{a})=\tilde{\mathcal{S}}(a)} \|\tilde{a} - a\| = \|\mathcal{R}(\tilde{\mathcal{S}}(a)) - \mathcal{R}(\mathcal{S}(a))\| =$$

$$\|D\mathcal{R}(\mathcal{S}(a))\left(\tilde{\mathcal{S}}(a) - \mathcal{S}(a)\right) + o\left(\|\tilde{\mathcal{S}}(a) - \mathcal{S}(a)\|\right)\| \leqslant_1 \|D\mathcal{R}(\mathcal{S}(a))\|\|\tilde{\mathcal{S}}(a) - \mathcal{S}(a)\|.$$

Le nombre $\|D\mathcal{R}(\mathcal{S}(a))\|$ est appelé le *conditionnement inverse du problème*. Il ne dépend lui aussi que du problème $(\mathcal{S})$ et de l'instance considérée (a) mais pas des erreurs de calcul $\tilde{\mathcal{S}}(a) - \mathcal{S}(a)$.

5.3 LE THÉORÈME DES FONCTIONS IMPLICITES

Bien souvent l'application solution associée à un problème donné n'est pas connue de façon explicite mais au travers d'une « équation définissante ». C'est le cas, par exemple, pour le problème des valeurs propres ou bien pour les racines d'un polynôme

$$f(z) = \sum_{k=0}^{d} a_k z^k = 0.$$

D'une façon générale on dispose d'une application de classe C^1

$$F : \mathbb{E} \times \mathbb{F} \to \mathbb{G}$$

et le problème est décrit par l'équation $F(x, y) = 0$. Par exemple, dans le cas des équations polynomiales, F est la fonction d'évaluation

$$F : \mathcal{P}_d(\mathbb{C}) \times \mathbb{C} \to \mathbb{C}, \quad F(f, z) = f(z),$$

et dans le cas du problème symétrique des valeurs propres

$$F : \mathcal{S}_n(\mathbb{R}) \times (\mathbb{R}^n \times \mathbb{R}) \to \mathbb{R}^n \times \mathbb{R}, \quad F(A, (x, \lambda)) = \begin{pmatrix} (\lambda I_n - A)x \\ \frac{1}{2}(\|x\|_2^2 - 1) \end{pmatrix}.$$

Le théorème des fonctions implicites est adapté à l'analyse de telles situations :

Théorème 5.1 *Soit $F : \mathbb{E} \times \mathbb{F} \to \mathbb{G}$ une application de classe C^1 où $\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ sont des espaces de Banach. Supposons que $F(a,x) = 0$ et que $D_2F(a,x) : \mathbb{F} \to \mathbb{G}$ soit un isomorphisme. Sous ces hypothèses, il existe un voisinage ouvert V_a de a dans $\mathbb{E}$ et une unique fonction S définie et de classe C^1 sur V_a, à valeurs dans un voisinage ouvert V_x de x dans $\mathbb{F}$ et telle que*

$$S(a) = x \quad et \quad F(a', S(a')) = 0$$

pour tout $a' \in V_a$. De plus,

$$DS(a) = -D_2F(a,x)^{-1} D_1F(a,x).$$

Nous allons illustrer ce théorème à l'aide de l'exemple suivant :

5.3.1 Équations polynomiales : conditionnement

Avant tout, introduisons une structure hermitienne sur $\mathcal{P}_d(\mathbb{C})$. Son produit hermitien est défini par

$$\langle f, g \rangle = \sum_{k=0}^{d} \binom{d}{k}^{-1} a_k \overline{b}_k$$

avec $f(z) = \sum_{k=0}^{d} a_k z^k$ et $g(z) = \sum_{k=0}^{d} b_k z^k$. Soit $x \in \mathbb{C}$ donné. Notons

$$p_x(z) = (1 + \overline{x}z)^d \in \mathcal{P}_d.$$

On vérifie facilement que

$$f(x) = \langle f, p_x \rangle$$

ce qui implique, par l'inégalité de Cauchy-Schwarz,

$$|f(x)| \leqslant \|f\| \, \|p_x\|.$$

En prenant $f = p_x$ on obtient

$$p_x(x) = \langle p_x, p_x \rangle = \|p_x\|^2 = (1 + |x|^2)^d$$

de sorte que

$$|f(x)| \leqslant \|f\| \, (1 + |x|^2)^{d/2}.$$

Calculons maintenant le conditionnement du calcul des racines d'un polynôme. Soit

$$F : \mathcal{P}_d(\mathbb{C}) \times \mathbb{C} \to \mathbb{C}, \quad F(f,z) = f(z).$$

Les dérivées partielles de F sont données par :

$$D_1F(f,z) : \mathcal{P}_d(\mathbb{C}) \to \mathbb{C}, \quad D_1F(f,z)\dot{f} = \dot{f}(z)$$

et

$$D_2F(f,z) = f'(z).$$

Donnons-nous f et x tels que $f(x) = 0$. Le théorème des fonctions implicites ne s'applique que si $D_2F(f,x)$ est un isomorphisme c'est-à-dire, dans ce contexte, si $f'(x) \neq 0$. Cela signifie que x est une racine simple de f. Sous cette hypothèse, il existe une application solution $\mathcal{S}$ définie dans un voisinage de f, à valeurs dans un voisinage de x, telle que $\mathcal{S}(f) = x$ et $g(\mathcal{S}(g)) = 0$ pour tout polynôme g dans ce voisinage. On a, au premier ordre, c'est-à-dire à un terme en $o(f-g)$ près,

$$\mathcal{S}(g) \approx \mathcal{S}(f) + D\mathcal{S}(f)(g-f)$$

c'est-à-dire

$$\mathcal{S}(g) \approx x - \frac{(g-f)(x)}{f'(x)}$$

de sorte que

$$|\mathcal{S}(g) - x| \approx \frac{|g(x)-f(x)|}{|f'(x)|} \leqslant \|f-g\| \frac{(1+|x|^2)^{d/2}}{|f'(x)|}.$$

Le conditionnement du problème associé à cette norme est donné par

$$\mathrm{cond}(f,x) = \frac{(1+|x|^2)^{d/2}}{|f'(x)|}.$$

Noter que ce conditionnement est d'autant plus « mauvais » (c'est-à-dire grand) que $f'(x)$ est petit c'est-à-dire lorsque x est « presque » une racine double.

5.3.2 Exemple numérique

Considérons le polynôme de Pochhammer : ses racines sont les entiers $1, 2, \ldots, 20$,

$$f(x) = (x-1)(x-2)\ldots(x-20).$$

La figure 5.1 montre les valeurs du conditionnement $\mathrm{cond}(f,x)$ pour les différentes racines. On remarque des valeurs importantes du conditionnement (valeurs de l'ordre de 10^{10}) à partir de la racine 12, avec un maximum pour la racine 16.

La figure 5.2 montre les racines dans le plan complexe du polynôme de Pochhammer dont les coefficients ont été perturbés par des valeurs aléatoires de distributions gaussiennes centrées et d'écarts-type 10^{-5}. Malgré les valeurs faibles de ces perturbations, on observe une modification importante des racines proches de 16 dont le conditionnement est élevé.

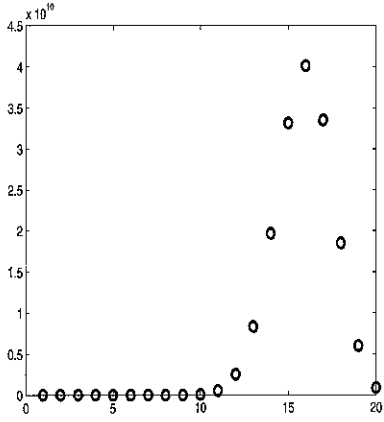


Figure 5.1 Valeur du conditionnement pour les racines du polynôme de Pochhammer. Les valeurs des racines sont portées en abscisse.

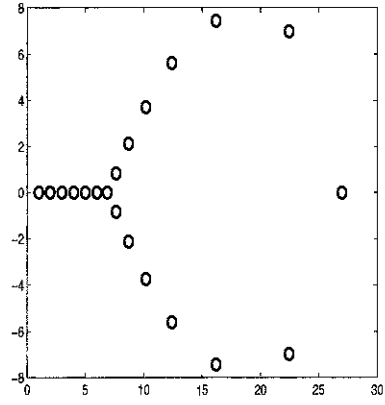


Figure 5.2 Racines du polynôme de Pochhammer perturbé. Les racines sont représentées dans le plan complexe.

5.3.3 Équations polynomiales : erreurs inverses

Supposons que $f(x) = 0$ et que l'on ait calculé une approximation x' de x. Quel est le polynôme $g \in \mathcal{P}_d$ qui vérifie $g(x') = 0$ et qui minimise la quantité $\|f - g\|$? Notons

$$\mathcal{H}_{x'} = \{g \in \mathcal{P}_d \ : \ g(x') = 0\}.$$

C'est un sous-espace vectoriel de $\mathcal{P}_d$ et, en vertu de l'égalité $h(x') = \langle h, p_{x'} \rangle$, c'est le sous-espace orthogonal à $p_{x'}$. On a

$$g \in \mathcal{H}_{x'} \ \text{ et } \ \|f - g\| = \min_{h \in \mathcal{H}_{x'}} \|f - h\|$$

lorsque g est la projection orthogonale de f sur $\mathcal{H}_{x'}$. La décomposition orthogonale de f est alors

$$f(z) = \lambda p_{x'}(z) + g(z) = \frac{f(x')}{(1 + |x'|^2)^d}(1 + \overline{x}'z)^d + g(z).$$

Cela résulte de l'égalité

$$f(x') = \lambda p_{x'}(x') + g(x') = \lambda p_{x'}(x') = \lambda(1 + |x'|^2)^d.$$

On en déduit que

$$\min_{h \in \mathcal{H}_{x'}} \|f - h\| = \|f - g\| = \left\| \frac{f(x')}{(1 + |x'|^2)^d}(1 + \overline{x}'z)^d \right\| = \frac{|f(x')|}{(1 + |x'|^2)^{d/2}}$$

qui est l'erreur inverse du problème.

L'application $\mathcal{R} : \mathbb{C} \to \mathcal{P}_d(\mathbb{C})$ qui à x' associe la solution optimale g est égale à

$$\mathcal{R}(x')(z) = f(z) - \frac{f(x')}{(1+|x'|^2)^d}(1+\overline{x}'z)^d.$$

Puisque $f(x) = 0$, la dérivée de $\mathcal{R}$ en x est donnée par le polynôme en z

$$D\mathcal{R}(x)(z) = \lim_{x' \to x} \frac{\mathcal{R}(x')(z) - \mathcal{R}(x)(z)}{x' - x} = -\frac{f'(x)}{(1+|x|^2)^d}(1+\overline{x}z)^d$$

dont la norme (c'est-à-dire le conditionnement inverse en (f, x)) vaut

$$\|D\mathcal{R}(x)\| = \frac{|f'(x)|}{(1+|x|^2)^{d/2}}.$$

5.4 LE CAS DES SYSTÈMES LINÉAIRES : CONDITIONNEMENT D'UNE MATRICE

Soit $A \in \mathbb{C}^{n\times n}$ inversible. Nous supposons que $\mathbb{C}^n$ est équipé d'une norme quelconque et $\mathbb{C}^{n\times n}$ d'une norme multiplicative et consistante avec la précédente.

À deux données voisines b et $b' \in \mathbb{C}^n$ correspondent deux solutions : $Ax = b$ et $Ax' = b'$ de ce système. L'erreur commise sur la solution est

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

de sorte que

$$\|x' - x\| \leqslant \|A^{-1}\|\|b' - b\|.$$

Ce résultat signifie que l'erreur absolue faite sur x est bornée par celle faite sur la donnée b multipliée par le facteur d'amplification $\|A^{-1}\|$. Ce nombre est appelé *conditionnement absolu* du système $Ax = b$. Comme

$$\|b\| \leqslant \|A\|\|x\|$$

on a aussi

$$\frac{\|x' - x\|}{\|x\|} \leqslant \|A\|\|A^{-1}\|\frac{\|b' - b\|}{\|b\|},$$

l'erreur relative faite sur x est bornée par l'erreur relative faite sur b multipliée par le facteur d'amplification $\|A\|\|A^{-1}\|$.

Définition 5.2 *On appelle conditionnement d'une matrice $A \in \mathbb{GL}_n$ le nombre* $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$. *Ce conditionnement dépend de la norme considérée sur*

$\mathbb{C}^{n\times n}$. *On note* cond_2, cond_1, cond_∞ *et* cond_F *le conditionnement associé à la norme spectrale, aux normes* $\|.\|_1$ *et* $\|.\|_\infty$ *décrites à l'exemple 3.1 et à la norme de Frobenius.*

Supposons maintenant que l'on commette à la fois une erreur sur le second membre b mais aussi sur la matrice A d'un système, quelle erreur commet-on sur sa solution ? Une réponse est donnée par le théorème suivant :

Théorème 5.3 *Etant donné des matrices* A *et* $E \in \mathbb{C}^{n\times n}$, A *inversible, avec* $\left\|A^{-1}\right\| \|E\| < 1$, *soient* b *et* $b' \in \mathbb{C}^n$ *et soient* x *et* $x' \in \mathbb{C}^n$ *tels que*

$$Ax = b \text{ et } (A+E)x' = b'.$$

Sous ces hypothèses

$$\frac{\|x'-x\|}{\|x\|} \leqslant \frac{1}{1-\|A^{-1}\|\,\|E\|}\left(\frac{\|b'-b\|}{\|b\|}\text{cond}(A) + \left\|A^{-1}\right\|\|E\|\right).$$

Démonstration. Notons que, par le corollaire 3.15, la matrice $A+E$ est inversible de sorte que x' existe bel et bien. On écrit que

$$x'-x = (I_n + A^{-1}E)^{-1}\left(A^{-1}(b'-b) - A^{-1}Ex\right)$$

puis on utilise la proposition 3.14 et l'inégalité $\|b\| \leqslant \|A\|\,\|x\|$.

Remarque 5.1. L'approche « calcul différentiel » donne le résultat suivant. Posons $\mathcal{S}(A,b) = A^{-1}b$. On a au premier ordre $x' - x \approx$

$$D\mathcal{S}(A,b)(E,b'-b) = -A^{-1}EA^{-1}b + A^{-1}(b'-b) = -A^{-1}Ex + A^{-1}(b'-b)$$

de sorte que

$$\frac{\|x'-x\|}{\|x\|} \leqslant_1 \frac{\|b'-b\|}{\|b\|}\text{cond}(A) + \left\|A^{-1}\right\|\|E\|$$

expression égale au premier ordre à celle du théorème 5.3. C'est rassurant !

Le conditionnement associé à la norme spectrale s'exprime bien à l'aide des valeurs singulières de la matrice :

Théorème 5.4 *Pour toute matrice* $A \in \mathbb{GL}_n$ *on a :*

1. $\text{cond}_2(A) = \text{cond}_2(A^{-1})$,
2. $\text{cond}_2(A) \geqslant 1$,
3. *Quelles que soient les matrices unitaires* U *et* $V \in \mathbb{U}_n$, $\text{cond}_2(UAV) = \text{cond}_2(A)$,

4. *Si* $\sigma_1 \geqslant \ldots \geqslant \sigma_n > 0$ *sont les valeurs singulières de* A *alors* $\mathrm{cond}_2(A) = \sigma_1/\sigma_n$.

Démonstration. 1 est évident. 2 est une conséquence de 4 ; 3 aussi parce que les valeurs singulières de A et celles de UAV sont les mêmes. Pour prouver 4 on utilise le théorème 3.9 qui donne $\|A\|_2 = \sigma_1$ et $\|A^{-1}\|_2 = \sigma_n^{-1}$ (voir l'exercice 4.7.)

Définition 5.5 *On dit qu'une matrice est mal conditionnée lorsque son conditionnement est grand, bien conditionnée lorsque son conditionnement est petit.*

Le conditionnement d'une matrice a une interprétation géométrique que nous allons décrire :

Théorème 5.6 (Eckart-Young, 1936) *Notons* Σ_n *l'ensemble des matrices* $n \times n$ *non-inversibles. On a :*

$$\mathrm{cond}_2(A) = \frac{\|A\|_2}{d_F(A, \Sigma_n)}$$

où $d_F(A, \Sigma_n) = \min_{B \in \Sigma_n} \|A - B\|_F$.

Démonstration. L'inégalité $d_F(A, \Sigma_n)^{-1} \leqslant \|A^{-1}\|_2$ résulte du fait suivant : si $\|S\|_2 < 1$ alors $I_n - S$ est inversible (proposition 3.14). Comme pour tout $B \in \Sigma_n$ on a aussi $A^{-1}B \in \Sigma_n$ on obtient

$$1 \leqslant \|I_n - A^{-1}B\|_2 = \|A^{-1}(A - B)\|_2 \leqslant$$

$$\|A^{-1}\|_2 \|A - B\|_2 \leqslant \|A^{-1}\|_2 \|A - B\|_F.$$

Pour prouver l'inégalité inverse considérons $x \in \mathbb{C}^n$ tel que $\|x\|_2 = 1$ et $\|A^{-1}\|_2 = \|A^{-1}x\|_2$. Un tel x existe puisque $\|A^{-1}\|_2 = \max_{\|x\|_2=1} \|A^{-1}x\|_2$. Posons $z = A^{-1}x/\|A^{-1}x\|_2^2$ et $B = A - xz^*$. La matrice B est singulière parce que

$$B(A^{-1}x) = AA^{-1}x - xz^*A^{-1}x = x - x\frac{(A^{-1}x)^*}{\|A^{-1}x\|_2^2}A^{-1}x = 0$$

alors que $A^{-1}x \neq 0$. On en déduit que

$$d_F(A, \Sigma_n) \leqslant \|A - B\|_F = \|xz^*\|_F = \|x\|_2\|z\|_2 = \frac{1}{\|A^{-1}\|_2}.$$

Le résultat précédent signifie que les matrices mal conditionnées sont celles qui sont proches des matrices singulières.

5.5 LE CAS DES SYSTÈMES LINÉAIRES : ERREURS INVERSES

Donnons-nous une matrice $A \in \mathbb{GL}_n$, un vecteur $b \in \mathbb{C}^n$, la solution $x = A^{-1}b$ du système $Ax = b$ et une approximation x' de x. Quelle est la plus petite matrice $E \in \mathbb{C}^{n\times n}$ telle que $(A + E)x' = b$. Plus petite est ici à prendre au sens d'une norme. Cette question est traitée dans le théorème suivant :

Théorème 5.7 (Rigal-Gaches, 1967) *Pour tout $x' \in \mathbb{C}^n$ non nul on a*

$$\min_{\substack{E \in \mathbb{C}^{n\times n} \\ (A+E)x' = b}} \|E\|_2 = \frac{\|A(x' - x)\|_2}{\|x'\|_2}.$$

Le minimum est atteint pour

$$E = \frac{A(x - x')x'^*}{\|x'\|_2^2}.$$

Démonstration. Si $(A + E)x' = b$ et $Ax = b$ alors $A(x' - x) = -Ex'$ et donc $\|A(x' - x)\|_2 \leqslant \|E\|_2\|x'\|_2$. Ceci prouve que

$$\frac{\|A(x' - x)\|_2}{\|x'\|_2} \leqslant \|E\|_2$$

et il y a égalité lorsque

$$E = \frac{A(x - x')x'^*}{\|x'\|_2^2}$$

(voir l'exercice 3.12).

Ce théorème prouve que l'erreur inverse commise est égale à

$$\|E\|_2 = \frac{\|A(x' - x)\|_2}{\|x'\|_2} \leqslant \|A\|_2 \frac{\|x' - x\|_2}{\|x'\|_2}.$$

5.6 PRÉCONDITIONNEMENT D'UN SYSTÈME LINÉAIRE

L'objectif du préconditionnement d'un système linéaire est de diminuer la valeur du conditionnement de la matrice du système. Pour cela on remplace le système par un système équivalent. On distingue trois types de préconditionnements :

- Le préconditionnement à gauche consiste à remplacer le système $Ax = b$ par le système $CAx = Cb$ où la matrice C est inversible.

- Le préconditionnement à droite est obtenu en considérant le système $ADy = b$ où D est inversible. La solution du système initial est alors donnée par $x = Dy$.

- Le préconditionnement à gauche et à droite qui combine les deux précédents : $CADy = Cb$.

Les matrices C et D sont appelées matrices de préconditionnement. Dans tous les cas, on cherche à diminuer la valeur du conditionnement :

$$\text{cond}(CA), \text{cond}(AD) \text{ et } \text{cond}(CAD) \leqslant \text{cond}(A).$$

On a bien sûr un choix optimal de préconditionnement en prenant $C = A^{-1}$ dans le cas du préconditionnement à gauche (ou à droite), ce qui veut dire qu'on a résolu le problème ! Bien entendu ce choix ne présente aucun intérêt. Dans la pratique, on recherche des matrices C qui permettent d'une part d'obtenir des valeurs faibles de $\text{cond}(CA)$ et d'autre part de ne pas augmenter de manière significative la complexité du calcul de la solution du système.

Un exemple classique de préconditionnement qui permet dans certains cas de diminuer sensiblement le conditionnement du système est obtenu en normalisant les lignes ou les colonnes du système. Dans le cas des colonnes, cela revient à considérer le préconditionnement à droite avec la matrice diagonale $D = \text{diag}(1/\|c_1\|_2, \ldots, 1/\|c_n\|_2)$ où c_i sont les vecteurs-colonne de la matrice A.

Voici un exemple de préconditionnement à droite obtenu en normalisant les colonnes de la matrice de Vandermonde (paragraphe 16.4) $A \in \mathbb{R}^{12\times 12}$ basée sur les points $x_i = 10(i+1), i = 0, \ldots, 11,$. On obtient les valeurs $\text{cond}_2(A) = 1.3765\,10^{26}$ et $\text{cond}_2(AD) = 2.4457\,10^9$.

Nous verrons au paragraphe 10.6 que le préconditionnement est également utilisé dans les méthodes itératives car il permet d'accélérer la convergence des suites.

5.7 NOTES ET RÉFÉRENCES

Deux articles fondent véritablement le sujet abordé dans ce chapitre dans l'immédiat après-guerre. Ce sont : « Numerical Inverting of Matrices of High Order (1947) » [13] par H. Goldstine (1913-2004) et J. Von Neumann (1903-1957) et « Rounding-off errors in matrix processes (1948) » [35] par A. Turing (1912-1954). Mais c'est à J. Wilkinson (1919-1986), qui fut assistant de Turing, que l'on doit d'avoir approfondi cette question dans les deux ouvrages : *Rounding Errors in Algebraic Processes* (1963) [36] et *The Algebraic Eigenvalue Problem* (1965) [37]. Nous recommandons la lecture du livre de Stewart-Sun *Matrix Perturbation Theory* [34] qui est désormais un classique ! Parmi les ouvrages récents citons ceux de F. Chaitin-Chatelin et V. Frayssé *Lectures on Finite*

Precision Computations (1996) [8], N. Higham *Accuracy and Stability of Numerical Algorithms* (2002) [18] qui traite en détail les problèmes issus de l'algèbre linéaire et enfin *Complexity and Real Computation* (1997) [6] par L. Blum, F. Cucker, M. Shub et S. Smale plus versé vers les problèmes polynomiaux.

EXERCICES

Exercice 5.1

Calculer la solution des systèmes suivants $AX = B_1$ et $AX = B_2$ où :

$$A = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix}, \; B_1 = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}, \; B_2 = \begin{pmatrix} 0.216999 \\ 0.254 \end{pmatrix}$$

et calculer $\text{cond}_2(A)$.

Exercice 5.2

Montrer que pour toutes matrices A et $B \in \mathbb{C}^{n\times n}$, A inversible et $B \neq A^{-1}$, on a

$$\frac{||AB - I_n||_2}{||BA - I_n||_2} \leqslant \text{cond}_2(A).$$

Exercice 5.3

Soit $A \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie positive. On va montrer que, pour tout $x \in \mathbb{C}^n$, $x \neq 0$,

$$1 \leqslant \frac{< Ax, x >< A^{-1}x, x >}{< x, x >^2} \leqslant \frac{\left((\text{cond}_2 A)^{1/2} + (\text{cond}_2 A)^{-1/2}\right)^2}{4}$$

(inégalité de Kantorovitch).

1. Notons $\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_n > 0$ les valeurs propres de A. Montrer que l'on peut écrire
$$\frac{< Ax, x >< A^{-1}x, x >}{< x, x >^2} = \left(\sum_{i=1}^{n} \alpha_i \lambda_i\right)\left(\sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i}\right)$$
pour des $\alpha_i \geqslant 0$ tels que $\sum \alpha_i = 1$. Indication : diagonaliser A en base orthonormée.
2. Montrer que
$$\frac{1}{\sum \alpha_i \lambda_i} \leqslant \sum \frac{\alpha_i}{\lambda_i} \leqslant \frac{\lambda_1 + \lambda_n - \sum \alpha_i \lambda_i}{\lambda_1 \lambda_n}.$$
Indication : utiliser un argument de convexité.
3. Calculer
$$\max_{\lambda_n \leqslant \lambda \leqslant \lambda_1} \lambda \frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n}$$
et conclure.

Exercice 5.4

Soit

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

Donner un minorant du conditionnement de A à l'aide de l'inégalité de Kantorovitch.

On prendra $x = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Exercice 5.5

Calculer $\text{cond}_2 A(a, b, c)$ avec $a \in \mathbb{R}$ et $c = \overline{b}$, et où $A(a, b, c)$ est la matrice décrite à l'exercice 1.13.

Chapitre 6

Pivot de Gauss et décomposition LU

6.1 RÉSOLUTION DES SYSTÈMES TRIANGULAIRES

Soient $A \in \mathbb{C}^{n\times n}$ triangulaire supérieure et $b \in \mathbb{C}^n$. On suppose que la matrice A est inversible c'est-à-dire que $a_{ii} \neq 0$ pour tout i. Le système $Ax = b$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n-1} & a_{1n} \\ 0 & a_{22} & \dots & a_{2n-1} & a_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & a_{n-1n-1} & a_{n-1n} \\ 0 & 0 & \dots & 0 & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}$$

se résout de la façon suivante (la notation $i = n-1 : -1 : 1$ indique que l'indice i décroît de $n-1$ à 1 avec un pas égal à -1) :

Algorithme de résolution d'un système triangulaire

$x_n = b_n/a_{nn}$
pour $i = n-1 : -1 : 1$
    $x_i = b_i$
    pour $k = i+1 : n$
        $x_i = x_i - a_{ik}x_k$
    fin
    $x_i = x_i/a_{ii}$
fin

Cette méthode requiert n^2 opérations arithmétiques. Noter que les erreurs commises à une étape du calcul se propagent aux étapes suivantes.

6.2 L'ÉLIMINATION DE GAUSS

Le procédé d'élimination de Gauss pour résoudre un système linéaire $n \times n$ consiste à utiliser la première équation pour exprimer la première inconnue x_1 en fonction des autres $x_2, \ldots, x_n$ puis à reporter la valeur ainsi trouvée dans les équations suivantes. On obtient alors un système $n-1 \times n-1$ en les inconnues $x_2, \ldots, x_n$ auquel on applique la même méthode. Ce procédé permet d'obtenir, après $n-1$ telles étapes, un nouveau système qui est triangulaire et équivalent au premier : tous deux ont les mêmes inconnues et leurs solutions respectives sont les mêmes.

Etudions un exemple :

$$\begin{array}{lcrcrcrcl} E_1 & : & 3x_1 & + & 2x_2 & + & x_3 & = & 1. \\ E_2 & : & x_1 & + & 3x_2 & + & 2x_3 & = & 2, \\ E_3 & : & 2x_1 & + & 4x_2 & + & 6x_3 & = & 3. \end{array}$$

Calculons x_1 en fonction de x_2 et x_3 en utilisant la première ligne et reportons dans E_2 et E_3. Ceci revient à remplacer E_2 par $E_2' = E_2 - E_1/3$ et E_3 par $E_3' = E_3 - 2E_1/3$. On obtient

$$\begin{array}{lcrcrcrcl} E_1 & : & 3x_1 & + & 2x_2 & + & x_3 & = & 1, \\ E_2' & : & & & \frac{7}{3}x_2 & + & \frac{5}{3}x_3 & = & \frac{5}{3}, \\ E_3' & : & & & \frac{8}{3}x_2 & + & \frac{16}{3}x_3 & = & \frac{7}{3}. \end{array}$$

Dans l'étape suivante, on calcule x_2 en fonction de x_3 en utilisant la seconde ligne et on le reporte dans E_3'. On remplace E_3' par $E_3'' = E_3' - 8E_2'/7$ ce qui conduit à

$$\begin{array}{lcrcrcrcl} E_1 & : & 3x_1 & + & 2x_2 & + & x_3 & = & 1, \\ E_2' & : & & & \frac{7}{3}x_2 & + & \frac{5}{3}x_3 & = & \frac{5}{3}, \\ E_3'' & : & & & & & \frac{24}{7}x_3 & = & \frac{3}{7}. \end{array}$$

Ce dernier système est triangulaire et sa solution est $x_3 = 1/8$, $x_2 = 5/8$. $x_1 = -1/8$.

Plus généralement, pour une matrice $A \in \mathbb{C}^{n \times n}$ et un second membre $b \in \mathbb{C}^n$, l'algorithme suivant retourne, lorsque les divisions par zéro n'apparaissent pas, un système triangulaire équivalent à $Ax = b$:

Algorithme d'élimination de Gauss

pour $i = 1 : n - 1$
  pour $j = i + 1 : n$
    $a_{ji} = 0$
    $b_j = b_j - \frac{a_{ji}}{a_{ii}} b_i$
    pour $k = i + 1 : n$
      $a_{jk} = a_{jk} - \frac{a_{ji}}{a_{ii}} a_{ik}$
    fin
  fin
fin

6.3 DÉCOMPOSITION LU

Le calcul de base du procédé d'élimination de Gauss est l'addition à une ligne de A d'une autre ligne multipliée par un scalaire. Cette opération peut se décrire comme un produit matriciel.

Définition 6.1 *Étant donnés deux entiers distincts* $1 \leqslant i, j \leqslant n$ *et un scalaire* $\lambda \in \mathbb{C}$, *la matrice élémentaire* $E(i, j, \lambda) \in \mathbb{C}^{n \times n}$ *a pour coefficients* $e_{kk} = 1$ *pour tout* $k = 1 \ldots n$, $e_{ij} = \lambda$ *et* $e_{kl} = 0$ *pour les autres entrées de la matrice.*

$$E(i,j,\lambda) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & \ddots & & & \\ & & \lambda & & 1 & & i \\ & & \vdots & & & \ddots & \\ & & 0 & & & & 1 \end{pmatrix}$$
(avec j au-dessus de la colonne de λ)

Les propriétés attendues des matrices élémentaires sont données par la proposition suivante dont la démonstration est laissée à titre d'exercice :

Proposition 6.2

1. *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ dont les lignes sont notées L_k, $1 \leqslant k \leqslant n$, les lignes de la matrice $E(i,j,\lambda)A$ sont L_k si $k \neq i$, $L_i + \lambda L_j$ si $k = i$,*
2. *Notons C_k, $1 \leqslant k \leqslant n$, les colonnes de A. Les colonnes de la matrice $AE(i,j,\lambda)$ sont C_k si $k \neq j$, $C_j + \lambda C_i$ si $k = j$,*
3. *Lorsque $i > j$ la matrice $E(i,j,\lambda)$ est triangulaire inférieure,*
4. *$E(i,j,\lambda)$ est inversible, son déterminant est égal à 1 et $E(i,j,\lambda)^{-1} = E(i,j,-\lambda)$.*

Algorithme d'élimination de Gauss (matrices élémentaires)

pour $i = 1 : n-1$
  pour $j = i+1 : n$
    $A = E(j,i,-\frac{a_{ji}}{a_{ii}})A$
    $b = E(j,i,-\frac{a_{ji}}{a_{ii}})b$
  fin
fin

L'effet de la boucle j est de multiplier à gauche la matrice courante A par le produit de matrices élémentaires

$$E(n,i,-\frac{a_{ni}}{a_{ii}})\dots E(i+1,i,-\frac{a_{i+1\,i}}{a_{ii}}).$$

Un tel produit de matrices élémentaires est égal à la matrice

$$E(n,i,\lambda_n)\dots E(i+1,i,\lambda_{i+1}) = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \lambda_{i+1} & \ddots & \\ & & \vdots & & \ddots \\ & & \lambda_n & & 1 \end{pmatrix}.$$

Définition 6.3 *La matrice précédente est appelée matrice d'élimination. Elle se note*

$$E(n,i,\lambda_n)\dots E(i+1,i,\lambda_{i+1}) = E(i,\lambda_{i+1},\dots,\lambda_n).$$

Ces matrices ont les propriétés suivantes données sans démonstration :

Proposition 6.4

1. $E(i, \lambda_{i+1}, \ldots, \lambda_n)$ *est triangulaire inférieure à diagonale unité,*
2. $\det E(i, \lambda_{i+1}, \ldots, \lambda_n) = 1$,
3. $E(i, \lambda_{i+1}, \ldots, \lambda_n)$ *est inversible et* $E(i, \lambda_{i+1}, \ldots, \lambda_n)^{-1} = E(i, -\lambda_{i+1}, \ldots, -\lambda_n)$.

L'algorithme d'élimination de Gauss s'écrit désormais :

Algorithme d'élimination de Gauss (matrices d'élimination)

pour $i = 1 : n-1$
$\quad A = E(i, -\frac{a_{i+1\,i}}{a_{ii}}, \ldots, -\frac{a_{ni}}{a_{ii}})A$
$\quad b = E(i, -\frac{a_{i+1\,i}}{a_{ii}}, \ldots, -\frac{a_{ni}}{a_{ii}})b$
fin

Il fournit une matrice triangulaire supérieure U en multipliant A à gauche par $n-1$ matrices d'élimination. Leur produit est une matrice triangulaire inférieure à diagonale unité que nous notons L^{-1}. On a donc décomposé $A = LU$ avec U triangulaire supérieure et L triangulaire inférieure à diagonale unité.

Définition 6.5 *On appelle décomposition LU d'une matrice* $A \in \mathbb{C}^{n\times n}$ *toute identité* $A = LU$ *avec* U *triangulaire supérieure et* L *triangulaire inférieure à diagonale unité.*

Pour obtenir cette décomposition nous avons effectué un certain nombre d'opérations du type a_{ji}/a_{ii}. Nous devons nous assurer que nous ne divisons pas par 0 ! L'algorithme serait alors en défaut comme par exemple pour la matrice

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

Il « se bloque » juste après la première étape puisque l'on obtient la matrice

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & -2 \\ 0 & -1 & -1 \end{pmatrix}$$

dont le coefficient a_{22} est nul. De façon plus précise :

Théorème 6.6 *Pour toute matrice* $A \in \mathbb{GL}_n$

1. *Une décomposition LU de A existe si et seulement si* $\det A(1:k, 1:k) \neq 0$ *pour tout* $k = 1 \ldots n$.

2. *Lorsqu'elle existe, la décomposition LU est unique.*

Démonstration. Prouvons la première assertion. Si $A = LU$ il est facile de voir que $A(1:k,1:k) = L(1:k,1:k)U(1:k,1:k)$ de sorte que $\det A(1:k,1:k) = u_{11}\dots u_{kk}$. Puisque $A \in \mathbb{GL}_n$ on a $\det A = u_{11}\dots u_{nn} \neq 0$ et donc $\det A(1:k,1:k) \neq 0$. Ceci prouve que la condition $\det A(1:k,1:k) \neq 0$ est nécessaire.

Pour voir qu'elle est suffisante nous allons raisonner par récurrence. La première étape d'élimination peut être effectuée parce que $a_{11} = \det A(1:1,1:1) \neq 0$. Supposons avoir réalisé $k-1$ étapes avec succès ce qui nous a conduit à la décomposition

$$A = L \begin{pmatrix} u_{11} & \dots & u_{1k} & u_{1\,k+1} & \dots & u_{1n} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & u_{kk} & u_{k\,k+1} & \dots & u_{kn} \\ 0 & \dots & 0 & v_{k+1\,k+1} & \dots & v_{k+1\,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & v_{n\,k+1} & \dots & v_{nn} \end{pmatrix}$$

où L est une matrice triangulaire inférieure à diagonale unité. Pour poursuivre l'algorithme, il faut être assuré que $v_{k+1\,k+1} \neq 0$. Notons que

$$A(1:k+1,1:k+1) = L(1:k+1,1:k+1) \begin{pmatrix} u_{11} & \dots & u_{1k} & u_{1\,k+1} \\ \vdots & \ddots & \vdots & \\ 0 & \dots & u_{kk} & u_{k\,k+1} \\ 0 & \dots & 0 & v_{k+1\,k+1} \end{pmatrix}$$

de sorte que $\det A(1:k+1,1:k+1) = u_{11}\dots u_{kk}v_{k+1\,k+1}$. Comme $\det A(1:k+1,1:k+1) \neq 0$ il en est de même pour $v_{k+1\,k+1}$.

Reste à prouver que la décomposition LU est unique. Si $A = L_1U_1 = L_2U_2$, puisque A est inversible les matrices L_i et U_i le sont aussi et $L_2^{-1}L_1 = U_2U_1^{-1}$. On reconnaît à gauche une matrice triangulaire inférieure à diagonale unité et à droite une matrice triangulaire supérieure ; la seule matrice ayant ces deux vertus est I_n d'où $L_1 = L_2$ et $U_1 = U_2$.

6.4 PIVOT PARTIEL, PIVOT TOTAL

La première étape de la méthode d'élimination de Gauss fait jouer un rôle particulier au coefficient a_{11} de la matrice A. On l'appelle *pivot de la méthode* et, pour cette raison, on parle souvent de la *méthode du pivot de Gauss*. Le choix de a_{11} comme

pivot n'est pas le seul possible : on obtient un système équivalent en permutant entre-elles les différentes équations ou bien en prenant les inconnues dans un ordre différent et, à ces nouveaux systèmes, on peut aussi appliquer la méthode du pivot de Gauss. Ces différentes stratégies ne sont pas sans influence du point de vue du calcul des erreurs.

6.4.1 Etude d'un exemple

Considérons le système suivant :

$$\begin{cases} 10^{-3}x + y &= b_1 \\ x + 10y &= b_2 \end{cases}$$

Il est équivalent au système

$$\begin{cases} x + 10y &= b_2 \\ 10^{-3}x + y &= b_1 \end{cases}$$

obtenu en permutant les deux équations ainsi qu'au système

$$\begin{cases} 10y + x &= b_2 \\ y + 10^{-3}x &= b_1 \end{cases}$$

où l'on a permuté l'ordre des inconnues.

Notons A la matrice du premier système et $U(A)$ la matrice triangulaire supérieure obtenue par la méthode du pivot de Gauss. On a :

$$A = \begin{pmatrix} 10^{-3} & 1 \\ 1 & 10 \end{pmatrix}, \quad U(A) = \begin{pmatrix} 10^{-3} & 1 \\ 0 & -990 \end{pmatrix}.$$

Notons B la matrice du second système et $U(B)$ la matrice triangulaire supérieure obtenue par la méthode du pivot de Gauss. On a :

$$B = \begin{pmatrix} 1 & 10 \\ 10^{-3} & 1 \end{pmatrix}, \quad U(B) = \begin{pmatrix} 1 & 10 \\ 0 & 0.99 \end{pmatrix}.$$

Le troisième système conduit de même aux matrices

$$C = \begin{pmatrix} 10 & 1 \\ 1 & 10^{-3} \end{pmatrix}, \quad U(C) = \begin{pmatrix} 10 & 1 \\ 0 & -0.099 \end{pmatrix}.$$

Quoique tous ces systèmes soient équivalents ils ont des comportements très différents quant au conditionnement. En effet :

$$\begin{aligned} \text{cond}_2(A) = \text{cond}_2(B) = \text{cond}_2(C) &= 103.0205972, \\ \text{cond}_2(U(A)) &= 990001.0100, \\ \text{cond}_2(U(B)) &= 103.0004933, \\ \text{cond}_2(U(C)) &= 102.0203000. \end{aligned}$$

Dans le passage de A à $U(A)$ on constate que le conditionnement de A a été détruit, restauré avec $U(B)$ et amélioré avec $U(C)$.

6.4.2 Matrices de transposition

Définition 6.7 *Pour deux entiers $1 \leqslant i, j \leqslant n$ distincts on note $P(i,j) \in \mathbb{C}^{n\times n}$ la matrice dont les entrées sont $p_{kk} = 1$ lorsque $k \neq i$ et $k \neq j$, $p_{ij} = p_{ji} = 1$, $p_{kl} = 0$ dans les autres cas. Cette matrice est appelée matrice de transposition.*

$$P(i,j) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & 1 & & i \\ & & & \ddots & & & \\ & & 1 & & 0 & & j \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

Ces matrices servent à permuter deux lignes ou deux colonnes dans une matrice donnée. De façon plus précise, on a :

Proposition 6.8

1. *$P(i,j)$ est orthogonale et symétrique,*
2. *$P(i,j)A$ est la matrice dont les lignes sont celles de A, les lignes i et j étant permutées, les autres inchangées,*
3. *$AP(i,j)$ est la matrice dont les colonnes sont celles de A, les colonnes i et j étant permutées, les autres inchangées.*

Définition 6.9 *On appelle matrice de permutation un produit de matrices de transposition.*

Si P est une telle matrice, PA (resp. AP) est une matrice dont les lignes (resp. des colonnes) sont celles de A prises dans un ordre différent. Noter que toutes les permutations possibles des lignes (resp. des colonnes) peuvent être ainsi obtenues parce que toute permutation est un produit de transpositions.

6.4.3 Pivot partiel

La méthode du *pivot partiel* consiste à choisir pour pivot le coefficient de plus grand module de la première colonne de A :

$$|a_{i1}| = \max_{1 \leqslant k \leqslant n} |a_{k1}| .$$

Notons que $a_{i1} \neq 0$ puisque l'on a supposé que A est inversible. On pourra donc diviser par a_{i1}. On obtient un nouveau système en permutant dans A les lignes 1 et i et en laissant les autres inchangées. Puis on applique une étape d'élimination à ce nouveau système. En termes de produits matriciels on a commencé par multiplier A à gauche par la matrice de transposition $P(1,i)$ puis par une matrice d'élimination :

$$E(1, -\frac{a_{21}}{a_{i1}}, \ldots, -\frac{a_{n1}}{a_{i1}})P(1,i)A$$

que l'on note plus simplement $E_1 P_1 A$. Les autres étapes sont identiques à la première : recherche d'un nouveau pivot puis élimination pour obtenir

$$E_{n-1} P_{n-1} \ldots E_2 P_2 E_1 P_1 A = U$$

où U est triangulaire supérieure.

6.4.4 Pivot total

La méthode du *pivot total* prend pour pivot a_{ij} tel que

$$|a_{ij}| = \max_{1 \leqslant k,l \leqslant n} |a_{kl}| .$$

On doit donc permuter les lignes 1 et i ainsi que les colonnes 1 et j. Ceci revient à considérer la matrice $P(1,i)AP(1,j)$ au lieu de A. Le système à résoudre est donc $P(1,i)AP(1,j)y = P(1,i)b$ avec $y = P(1,j)x$ au lieu de $Ax = b$ puis on effectue une étape d'élimination. Après $n-1$ telles opérations on a :

$$E_{n-1} P_{n-1} \ldots E_2 P_2 E_1 P_1 A P_1 P_2 \ldots P_{n-1} = U$$

où U est triangulaire supérieure.

6.4.5 Une justification

Le choix de ces stratégies est motivé par l'analyse que l'on fait des erreurs dans une division ou dans le calcul de l'inverse d'un nombre réel. Soient x et $h \in \mathbb{R}$ avec x et $x + h \neq 0$. On a, au premier ordre,

$$\frac{1}{x+h} - \frac{1}{x} \approx -\frac{h}{x^2}$$

ce qui prouve que l'erreur commise dans le calcul de l'inverse est d'autant plus grande que le nombre par lequel on divise est petit et d'autant plus petite que le diviseur est grand. Les choix du pivot partiel et celui du pivot total sont motivés par ces considérations : on minimise les erreurs d'arrondi en divisant par le pivot de plus grand module.

6.4.6 Exemple numérique

Considérons le système $Ax = b$ où A est une matrice de Vandermonde (voir paragraphe 16.4) 8×8 définie par des points choisis aléatoirement. Posons $x = (1, \ldots, 1)^T$ et prenons pour second membre $b = Ax = A(1, \ldots, 1)^T$. La solution exacte de ce système est bien sûr $x = (1, \ldots, 1)^T$. La solution calculée par la méthode d'élimination de Gauss donne les normes d' erreurs suivantes :

0.0032 pour la méthode sans stratégie de pivot,

$1.2279\ 10^{-13}$ pour la méthode avec stratégie de pivot partiel,

$3.8599\ 10^{-14}$ pour la méthode avec stratégie de pivot total.

La méthode d'élimination de Gauss sans stratégie de pivotage peut donc détériorer la solution du système. Dans la plupart des cas on observe que la stratégie de pivot partiel est suffisante pour obtenir des résultats satisfaisants.

6.4.7 Décompositions PLU et LUP

Une conséquence des méthodes du pivot partiel ou du pivot total est résumée dans le théorème suivant :

Théorème 6.10 *Pour toute matrice $A \in \mathbb{GL}_n$ il existe une matrice de permutation P, une matrice triangulaire inférieure à diagonale unité L et une matrice triangulaire supérieure U telles que $PA = LU$.*

Démonstration. Elle est basée sur la méthode du pivot partiel qui permet d'écrire $E_{n-1}P_{n-1}\ldots E_2P_2E_1P_1A = U$ que l'on écrit de façon plus compliquée

$$E_{n-1}E'_{n-2}\ldots E'_1PA = U$$

avec $E'_k = P_{n-1}P_{n-2}\ldots P_{k+1}E_kP_{k+1}\ldots P_{n-2}P_{n-1}$ et $P = P_{n-1}P_{n-2}\ldots P_1$ (noter que $P_i^2 = I_n$ pour toute matrice de transposition). Il faut, pour conclure, se convaincre que les matrices E'_k sont encore des matrices d'élimination (examiner le cas $P_{k+1}E_kP_{k+1}$ et continuer par récurrence).

Une autre possibilité est de rechercher le premier pivot dans la première ligne de A, ce qui revient à multiplier A à droite par une matrice de permutation, puis à effectuer une étape d'élimination de Gauss. La nouvelle matrice est du type $E_1 A P_1$ et, après $n-1$ telles étapes, on obtient

$$E_{n-1} \dots E_2 E_1 A P_1 P_2 \dots P_{n-1} = U.$$

On a prouvé que :

Théorème 6.11 *Pour toute matrice $A \in \mathbb{GL}_n$ il existe une matrice de permutation P, une matrice triangulaire inférieure à diagonale unité L et une matrice triangulaire supérieure U telles que $AP = LU$.*

6.4.8 Le cas des matrices rectangulaires

Le procédé d'élimination de Gauss s'applique aussi aux systèmes d'équations linéaires dont le nombre d'équations et celui d'inconnues sont différents. Il permet, dans le cas le plus général, d'obtenir un système échelonné équivalent.

Définition 6.12 *Une matrice $E \in \mathbb{C}^{m\times n}$, $E \neq 0$, est échelonnée lorsqu'il existe des entiers $1 \leqslant r \leqslant \min(m,n)$ et $1 \leqslant n_1 < n_2 < \dots < n_r \leqslant n$ tels que :*

1. *Les lignes 1 à r de E ont la structure suivante : pour $1 \leqslant i \leqslant r$ et $1 \leqslant j < n_i$ on a $e_{ij} = 0$ et $e_{in_i} \neq 0$.*
2. *Les lignes $r+1$ à m sont nulles : pour $r+1 \leqslant i \leqslant m$ et $1 \leqslant j \leqslant n$ on a $e_{ij} = 0$.*

Un système linéaire est échelonné lorsque c'est le cas de la matrice de ce système.

$$E = \left(\overset{}{\begin{bmatrix} 0 & \dots & 0 \\ & & \\ \vdots & & \vdots \\ & & \\ 0 & \dots & 0 \end{bmatrix}} \overset{n_1}{\begin{bmatrix} \neq 0 & \times & \dots & \times \\ 0 & 0 & \dots & 0 \\ & & & \\ \vdots & \vdots & & \vdots \\ & & & \\ 0 & 0 & \dots & 0 \end{bmatrix}} \quad \dots \quad \overset{n_r}{\begin{bmatrix} \times & \times & \dots & \times \\ \vdots & & & \vdots \\ \neq 0 & \times & \dots & \times \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}} \right)$$

Remarque 6.1.

1. Les entrées $e_{in_i} \neq 0$, $1 \leqslant i \leqslant r$, sont appelées les *pivots* de la matrice E.
2. Le nombre r de pivots est égal au rang de E.

Le calcul de la forme échelonnée d'une matrice $B \in \mathbb{C}^{m\times n}$ se mène par la méthode du pivot partiel. Si $B = \begin{pmatrix} 0 & \dots & 0 & B_1 \end{pmatrix}$ où la première colonne non nulle de B

est b_{n_1} (c'est aussi la première colonne de B_1), par multiplication par une matrice de transposition P_1 puis par une matrice d'élimination E_1, on obtient

$$E_1P_1B = \left(\begin{bmatrix} 0 & \dots & 0 \\ & & \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \neq 0 & \dots & \times \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \times & \dots & \times \\ & & \\ & B_2 & \\ & & \end{bmatrix} \right)$$

où la première colonne de B_2 est $b_{n_2} \in \mathbb{C}^{m-1}$, $b_{n_2} \neq 0$. L'étape suivante poursuivra la construction de la forme échelonnée sans changer la structure déjà acquise jusqu'à obtenir la matrice échelonnée $E_rP_r \dots E_1P_1B = E$. On obtient ainsi, en suivant la preuve du théorème 6.10, le

Théorème 6.13 *Pour toute matrice $B \in \mathbb{C}^{m\times n}$ il existe une matrice de permutation $P \in \mathbb{C}^{m\times m}$, une matrice triangulaire inférieure à diagonale unité $L \in \mathbb{C}^{m\times m}$ et une matrice échelonnée $E \in \mathbb{C}^{m\times n}$ telles que $PB = LE$.*

6.5 COMPLEXITÉ

Chaque étape d'élimination requiert le calcul de $n - i$ divisions, $(n - i)^2$ additions et $(n - i)^2$ multiplications avec $1 \leqslant i \leqslant n - 1$. Pour l'ensemble des étapes on obtient :

- $1 + \dots + n - 1 = \dfrac{n(n-1)}{2}$ divisions,
- $1^2 + \dots + (n-1)^2 = \dfrac{n(n-1)(2n-1)}{6}$ additions,
- $1^2 + \dots + (n-1)^2 = \dfrac{n(n-1)(2n-1)}{6}$ multiplications,

donc un total de

$$\frac{n(n-1)(2n-1)}{3} + \frac{n(n-1)}{2} \approx \frac{2n^3}{3}$$

opérations arithmétiques.

Nous n'avons pas tenu compte des opérations sur le second membre b du système $Ax = b$. Leur ordre de grandeur est $O(n)$.

On déduit de ce compte que

- Le calcul de la solution d'un système linéaire $n \times n$ peut se faire en $O(n^3)$ opérations arithmétiques ($O(n^3)$ pour la décomposition LU avec ou sans permutations et $O(n^2)$ pour la solution du système triangulaire obtenu),

- Le calcul de l'inverse d'une matrice requiert $O(n^3)$ opérations arithmétiques : ce calcul se ramène au précédent pour un second membre arbitraire. Plus précisément, ce sont $\approx \frac{2n^3}{3} + n^3$ opérations qu'il faut exécuter. Ce calcul peut être évité dans la plupart des expressions utilisant la matrice inverse comme par exemple le scalaire $a^* A^{-1} b$ où a et b sont des vecteurs. Dans ce cas $A^{-1}b$ est obtenu en résolvant un seul système linéaire sans calcul explicite de la matrice A^{-1}.

- Le calcul du déterminant de A se fait en $O(n^3)$ opérations arithmétiques : si $A = LU$ alors $\det A = \det U = u_{11} \ldots u_{nn}$. Si des permutations sont effectuées, $\det A = \pm \det U$, chaque permutation de deux lignes ou de deux colonnes changeant le signe du déterminant. C'est la méthode utilisée par le logiciel Matlab.

Il faut rapprocher ces résultats de complexité avec ceux que l'on obtient en utilisant les formules de Cramer. Un déterminant y est décrit comme une somme de $n!$ monômes (chacun d'eux de degré n en les entrées de la matrice) d'où une complexité en $O(n! \times n)$ pour un calcul brutal !

6.6 CONDITIONNEMENT DE LA DÉCOMPOSITION LU

Nous allons analyser, via le calcul différentiel, les variations au premier ordre de la décomposition LU d'une matrice en fonction des variations de cette matrice. Nous renvoyons le lecteur à son ouvrage préféré de calcul différentiel pour le théorème de dérivation des fonctions inverses qui est utilisé dans ce paragraphe. Nous utilisons les notations suivantes :

- $\mathcal{L}_n$ est l'espace vectoriel des matrices $L \in \mathbb{C}^{n \times n}$ qui sont triangulaires inférieures et qui ont une diagonale nulle,

- $\mathcal{U}_n$ est l'espace vectoriel des matrices $U \in \mathbb{C}^{n \times n}$ qui sont triangulaires supérieures,

- $\mathcal{GU}_n$ est le sous-ensemble de $\mathcal{U}_n$ constitué des matrices triangulaires supérieures et inversibles,

- $\mathcal{LU}$ est le sous-ensemble de $\mathbb{GL}_n$ constitué par les matrices qui possèdent une décomposition LU (voir la caractérisation donnée au théorème 6.6).

Soit $A \in \mathbb{GL}_n$. Ecrivons la décomposition LU de A sous la forme

$$A = (I_n + L)U$$

avec $U \in \mathcal{GU}_n$ et $L \in \mathcal{L}_n$. Nous souhaitons calculer la dérivée de l'application $A \to (L, U)$ ainsi que la norme de cette dérivée. Pour ce faire, nous allons calculer la dérivée de $(L, U) \to A$, ce qui est très facile, puis utiliser le théorème de dérivation des fonctions inverses. Notons

$$\mathcal{P} : \mathcal{L}_n \times \mathcal{GU}_n \to \mathcal{LU}, \ \mathcal{P}(L, U) = (I_n + L)U.$$

Par le théorème 6.6 c'est une bijection entre $\mathcal{L}_n \times \mathcal{GU}_n$ et $\mathcal{LU}$. La bijection inverse associe à $A \in \mathcal{LU}$ les matrices $L \in \mathcal{L}_n$ et $U \in \mathcal{GU}_n$ telles que $A = (I_n + L)U$. On note

$$\mathcal{L} : \mathcal{LU} \to \mathcal{L}_n, \ \mathcal{L}(A) = L$$

et

$$\mathcal{U} : \mathcal{LU} \to \mathcal{GU}_n, \ \mathcal{U}(A) = U.$$

On équipe l'espace des applications linéaires $\mathcal{M} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ de la norme

$$\|\mathcal{M}\|_{FF} = \sup_{\substack{B \in \mathbb{C}^{n\times n} \\ B \neq 0}} \frac{\|\mathcal{M}(B)\|_F}{\|B\|_F}.$$

Théorème 6.14 *$\mathcal{LU}$ et $\mathcal{GU}_n$ sont des sous-ensembles ouverts de $\mathbb{C}^{n\times n}$, $\mathcal{L}$ et $\mathcal{U}$ sont des applications de classe $\mathbb{C}^\infty$ sur $\mathcal{LU}$ et, pour toute matrice $A \in \mathcal{LU}$, $A = (I_n + L)U$,*

$$\|D\mathcal{L}(A)\|_{FF} \leqslant \mathrm{cond}_2(I_n + L) \left\|U^{-1}\right\|_2$$

et

$$\|D\mathcal{U}(A)\|_{FF} \leqslant \mathrm{cond}_2(U) \left\|(I_n + L)^{-1}\right\|_2.$$

Démonstration. Les étapes principales de cette démonstration sont les suivantes. Nous laissons le lecteur en vérifier les détails à titre d'exercice.

1. $\mathcal{LU}$ est ouvert dans $\mathbb{C}^{n\times n}$ (par le théorème 6.6) et $\mathcal{GU}_n$ est ouvert dans $\mathcal{U}_n$ (par le corollaire 3.15).
2. L'application $\mathcal{P}$ est de classe C^∞ et sa dérivée en $(L, U) \in \mathcal{L}_n \times \mathcal{GU}_n$ est donnée par :

$$D\mathcal{P}(L, U) : \mathcal{L}_n \times \mathcal{U}_n \to \mathbb{C}^{n\times n}, \ D\mathcal{P}(L, U)(M, V) = (I_n + L)V + MU.$$

3. $D\mathcal{P}(L, U)$ est un isomorphisme ($D\mathcal{P}(L, U)$ est une injection entre deux espaces de même dimension).

4. On utilise le théorème d'inversion locale pour en déduire que les applications $\mathcal{L} : \mathcal{LU} \to \mathcal{L}_n$ et $\mathcal{U} : \mathcal{LU} \to \mathcal{GU}_n$ sont elles aussi C^∞ et que, pour tout $B \in \mathbb{C}^{n\times n}$

$$D\mathcal{L}(A)B = (I_n + L)\Pi_{\mathcal{L}_n}\left((I_n + L)^{-1}BU^{-1}\right),$$

$$D\mathcal{U}(A)B = [\Pi_{\mathcal{U}_n}\left((I_n + L)^{-1}BU^{-1}\right)]\,U$$

et où, pour toute matrice $M \in \mathbb{C}^{n\times n}$, $\Pi_{\mathcal{L}_n}(M)$ et $\Pi_{\mathcal{U}_n}(M)$ sont les parties triangulaires inférieure stricte et supérieure de M.

5. On calcule les normes de ces dérivées.

Remarque 6.2. Ce théorème montre que les variations de la décomposition LU de la matrice A dépendent non pas du conditionnement de A (on verra que c'est le cas pour la décomposition de Cholesky et pour la décomposition QR) mais des conditionnements ou, plus précisemment, des plus grandes et des plus petites valeurs singulières des matrices $I_n + L$ et U telles que $A = (I_n + L)U$. Nous avons vu, sur des exemples, que $\mathrm{cond}_2(U)$ peut être bien plus grand que $\mathrm{cond}_2(A)$ ce qui peut conduire à une vision pessimiste de la décomposition LU et à son utilité quant à la résolution de systèmes linéaires. Un exemple classique, dû à Wilkinson, donne une détérioration exponentielle en le nombre de variables du conditionnement de U ! Mais le cas moyen est par contre extrêmement stable ; voir à ce sujet l'article de Schreiber-Trefethen [31] et les ouvrages de Wilkinson et Higham cités en bibliographie.

6.7 NOTES ET RÉFÉRENCES

LU vient de l'anglais Lower-Upper (triangular matrices).

Une idée récurrente pour résoudre un système d'équations linéaires est de se ramener via une décomposition matricielle adaptée à un système triangulaire, celui-ci étant facile à résoudre. Les décompositions LU, QR et de Cholesky permettent d'y arriver. La décomposition LU, due à Gauss, est la plus célèbre et la plus ancienne de ces méthodes. Gauss naquit à Brunswick en 1777 et mourut à Göttingen en 1855.

Un des intérêts de la décomposition LU, pour une matrice à coefficients dans un anneau commutatif quelconque, est que cette décomposition s'effectue dans le corps des fractions associé. Ceci en fait un outil de choix pour le calcul formel.

Le livre de Stewart [32] est extrêmement documenté sur l'algorithmique de la décomposition LU. Pour l'étude de la stabilité de l'élimination de Gauss nous renvoyons aux ouvrages de Wilkinson [36] et [37] ainsi qu'à celui de Higham [18].

EXERCICES

Exercice 6.1

On suppose que $A \in \mathbb{GL}_n$ admet une décomposition LU. En calculant formellement le produit $A = LU$, donner un algorithme de calcul des coefficients u_{ij} de U et l_{ij} de L à partir des coefficients a_{ij} de A (algorithme de Crout et Doolittle). Indication : calculer la première ligne de U puis la première colonne de L et ainsi de suite. Application numérique : donner la décomposition LU de la matrice

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

Exercice 6.2

On utilise la méthode d'élimination de Gauss pour trianguler le système :

$$(S) \begin{cases} 10^{-4}x + y = 1 \\ x + y = 2 \end{cases}$$

1. Calculer le conditionnement associé à la norme $\| \cdot \|_\infty$ de la matrice A de ce système et de la matrice U_1 du système triangulaire obtenu par la décomposition LU.
2. Même question si l'on permute d'abord les deux lignes de A.

Exercice 6.3

Une matrice d'élimination $E(i, \lambda_{i+1}, \ldots, \lambda_n)$ s'écrit sous la forme

$$E(i, \lambda_{i+1}, \ldots, \lambda_n) = I_n + l\, e_i^T,$$

où $l = (0, \ldots, 0, \lambda_{i+1}, \ldots, \lambda_n)^T$ est un vecteur dont les i premières coordonnées sont nulles et e_i est le i-ième vecteur de la base canonique de $\mathbb{R}^n$.

1. Pour $i < j$, calculer le produit de deux matrices d'élimination $E(i, \lambda_{i+1}, \ldots, \lambda_n)$ $E(j, \lambda'_{j+1}, \ldots, \lambda'_n)$. Généraliser le résultat au produit de k matrices d'élimination ordonnées par ordre croissant d'indice $i_1 < i_2 < \ldots < i_k$.
2. Soit A inversible admettant la décomposition LU. Déduire du calcul précédent l'expression de la matrice L à partir des matrices d'élimination $E(i, \lambda_{i+1}, \ldots, \lambda_n)$ intervenant dans l'algorithme d'élimination de Gauss.

Exercice 6.4

On considère deux vecteurs $x, y \in \mathbb{C}^n$, la matrice $M = I_n + xy^*$ et l'on suppose que $\delta_k = 1 + \sum_{l=1}^{k} x_l \overline{y}_l \neq 0$, pour tout $k = 0, \ldots, n$ (par convention, $\delta_0 = 1$).

1. Montrer que $\det(M) = \delta_n$.
2. Montrer que M admet une décomposition LU.
3. Montrer, par récurrence, l'égalité

$$\sum_{l=1}^{k} \frac{x_l \overline{y}_l}{\delta_l \delta_{l-1}} = \frac{\delta_k - 1}{\delta_k}$$

pour tout $k = 1, \ldots, n$.

4. On décompose M sous la forme $M = D+E+F$, avec D diagonale, E triangulaire strictement inférieure et F triangulaire strictement supérieure. On considère les matrices diagonales $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_n)$, $\Delta_+ = \mathrm{diag}(1, \delta_1, \ldots, \delta_{n-1})$ et les matrices triangulaires $L = (E + \Delta)\Delta^{-1}$ et $U = \Delta_+^{-1}(F + \Delta)$. Montrer que $M = LU$ (calculer les coefficients de la matrice LU, considérer pour cela les cas $i < j$, $i > j$ et $i = j$).
5. Soit $A \in \mathbb{GL}_n$ et $u, v \in \mathbb{C}^n$. On suppose que la matrice A admet une décomposition $A = LU$. Donner une condition suffisante portant sur u, v, L et U pour que la matrice $B = A + uv^*$ possède une décomposition LU. Utiliser les résultats précédents pour exprimer la décomposition LU de B à partir de la décomposition LU de A (formule de mise à jour de la décomposition LU).

Exercice 6.5

Soit $A \in \mathbb{C}^{n \times n}$ une *matrice bande* inversible de largeur $2p + 1$ c'est-à-dire telle que $a_{ij} = 0$ pour $| i - j | > p$. On suppose que A admet une décomposition LU. Montrer que les matrices L et U ont également une structure bande de largeur $2p + 1$.

Chapitre 7

Matrices définies positives et décomposition de Cholesky

7.1 MATRICES DÉFINIES POSITIVES

Définition 7.1 *Une matrice $A \in \mathbb{C}^{n\times n}$ est semi-définie positive lorsqu'elle est hermitienne et que $x^*Ax \geqslant 0$ pour tout $x \in \mathbb{C}^n$.*
Une matrice $A \in \mathbb{C}^{n\times n}$ est définie positive lorsqu'elle est hermitienne et que $x^*Ax > 0$ pour tout $x \in \mathbb{C}^n$, $x \neq 0$.

Soit $A \in \mathbb{C}^{n\times n}$, x et $y \in \mathbb{C}^n$. Notons

$$\langle x, y\rangle_A = y^*Ax = \sum_{i,j=1}^{n} a_{ij}x_j\overline{y}_i.$$

Il est facile de voir que la forme $\langle .,.\rangle_A$ est un produit hermitien sur $\mathbb{C}^n$ si et seulement si A est une matrice définie positive. D'autres caractérisations des matrices définies positives sont données par le théorème suivant :

Théorème 7.2 *Pour une matrice $A \in \mathbb{C}^{n\times n}$, il y a équivalence entre :*

1. *A est définie positive,*
2. *Il existe un espace hermitien $\mathbb{E}$ et n vecteurs indépendants $e_i \in \mathbb{E}$ tels que $a_{ij} = \langle e_j, e_i\rangle_{\mathbb{E}}$,*
3. *A est hermitienne et ses valeurs propres sont positives,*
4. *Il existe une matrice $M \in \mathbb{C}^{m\times n}$ de rang n telle que $A = M^*M$.*

Démonstration. $1 \Rightarrow 2$. Cela a déjà été quasiment démontré : on peut prendre $\mathbb{E} = \mathbb{C}^n$ équipé du produit scalaire $\langle x, y\rangle_A = y^*Ax$. Pour les vecteurs e_i de la base canonique on a bien $\langle e_j, e_i\rangle_A = e_i^*Ae_j = a_{ij}$.

$2 \Rightarrow 3$. Si $a_{ij} = \langle e_j, e_i\rangle_{\mathbb{E}}$, il est clair que A est hermitienne. Si $Ax = \lambda x$ avec $x \neq 0$ alors $x^*Ax = \lambda \|x\|_2^2$ et

$$x^*Ax = \sum_{i,j=1}^{n} \langle e_j, e_i\rangle_{\mathbb{E}} x_j \overline{x}_i = \left\langle \sum_{j=1}^{n} x_j e_j, \sum_{i=1}^{n} x_i e_i \right\rangle_{\mathbb{E}} = \left\| \sum_{i=1}^{n} x_i e_i \right\|_{\mathbb{E}}^2.$$

Puisque $x \neq 0$ et que les e_i sont indépendants, on a $\left\|\sum_{i=1}^n x_i e_i\right\|_{\mathbb{E}} > 0$ et $\|x\|_2 > 0$ de sorte que $\lambda > 0$.

$3 \Rightarrow 4$. Puisque A est hermitienne et à valeurs propres positives, nous pouvons diagonaliser $A = UDU^*$ avec $U \in \mathbb{U}_n$, $D = \mathrm{diag}(\lambda_i)$ et $\lambda_i > 0$ (théorème 1.6). Prenons $M = \mathrm{diag}(\sqrt{\lambda_i})U^*$. Il est clair que $M^*M = A$ et que $M \in \mathbb{C}^{n\times n}$ est inversible donc de rang n.

$4 \Rightarrow 1$. Notons que si $A = M^*M$ alors A est hermitienne. De plus, pour tout $x \in \mathbb{C}^n$, $x \neq 0$, nous avons :

$$x^*Ax = x^*M^*Mx = \|Mx\|_2^2 \geqslant 0.$$

Si cette quantité était nulle, cela signifierait que $Mx = 0$ avec $x \neq 0$. Les n colonnes de M seraient donc dépendantes et M ne serait pas de rang n. Ainsi $x^*Ax = \|Mx\|_2^2 > 0$ et A est définie positive.

Remarque 7.1.

1. Une matrice du type $(\langle e_j, e_i\rangle_{\mathbb{E}})$ est appelée *matrice de Gram* du système de vecteurs (e_i). Toute matrice semi-définie positive est une matrice de Gram (théorème 7.3). Une matrice de Gram est inversible si et seulement si les vecteurs qui la définissent sont indépendants (théorème 7.2).
2. Soit $A \in \mathbb{R}^{n\times n}$ une matrice définie positive. L'ensemble

$$\mathcal{E}_A = \{x \in \mathbb{R}^n \ : \ x^*Ax = 1\}$$

est un *ellipsoïde*. Dans une base orthonormée (u_i) de vecteurs propres de A son équation est

$$\mathcal{E}_A = \{x = v_1u_1 + \ldots + v_nu_n \ : \ \lambda_1 v_1^2 + \ldots + \lambda_n v_n^2 = 1\}$$

où les $\lambda_i > 0$ sont les valeurs propres de A.

À titre d'exemple, montrons que la *matrice de Hilbert*

$$H_n = \left(\frac{1}{i+j-1}\right)_{i,j=1\ldots n}$$

est définie positive. On a

$$\frac{1}{i+j-1} = \int_0^1 x^{i-1}x^{j-1}dx$$

qui peut être vu comme le produit scalaire des monômes (linéairement indépendants) x^{i-1}, $i = 1, \ldots, n$, pour le produit scalaire

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx$$

dans $C[0, 1]$. On conclut à l'aide du théorème 7.2-2.

Les matrices semi-définies positives ont une caractérisation similaire. Nous n'en donnons pas la preuve qui est, *mutatis mutandis*, similaire à la précédente :

Théorème 7.3 *Pour une matrice $A \in \mathbb{C}^{n\times n}$ il y a équivalence entre*

1. *A est semi-définie positive,*
2. *Il existe un espace hermitien $\mathbb{E}$ et n vecteurs $e_i \in \mathbb{E}$ tels que $a_{ij} = \langle e_j, e_i\rangle_{\mathbb{E}}$,*
3. *A est hermitienne et ses valeurs propres sont positives ou nulles,*
4. *Il existe une matrice $M \in \mathbb{C}^{m\times n}$ telle que $A = M^*M$.*

Voici quelques conséquences du théorème de caractérisation des matrices définies positives

Corollaire 7.4 *Supposons que $A \in \mathbb{C}^{n\times n}$ soit définie positive (resp. semi-définie positive) alors :*

1. *$\det A > 0$ (resp. $\geqslant 0$) et $\operatorname{trace} A > 0$ (resp. $\geqslant 0$),*
2. *Pour tout $i = 1 \ldots n$, $a_{ii} > 0$ (resp. $\geqslant 0$),*
3. *Toute matrice obtenue en supprimant dans A les lignes $L_{i_1}, \ldots, L_{i_p}$ et les colonnes $C_{i_1}, \ldots, C_{i_p}$ est définie positive (resp. semi-définie positive).*

Démonstration. 1 provient du théorème 7.2-3, 2 découle de 3 (supprimer $n-1$ lignes et les $n-1$ colonnes de même indice), 3 est une conséquence du théorème 7.2-2.

La caractérisation suivante des matrices définies positives ne s'étend pas aux matrices semi-définies positives :

Théorème 7.5 *Pour une matrice $A \in \mathbb{C}^{n\times n}$, il y a équivalence entre*

1. *A est définie positive.*
2. *A est hermitienne et les matrices $A(1:k, 1:k)$, $1 \leqslant k \leqslant n$, ont un déterminant positif.*

Démonstration. Le corollaire 7.4-3 et 1 prouve que la condition est nécessaire. Pour montrer qu'elle est suffisante, nous raisonnons par récurrence. Le cas $n = 1$ est évident. Supposons la propriété établie pour $n-1$. Ainsi $A(1:n-1, 1:n-1)$ est définie positive et il existe une matrice $C \in \mathbb{GL}_{n-1}$ telle que $A(1:n-1, 1:n-1) = CC^*$. Écrivons

$$A = \begin{pmatrix} A(1:n-1, 1:n-1) & a \\ a^* & a_{nn} \end{pmatrix} \text{ avec } a = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{n-1\,n} \end{pmatrix}.$$

Notons que

$$\begin{pmatrix} C & 0 \\ u^* & \alpha \end{pmatrix} \begin{pmatrix} C^* & u \\ 0 & \beta \end{pmatrix} = \begin{pmatrix} A(1:n-1, 1:n-1) & Cu \\ u^*C^* & u^*u + \alpha\beta \end{pmatrix}$$

de sorte que

$$A = \begin{pmatrix} C & 0 \\ u^* & \alpha \end{pmatrix} \begin{pmatrix} C^* & u \\ 0 & \beta \end{pmatrix}$$

si l'on prend $Cu = a$ c'est-à-dire $u = C^{-1}a$ et $u^*u + \alpha\beta = a_{nn}$. On aura prouvé que A est définie positive si l'on peut prendre $\alpha = \beta > 0$ (appliquer le théorème 7.2-4). Il reste donc à prouver que $\alpha\beta > 0$. L'égalité ci-dessus prouve que

$$\det A = \det C\ \alpha\ \det C^*\ \beta = |\det C|^2\ \alpha\beta;$$

comme, par hypothèse, $\det A > 0$ on a bien $\alpha\beta > 0$.

Corollaire 7.6 *Toute matrice définie positive possède une décomposition LU.*

Démonstration. C'est une conséquence des théorèmes 6.6 et 7.5.

L'exemple de la matrice $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ montre que le théorème 7.5 n'a pas d'extension immédiate au cas semi-défini positif : elle n'est pas semi-define positive alors que ses déterminants principaux sont $\geqslant 0$.

7.2 QUADRIQUES ET OPTIMISATION

Étant donné une matrice symétrique $A \in \mathbb{R}^{n\times n}$, un vecteur $b \in \mathbb{R}^n$ et un nombre $\alpha \in \mathbb{R}$, nous allons décrire les minimums et les maximums de la *quadrique*

$$q(x) = \frac{1}{2}x^T Ax - b^T x + \alpha = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_i x_j - \sum_{i=1}^{n} b_i x_i + \alpha.$$

Puisque A est symétrique ses valeurs propres sont réelles et notées

$$\lambda_1 \geqslant \ldots \geqslant \lambda_n.$$

Théorème 7.7

1. *Si* $\lambda_1 > 0$ *et* $\lambda_n < 0$ *alors*

$$\inf_{x\in\mathbb{R}^n} q(x) = -\infty \text{ et } \sup_{x\in\mathbb{R}^n} q(x) = \infty.$$

2. *Si* $b \notin \operatorname{Im} A$ *alors*

$$\inf_{x\in\mathbb{R}^n} q(x) = -\infty \text{ et } \sup_{x\in\mathbb{R}^n} q(x) = \infty.$$

3. *Les points stationnaires de* q *sont les* $\overline{x} \in \mathbb{R}^n$ *tels que* $A\overline{x} = b$. *Lorsque* $b \in \operatorname{Im} A$ *et que* $A\overline{x} = b$:

   a) Si $\lambda_1 \geqslant \ldots \geqslant \lambda_n \geqslant 0$ *alors* $q(\overline{x}) = \min_{x\in\mathbb{R}^n} q(x)$,

   b) Si $0 \geqslant \lambda_1 \geqslant \ldots \geqslant \lambda_n$ *alors* $q(\overline{x}) = \max_{x\in\mathbb{R}^n} q(x)$.

Démonstration. 1. Soit u_1 un vecteur propre unitaire associé à λ_1. On a $Au_1 = \lambda_1 u_1$ et $\|u_1\|_2 = 1$ de sorte que

$$q(\mu u_1) = \frac{1}{2}\mu^2 u_1^T Au_1 - \mu b^T u_1 + \alpha = \frac{1}{2}\mu^2\lambda_1 - \mu b^T u_1 + \alpha \to \infty$$

lorsque $\mu \to \infty$. De la même manière, prenons un vecteur propre unitaire u_n associé à λ_n :

$$q(\mu u_n) = \frac{1}{2}\mu^2\lambda_n - \mu b^T u_n + \alpha \to -\infty.$$

2. Puisque A est symétrique on a $\operatorname{Ker} A = (\operatorname{Im} A)^\perp$. En effet, si $x \in \operatorname{Ker} A$ on a $Ax = 0$ d'où $\langle Ax, u\rangle = 0$ pour tout $u \in \mathbb{R}^n$ et, puisque A est symétrique, $\langle x, Au\rangle = 0$ pour tout u c'est-à-dire $x \in (\operatorname{Im} A)^\perp$. On a prouvé l'inclusion $\operatorname{Ker} A \subset (\operatorname{Im} A)^\perp$. Il y a égalité parce que

$$\dim \operatorname{Ker} A = n - \dim \operatorname{Im} A = \dim(\operatorname{Im} A)^\perp.$$

Supposons désormais que $\operatorname{Ker} A \subset b^{\perp}$. On a donc $(\operatorname{Im} A)^{\perp} \subset b^{\perp}$ et par passage aux orthogonaux

$$b \in (b^{\perp})^{\perp} \subset ((\operatorname{Im} A)^{\perp})^{\perp} = \operatorname{Im} A$$

ce qui est contraire à l'hypothèse. Ceci prouve qu'il existe $\overline{x} \in \operatorname{Ker} A$ tel que $b^T \overline{x} \neq 0$. On a alors

$$q(\mu \overline{x}) = \frac{1}{2}\mu^2 \overline{x}^T A \overline{x} - \mu b^T \overline{x} + \alpha = -\mu b^T \overline{x} + \alpha \to \pm\infty$$

lorsque $\mu \to \pm\infty$ en fonction du signe de $b^T \overline{x}$.

3. Un point stationnaire $\overline{x}$ de q vérifie, par définition, $Dq(\overline{x})h = 0$ pour tout h. Par un calcul facile, on a

$$Dq(\overline{x})h = \lim_{\varepsilon \to 0} \frac{q(\overline{x} + \varepsilon h) - q(\overline{x})}{\varepsilon} = h^T(A\overline{x} - b)$$

de sorte que $Dq(\overline{x})h = 0$ pour tout h si et seulement si $A\overline{x} = b$.

Si $b \in \operatorname{Im} A$, $b = A\overline{x}$, on a

$$q(x) = \frac{1}{2}(x - \overline{x})^T A(x - \overline{x}) + \alpha - \frac{1}{2}\overline{x}^T A \overline{x}.$$

Utilisons la décomposition $A = UDU^T$. On obtient

$$q(x) = \frac{1}{2}(U^T(x - \overline{x}))^T DU^T(x - \overline{x}) + \alpha - \frac{1}{2}\overline{x}^T A\overline{x} = \frac{1}{2}y^T Dy + \beta = Q(y)$$

avec $y = U^T(x - \overline{x})$ et $\beta = \alpha - \frac{1}{2}\overline{x}^T A\overline{x}$. Remarquons que

$$\inf_{y \in \mathbb{R}^n} Q(y) = \inf_{x \in \mathbb{R}^n} q(x)$$

et que si $\tilde{y}$ est un minimum pour $Q(y)$ alors $\tilde{x} = U\tilde{y} + \overline{x}$ est un minimum pour $q(x)$. Idem pour le sup et les maxima. On remarque maintenant que

$$Q(y) = \frac{1}{2}y^T Dy + \beta = \sum_{i=1}^{n} \lambda_i y_i^2 + \beta.$$

Si tous les λ_i sont $\geqslant 0$ cette quantité a pour minimum $\tilde{y} = 0$ d'où la solution $\tilde{x} = \overline{x}$ et l'égalité $q(\overline{x}) = \min q(x)$. Lorsque tous les λ_i sont $\leqslant 0$, $\tilde{y} = 0$ est un maximum d'où $q(\overline{x}) = \max q(x)$.

Remarque 7.2.

- La démonstration de ce théorème prouve que, pour une quadrique, tout minimum (maximum) local est global.
- Lorsque A est inversible, la quadrique $q(x)$ possède un unique minimum si et seulement si A est définie positive. Il est alors donné par $\overline{x} = A^{-1}b$. Le graphe de la quadrique correspondante est un paraboloïde elliptique.

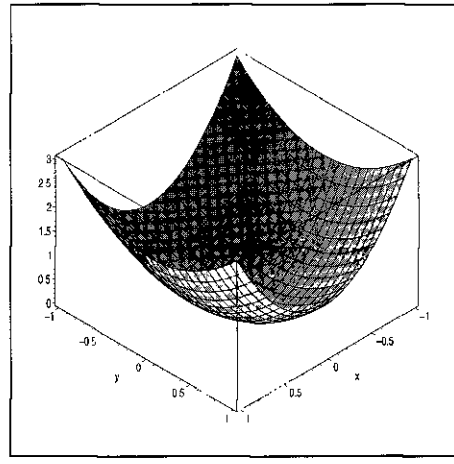


Tableau 7.1 Paraboloïde elliptique d'équation $z = 2x^2 + y^2$.

- Lorsque A est seulement semi-définie positive, le graphe de la quadrique $q(x)$ est un paraboloïde cylindrique. Suivant que b possède ou non une composante le long du noyau de A, la « gouttière » est « inclinée » et il n'y a pas de minimum ou bien « horizontale » et il y en a une infinité.

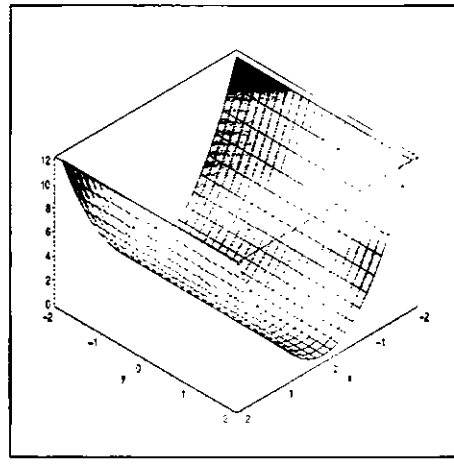


Tableau 7.2 Paraboloïde cylindrique d'équation $z = 3x^2$.

- Lorsque A est *indéfinie* c'est-à-dire lorsqu'elle possède des valeurs propres de signes différents. la quadrique $q(x)$ ne possède ni minimum ni maximum. Les points stationnaires sont des *points-selle* (on dit encore des *cols*). Dans ce cas le graphe de $q(x)$ est un paraboloïde hyperbolique.

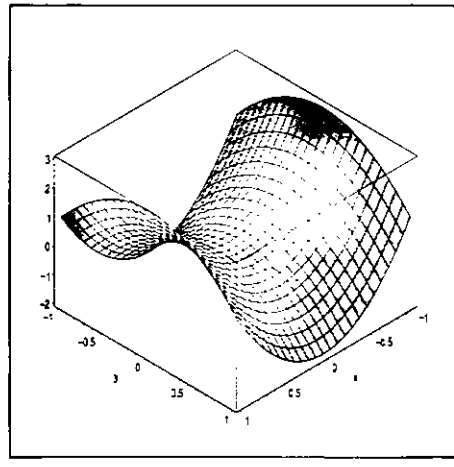


Tableau 7.3 Paraboloïde hyperbolique d'équation $z = 3x^2 - 2y^2$.

- $q(x)$ possède un unique maximum si et seulement si A est *définie négative* c'est-à-dire lorsque A est symétrique et que ses valeurs propres sont < 0.

7.3 RACINE CARRÉE D'UNE MATRICE, DÉCOMPOSITION POLAIRE

À la différence de l'équation scalaire $x^2 = a$, l'équation $X^2 = A$, lorsque A et X sont des matrices carrées, possède toutes sortes de solutions. L'exemple $X^2 = 0$ avec $X \in \mathbb{R}^{2\times 2}$ est déjà parlant ! Mais comme pour les équations scalaires, la définie positivité permet d'isoler une solution particulière.

Théorème 7.8 *Pour toute matrice semi-définie positive $A \in \mathbb{C}^{n\times n}$ il existe une et une seule matrice $B \in \mathbb{C}^{n\times n}$ qui soit semi-définie positive et qui vérifie $B^2 = A$. On l'appelle la racine carrée de A et on la note $A^{1/2}$.*

Démonstration. L'existence d'une racine carrée de A est donnée par $A = UDU^*$ avec U unitaire, $D = \mathrm{diag}(\lambda_i)$, $\lambda_i \geqslant 0$ et $A^{1/2} = UD^{1/2}U^*$ où $D^{1/2} = \mathrm{diag}(\lambda_i^{1/2})$. L'unicité est plus délicate à établir. Soit B une autre racine carrée de A : B est semi-définie positive et $B^2 = A$. Comme B et $A^{1/2}$ ont les mêmes valeurs propres (elles sont positives ou nulles et leurs carrés sont les valeurs propres de A), on peut écrire que $B = VD^{1/2}V^*$ pour une matrice unitaire $V \in \mathbb{U}_n$. On a ainsi, en élevant au carré,

$$UDU^* = VDV^*$$

ou encore

$$(V^*U)D(V^*U)^* = D$$

que nous écrivons

$$WDW^* = D$$

avec $W = V^*U$. La question est de savoir si

$$WD^{1/2}W^* = D^{1/2}$$

et donc si

$$UD^{1/2}U^* = VD^{1/2}V^*.$$

Supposons que les valeurs propres de A soient ordonnées en décroissant :

$$\lambda_1 = \ldots = \lambda_{n_1} > \lambda_{n_1+1} = \ldots = \lambda_{n_1+n_2} > \ldots >$$

$$\lambda_{n_1+\ldots+n_{p-1}+1} = \ldots = \lambda_{n_1+\ldots+n_{p-1}+n_p}$$

avec $n_1 + \ldots + n_{p-1} + n_p = n$. L'équation $WDW^* = D$ montre que les colonnes $w_1, \ldots, w_{n_1}$ de W constituent une base orthonormée de l'espace engendré par les vecteurs $e_1 \ldots, e_{n_1}$ de la base canonique de $\mathbb{C}^n$. Idem pour

$w_{n_1+1}, \ldots, w_{n_1+n_2}$ et cætera. Ainsi W est une matrice unitaire diagonale par blocs $W = \mathrm{diag}(W_i,\ 1 \leqslant i \leqslant p)$ avec $W_i \in \mathbb{U}_{n_i}$. Mais alors

$$WD^{1/2}W^* = \mathrm{diag}(W_i)\mathrm{diag}(\lambda_i^{1/2} I_{n_i})\mathrm{diag}(W_i)^* =$$

$$\mathrm{diag}(W_i \lambda_i^{1/2} I_{n_i} W_i^*) = \mathrm{diag}(\lambda_i^{1/2}) I_{n_i}) = D^{1/2}.$$

Corollaire 7.9 *Pour toute matrice $A \in \mathbb{GL}_n$ il existe des matrices $U \in \mathbb{U}_n$ et P définie positive telles que $A = PU$. Cette décomposition est unique et s'appelle la décomposition polaire de A.*

Démonstration. Si $A = PU$ comme ci-dessus alors $AA^* = PUU^*P^* = P^2$ et donc $P = (AA^*)^{1/2}$ (AA^* est définie positive par le théorème 7.2) et $U = P^{-1}A$. Il est clair que $P^{-1}A \in \mathbb{U}_n$ d'où l'existence et l'unicité.

7.4 LA DÉCOMPOSITION DE CHOLESKY

Théorème 7.10 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice définie positive. Il existe une unique matrice $L \in \mathbb{C}^{n\times n}$ triangulaire inférieure telle que $l_{ii} > 0$ pour tout i et $A = LL^*$ (décomposition de Cholesky).*

Démonstration. Supposons que $LL^* = MM^*$, que $l_{ii} > 0$ et que $m_{ii} > 0$ pour tout i. On a $M^{-1}L = M^*L^{-*}$ qui est à la fois triangulaire inférieure (à gauche) et triangulaire supérieure (à droite). C'est donc une matrice diagonale. Les entrées diagonales valent $m_{ii}^{-1}l_{ii} = m_{ii}l_{ii}^{-1}$ et donc sont égales à 1 par la condition de positivité. Ainsi $M^{-1}L = M^*L^{-*} = I_n$ c'est-à-dire $L = M$.

L'existence de cette décomposition se prouve par récurrence sur n. Pour $n = 1$ on prend $L = (\sqrt{a_{11}})$. Supposons que la décomposition de Cholesky existe pour toute matrice définie positive $n-1 \times n-1$. Écrivons

$$A = \begin{pmatrix} A_{n-1} & a \\ a^* & a_{nn} \end{pmatrix} \text{ avec } a = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{n-1\,n} \end{pmatrix}.$$

Notons $A_{n-1} = L_{n-1}L_{n-1}^*$ la décomposition de Cholesky de A_{n-1}. On a :

$$A = \begin{pmatrix} L_{n-1} & 0 \\ u^* & \alpha \end{pmatrix} \begin{pmatrix} L_{n-1}^* & u \\ 0 & \beta \end{pmatrix}$$

en prenant $L_{n-1}u = a$ et $u^*u + \alpha\beta = a_{nn}$. On obtiendra la décomposition de Cholesky de A si l'on peut prendre $\alpha = \beta > 0$ ce qui sera possible si $\alpha\beta > 0$. L'égalité ci-dessus prouve que

$$\det A = \det L_{n-1}\ \alpha\ \det L_{n-1}^*\ \beta.$$

Comme, par hypothèse, $\det A > 0$ et $\det L_{n-1} > 0$ on a bien $\alpha\beta > 0$.

Remarque 7.3. Lorsque A est définie positive, la méthode de Cholesky pour la résolution du système $Ax = LL^*x = b$ consiste à étudier les deux systèmes triangulaires $Ly = b$ et $L^*x = y$.

7.5 COMPLEXITÉ DE LA DÉCOMPOSITION

La décomposition de Cholesky s'obtient par l'algorithme suivant qui décrit l'identification des deux membres de l'équation matricielle :

$$\begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & \bar{l}_{21} & \dots & \bar{l}_{n1} \\ & l_{22} & \dots & \bar{l}_{n2} \\ & & \ddots & \vdots \\ & & & l_{nn} \end{pmatrix} = A.$$

On obtient :

Algorithme de décomposition de Cholesky

$l_{11} = \sqrt{a_{11}}$
pour $j = 2 : n$
  $l_{j1} = a_{j1}/l_{11}$
fin
pour $i = 2 : n$
  $l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} |l_{ik}|^2}$
  pour $j = i + 1 : n$
    $l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}\bar{l}_{ik}\right)/l_{ii}$
  fin
fin

Chaque étape de cet algorithme requiert $(2i-1)(n-i+1)$ opérations arithmétiques : $+, -, \times, /, |.|^2$ et $\sqrt{.}$. Puisque $1 \leqslant i \leqslant n$, le compte total est de $(n-1)n(n+1)/3 \approx n^3/3$ opérations arithmétiques soit moitié moins que pour la décomposition LU.

7.6 CONDITIONNEMENT DE LA DÉCOMPOSITION DE CHOLESKY

Comment varie la décomposition de Cholesky d'une matrice définie positive ? Nous allons répondre à cette question en utilisant le calcul différentiel suivant le schéma utilisé pour la décomposition LU. Nous noterons :

- $\mathcal{L}_r$ l'espace vectoriel des matrices $L \in \mathbb{C}^{n\times n}$ qui sont triangulaires inférieures et qui ont une diagonale réelle,
- $\mathcal{PL}_r$ le sous-ensemble de $\mathcal{L}_r$ constitué par les matrices à diagonale positive,
- $\mathcal{H}_n$ l'espace vectoriel des matrices hermitiennes $A \in \mathbb{C}^{n\times n}$,
- $\mathcal{PH}_n$ le sous-ensemble de $\mathcal{H}_n$ constitué par les matrices définies positives.

Par les théorèmes 7.2 et 7.10 l'application

$$\mathcal{P} : \mathcal{PL}_r \to \mathcal{PH}_n, \quad \mathcal{P}(L) = LL^*$$

est une bijection de $\mathcal{PL}_r$ sur $\mathcal{PH}_n$. La bijection réciproque,

$$\mathcal{C} : \mathcal{PH}_n \to \mathcal{PL}_r$$

est l'application qui à $A \in \mathcal{PH}_n$ associe sa décomposition de Cholesky $\mathcal{C}(A) = L \in \mathcal{PL}_r$.

On équipe l'espace des applications linéaires $\mathcal{M} : \mathcal{H}_n \to \mathcal{L}_r$ de la norme

$$\|\mathcal{M}\|_{FF} = \sup_{\substack{B \in \mathcal{H}_n \\ B \neq 0}} \frac{\|\mathcal{M}(B)\|_F}{\|B\|_F}.$$

Nous allons voir que l'application « Cholesky » est différentiable et nous allons estimer la norme de sa dérivée. On a :

Théorème 7.11 *L'application « Cholesky » $\mathcal{C} : \mathcal{PH}_n \to \mathcal{PL}_r$ est de classe C^∞ et, pour toute matrice $A \in \mathcal{PH}_n$,*

$$\|D\mathcal{C}(A)\|_{FF} \leqslant \frac{1}{\sqrt{2}} \frac{\mathrm{cond}_2(A)}{\|A\|_2^{1/2}}.$$

Démonstration. Nous n'en décrivons que les étapes principales, laissant au lecteur le soin d'en compléter les détails à titre d'exercice.

1. $\mathcal{L}_r$ et $\mathcal{H}_n$ sont des espaces de même dimension (réelle) égale à n^2, $\mathcal{PL}_r$ est un ensemble ouvert dans $\mathcal{L}_r$ et $\mathcal{PH}_n$ est un ensemble ouvert dans $\mathcal{H}_n$ (utiliser le théorème 7.5). Comme ces ensembles sont ouverts on peut envisager de dériver $\mathcal{P}$ sur $\mathcal{PL}_r$ et $\mathcal{C}$ sur $\mathcal{PH}_n$.
2. $\mathcal{P}$ est de classe C^∞ et sa dérivée en $L \in \mathcal{PL}_r$ est donnée par
$$D\mathcal{P}(L) : \mathcal{L}_r \to \mathcal{H}_n,\ D\mathcal{P}(L)M = LM^* + ML^*.$$
3. $D\mathcal{P}(L) : \mathcal{L}_r \to \mathcal{H}_n$ est un isomorphisme (comme ces deux espaces ont même dimension il suffit de prouver que $D\mathcal{P}(L)$ est injective. Pour ce faire on écrira que $LM^* + ML^* = 0$ et on considèrera la première colonne de cette matrice).
4. Par le théorème de dérivation des fonctions inverses, $\mathcal{C}$ est aussi de classe C^∞ et si $L \in \mathcal{PL}_r$ est la décomposition de Cholesky de $A \in \mathcal{PH}_n$ alors
$$D\mathcal{C}(A) = D\mathcal{P}(L)^{-1}.$$
5. Pour tout $L \in \mathcal{PL}_r$ et $M \in \mathcal{L}_r$ on a :
$$\left\|L^{-1}M + M^*L^{-*}\right\|_F \geqslant \sqrt{2}\,\|M\|_F / \|L\|_2 .$$
En effet, puisque $L^{-1}M$ est triangulaire inférieure à diagonale réelle et que $(L^{-1}M)^* = M^*L^{-*}$, on a
$$\left\|L^{-1}M + M^*L^{-*}\right\|_F^2 = 2\|L^{-1}M\|_F^2 + 2\sum_{i=1}^n |l_{ii}^{-1}m_{ii}|^2 \geqslant 2\|L^{-1}M\|_F^2 \geqslant 2\|M\|_F^2 / \|L\|_2^2$$
par la proposition 3.12-4. Il est de plus clair que
$$\left\|L^{-1}M + M^*L^{-*}\right\|_F \leqslant \left\|L^{-1}\right\|_2^2 \|D\mathcal{P}(L)M\|_F .$$
6. On en déduit que, pour $A = LL^*$, on a
$$\|D\mathcal{C}(A)\|_{FF} \leqslant \frac{1}{\sqrt{2}}\|L\|_2 \left\|L^{-1}\right\|_2^2 .$$
7. On montre enfin que $\mathrm{cond}_2(A) = \mathrm{cond}_2(L)^2$ d'où l'inégalité
$$\|D\mathcal{C}(A)\|_{FF} \leqslant \frac{1}{\sqrt{2}} \frac{\mathrm{cond}_2(A)}{\|A\|_2^{1/2}}.$$

7.7 NOTES ET RÉFÉRENCES

André-Louis Cholesky (15 octobre 1875 - 31 août 1918) était un mathématicien et un militaire français. Il a effectué sa carrière dans les services géographiques et topographiques de l'armée. C'est pour résoudre des problèmes de moindres carrés posés par la géodésie que Cholesky inventa la décomposition qui porte aujourd'hui son nom.

EXERCICES

Exercice 7.1

On considère la matrice $A_n = (a_{ij}) \in \mathbb{R}^{n\times n}$ définie par $a_{ij} = \min(i, j)$.

1. Montrer que pour tout n, $\det(A_n) = 1$. En déduire que A_n est définie positive.
2. Une autre façon de prouver que A_n est définie positive utilise un argument d'analyse. Posons $x_+^0 = 1$ si $x \geqslant 0$ et 0 sinon. Montrer que

$$\min(i, j) = \int_0^\infty (i - y)_+^0 (j - y)_+^0 \, dy$$

et que les fonctions $y \to (i - y)_+^0$, $1 \leqslant i \leqslant n$, sont indépendantes. En déduire que A_n est définie positive.

3. Déterminer la décomposition de Cholesky de A_n : $A_n = CC^T$. Calculer C^{-1} et en déduire A_n^{-1}.

Exercice 7.2

Soit $A \in \mathbb{C}^{n\times n}$ définie positive, $A = LU$ sa décomposition LU et soit $D = \text{diag}(u_{ii})$.

1. Montrer que $u_{ii} > 0$ et que $\left(LD^{1/2}\right)^* = D^{-1/2}U$. En déduire que $\left(LD^{1/2}\right)\left(LD^{1/2}\right)^*$ est la décomposition de Cholesky de A.
2. On considère la matrice $B = A + xx^*$ où $x \in \mathbb{C}^n$ est donné. En utilisant l'exercice 6.4 donner la décomposition de Cholesky de la matrice B à l'aide de celle de A.

Exercice 7.3 Matrice de Cauchy

Soit $a_i \in \mathbb{C}$, $i = 1, \ldots, n$, tels que $\Re(a_i) > 0$ pour tout i. La matrice

$$C_a = \left(\frac{1}{a_i + \overline{a}_j}\right)_{i,j=1\ldots n}$$

est appelée *matrice de Cauchy*. La matrice de Hilbert H_n considérée au paragraphe 7.1 est ainsi une matrice de Cauchy (prendre $a_i = i - \frac{1}{2}$). On considère les fonctions $\varphi_i(x) = x^{a_i - \frac{1}{2}}$. L'espace des fonctions de carré intégrable sur $]0, 1[$ est équipé du produit hermitien

$$\langle f, g\rangle = \int_0^1 f(x)\overline{g}(x)\, dx.$$

1. Montrer que, pour tout i, les fonctions φ_i sont de carré intégrable sur $]0, 1[$ et que

$$\langle \varphi_i, \varphi_j \rangle = \frac{1}{a_i + \overline{a}_j}.$$

En déduire que la matrice C_a est semi-définie positive.

2. On veut montrer que, si les coefficients a_i sont distincts, les fonctions $\varphi_i(x)$ sont linéairement indépendantes sur l'intervalle $]0, 1[$. Pour cela, on considère une combinaison linéaire telle que $\sum_{i=1}^{n} \alpha_i \varphi_i(x) = 0$, pour tout $x \in]0, 1[$. Montrer que cette égalité implique

$$\sum_{i=1}^{n} \alpha_i a_i^k = 0,$$

pour tout entier $k = 0, \ldots, n-1$. En déduire que les coefficients α_i sont tous nuls et que, lorsque les a_i sont distincts, la matrice C_a est définie positive.

Exercice 7.4

Montrer que l'application exponentielle (exercice 3.14) est une bijection des matrices hermitiennes sur les matrices définies positives. Pour prouver l'injectivité on raisonnera comme dans la démonstration du théorème 7.8.

Exercice 7.5

Résoudre par la méthode de Cholesky le système :

$$\begin{cases} 9x + 6y + 3z = 39 \\ 6x + 20y + 6z = 86 \\ 3x + 6y + 3z = 27 \end{cases}$$

Exercice 7.6

Résoudre via la décomposition de Cholesky le système linéaire $Ax = b$ avec

$$A = \begin{pmatrix} 4 & -2 & 2 & 0 \\ -2 & 2 & -2 & 1 \\ 2 & -2 & 6 & 3 \\ 0 & 1 & 3 & 6 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ -4 \\ 8 \\ 3 \end{pmatrix}.$$

Exercice 7.7

Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2\times 2}$ non nécessairement symétrique.

1. a) Écrire des conditions nécessaires et suffisantes que doivent vérifier a, b, c, d pour que l'on ait $x^T Ax > 0$ pour tout $x \in \mathbb{R}^2$, $x \neq 0$.

   b) Montrer que ces conditions impliquent $a > 0$ et $\det A > 0$ mais que la réciproque est fausse.

2. On étudie désormais le cas général. On dit qu'une matrice $A \in \mathbb{R}^{n\times n}$ appartient à $\mathcal{P}$ si $x^T Ax > 0$ pour tout $x \in \mathbb{R}^n$, $x \neq 0$. On désigne par $\mathcal{L}_1$ l'ensemble des matrices triangulaires inférieures et à diagonale unité ; $\mathcal{L}_1$ est un groupe multiplicatif. Soit $Q \in \mathbb{R}^{n\times n}$ une matrice inversible. Montrer que

$$A \in \mathcal{P} \Leftrightarrow QAQ^T \in \mathcal{P}$$

3. Soit $A \in \mathcal{P}$. Montrer que $a_{ii} > 0$ pour tout $i = 1, \cdots, n$.

4. Soit $A \in \mathcal{P}$.

   a) Montrer qu'il existe $L \in \mathcal{L}_1$ telle que $A^{(1)} = LAL^T$ se mette sous la forme

$$A^{(1)} = \begin{pmatrix} a_{11} & v^T \\ 0 & B \end{pmatrix} \text{ avec } v \in \mathbb{R}^{n-1} \text{ et } B \in \mathbb{R}^{(n-1)\times(n-1)}$$

   .

   b) En déduire qu'il existe $L \in \mathcal{L}_1$ telle que

$$LAL^T = U$$

   avec $U \in \mathcal{P}$ triangulaire supérieure.

5. a) Montrer que toute matrice $A \in \mathcal{P}$ se décompose sous la forme $A = MUM^T$ avec $M \in \mathcal{L}_1$ et $U \in \mathcal{P}$, triangulaire supérieure et que cette décomposition est unique.

   b) Lorsque A est définie positive, montrer que l'on retrouve la décomposition de Cholesky.

Exercice 7.8 Calcul de la racine carrée d'une matrice définie positive

1. Montrer que pour tout nombre réel $a > 0$ la suite définie par

$$x_0 = 1, \quad x_{p+1} = \frac{1}{2}\left(x_p + \frac{a}{x_p}\right),$$

converge vers $\sqrt{a}$. Cette méthode est attribuée à Héron d'Alexandrie (premier siècle de notre ère).

2. On se donne maintenant une matrice $A \in \mathbb{C}^{n\times n}$ définie positive. Montrer que la suite de matrices définie par

$$X_0 = I_n, \quad X_{p+1} = \frac{1}{2}\left(X_p + AX_p^{-1}\right)$$

est bien définie et converge vers $A^{1/2}$.

Exercice 7.9 Dérivée de la racine carrée d'une matrice définie positive

Notons $\sqrt{\ } : \mathcal{PH}_n \to \mathcal{PH}_n$ l'application qui à la matrice A définie positive associe sa racine carrée $B = \sqrt{A}$.

1. Montrer que pour toute matrice $K \in \mathbb{C}^{n\times n}$ l'intégrale

$$\int_0^\infty \exp(-tB)K\exp(-tB)dt$$

est absolument convergente (utiliser l'exercice 3.14).

2. Montrer que l'application $C : \mathcal{PH}_n \to \mathcal{PH}_n$ définie par $C(B) = B^2$ est de classe C^∞ et que

$$DC(B)H = BH + HB$$

pour tout $H \in \mathcal{H}_n$ (espace des matrices $n \times n$ hermitiennes).

3. Montrer que $\sqrt{\ }$ est de classe C^∞ et que

$$D_{\sqrt{\ }}(A) = DC(B)^{-1}$$

lorsque $B = \sqrt{A}$.

4. En déduire que

$$D_{\sqrt{\ }}(A)K = \int_0^\infty \exp(-tB)K\exp(-tB)dt,$$

pour toute matrice hermitienne $H \in \mathcal{H}_n$, et que

$$\left\|D_{\sqrt{\ }}(A)K\right\|_2 \leqslant \frac{1}{2}\left\|B^{-1}\right\|_2 \|K\|_2.$$

Exercice 7.10

On veut calculer l'inverse de la matrice définie positive $A_2 \in \mathbb{R}^{n\times n}$

$$A_2 = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

On sait qu'elle est associée à la discrétisation de la dérivée seconde dans un schéma de différences finies (paragraphe 16.1). Considérons d'abord la matrice B_2

$$B_2 = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

qui diffère de A_2 par une matrice de rang 1 :

$$A_2 = B_2 + e_1 e_1^T, \tag{7.1}$$

où $e_1 \in \mathbb{R}^n$, $e_1 = (1, 0, \ldots, 0)^T$ est le premier vecteur de la base canonique de $\mathbb{R}^n$.

1. Montrer par récurrence sur n que $\det B_2 = 1$. En déduire que B_2 est définie positive et donner sa décomposition de Cholesky $B_2 = CC^T$. Calculer B_2^{-1} à l'aide de cette décomposition.
2. À partir de l'égalité (7.1) montrer que

$$A_2^{-1} = \left(\min(i, j) - \frac{ij}{n+1}\right)$$

en utilisant la formule de Sherman-Morrison-Woodbury (exercice 1.9).

Exercice 7.11

Soit A une matrice définie positive ayant une structure bande de largeur $2p + 1$. Montrer que la matrice L triangulaire inférieure donnée par la décomposition de Cholesky possède la même structure bande. On a vu que cette même propriété est aussi vérifiée pour la décomposition LU (exercice 6.5).

Chapitre **8**

La décomposition QR

8.1 MATRICES DE STIEFEL

Nous verrons, lors de l'étude de la méthode des moindres carrés, qu'il est utile de considérer la décomposition QR d'une matrice rectangulaire. Une telle décomposition requiert des matrices unitaires « incomplètes » que nous introduisons ici.

Définition 8.1 *On appelle matrice de Stiefel toute matrice $S \in \mathbb{C}^{m\times n}$ $(m \geqslant n)$ dont les n vecteurs-colonne constituent une famille orthonormée dans $\mathbb{C}^m$. On note $\mathbb{S}t_{mn}$ l'ensemble de ces matrices.*

Remarque 8.1. Une matrice unitaire est une matrice de Stiefel carrée : $\mathbb{S}t_{nn} = \mathbb{U}_n$.

Proposition 8.2 *$S \in \mathbb{C}^{m\times n}$ est une matrice de Stiefel si et seulement si $S^*S = I_n$. Dans ce cas, SS^* est la projection orthogonale de $\mathbb{C}^m$ sur le sous-espace* Im S *engendré par les vecteurs-colonne de S.*

Démonstration. La première assertion est évidente : S^*S est en effet la matrice dont les entrées sont les produits scalaires $\langle s_j, s_i\rangle$ où les s_i sont les colonnes de S. Pour prouver la seconde assertion (voir paragraphe 1.13 sur les projections orthogonales) on note que $SS^*s_i = s_i$ pour toute colonne s_i de S et que $SS^*x = 0$ pour tout $x \in (\text{Im } S)^\perp$. En effet cette dernière condition est équivalente à $S^*x = 0$.

8.2 DÉCOMPOSITION QR

Dans tout ce chapitre nous supposons que $m \geqslant n$.

Définition 8.3 *On appelle décomposition QR d'une matrice $A \in \mathbb{C}^{m\times n}$ avec $m \geqslant n$ une identité $A = QR$ où $Q \in \mathbb{S}t_{mn}$ est une matrice de Stiefel et où $R \in \mathbb{C}^{n\times n}$ est triangulaire supérieure.*

Il n'y a pas unicité de ce type de décomposition puisque, par exemple, $1 = z\bar{z}$ pour tout nombre complexe z de module 1. Toutefois on a :

Proposition 8.4 *Toute matrice $A \in \mathbb{C}^{m\times n}$ de rang n possède une et une seule décomposition $A = QR$ avec $r_{ii} > 0$ pour tout $i = 1 \ldots n$.*

Démonstration. Puisque A est de rang n la matrice $A^*A \in \mathbb{C}^{n\times n}$ est définie positive (théorème 7.2). Elle possède donc une décomposition de Cholesky $A^*A = R^*R$ (théorème 7.10) où $R \in \mathbb{C}^{n\times n}$ est triangulaire supérieure et à diagonale positive. Prenons $Q = AR^{-1}$. On a :

$$Q^*Q = R^{-*}A^*AR^{-1} = R^{-*}(R^*R)R^{-1} = I_n$$

ce qui prouve que $Q \in \mathbb{S}t_{mn}$ et que $A = QR$.
Pour prouver l'unicité on part d'une seconde décomposition : $A = Q'R'$. On a $A^*A = R'^*R' = R^*R$ donc $R' = R$ puisque la décomposition de Cholesky est unique d'où $Q = Q'$.

Remarque 8.2. On rencontre, dans la littérature consacrée à l'algèbre linéaire, deux définitions de la décomposition QR. La première est celle donnée ci-dessus $A = Q_1R_1$ avec $Q_1 \in \mathbb{S}t_{mn}$ et $R_1 \in \mathbb{C}^{n\times n}$ triangulaire supérieure. Une seconde définition est $A = Q_2R_2$ avec $Q_2 \in \mathbb{U}_m$ et $R_2 \in \mathbb{C}^{m\times n}$ triangulaire supérieure. Dans le premier cas Q_1 est rectangulaire et R_1 carrée, dans le second cas c'est l'inverse.

1. Les deux définitions coïncident lorsque $m = n$ c'est-à-dire lorsque A est carrée,
2. Le lien entre ces deux définitions est le suivant : $Q_1 = Q_2(1 : m, 1 : n)$ et $R_1 = R_2(1 : n, 1 : n)$,
3. Seule la première définition assure l'unicité de la décomposition.

Cette première définition apparaît naturellement avec la méthode de Gram-Schmidt, la seconde avec Givens et Householder. C'est la raison pour laquelle nous les conservons toutes deux.

Un des intérêts de la décomposition QR est que le conditionnement de la matrice A n'est pas détruit comme cela peut être le cas pour LU. Si $A = QR$ le système linéaire

$Ax = b$ est équivalent à $Rx = Q^*b$ qui est un système triangulaire. Par le théorème 5.4 A et R ont le même conditionnement $\text{cond}_2(A) = \text{cond}_2(R)$.

Nous allons maintenant décrire plusieurs procédés pour calculer cette décomposition.

8.3 L'ORTHONORMALISATION DE GRAM-SCHMIDT

8.3.1 Description du procédé

Soient $a_1, \ldots, a_n$ des vecteurs linéairement indépendants d'un espace hermitien (ou préhilbertien) $\mathbb{E}$. Le procédé d'*orthonormalisation de Gram-Schmidt* calcule n vecteurs $q_1, \ldots, q_n$ qui ont les deux vertus suivantes :

- Ils sont orthonormés,
- Pour tout i, les sous-espaces de $\mathbb{E}$ engendrés par $a_1, \ldots, a_i$ et $q_1, \ldots, q_i$ sont les mêmes.

On obtient un tel résultat par l'algorithme suivant :

Algorithme de Gram-Schmidt

$q_1 = a_1 / \|a_1\|_2$
pour $i = 1 : n - 1$
$\quad p_{i+1} = a_{i+1} - \sum_{k=1}^{i} \langle a_{i+1}, q_k \rangle q_k$
$\quad q_{i+1} = p_{i+1} / \|p_{i+1}\|_2$
fin

Ce procédé est à la base de la construction des polynômes orthogonaux. Notons que la base orthonormée obtenue dépend de l'ordre dans lequel sont pris les vecteurs a_i. Voir l'exercice 8.9 pour un procédé d'orthonormalisation indépendant de l'ordre des a_i.

Soit $A = (a_1 \ldots a_n) \in \mathbb{C}^{m \times n}$ une matrice de rang n dont les a_i sont ses vecteurs-colonne. L'orthonormalisation de Gram-Schmidt appliquée aux a_i donne la matrice $Q = (q_1 \ldots q_n) \in \mathbb{S}t_{mn}$ et la matrice $R \in \mathbb{C}^{n \times n}$ triangulaire supérieure à diagonale positive telles que $A = QR$ par l'algorithme qui suit :

Décomposition QR par l'algorithme de Gram-Schmidt

$r_{11} = \|a_1\|_2$
$q_1 = a_1 / \|a_1\|_2$
pour $i = 1 : n - 1$
    $p_{i+1} = a_{i+1}$
    pour $k = 1 : i$
        $r_{k\,i+1} = \langle a_{i+1}, q_k \rangle$
        $p_{i+1} = p_{i+1} - r_{k\,i+1} q_k$
    fin
    $r_{i+1\,i+1} = \|p_{i+1}\|_2$
    $q_{i+1} = p_{i+1} / r_{i+1\,i+1}$.
fin

8.3.2 Interprétation géométrique de G-S

Notons $Q_i \in \mathbb{S}t_{mi}$ la matrice de Stiefel constituée des i premières colonnes de Q : $q_1, \ldots, q_i$. L'identité

$$p_{i+1} = a_{i+1} - \sum_{k=1}^{i} \langle a_{i+1}, q_k \rangle \, q_k$$

montre que $p_{i+1} - a_{i+1} \in \text{Im } Q_i$ et que $p_{i+1} \in (\text{Im } Q_i)^{\perp}$, autrement dit p_{i+1} est la projection orthogonale de a_{i+1} sur $(\text{Im } Q_i)^{\perp}$. Ainsi

$$p_{i+1} = P_i a_{i+1}$$

où $P_i = I_m - Q_i Q_i^*$ est la matrice de la projection orthogonale sur $(\text{Im } Q_i)^{\perp}$ (voir le paragraphe 1.13). L'algorithme de Gram-Schmidt s'écrit alors

Algorithme de Gram-Schmidt (formulation géométrique)

$q_1 = a_1 / \|a_1\|_2$
pour $i = 1 : n - 1$
    $p_{i+1} = P_i a_{i+1}$
    $q_{i+1} = p_{i+1} / \|p_{i+1}\|_2$
fin

8.3.3 Complexité de G-S

La complexité de cet algorithme est donnée par le compte suivant du nombre d'opérations arithmétiques :

- $2m$ pour le calcul de r_{11},
- m pour le calcul de q_1,
- $2mi$ pour le calcul de $r_{k\,i+1}$,
- $2mi$ pour le calcul de p_{i+1},
- $2m$ pour le calcul de $r_{i+1\,i+1}$,
- m pour le calcul de q_{i+1}.

On obtient un total de

$$3m + \sum_{i=1}^{n-1}(4mi + 3m) \approx 2mn^2 \text{ opérations arithmétiques.}$$

8.3.4 Instabilité numérique de G-S

Donnons un exemple numérique. On prend une matrice A de Vandermonde 10×10 (voir le paragraphe 16.4) définie à partir de 10 points aléatoires. Le conditionnement de A est important : $\text{cond}_2(A) = 1.0022\,10^8$. Soit $b \in \mathbb{R}^{10}$ un vecteur aléatoire. On résout le système $Ax = b$ grâce à la décomposition QR de A : $x = R^{-1}Q^*b$. Les résultats numériques obtenus par l'algorithme de Gram-Schmidt sont les suivants ($\overline{x}$ désigne la solution approchée du système) :

- défaut d'orthonormalité $\|Q^*Q - I_{10}\|_2 = 7.1789\,10^{-4}$,
- norme de l'erreur $\|x - \overline{x}\|_2 = 71.7977$.

L'égalité $A = QR$ est par ailleurs satisfaite à l'ordre de l'unité d'arrondi $\approx 10^{-16}$. On constate une perte significative d'orthogonalité des vecteurs q_i et la solution du système calculée grâce à cette décomposition est très différente de la solution exacte.

> ***Remarque 8.3.*** La raison de cette instabilité numérique peut être comprise en raisonnant sur deux vecteurs a_1 et a_2. Le procédé de Gram-Schmidt va générer $q_1 = a_1/\|a_1\|_2$ puis le vecteur $p_2 = a_2 - \langle a_2, q_1\rangle q_1$ et enfin $q_2 = p_2/\|p_2\|_2$. Lorsque a_1 et a_2 sont deux vecteurs quasi proportionnels, les vecteurs a_2 et $\langle a_2, q_1\rangle q_1$ sont à peu près égaux, leur différence sera donc petite et q_2 sera obtenu comme quotient de deux petites quantités d'où des instabilités numériques.

Prenons $a_1 = (-0.88061, -0.47384)^T$, $a_2 = (-0.881, -0.474)^T$ et calculons avec cinq décimales. On obtient : $q_1 = a_1$, $\langle a_2, q_1\rangle = 1.0004$, $p_2 = (-0.00004, 0.00003)^T$ ce qui conduit à un angle $(\widehat{q_1,\ p_2}) = 66.4$ degrés bien loin des 90 degrés souhaités.

8.3.5 L'algorithme de Gram-Schmidt modifié

On peut remédier en partie à ce défaut en ordonnant différemment le calcul des vecteurs p_i. On décompose l'opérateur de projection orthogonale P_i de la manière suivante :

$$P_i = I_m - Q_i Q_i^* = (I_m - q_i q_i^*) \ldots (I_m - q_1 q_1^*).$$

Il est facile de prouver que ces deux expressions de P_i sont égales. Cependant, du fait des erreurs d'arrondi, les deux formes ne sont pas numériquement équivalentes. Ce nouveau calcul donne lieu à l'algorithme de Gram-Schmidt modifié :

Algorithme de Gram-Schmidt modifié

$q_1 = a_1 / \|a_1\|_2$
pour $i = 1 : n - 1$
    $z = a_{i+1}$
    pour $k = 1 : i$
        $z = z - \langle z, q_k \rangle q_k$
    fin
    $q_{i+1} = z / \|z\|_2$
fin

L'algorithme de Gram-Schmidt modifié appliqué au système considéré au paragraphe 8.3.4 donne le résultat suivant :

- défaut d'orthonormalité $\|Q^* Q - I_{10}\|_2 = 1.5229\, 10^{-9}$,
- norme de l'erreur $\|x - \overline{x}\|_2 = 0.0970$.

L'orthogonalité des vecteurs est mieux satisfaite que par l'algorithme de Gram-Schmidt et la solution du système meilleure, bien qu'encore assez médiocre. On verra au paragraphe 8.5.3 que la décomposition QR obtenue grâce à la méthode de Householder donne de meilleurs résultats.

8.3.6 Procédé de réorthogonalisation

Une autre méthode fréquemment utilisée pour améliorer l'orthogonalité des colonnes de Q consiste à appliquer une seconde fois l'algorithme de Gram-Schmidt à la matrice Q que l'on vient de calculer et dont on a vu qu'elle ne vérifiait qu'imparfaitement l'égalité $Q^* Q = I_n$. On obtient la décomposition $Q = Q' R'$ et donc $A = Q'(R' R)$. La matrice $R' R$ est triangulaire supérieure et Q' est la matrice de Stiefel de cette

nouvelle décomposition QR de A. En général les colonnes de Q' vérifient les relations d'orthogonalité avec une bonne précision et il n'est par nécessaire de poursuivre au-delà ce processus.

La complexité des calculs de cette méthode a plus que doublé par rapport à l'algorithme de Gram-Schmidt puisqu'il faut aussi considérer le produit des matrices triangulaires supérieures $R'R$.

Donnons une illustration de l'efficacité de cette méthode pour le système numérique déjà considéré :

- défaut d'orthonormalité $\|Q'^* Q' - I_{10}\|_2 = 6.7759\, 10^{-16}$,
- norme de l'erreur $\|x - \overline{x}\|_2 = 1.1819\, 10^{-11}$.

Les résultats sont meilleurs que ceux obtenus par les transformations de Householder (voir paragraphe 8.5.3).

8.4 ROTATIONS DE GIVENS

Nous n'allons considérer, dans ce paragraphe, que des matrices réelles, le cas complexe fait l'objet de l'exercice 8.8. Une *rotation de Givens* consiste en une rotation d'angle $-\theta$ dans le plan $Ox_i x_j$. Elle est donnée par $G(i, j, \theta)x = y$ avec

$$G(i,j,\theta) = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\theta & & \sin\theta & & & \\ & & & & 1 & & & & \\ & & & \vdots & & \ddots & & \vdots & \\ & & & & & & 1 & & \\ & & & -\sin\theta & & \cos\theta & & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow i \\ \\ \\ \\ \leftarrow j \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & & i & & & j & & \\ \end{matrix}$$

c'est-à-dire

$$y_k = \begin{cases} x_k & k \neq i, j, \\ x_i \cos\theta + x_j \sin\theta & k = i, \\ -x_i \sin\theta + x_j \cos\theta & k = j. \end{cases}$$

Il est clair qu'il s'agit là d'une transformation orthogonale. On peut forcer y_j à être nul en prenant

$$\cos\theta = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} \text{ et } \sin\theta = \frac{x_j}{\sqrt{x_i^2 + x_j^2}}.$$

Une succession de telles rotations permet d'appliquer un vecteur $a \in \mathbb{R}^m$ sur le vecteur

$$\begin{pmatrix} \pm \|a\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Il suffit, pour ce faire, d'effectuer le produit suivant de rotations de Givens :

$$G(1,2,\theta_2)G(2,3,\theta_3)\ldots G(m-1,m,\theta_m)a$$

l'angle θ_k est choisi de façon à annuler la k-ième composante du vecteur

$$G(k,k+1,\theta_{k+1})\ldots G(m-1,m,\theta_n)a$$

comme indiqué dans l'algorithme suivant :

Algorithme de Givens pour la transformation d'un vecteur

pour $i = m : -1 : 2$
  $r = \sqrt{a_{i-1}^2 + a_i^2}$
  $a_{i-1} = r$
  $a_i = 0$
fin

Notons enfin que l'on peut contrôler le signe de la première coordonnée $\pm \|a\|_2$ en remplaçant θ par $\pi + \theta$.

Pour obtenir la décomposition QR d'une matrice $A \in \mathbb{R}^{m\times n}$ on applique la méthode précédente aux différentes colonnes de A comme indiqué dans le schéma suivant :

$$\begin{pmatrix} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{pmatrix} \xrightarrow{G(3,4)} \begin{pmatrix} \times & \times \\ \times & \times \\ \times & \times \\ 0 & \times \end{pmatrix} \xrightarrow{G(2,3)} \begin{pmatrix} \times & \times \\ \times & \times \\ 0 & \times \\ 0 & \times \end{pmatrix} \xrightarrow{G(1,2)} \begin{pmatrix} \times & \times \\ 0 & \times \\ 0 & \times \\ 0 & \times \end{pmatrix} \xrightarrow{G(3,4)}$$

$$\begin{pmatrix} \times & \times \\ 0 & \times \\ 0 & \times \\ 0 & 0 \end{pmatrix} \overset{G(2,3)}{\rightarrow} \begin{pmatrix} \times & \times \\ 0 & \times \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Noter que l'ordre choisi dans la composition des rotations de Givens ne perturbe pas les zéros déjà acquis. L'algorithme correspondant est le suivant :

Algorithme de Givens

pour $j = 1 : n - 1$
  pour $i = m : -1 : j - 1$
    $r = \sqrt{a_{i-1\,j}^2 + a_{ij}^2}$
    $c = a_{i-1\,j}/r$
    $s = a_{ij}/r$
    $a_{i-1\,j} = r$
    $a_{ij} = 0$
    pour $k = j + 1 : n$
      $a_{i-1\,k} = ca_{i-1\,k} + sa_{ik}$
      $a_{ik} = -sa_{i-1\,k} + ca_{ik}$
    fin
  fin
fin

Ceci montre que

Théorème 8.5 *Toute matrice $A \in \mathbb{R}^{m\times n}$ possède une décomposition $A = QR$ avec $R \in \mathbb{R}^{m\times n}$ triangulaire supérieure et $Q \in \mathbb{O}_m$, produit d'au plus $n(2m - n - 1)/2$ rotations de Givens. L'algorithme de Givens calcule la matrice R en $\approx 3mn^2 - n^3$ opérations arithmétiques.*

Démonstration. Le calcul de r, c et s compte pour 6 opérations, puis 6 autres opérations dans la boucle k d'où un total de $6(n - j) + 6$ ops. La boucle sur i puis celle sur j conduisent à un total de

$$\sum_{j=1}^{n-1} (6(n - j) + 6)(m - j + 2) \approx 3mn^2 - n^3.$$

8.5 LA MÉTHODE DE HOUSEHOLDER

8.5.1 Symétries orthogonales

Définition 8.6 *Soit $w \in \mathbb{C}^m$ un vecteur unitaire c'est-à-dire tel que*

$$\|w\|_2^2 = \langle w, w \rangle = w^* w = 1.$$

On appelle matrice de Householder associée à w

$$H_w = I_m - 2ww^*.$$

Théorème 8.7 *H_w est la symétrie orthogonale par rapport à l'hyperplan*

$$\mathcal{H}_w = \{u \in \mathbb{C}^m \ : \ \langle u, w \rangle = 0\}.$$

Elle est hermitienne et unitaire : $H_w = H_w^* = H_w^{-1}$.

Démonstration. $H_w(w) = (I_m - 2ww^*)w = w - 2w(w^*w) = -w$ parce que $w^*w = 1$. De plus, pour tout $u \in \mathcal{H}_w$, $H_w(u) = (I_m - 2ww^*)u = u - 2w(w^*u) = u$ puisque $w^*u = \langle u, w \rangle = 0$. Ceci prouve que H_w est bien la symétrie orthogonale par rapport à l'hyperplan $\mathcal{H}_w$. Notons enfin que $H_w^* = (I_m - 2ww^*)^* = I_m - 2ww^* = H_w$ et $H_w^* H_w = (I_m - 2ww^*)^2 = I_m - 4ww^* + 4ww^*ww^* = I_m$ parce que $w^*w = 1$. Ceci prouve que la symétrie orthogonale par rapport à l'hyperplan $\mathcal{H}_w$ est hermitienne et unitaire.

Étant donné deux vecteurs dans $\mathbb{R}^m$ de même longueur, on peut les rabattre l'un sur l'autre par une symétrie orthogonale. En termes plus imagés, l'un est « l'image miroir » de l'autre. Pour des vecteurs complexes cela n'est vrai qu'à un changement d'argument près :

Théorème 8.8 *Soient u_1 et $u_2 \in \mathbb{C}^m$ de mêmes normes et indépendants. Il existe $w \in \mathbb{C}^m$, unitaire, et un nombre complexe θ de module 1 tels que*

$$H_w(u_1) = \theta u_2.$$

Démonstration. On prend

$$\theta = \exp(-i \arg(u_1^* u_2))$$

et

$$w = \frac{u_1 - \theta u_2}{\|u_1 - \theta u_2\|_2}.$$

L'indépendance de u_1 et u_2 fait que $u_1 - \theta u_2 \neq 0$. On a : $H_w(u_1) =$

$$u_1 - 2\frac{(u_1 - \theta u_2)(u_1 - \theta u_2)^*}{(u_1 - \theta u_2)^*(u_1 - \theta u_2)}u_1 = u_1 - (u_1 - \theta u_2)\frac{u_1^* u_1 - \bar{\theta} u_2^* u_1}{u_1^* u_1 - \Re(\theta u_1^* u_2)}.$$

Le choix de θ fait de $\theta u_1^* u_2$ un nombre réel positif donc

$$\Re(\theta u_1^* u_2) = \theta u_1^* u_2 = \bar{\theta} u_2^* u_1$$

de sorte que

$$H_w(u_1) = u_1 - (u_1 - \theta u_2) = \theta u_2.$$

Lorsque les vecteurs u_1 et u_2 ne sont pas de même norme, il suffit de considérer le théorème précédent avec $u_1/\|u_1\|_2$ et $u_2/\|u_2\|_2$. On obtient le résultat suivant

Corollaire 8.9 *Soient u_1 et $u_2 \in \mathbb{C}^m$ indépendants. Il existe $w \in \mathbb{C}^m$ unitaire et un nombre complexe $k \neq 0$ tels que*

$$H_w(u_1) = ku_2.$$

Le cas réel se traite de façon similaire :

Théorème 8.10 *Soient u_1 et $u_2 \in \mathbb{R}^m$ de mêmes normes et indépendants. Posons*

$$w_+ = \frac{u_1 + u_2}{\|u_1 + u_2\|_2}$$

et

$$w_- = \frac{u_1 - u_2}{\|u_1 - u_2\|_2}.$$

Alors

$$H_{w_-}(u_1) = u_2 \text{ et } H_{w_+}(u_1) = -u_2.$$

Corollaire 8.11 *Soient u_1 et $u_2 \in \mathbb{R}^m$ indépendants. Il existe $w \in \mathbb{R}^m$, unitaire, et un nombre réel positif (resp. négatif) k tels que*

$$H_w(u_1) = ku_2.$$

8.5.2 QR via Householder

La méthode de Householder pour calculer la décomposition QR d'une matrice $A \in \mathbb{C}^{m\times n}$ repose sur la construction suivante :

Théorème 8.12 *Il existe une matrice unitaire $Q \in \mathbb{U}_m$ produit d'au plus n matrices de Householder ($n - 1$ matrices si $m = n$) et une matrice triangulaire supérieure $R \in \mathbb{C}^{m\times n}$ à diagonale positive ou nulle telles que $A = QR$.*

Démonstration. Soit a_1 le premier vecteur-colonne de A et soit e_1 le premier vecteur de la base canonique de $\mathbb{C}^m$: $e_1 = (1, 0, \ldots, 0)^T$. Si $a_1 = ke_1$ il n'y a rien à faire et l'on pose $H_1 = I_m$. Sinon, en vertu du corollaire 8.9, il existe une matrice de Householder H_1 et un scalaire $k_1 \neq 0$ tels que

$$H_1 A = \begin{pmatrix} k_1 & a'_{12} & \ldots & a'_{1n} \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}.$$

Après j telles étapes on obtient

$$H_j \ldots H_1 A = \begin{pmatrix} k_1 & & & \ldots \\ 0 & \ddots & & \\ & & k_j & \ldots \\ \vdots & & 0 & \\ & & \vdots & A_j \\ 0 & & 0 & \end{pmatrix}$$

où A_j est une matrice $(m-j)\times(n-j)$. Appliquons à A_j la même procédure que pour A : notons $\tilde{a}_{j+1} \in \mathbb{C}^{m-j}$ la première colonne de A_j et $\tilde{e}_1 \in \mathbb{C}^{m-j}$ le premier vecteur de la base canonique de $\mathbb{C}^{m-j}$. Si $\tilde{a}_{j+1}$ est proportionnel à $\tilde{e}_1$ on pose $H_{j+1} = I_m$ et on passe à l'étape suivante. Sinon, on calcule une matrice de Householder $\tilde{H}_{j+1} \in \mathbb{C}^{(m-j)\times(m-j)}$ et un scalaire $k_{j+1} \in \mathbb{C}$ tels que

$$\tilde{H}_{j+1}\tilde{a}_{j+1} = k_{j+1}\tilde{e}_1.$$

À la matrice de Householder $\tilde{H}_{j+1} \in \mathbb{C}^{(m-j)\times(m-j)}$ nous associons la matrice $H_{j+1} \in \mathbb{C}^{m\times m}$ donnée par

$$H_{j+1} = \begin{pmatrix} I_j & 0 \\ 0 & \tilde{H}_{j+1} \end{pmatrix}.$$

Il est facile de voir que lorsque $\tilde{H}_{j+1}$ est la matrice de Householder associée au vecteur $\tilde{w} \in \mathbb{C}^{m-j}$, H_{j+1} est la matrice de Householder associée au vecteur $w \in \mathbb{C}^m$ tel que

$$w = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{w} \end{pmatrix}.$$

La multiplication à gauche de $H_j \dots H_1 A$ par H_{j+1} ne modifie que le bloc A_j, qui est transformé en $\tilde{H}_{j+1} A_j$, et laisse inchangé le reste de la structure d'où

$$H_{j+1} H_j \dots H_1 A = \begin{pmatrix} k_1 & & & \dots \\ 0 & \ddots & & \\ & & k_{j+1} & \dots \\ \vdots & & 0 & \\ & & \vdots & A_{j+1} \\ 0 & & 0 & \end{pmatrix}.$$

Le processus s'arrête après n telles étapes. On obtient alors une matrice triangulaire supérieure $H_n \dots H_1 A = R$ d'où $A = QR$ avec $Q = H_1 \dots H_n$. Lorsque $m = n$ la dernière étape n'est plus nécessaire parce que la matrice A_n est de taille 1×1 donc déjà triangulaire supérieure ; $n - 1$ matrices de Householder suffisent.

Remarque 8.4. Par construction des matrices de Householder H_j, les différentes colonnes q_j de la matrice $Q = H_1 \dots H_n$ sont données par $q_j = Qe_j = H_1 \dots H_j e_j$.

8.5.3 L'algorithme et sa complexité

La triangulation d'un système via la méthode de Householder est résumée dans l'algorithme suivant que nous donnons pour des matrices réelles. La matrice R obtenue a une diagonale positive ou nulle.

Décomposition QR par la méthode de Householder

pour $j = 1 : n$
  $\alpha^2 = \sum_{i=j+1}^{m} |a_{ij}|^2$
  si $\alpha \neq 0$
    $\beta = \sqrt{|a_{jj}|^2 + \alpha^2}$
    $w_j = a_{jj} - \beta$
    pour $i = j+1 : m$
      $w_i = a_{ij}$
    fin
    $H = I_{m-j+1} - 2\dfrac{ww^T}{\|w\|_2^2}$
    $A(j : m, j : n) = HA(j : m, j : n)$
  fin
fin

Le calcul du produit de matrices $HA(j : m, j : n)$ demande en principe $2(m - j + 1)^2(n - j + 1)$ opérations arithmétiques. Compte tenu de la structure particulière de H, on peut l'effectuer en $\approx 4(m - j + 1)(n - j + 1)$ opérations. Il suffit de procéder de la façon suivante :

1. On calcule le produit vecteur ligne-matrice $w^T A(j : m, j : n)$: ceci nécessite $\approx 2(m - j + 1)(n - j + 1)$ opérations,
2. On évalue $v = -2\frac{w}{\|w\|^2}$ puis $v\left(w^T A(j : m, j : n)\right)$ ce qui demande $\approx (m - j + 1)(n - j + 1)$,
3. On ajoute le résultat obtenu à $A(j : m, j : n)$ d'où $(m - j + 1)(n - j + 1)$ opérations supplémentaires,
4. Comme $1 \leqslant j \leqslant n$ on obtient au total (les opérations oubliées dans ce compte ont un ordre de grandeur inférieur)

$$\sum_{j=1}^{n} 4(m - j + 1)(n - j + 1) \approx 2mn^2 - \frac{2}{3}n^3 \text{ opérations arithmétiques.}$$

Cet algorithme est donc moins coûteux que l'algorithme de Gram-Schmidt ($2mn^2$) mais il faut noter que ce dernier calcule les matrices Q et R à la fois alors que les algorithmes de Givens et de Householder ne fournissent que la matrice R.

Une évaluation possible de Q utilise la remarque 8.4 : les différentes colonnes q_j de la matrice $Q = H_1 \ldots H_n$ sont données par $q_j = Qe_j = H_1 \ldots H_j e_j$. Ce calcul

requiert $\approx 2mn^2 - \frac{2}{3}n^3$ opérations arithmétiques, mais il suppose que l'on stocke les matrices $H_1 \ldots H_n$.

Une dernière possibilité consiste à effectuer le produit $Q^* = H_n \ldots H_1$ ce qui évite de conserver les différentes matrices de Householder mises en oeuvre. Cette stratégie coûte $\approx 4(m^2n - mn^2 + n^3/3)$ opérations arithmétiques.

Considérons à nouveau l'exemple du paragraphe 8.3.4 en utilisant la décomposition QR par les matrices de Householder (fonction qr de Matlab). On obtient le résultat suivant :

- défaut d'orthonormalité $\|Q^*Q - I_{10}\|_2 = 1.1917\,10^{-15}$,
- norme de l'erreur $\|x - \overline{x}\|_2 = 1.4980\,10^{-8}$.

On constate l'excellente précision numérique des calculs obtenus avec la méthode de Householder.

8.6 RÉDUCTION À LA FORME DE HESSENBERG

Définition 8.13 *Une matrice $H \in \mathbb{C}^{n\times n}$ est dite de Hessenberg lorsque $h_{ij} = 0$ pour tout $i > j + 1$.*

Une matrice de Hessenberg H est de la forme

$$H = \begin{pmatrix} \times & \cdots & \cdots & \cdots & \times \\ \times & \ddots & & & \vdots \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \times & \times \end{pmatrix}. \tag{8.1}$$

On va montrer que toute matrice A est unitairement semblable à une matrice de Hessenberg. Cette décomposition peut être obtenue grâce aux matrices de Householder et de Givens déjà considérées pour la décomposition QR. Nous donnons ici la décomposition par les matrices de Householder.

Théorème 8.14 *Étant donné une matrice $A \in \mathbb{C}^{n\times n}$, il existe une matrice unitaire Q produit d'au plus $n-2$ matrices de Householder telle que Q^*AQ soit une matrice de Hessenberg.*

Démonstration. Si la première colonne a_1 de A est du type $a_1 = (a_{11}, a_{21}, 0, \ldots, 0)^T$ alors on pose $H_2 = I_n$. Sinon, les vecteurs $\tilde{a}_2 = (a_{21}, \ldots, a_{n1})^T$ et $\tilde{e}_1 = (1, 0, \ldots, 0)^T \in \mathbb{C}^{n-1}$ sont indépendants

donc, par le corollaire 8.9, il existe une matrice de Householder $\tilde{H}_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ et un scalaire k_1 tels que

$$\tilde{H}_2 \tilde{a}_2 = k_1 \tilde{e}_1.$$

On pose

$$H_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_2 \end{pmatrix}$$

qui, comme nous l'avons vu, est aussi une matrice de Householder. On a

$$H_2 A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ k_1 & & & & \\ 0 & & & & \\ \vdots & & & A_1 & \\ 0 & & & & \end{pmatrix}.$$

La multiplication à droite de $H_2 A$ par H_2^* ne modifie pas la première colonne de $H_2 A$. On obtient la matrice

$$H_2 A H_2^* = \begin{pmatrix} a_{11} & \tilde{a}_{12} & \tilde{a}_{13} & \dots & \tilde{a}_{1n} \\ k_1 & & & & \\ 0 & & & & \\ \vdots & & & \tilde{A}_1 & \\ 0 & & & & \end{pmatrix}$$

On recommence ce même procédé sur la matrice $\tilde{A}_1$ et après $n-2$ étapes on obtient

$$H_{n-1} \dots H_2 A H_2^* \dots H_{n-1}^* = H$$

où H est de Hessenberg (à ne pas confondre avec les matrices de Householder H_j). Il suffit de poser $Q = H_2^* \dots H_{n-1}^* = H_2 \dots H_{n-1}$ pour avoir la décomposition de Hessenberg $A = QHQ^*$.

La forme algorithmique du théorème 8.14 est donnée ici pour des matrices réelles :

Forme Hessenberg par la méthode de Householder

pour $j = 1 : n-2$
  $\alpha^2 = \sum_{i=j+2}^{n} |a_{ij}|^2$
  si $\alpha \neq 0$
    $\beta = \sqrt{|a_{j+1\,j}|^2 + \alpha^2}$
    $w_{j+1} = a_{j+1\,j} - \beta$
    pour $i = j+2 : n$
      $w_i = a_{ij}$
    fin
    $H = I_{n-j} - 2\dfrac{ww^T}{\|w\|_2^2}$
    $A(j+1 : n, j : n) = HA(j+1 : n, j : n)$
    $A(1 : n, j+1 : n) = A(1 : n, j+1 : n)H$
  fin
fin

Le décompte des opérations est le suivant :

- $\approx \frac{4}{3}n^3$ opérations pour le produit des matrices $HA(j+1 : n, j : n)$ (voir paragraphe 8.5.3),

- $\sum_{j=1}^{n-2} 4n(n-j) \approx 2n^3$ opérations pour le produit des matrices $A(1 : n, j+1 : n)H$,

soit $\approx \frac{10}{3}n^3$ opérations pour le calcul de la matrice de Hessenberg. Il faut ajouter à cela $\approx \frac{4}{3}n^3$ opérations pour le calcul de la matrice Q (voir paragraphe 8.5.3).

Remarque 8.5.

1. La matrice obtenue par cet algorithme est du type Hessenberg et, lorsque le cas $\alpha = 0$ ne s'est jamais présenté dans son déroulement, les entrées $h_{i+1\,i}$ sont toutes positives.
2. La décomposition de Hessenberg peut aussi s'obtenir à l'aide des matrices de Givens.

8.7 TRIDIAGONALISATION D'UNE MATRICE HERMITIENNE

Définition 8.15 *Une matrice T est tridiagonale* [1] *si $T_{ij} = 0$ lorsque $|i - j| > 1$.*

La décomposition de Hessenberg d'une matrice hermitienne conduit à une matrice (hermitienne) tridiagonale. Nous donnons ici le procédé de tridiagonalisation qui utilise les matrices de Householder.

Théorème 8.16 *Étant donné une matrice $A \in \mathbb{C}^{n\times n}$ hermitienne, il existe une matrice unitaire Q produit d'au plus $n-2$ matrices de Householder telle que QAQ^* soit une matrice tridiagonale hermitienne.*

Démonstration. On procède comme dans la preuve du théorème 8.14. Le résultat de la multiplication à gauche de A par la matrice de Householder H_2 ne modifie pas la première ligne de A. Puisque la matrice A est hermitienne cette ligne est la conjuguée de la première colonne de A et $a_{11} = \overline{a}_{11}$. On a

$$H_2A = \begin{pmatrix} a_{11} & \overline{a}_{21} & \overline{a}_{31} & \dots & \overline{a}_{n1} \\ k_1 & & & & \\ 0 & & & & \\ \vdots & & & A_1 & \\ 0 & & & & \end{pmatrix}.$$

En multipliant à droite par $H_2^* = H_2$ on a donc

$$H_2AH_2^* = \begin{pmatrix} a_{11} & \overline{k_1} & 0 & \dots & 0 \\ k_1 & & & & \\ 0 & & & & \\ \vdots & & & \tilde{A}_1 & \\ 0 & & & & \end{pmatrix}$$

où $\tilde{A}_1$ est elle-même hermitienne. On recommence avec $\tilde{A}_1$ ce qui vient d'être fait avec A et après $n-2$ telles étapes on a :

$$H_{n-2}\dots H_2AH_2^*\dots H_{n-2}^* = T$$

tridiagonale hermitienne. Il suffit de poser $Q = H_2^*\dots H_{n-2}^* = H_2\dots H_{n-2}$ pour avoir la décomposition $A = QTQ^*$.

Cette décomposition s'obtient également grâce aux transformations de Givens.

1. D'après la définition générale (voir exercice 6.5) une matrice tridiagonale est une matrice bande de largeur 3.

8.8 L'ALGORITHME D'ARNOLDI

La décomposition $A = QHQ^*$ sous la forme de Hessenberg n'est pas unique. Déterminons le nombre de variables réelles indépendantes d'une matrice $Q \in \mathbb{U}_n$ unitaire et d'une matrice $H \in \mathbb{C}^{n\times n}$ de Hessenberg.
Pour $Q \in \mathbb{U}_n$ nous obtenons le décompte suivant :

- $\dim_{\mathbb{R}} \mathbb{C}^{n\times n} = 2n^2$,
- $n(n-1)/2$ conditions d'orthogonalité et donc $n(n-1)$ équations réelles,
- n conditions de normalité et donc n équations réelles.

Cette heuristique « montre que » $\dim_{\mathbb{R}} \mathbb{U}_n = 2n^2 - n(n-1) - n = n^2$.[2] Pour une matrice $H \in \mathbb{C}^{n\times n}$ de Hessenberg, on a $(1+2+\ldots+n)+(n-1)$ coefficients complexes et donc n^2+3n-2 coefficients réels. Sachant que $A \in \mathbb{C}^{n\times n}$ admet $2n^2$ coefficients réels, nous disposons donc de $3n-2$ paramètres réels arbitraires. Cette possibilité est utilisée par l'algorithme d'Arnoldi qui calcule une décomposition de Hessenberg de A ayant la première colonne de Q fixée ($2n-1$ paramètres réels) et les arguments des coefficients $h_{j+1\,j}$ fixés, par exemple $h_{j+1\,j} > 0$ ($n-1$ paramètres réels).

Considérons un vecteur unitaire $q_1 \in \mathbb{C}^n$. Nous souhaitons construire une matrice unitaire $Q = (q_1 q_2 \ldots q_n)$ ayant q_1 pour première colonne et une matrice H de Hessenberg à sous-diagonale positive ($h_{i+1\,i} > 0$ pour tout $i = 1 \ldots n-1$) telles que $A = QHQ^*$, c'est-à-dire

$$AQ = QH. \tag{8.2}$$

La première colonne de (8.2) montre que

$$Aq_1 = h_{11}q_1 + h_{21}q_2$$

on a donc $h_{21}q_2 = Aq_1 - h_{11}q_1$. La condition d'orthogonalité $\langle q_1, q_2\rangle = 0$ permet de déterminer

$$h_{11} = \langle Aq_1, q_1\rangle = q_1^* Aq_1.$$

Par ailleurs, la condition $\|q_2\|_2 = 1$ implique

$$|h_{21}| = \|Aq_1 - h_{11}q_1\|_2.$$

Si $Aq_1 - h_{11}q_1 \neq 0$ est distinct de zéro, on peut alors choisir

$$h_{21} = \|Aq_1 - h_{11}q_1\|_2 > 0$$

2. Il s'agit là de la dimension de $\mathbb{U}_n$ en tant que sous-variété différentiable de $\mathbb{R}^{2n^2}$. Nous admettons ici que les relations d'orthonormalité $Q^*Q = I_n$ sont indépendantes.

et

$$q_2 = (Aq_1 - h_{11}q_1)/h_{21}.$$

On aurait pu tout aussi bien prendre $h_{21} = z\|Aq_1 - h_{11}q_1\|_2$ avec $z \in \mathbb{C}$ de module 1.

Supposons avoir construit les j premières colonnes de Q et de H avec les conditions requises. La relation (8.2) appliquée à la colonne j donne

$$Aq_j = \sum_{i=1}^{j+1} h_{ij}q_i$$

et l'on obtient

$$h_{j+1\,j}\,q_{j+1} = Aq_j - \sum_{i=1}^{j} h_{ij}q_i. \tag{8.3}$$

Puisque les vecteurs q_i $(i = 1, \ldots, j)$ sont orthonormés, les coefficients h_{ij} $(i = 1, \ldots, j)$ sont déterminés par les contraintes d'orthogonalité $\langle q_{j+1}, q_i\rangle = 0$, ce qui donne

$$h_{ij} = \langle Aq_j, q_i\rangle = q_i^* Aq_j.$$

Si $Aq_j - \sum_{i=1}^{j} h_{ij}q_i \neq 0$ alors on choisit

$$h_{j+1\,j} = \|Aq_j - \sum_{i=1}^{j} h_{ij}q_i\|_2 > 0$$

et

$$q_{j+1} = (Aq_j - \sum_{i=1}^{j} h_{ij}q_i)/h_{j+1\,j}.$$

Si le processus se poursuit jusqu'à l'étape $n-1$, c'est-à-dire si $Aq_j - \sum_{i=1}^{j} h_{ij}q_i \neq 0$ pour $j = 1, \ldots, n-1$, alors $(q_1, \ldots, q_n)$ est une base orthonormée de $\mathbb{C}^n$. Dans cette base, écrivons $Aq_n = \sum_{i=1}^{n} h_{in}q_i$. Les coefficients h_{in} sont encore donnés par $h_{in} = q_i^* Aq_n$ et la matrice H est enfin déterminée. On a ainsi défini l'algorithme d'Arnoldi :

Algorithme d' Arnoldi

Entrée : $A, H \in \mathbb{C}^{n\times n}$, $H = 0$, $q_1 \in \mathbb{C}^n$ de norme 1

pour $j = 1 : n-1$
  $z = Aq_j$
  $p_{j+1} = z$
  pour $i = 1 : j$
    $h_{ij} = q_i^* z$
    $p_{j+1} = p_{j+1} - h_{ij} q_i$
  fin
  $h_{j+1\,j} = \|p_{j+1}\|_2$
  si $h_{j+1\,j} = 0$
    $k = j$
    stop
  fin
  $q_{j+1} = p_{j+1}/h_{j+1\,j}$
fin
  $k = n$
pour $i = 1 : n$
  $h_{in} = q_i^* A q_n$
fin

Sortie : k, $Q_k = (q_1 \dots q_k) \in \mathbb{S}t_{nk}$, $H_k \in C^{k\times k}$ de Hessenberg avec $h_{j+1\,j} > 0$.

D'une manière générale, lorsque l'algorithme s'arrête

- Si $k \leqslant n-1$ et $h_{k+1\,k} = 0$, on a $AQ_k = Q_k H_k$ avec $Q_k \in \mathbb{S}t_{nk}$, $H_k \in \mathbb{C}^{k\times k}$ de Hessenberg et $AQ_j = Q_{j+1}\tilde{H}_j$ pour tout $j < k$ avec $\tilde{H}_j = H_k(1 : j+1, 1 : j)$. Les colonnes de Q_k engendrent un sous-espace de $\mathbb{C}^n$, de dimension k, invariant par A.

- Si $k = n-1$ et $h_{k+1\,k} \neq 0$, on a obtenu à la fin de la boucle k l'égalité $AQ_{n-1} = Q_n \tilde{H}_{n-1}$ avec $\tilde{H}_{n-1} \in \mathbb{C}^{n\times(n-1)}$ que l'on transforme en $AQ_n = Q_n H_n$ au travers de la dernière boucle de l'algorithme.

Nous résumons les propriétés que nous venons de décrire dans la

Proposition 8.17 *L'algorithme d'Arnoldi calcule un entier $k \leqslant n$, une matrice de Stiefel $Q_k \in \mathbb{S}t_{nk}$ et une matrice de Hessenberg $H_k \in \mathbb{C}^{k\times k}$ telles que $AQ_k = Q_k H_k$*

et $h_{j+1\,j} > 0$ *pour tout* $j = 1, \ldots, k-1$. *Les colonnes* q_j *de* Q_k *engendrent un sous-espace de dimension* k *de* $\mathbb{C}^n$ *invariant par* A. *On a les égalités*

$$\begin{aligned} AQ_j &= Q_{j+1}\tilde{H}_j \\ AQ_j &= Q_jH_j + h_{j+1\,j}\,q_{j+1}e_j^T \\ Q_j^*AQ_j &= H_j \end{aligned} \tag{8.4}$$

pour tout $j < k$, *en notant* $Q_j = Q_k(1:n, 1:j)$, $H_j = H_k(1:j, 1:j)$, $\tilde{H}_j = H_k(1:j+1, 1:j)$ *et* e_j *le* j*-ième vecteur de base de* $\mathbb{R}^j$.

Cet algorithme est un cas particulier d'algorithme de Gram-Schmidt : il s'agit de l'orthonormalisation de la base $(q_1, Aq_1, Aq_2, \ldots, Aq_{k-1})$ où chaque q_j est construit grâce aux vecteurs $q_1, \ldots, q_{j-1}$ et Aq_{j-1}. À ce titre on observe les mêmes défauts numériques que dans l'algorithme de Gram-Schmidt : lorsque j croît, les vecteurs q_j calculés ne vérifient pas la contrainte d'orthogonalité de manière satisfaisante. On modifie les calculs de la même façon que dans l'algorithme de Gram-Schmidt (paragraphe 8.3.5). L'égalité (8.3) s'écrit

$$h_{j+1\,j}\,q_{j+1} = P_jAq_j$$

où $P_j = (I_n - Q_jQ_j^*)$ et $Q_j = (q_1 \ldots q_j) \in \mathbb{S}t_{nj}$. Le projecteur orthogonal P_j admet la décomposition

$$P_j = \Pi_{i=1}^{j}(I_n - q_iq_i^*)$$

et les matrices $(I_n - q_iq_i^*)$ commutent entre elles. Nous obtenons ainsi l'algorithme d'Arnoldi modifié.

Algorithme d' Arnoldi modifié

Entrée : $A, H \in \mathbb{C}^{n\times n}$, $H = 0$, $q_1 \in \mathbb{C}^n$ de norme 1

pour $j = 1 : n-1$
    $z = Aq_j$
    pour $i = 1 : j$
        $h_{ij} = q_i^* z$
        $z = z - h_{ij} q_i$
    fin
    $h_{j+1\,j} = \|z\|_2$
    si $h_{j+1\,j} = 0$
        $k = j$
        stop
    fin
    $q_{j+1} = z/h_{j+1\,j}$
fin
    $k = n$
pour $i = 1 : n$
    $h_{in} = q_i^* A q_n$
fin

Sortie : k, $Q_k = (q_1 \ldots q_k) \in \mathbb{S}t_{nk}$, $H_k \in C^{k\times k}$ de Hessenberg avec $h_{j+1\,j} > 0$.

Définition 8.18 *Une matrice de Hessenberg H dont les coefficients $h_{j+1\,j}$ sont distincts de zéro est dite non réduite. La matrice H_k donnée par l'algorithme d'Arnoldi en est un exemple.*

La complexité de l'algorithme d'Arnoldi est celle de l'algorithme de Gram-Schmidt auquel il faut ajouter le coût des produits Aq_j, $j = 1, \ldots, k$. Chaque produit nécessite $2n^2$ opérations. L'algorithme d'Arnoldi requiert donc $\approx 2nk^2 + 2n^2k$ opérations arithmétiques.

8.9 L'ALGORITHME DE LANCZOS

Lorsque la matrice A est hermitienne, la décomposition de Hessenberg $A = QHQ^*$ montre que la matrice H est également hermitienne et donc tridiagonale. Notons T

cette matrice

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \overline{\beta}_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \overline{\beta}_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ & & & \overline{\beta}_{n-1} & \alpha_n \end{pmatrix}.$$

Nous allons suivre la même démarche qu'au paragraphe précédent en considérant les relations existant entre les vecteurs-colonne des matrices A et Q déduites de la relation

$$AQ = TQ \tag{8.5}$$

et les propriétés d'orthogonalité des colonnes de Q.

On suppose que le vecteur unitaire q_1 est donné. La première colonne de l'égalité (8.5) donne

$$Aq_1 = \alpha_1 q_1 + \overline{\beta}_1 q_2$$

d'où l'on déduit $\overline{\beta}_1 q_2 = Aq_1 - \alpha_1 q_1$. Le vecteur $Aq_1 - \alpha_1 q_1$ colinéaire à q_2 doit être perpendiculaire à q_1. Le choix du coefficient α_1 est imposé par cette contrainte : on a $\langle Aq_1 - \alpha_1 q_1, q_1 \rangle = 0$ et donc $\alpha_1 = q_1^* Aq_1$. Par ailleurs, $\|q_2\|_2 = 1$ implique $|\overline{\beta}_1| = \|Aq_1 - \alpha_1 q_1\|_2$. Si $\|Aq_1 - \alpha_1 q_1\|_2$ est distinct de zéro, on peut alors choisir $\beta_1 = \|Aq_1 - \alpha_1 q_1\|_2 > 0$ et $q_2 = (Aq_1 - \alpha_1 q_1)/\beta_1$. De cette dernière égalité on déduit aussi que $q_2^* q_2 = 1 = q_2^*(Aq_1 - \alpha_1 q_1)/\beta_1 = q_2^* Aq_1/\beta_1$ et donc $\beta_1 = q_2^* Aq_1 = q_1^* Aq_2$ puisque A est hermitienne et β_1 réel.

Plus généralement, pour les colonnes successives $j = 2, \ldots, n-1$, on a

$$Aq_j = \beta_{j-1} q_{j-1} + \alpha_j q_j + \overline{\beta}_j q_{j+1},$$

que l'on écrit sous forme d'une récurrence à trois termes :

$$\overline{\beta}_j q_{j+1} = Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j. \tag{8.6}$$

Supposons que les vecteurs $q_i, i = 1, \ldots, j$ soient orthonormés et que $\beta_{j-1} \in \mathbb{R}$ vérifie $\beta_{j-1} = q_j^* Aq_{j-1} = q_{j-1}^* Aq_j$. Le produit scalaire par q_j des deux membres de l'égalité (8.6) donne $\alpha_j = q_j^* Aq_j$. Le produit par q_{j-1} donne $\beta_{j-1} = q_{j-1}^* Aq_j$ qui est déjà vérifié par hypothèse. En considérant la norme, on obtient

$$|\beta_j| = \|Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j\|_2.$$

Si $\|Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j\|_2 \neq 0$ on pose $\beta_j = \|Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j\|_2$ et

$$q_{j+1} = (Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j)/\beta_j.$$

Cette égalité donne en outre $\beta_j = q_{j+1}^* A q_j = q_j^* A q_{j+1}$. On poursuit ainsi le calcul tant que β_j est distinct de zéro.
De cette récurrence, nous déduisons l'algorithme de Lanczos qui calcule les vecteurs q_j à partir d'une vecteur q_1.

Algorithme de Lanczos

Entrée : $A, T \in \mathbb{C}^{n\times n}$, $T = 0$, $q_1 \in \mathbb{C}^n$ de norme 1, $q_0 = 0$, $\beta_0 = 0$

pour $j = 1 : n-1$
  $z = Aq_j$
  $\alpha_j = q_j^* z$
  $z = z - \alpha_j q_j - \beta_{j-1} q_{j-1}$
  $\beta_j = \|z\|_2$
  si $\beta_j = 0$
    $k = j$
    stop
  fin
  $q_{j+1} = z/\beta_j$
fin
$k = n$
$\alpha_n = q_n^* A q_n$
fin

Sortie : k, $Q_k = (q_1 \dots q_k) \in \mathbb{S}t_{nk}$, $T_k = T(1:k, 1:k) \in C^{k\times k}$ tridiagonale

La complexité de l'algorithme de Lanczos est dominée par les produits $Aq_j, j = 1, \dots, k$ ce qui donne $\approx 2n^2 k$ opérations.

Remarque 8.6. Une récurrence à trois termes se rencontre aussi dans la construction des polynômes orthogonaux tels que, par exemple, les polynômes de Legendre, de Chebyshev ou de Jacobi utilisés dans les formules de quadrature de Gauss. L'orthogonalité des polynômes y est donnée au sens du produit scalaire

$$\langle f, g\rangle = \int_I f(t)g(t)\omega(t)\,dt,$$

où I est un intervalle et $\omega : I \to]0, +\infty[$ une fonction poids.

8.10 CONDITIONNEMENT DE LA DÉCOMPOSITION QR

Comment les variations de la matrice A influent-elles sur sa décomposition QR ? Pour traiter cette question nous allons, suivant les principes exposés au chapitre 5, calculer la dérivée de l'application $A \to (Q, R)$. Nous avons vu à la proposition 8.4 que cette application est bien définie à condition de prendre pour Q une matrice de Stiefel et pour R une matrice à diagonale positive. Le calcul de sa dérivée repose sur le théorème de dérivation des fonctions inverses : on commence par dériver l'application $(Q, R) \to A$, ce qui est très facile, puis on calcule l'inverse de cette dérivée, ce qui l'est moins. On a ainsi obtenu la dérivée de l'application $A \to (Q, R)$. Mais une difficulté se présente. L'application $(Q, R) \to A$ est définie non pas sur un ouvert d'un espace vectoriel mais sur une variété différentiable : c'est donc de calcul différentiel sur les variétés (ici des sous-variétés) dont nous aurons besoin. Afin qu'un lecteur peu familier de ces notions puisse suivre le déroulement des calculs nous présentons brièvement, dans les lignes qui suivent, les outils nécessaires. On peut sans dommage éviter ce paragraphe, il ne sera pas utilisé par la suite.

8.10.1 Sous-variétés différentiables

Définition 8.19 (Sous-variétés) *Un sous-ensemble V de $\mathbb{R}^n$ est une sous-variété de classe C^k ($k \geqslant 1$) lorsque, pour tout $x \in V$, il existe un voisinage ouvert U de x dans $\mathbb{R}^n$ et une application $F : U \to \mathbb{R}^m$, m indépendant de $x \in V$, qui vérifie les trois conditions suivantes :*

1. *F est de classe C^k,*
2. *Pour tout $x \in V$, $\operatorname{rang} DF(x) = m$,*
3. *$V \cap U = \{y \in U : F(y) = 0\}$. Une telle application est appelée « équation locale » de V en x. La dimension de V est définie par $\dim V = n - m$.*

> ***Remarque 8.7.*** La condition de surjectivité pour les équations locales ($\operatorname{rang} DF(x) = m$ pour tout $x \in V$) peut être remplacée par une condition de rang constant : pour un même entier r, $0 \leqslant r \leqslant m$, pour tout $x \in V$, $\operatorname{rang} DF(y) = r$ pour tout $y \in U$. Dans ce cas $\dim V = n - r$.

Donnons quatre exemples de sous-variétés :

- Un ouvert Ω de $\mathbb{R}^n$. Une équation locale est donnée par l'application nulle $F : \Omega \to \mathbb{R}$, $F(x) = 0$. Ainsi $\dim \Omega = n$.

- Un sous-espace affine E de $\mathbb{R}^n$. Un tel sous-espace s'écrit $E = a + \operatorname{Ker} L$ avec $a \in E$ et $L : \mathbb{R}^n \to \mathbb{R}^m$ linéaire. Une équation locale est donnée par $F(x) = L(x - a)$, $x \in \mathbb{R}^n$, et on a $\dim E = n - \operatorname{rang} DF(x) = \dim \operatorname{Ker} L$.

- La sphère unité $\mathbb{S}^{n-1}$ dans $\mathbb{R}^n$. L'équation de la sphère, $F(x) = \|x\|_2^2 - 1$, est de rang maximum, égal à 1, pour tout $x \in \mathbb{R}^n$, $x \neq 0$, donc $\dim \mathbb{S}^{n-1} = n - 1$.

- Le groupe orthogonal $\mathbb{O}_n$. Dans ce dernier cas, on peut prendre $F : \mathbb{GL}_n \to \mathbb{R}^{n\times n}$ donnée par $F(A) = AA^T - I_n$. Comme rang $DF(A) = n(n+1)/2$ pour tout $A \in \mathbb{GL}_n$ on obtient $\dim \mathbb{O}_n = n^2 - n(n+1)/2 = n(n-1)/2$.

Définition 8.20 (Espace tangent) *Soit V une sous-variété de classe C^k et de dimension d de $\mathbb{R}^n$ et soit $x \in V$. À toute courbe (de classe C^1) contenue dans V et passant par x associons son vecteur-vitesse en x : si $\gamma :]-1, 1[\to V$ avec $\gamma(0) = x$ posons $\dot{x} = \frac{d}{dt}\gamma(t)\,|_{t=0}$. L'ensemble de ces vecteurs constitue un sous-espace vectoriel de dimension d de $\mathbb{R}^n$ appelé espace tangent en x à V et noté T_xV.*

Lorsque F est une équation locale de V en x on montre que

$$T_xV = \operatorname{Ker} DF(x).$$

Reprenons nos exemples :

- Pour un ouvert Ω de $\mathbb{R}^n$ on a $T_x\Omega = \mathbb{R}^n$,

- Pour un sous-espace affine $E = a + \operatorname{Ker} L$ on a $T_xE = \operatorname{Ker} L$,

- Pour la sphère unité, $T_x\mathbb{S}^{n-1} = \{y \in \mathbb{R}^n \ : \ \langle y, x\rangle = 0\}$

- Pour le groupe orthogonal, $T_Q\mathbb{O}_n = \left\{QV \ : \ V \in \mathbb{R}^{n\times n},\ V^T + V = 0\right\}$.

Définition 8.21 (Dérivée-1) *Soit $f : V \to \mathbb{R}^m$. On dit que f est de classe C^1 lorsque, pour tout $x \in V$ et pour toute courbe $\gamma :]-1, 1[\to V$ passant par x ($\gamma(0) = x$) l'application $f \circ \gamma :]-1, 1[\to \mathbb{R}^m$ est de classe C^1. Lorsque c'est le cas on définit la dérivée de f en x dans la direction $\dot{x} \in T_xV$ par*

$$Df(x)\dot{x} = \frac{d}{dt} f \circ \gamma(t)\,|_{t=0}\ ,\ \textit{lorsque}\ \dot{x} = \frac{d}{dt}\gamma(t)\,|_{t=0}\ .$$

Définition 8.22 (Dérivée-2) *Soient $V \subset \mathbb{R}^n$ et $W \subset \mathbb{R}^m$ deux sous-variétés de classe C^k et soit $f : V \to W$ de classe C^1 (au sens de la définition précédente). Il est clair que $Df(x)\dot{x} \in T_{f(x)}W$ pour tout $\dot{x} \in T_xV$. La dérivée de f en x est donc une application*

$$Df(x) : T_xV \to T_{f(x)}W.$$

Bien souvent f admet un prolongement naturel g défini sur un voisinage ouvert de V dans $\mathbb{R}^n$. Par exemple l'application $(Q, R) \to QR$ est définie quel que soit la paire (Q, R) et pas seulement lorsque Q est orthogonale et R triangulaire supérieure.

Lorsque le prolongement g de f est de classe C^1 au sens habituel du terme, la dérivée de f en $x \in V$ est égale à la restriction de $Dg(x)$ à T_xV :

$$D(g|_V)(x) = Dg(x)|_{T_xV} .$$

Nous sommes en mesure d'énoncer le théorème d'inversion locale dans le cadre des sous-variétés :

Théorème 8.23 (Inversion locale) *Soient $V \subset \mathbb{R}^n$ et $W \subset \mathbb{R}^m$ deux sous-variétés de classe C^k et de même dimension d et soit $x \in V$. Soit $f : V \to W$ de classe C^1 telle que $Df(x) : T_xV \to T_{f(x)}W$ soit un isomorphisme. Alors, il existe un ouvert V_x de V contenant x ainsi qu'un ouvert $W_{f(x)}$ de W contenant $f(x)$ tels que $f : V_x \to W_{f(x)}$ soit bijective. La bijection inverse $f^{-1} : W_{f(x)} \to V_x$ est de classe C^1 et*

$$Df^{-1}(f(x)) = (Df(x))^{-1} .$$

8.10.2 Calcul du conditionnement

Afin de simplifier un peu les calculs nous ne considérons que des matrices carrées et réelles. Nos espaces de travail sont :

- $\mathbb{GL}_n(\mathbb{R})$: matrices $n \times n$ réelles et inversibles,
- $\mathbb{O}_n$: matrices $n \times n$ orthogonales,
- $\mathcal{U}_n$: matrices $n \times n$ triangulaires supérieures,
- $\mathcal{PU}_n$: matrices $n \times n$ triangulaires supérieures à diagonale positive.

Toute matrice $B \in \mathbb{R}^{n\times n}$ est somme d'une matrice antisymétrique $\Pi_{as}(B)$ et d'une matrice triangulaire supérieure $\Pi_{up}(B)$ et cette décomposition est unique : si $B = E + D + F$ où D, E, F sont les parties diagonale, triangulaire supérieure stricte et triangulaire inférieure stricte, alors : $\Pi_{as}(B) = F - F^T$ et $\Pi_{up}(B) = E + D + F^T$. Remarquons que

$$\|\Pi_{as}(B)\|_F^2 \leqslant 2\,\|B\|_F^2$$

et que

$$\|\Pi_{up}(B)\|_F^2 \leqslant 2\,\|B\|_F^2 .$$

Nous considérons les applications suivantes :

- $\mathcal{P} : \mathbb{O}_n \times \mathcal{PU}_n \to \mathbb{GL}_n(\mathbb{R})$ définie par $\mathcal{P}(Q, R) = QR$
- $\mathcal{QR} : \mathbb{GL}_n(\mathbb{R}) \to \mathbb{O}_n \times \mathcal{PU}_n$, $\mathcal{QR}(A) = (Q, R)$. Elle se décompose en

- $\mathcal{Q} : \mathbb{GL}_n(\mathbb{R}) \to \mathbb{O}_n$, $\mathcal{Q}(A) = Q$, et
- $\mathcal{R} : \mathbb{GL}_n(\mathbb{R}) \to \mathcal{PU}_n$, $\mathcal{R}(A) = R$.

L'espace tangent en $Q \in \mathbb{O}_n$ au groupe orthogonal est donné par

$$T_Q\mathbb{O}_n = \left\{\dot{Q} = QV \in \mathbb{R}^{n\times n} \ : \ V \in \mathbb{R}^{n\times n},\ V^T + V = 0\right\}.$$

Nous allons prouver le

Théorème 8.24 *Soient $A \in \mathbb{GL}_n(\mathbb{R})$, $Q \in \mathbb{O}_n$ et $R \in \mathcal{PU}_n$ telles que $A = QR$. La dérivée de $\mathcal{Q}$ en A est donnée par :*

$$D\mathcal{Q}(A) : \mathbb{R}^{n\times n} \to T_Q\mathbb{O}_n,\ D\mathcal{Q}(A)\dot{A} = Q\Pi_{as}(Q^T\dot{A}R^{-1})$$

De plus

$$\left\|D\mathcal{Q}(A)\dot{A}\right\|_F \leqslant \sqrt{2}\ \left\|A^{-1}\right\|_2 \left\|\dot{A}\right\|_F.$$

La dérivée de $\mathcal{R}$ en A est donnée par :

$$D\mathcal{R}(A) : \mathbb{R}^{n\times n} \to \mathcal{U}_n,\ D\mathcal{R}(A)\dot{A} = \Pi_{up}(Q^T\dot{A}R^{-1})R$$

et

$$\left\|D\mathcal{R}(A)\dot{A}\right\|_F \leqslant \sqrt{2}\,\mathrm{cond}_2(A)\left\|\dot{A}\right\|_F.$$

Démonstration. L'application $\mathcal{P} : \mathbb{O}_n \times \mathcal{PU}_n \to \mathbb{GL}_n(\mathbb{R})$ est la restriction à $\mathbb{O}_n \times \mathcal{PU}_n$ de l'application

$$\mathcal{P} : \mathbb{R}^{n\times n} \times \mathcal{PU}_n \to \mathbb{R}^{n\times n} \text{ définie par } \mathcal{P}(Q, R) = QR.$$

Cette dernière est de classe C^∞ et sa dérivée en $(Q, R) \in \mathbb{R}^{n\times n} \times \mathcal{PU}_n$ est donnée par

$$D\mathcal{P}(Q, R) : \mathbb{R}^{n\times n} \times \mathcal{U}_n \to \mathbb{R}^{n\times n},\ D\mathcal{P}(Q, R)(\dot{Q}, \dot{R}) = \dot{Q}R + Q\dot{R}.$$

La dérivée de la restriction $\mathcal{P} : \mathbb{O}_n \times \mathcal{PU}_n \to \mathbb{GL}_n(\mathbb{R})$ est la restriction à $T_Q\mathbb{O}_n \times \mathcal{U}_n$ de la dérivée définie sur $\mathbb{R}^{n\times n} \times \mathcal{U}_n$. On obtient

$$D\mathcal{P}(Q, R) : T_Q\mathbb{O}_n \times \mathcal{U}_n \to \mathbb{R}^{n\times n},\ D\mathcal{P}(Q, R)(\dot{Q}, \dot{R}) = \dot{Q}R + Q\dot{R}.$$

Cette dérivée est un isomorphisme. En effet, $T_Q\mathbb{O}_n \times \mathcal{U}_n$ et $\mathbb{R}^{n\times n}$ ont même dimension n^2 et si $D\mathcal{P}(Q, R)(\dot{Q}, \dot{R}) = 0$ avec $\dot{Q} = QV$ et $V^T = -V$, on a $QVR + Q\dot{R} = 0$ d'où $V = -\dot{R}R^{-1}$. Cette matrice est triangulaire supérieure et antisymétrique ce qui implique $V = 0$ et $\dot{R} = 0$. Ainsi $\mathrm{Ker}\, D\mathcal{P}(Q, R) = 0$ et $D\mathcal{P}(Q, R)$ est un isomorphisme et on peut donc appliquer le théorème

d'inversion locale (théorème 8.23). La dérivée de $\mathcal{QR}$ en $A = QR$ est donnée par

$$D\mathcal{QR}(A) : \mathbb{R}^{n\times n} \to T_Q\mathbb{O}_n \times \mathcal{U}_n,\ D\mathcal{QR}(A) = D\mathcal{P}(Q,R)^{-1}.$$

On a donc pour tout $\dot{Q} \in T_Q\mathbb{O}_n$, $\dot{R} \in \mathcal{U}_n$ et $\dot{A} \in \mathbb{R}^{n\times n}$

$$D\mathcal{QR}(A)(\dot{A}) = (\dot{Q},\dot{R}) \text{ si et seulement si } D\mathcal{P}(Q,R)(\dot{Q},\dot{R}) = \dot{A}$$

c'est-à-dire si

$$\dot{Q}R + Q\dot{R} = \dot{A}.$$

Posons $\dot{Q} = QV$ avec $V^T = -V$ on a

$$V + \dot{R}R^{-1} = Q^T\dot{A}R^{-1}.$$

Il en résulte que

$$D\mathcal{Q}(A)\dot{A} = \dot{Q} = Q\Pi_{as}(Q^T\dot{A}R^{-1}),\ \ D\mathcal{R}(A)\dot{A} = \dot{R} = \Pi_{up}(Q^T\dot{A}R^{-1})R.$$

Prouvons les inégalités sur les normes. En utilisant les propositions 3.10, 3.12 et 3.13 on a :

$$\left\|\dot{Q}\right\|_F = \left\|Q\Pi_{as}(Q^T\dot{A}R^{-1})\right\|_F = \left\|\Pi_{as}(Q^T\dot{A}R^{-1})\right\|_F \leqslant \sqrt{2}\left\|Q^T\dot{A}R^{-1}\right\|_F$$

$$= \sqrt{2}\left\|\dot{A}R^{-1}\right\|_F \leqslant \sqrt{2}\left\|\dot{A}\right\|_F\left\|R^{-1}\right\|_2 = \sqrt{2}\left\|\dot{A}\right\|_F\left\|A^{-1}\right\|_2.$$

L'inégalité sur $\left\|\dot{R}\right\|_F$ se prouve de la même manière.

Remarque 8.8. Ce théorème, que l'on peut reformuler en

$$\left\|D\mathcal{Q}(A)\dot{A}\right\|_F \leqslant \sqrt{2}\,\mathrm{cond}_2(A)\frac{\left\|\dot{A}\right\|_F}{\|A\|_2} \text{ et } \left\|D\mathcal{R}(A)\dot{A}\right\|_F \leqslant \sqrt{2}\,\mathrm{cond}_2(A)\left\|\dot{A}\right\|_F$$

montre que l'erreur commise sur Q dépend de l'erreur relative $\left\|\dot{A}\right\|_F/\|A\|_2$ alors que l'erreur commise sur R dépend de l'erreur absolue $\left\|\dot{A}\right\|_F$. Le facteur amplificateur est à chaque fois proportionnel au conditionnement de A.

8.11 NOTES ET RÉFÉRENCES

Le sigle QR vient de l'anglais : Q pour « orthogonal » et R pour « upper triangular ». Clair comme de l'eau de roche ! La décomposition QR est une conséquence immédiate du procédé d'orthonormalisation dit « de Gram-Schmidt » introduit afin d'orthonormaliser des suites de fonctions. Cette dénomination nous renvoie à Jørgen Pedersen Gram (1850 - 1916) mathématicien danois spécialiste de théorie des nombres et à Erhard Schmidt (1876 - 1959) connu pour ses travaux sur l'analyse fonctionnelle et les équations intégrales.

Les transformations de Householder étaient utilisées dès le début du XX-ème siècle (Schur, 1909) pour établir qu'une matrice carrée pouvait être rendue triangulaire via une transformation unitaire. Mais c'est Alston S. Householder (1904 - 1993) qui a reconnu le premier les propriétés de stabilité numérique des transformations qui portent désormais son nom, voir [19].

Pour en savoir plus, voir les livres de Golub-van Loan [15] et de Stewart [32].

Les algorithmes d'Arnoldi et de Lanczos sont utilisés de manière importante dans les méthodes de projection dans les sous-espaces de Krylov (voir chapitre 11). L'algorithme de Lanczos est présenté dans un article publié en 1950 [22] qui traite du problème du calcul numérique des valeurs propres d'un opérateur. Cornelius Lanczos (1893-1974) est surtout connu pour ses travaux en physique mathématique, en théorie de la relativité et également en analyse numérique. Il est à l'origine de la transformée de Fourier rapide (voir paragraphe 16.5) bien avant les travaux de Cooley et Tuckey.

Walter Edwin Arnoldi (1917-1995) était ingénieur en mécanique. L'algorithme que nous présentons apparait dans un article publié en 1951 [2] dans lequel il reprend les idées de Lanczos et les généralise aux matrices non hermitiennes.

EXERCICES

Exercice 8.1

Donner la décomposition QR à diagonale positive de la matrice

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

en utilisant les méthodes de Gram-Schmidt, Givens et Householder.

Exercice 8.2

Résoudre le système linéaire suivant en calculant au préalable la décomposition QR de la matrice du système via la méthode de Householder :

$$\begin{cases} 2x_1 + x_2 + 2x_3 = 1 \\ x_1 + x_2 + 2x_3 = 1 \\ 2x_1 + x_2 + x_3 = 1 \end{cases}$$

Exercice 8.3

Même question que précédemment avec :

$$\begin{cases} 70x + 121y + 71z = 525 \\ -40x + 80y + 70z = 330 \\ -40x - 172y - 47z = -525 \end{cases}$$

Exercice 8.4

Quelles sont les valeurs propres d'une matrice de Householder, d'une rotation de Givens ?

Exercice 8.5

Soit A une matrice orthogonale, $A \in \mathbb{O}_n$. Montrer qu'il existe des entiers $0 \leqslant p, q, r \leqslant n$, des réels $\theta_1, \ldots, \theta_r$ et une matrice orthogonale Q tels que :

$$Q^T A Q = \begin{pmatrix} I_p & 0 & 0 & 0 & \cdots & 0 \\ 0 & -I_q & 0 & 0 & \cdots & 0 \\ 0 & 0 & A_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & A_r \end{pmatrix} \text{ où } A_i = \begin{pmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{pmatrix}.$$

Exercice 8.6

Montrer que toute transformation orthogonale dans $\mathbb{R}^n$ est le produit d'au plus n symétries orthogonales. Utiliser la méthode de Householder à la matrice A de la transformation.

Exercice 8.7

Montrer que toute transformation orthogonale dans $\mathbb{R}^n$ est le produit d'au plus $n-1$ rotations et d'une symétrie orthogonale.

Exercice 8.8 Rotations de Givens complexes

1. Montrer que le groupe unitaire $\mathbb{U}_2$ est l'ensemble des matrices
$$\begin{pmatrix} e^{i\sigma}\cos\alpha & e^{i\nu}\sin\alpha \\ -e^{i\tau}\sin\alpha & e^{i(\tau+\nu-\sigma)}\cos\alpha \end{pmatrix}$$
avec $0 \leqslant \sigma, \tau, \nu \leqslant 2\pi$, $0 \leqslant \alpha \leqslant \pi/2$.
2. Montrer que le groupe spécial unitaire $\mathbb{SU}_2$ est l'ensemble des matrices
$$\begin{pmatrix} e^{i\sigma}\cos\alpha & e^{-i\tau}\sin\alpha \\ -e^{i\tau}\sin\alpha & e^{-i\sigma}\cos\alpha \end{pmatrix}$$
avec $0 \leqslant \sigma, \tau \leqslant 2\pi$, $0 \leqslant \alpha \leqslant \pi/2$.
3. Étant donnés deux nombres complexes $x, y \in \mathbb{C}$ avec $r = |x|^2 + |y|^2 \neq 0$ montrer que la matrice
$$R = \frac{1}{r}\begin{pmatrix} \overline{x} & \overline{y} \\ -y & x \end{pmatrix}$$
vérifie $R \in \mathbb{SU}_2$ et $R\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$.
4. Étendre au cas complexe le calcul de la décomposition QR via les rotations de Givens.

Exercice 8.9 Cet exercice décrit un procédé d'orthonormalisation invariant par permutation des vecteurs de la base initiale. Soit $A \in \mathbb{C}^{m\times n}$ de rang n. Les colonnes a_i de A constituent une base de Im A.

1. Montrer que les colonnes q_i de $Q_A = A(A^*A)^{-1/2}$ constituent une base orthonormée de Im A. (Montrer que Im Q_A = Im A et que $Q_A^* Q_A = I_n$.)
2. Montrer que l'ensemble des vecteurs q_i, $1 \leqslant i \leqslant n$. est indépendant de l'ordre dans lequel on a rangé les vecteurs a_i (montrer que pour toute matrice de permutation $P \in \mathbb{C}^{n\times n}$ on a $Q_{AP} = Q_A P$).

Chapitre 9

Inverses généralisés et moindres carrés

9.1 INVERSES GÉNÉRALISÉS

Le problème des moindres carrés et l'inverse généralisé d'une application linéaire poursuivent un but commun : il s'agit dans le premier cas de résoudre un système linéaire dont la matrice n'est pas nécessairement carrée et dans le second cas, de calculer un « inverse » pour une application linéaire entre deux espaces de dimensions qui peuvent être différentes. Nous avons choisi d'introduire d'abord les inverses généralisés.

Soit $L : \mathbb{E} \to \mathbb{F}$ une application linéaire entre deux espaces hermitiens. Considérons les deux décompositions en sommes directes orthogonales suivantes : $\mathbb{E} = \text{Ker } L \oplus (\text{Ker } L)^\perp$ et $\mathbb{F} = \text{Im } L \oplus (\text{Im } L)^\perp$. Si nous notons $n = \dim \mathbb{E}$, $m = \dim \mathbb{F}$ et $r = \text{rang } L$, les sous-espaces précédents ont pour dimensions

$$\dim \text{Ker } L = n - r,\ \dim(\text{Ker } L)^\perp = r,\ \dim \text{Im } L = r,\ \dim(\text{Im } L)^\perp = m - r.$$

La restriction M de L à $(\text{Ker } L)^\perp$ est une bijection entre $(\text{Ker } L)^\perp$ et $\text{Im } L$. En effet ces deux espaces ont même dimension r et M est injective (si $M(x) = 0$ pour un $x \in (\text{Ker } L)^\perp$ on a aussi $x \in \text{Ker } L$ et donc $x = 0$). On peut donc inverser M et considérer le produit de composition suivant

$$\mathbb{F} \overset{\Pi_{\text{Im } L}}{\to} \text{Im } L \overset{M^{-1}}{\to} (\text{Ker } L)^\perp \overset{i_{(\text{Ker } L)^\perp}}{\to} \mathbb{E}$$

où $\Pi_{\text{Im } L} : \mathbb{F} \to \text{Im } L$ est la projection orthogonale sur l'image de L et $i_{(\text{Ker } L)^\perp}$ est l'injection canonique. Pour tout $y \in \mathbb{F}$ cette projection orthogonale vérifie

$$\Pi_{\text{Im } L}(y) \in \text{Im } L \text{ et } y - \Pi_{\text{Im } L}(y) \in (\text{Im } L)^\perp,$$

quant à l'injection canonique, il s'agit simplement de

$$i_{(\mathrm{Ker}\ L)^\perp}(x) = x \text{ pour tout } x \in (\mathrm{Ker}\ L)^\perp.$$

Définition 9.1 *On appelle inverse généralisé ou encore inverse de Moore-Penrose l'application linéaire*

$$L^\dagger : \mathbb{F} \to \mathbb{E},\ L^\dagger = i_{(\mathrm{Ker}\ L)^\perp} \circ M^{-1} \circ \Pi_{\mathrm{Im}\ L}.$$

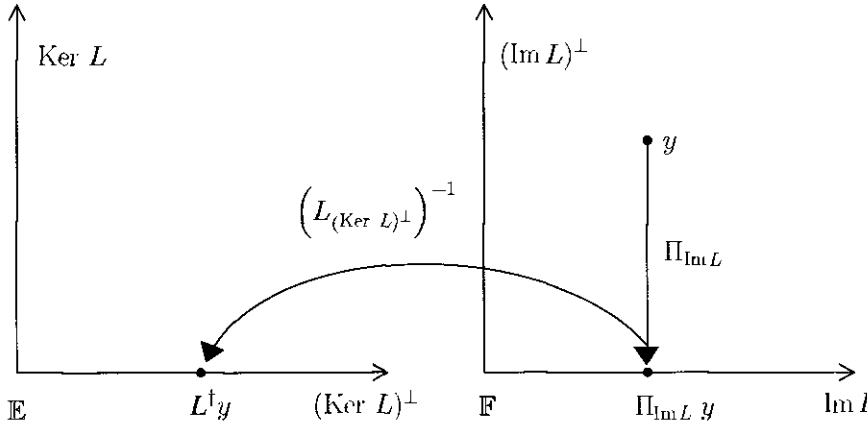


Figure 9.1 Inverse généralisé $L^\dagger$.

Comme le montre la définition et le schéma ci-dessus 9.1, l'inverse généralisé de $y \in \mathbb{F}$ est l'unique préimage dans $(\mathrm{Ker}\ L)^\perp$ de la projection orthogonale de y sur l'image de L. Ses propriétés principales sont résumées dans le théorème suivant :

Théorème 9.2 *L'inverse généralisé $L^\dagger$ vérifie*

1. $L^\dagger \circ L = \Pi_{(\mathrm{Ker}\ L)^\perp}$ *(projection orthogonale sur* $(\mathrm{Ker}\ L)^\perp$*),*
2. $L \circ L^\dagger = \Pi_{\mathrm{Im}\ L}$ *(projection orthogonale sur* $\mathrm{Im}\ L$*),*
3. $L^\dagger \circ L$ *et*
4. $L \circ L^\dagger$ *sont des applications hermitiennes,*
5. $L \circ L^\dagger \circ L = L$,
6. $L^\dagger \circ L \circ L^\dagger = L^\dagger$.

Démonstration. Soient $x \in \mathbb{E}$ et $x' = \Pi_{(\mathrm{Ker}\ L)^\perp}(x)$. Comme $L(x) = L(x')$ (que nous notons y) on a, par définition de l'inverse généralisé, $L^\dagger(y) = x'$. Aussi

$$L^\dagger \circ L(x) = L^\dagger(y) = x' = \Pi_{(\mathrm{Ker}\ L)^\perp}(x)$$

ce qui prouve la première assertion.

Passons à la seconde : soient $y \in \mathbb{F}$ et $z = \Pi_{\text{Im } L}(y)$. Par définition de l'inverse généralisé $L^\dagger(y) = x$ avec $x \in (\text{Ker } L)^\perp$ et $L(x) = z$. Ceci prouve que

$$L \circ L^\dagger(y) = L(x) = z = \Pi_{\text{Im } L}(y)$$

d'où la seconde assertion.

Les troisième et quatrième assertions sont des conséquences des deux premières : une projection orthogonale est une application hermitienne (exercice 1.13). On a donc

$$(L \circ L^\dagger)^* = \Pi^*_{\text{Im } L} = \Pi_{\text{Im } L} = L \circ L^\dagger.$$

Même chose pour $L^\dagger \circ L$.

On a enfin

$$L \circ L^\dagger \circ L = L \circ \Pi_{(\text{Ker } L)^\perp} = L$$

et

$$L^\dagger \circ L \circ L^\dagger = L^\dagger \circ \Pi_{\text{Im } L} = L^\dagger.$$

Corollaire 9.3 *Lorsque L est injective (resp. surjective) on a $L^\dagger \circ L = id_{\mathbb{E}}$ (resp. $L \circ L^\dagger = id_{\mathbb{F}}$) et lorsque L est bijective $L^\dagger = L^{-1}$.*

Démonstration. On applique le théorème 9.2 1 et 2. Lorsque L est injective $\text{Ker } L = 0$ et donc $\Pi_{(\text{Ker } L)^\perp} = \text{id}_{\mathbb{E}}$. Lorsque L est surjective $\text{Im } L = \mathbb{F}$ et donc $\Pi_{\text{Im } L} = \text{id}_{\mathbb{F}}$. Le cas bijectif s'ensuit.

Les propriétés 3, 4, 5 et 6 du théorème 9.2 caractérisent l'inverse généralisé comme le montre la proposition suivante :

Proposition 9.4 *Supposons qu'une application linéaire $P : \mathbb{F} \to \mathbb{E}$ satisfasse les propriétés suivantes :*

1. *$P \circ L$ et*
2. *$L \circ P$ sont des applications hermitiennes,*
3. *$L \circ P \circ L = L$,*
4. *$P \circ L \circ P = P$.*

Alors $P = L^\dagger$.

Démonstration. $P = P \circ L \circ P = P \circ L \circ L^\dagger \circ L \circ P = P \circ L \circ L^\dagger \circ L \circ L^\dagger \circ L \circ L^\dagger \circ L \circ P = L^* \circ P^* \circ L^* \circ L^{\dagger *} \circ L^\dagger \circ L^{\dagger *} \circ L^* \circ P^* \circ L^* = L^* \circ L^{\dagger *} \circ L^\dagger \circ L^{\dagger *} \circ L^* = L^\dagger \circ L \circ L^\dagger \circ L \circ L^\dagger = L^\dagger \circ L \circ L^\dagger = L^\dagger$.

Nous allons déduire de la proposition précédente les résultats suivants :

Proposition 9.5

1. $(L^*)^\dagger = (L^\dagger)^*$,
2. $(L^\dagger)^\dagger = L$,
3. *Lorsque L est injective, c'est-à-dire si* $\operatorname{rang} L = \dim \mathbb{E}$, *on a* $L^\dagger = (L^* \circ L)^{-1} \circ L^*$,
4. *Lorsque L est surjective, c'est-à-dire si* $\operatorname{rang} L = \dim \mathbb{F}$, *on a* $L^\dagger = L^* \circ (L \circ L^*)^{-1}$.

Démonstration. Pour la première assertion on remplace L par L^* et on prend $P = (L^\dagger)^*$ dans la proposition 9.4. On a

1. $P \circ L^* = (L^\dagger)^* \circ L^* = (L \circ L^\dagger)^*$ qui est hermitienne par le théorème 9.2,
2. Idem pour $L^* \circ P$,
3. $L^* \circ P \circ L^* = L^* \circ (L^\dagger)^* \circ L^* = (L \circ L^\dagger \circ L)^* = L^*$ par le théorème 9.2,
4. On montre de même que $P \circ L^* \circ P = P$.

La proposition 9.4 montre qu'alors $P = (L^*)^\dagger$.

On procède de façon similaire pour les trois autres assertions. Notons simplement que l'hypothèse « L injective » faite en 3 implique l'inversibilité de $L^* \circ L$. De même, l'hypothèse « L surjective » faite en 4 implique l'inversibilité de $L \circ L^*$.

Pour deux opérateurs linéaires inversibles L et M on a :

$$(L \circ M)^{-1} = M^{-1} \circ L^{-1}.$$

Cette propriété ne s'étend pas aux inverses généralisés, même si l'une des deux applications est inversible. Un exemple est donné à l'exercice 9.1. Mais les choses se passent bien si l'une des applications est unitaire :

Proposition 9.6 *Soient $U : \mathbb{E} \to \mathbb{E}$ et $V : \mathbb{F} \to \mathbb{F}$ des applications unitaires. On a :*

$$(V \circ L \circ U)^\dagger = U^* \circ L^\dagger \circ V^*.$$

Démonstration. On remplace L par $V \circ L \circ U$ et P par $U^* \circ L^\dagger \circ V^*$ dans la proposition 9.4

Le résultat précédent permet de calculer facilement l'inverse généralisé d'une matrice lorsque l'on en connait une décomposition en valeurs singulières (voir théorème 4.2).

Théorème 9.7 *Soit $A \in \mathbb{C}^{m\times n}$ de rang r qui possède la décomposition en valeurs singulières $A = V\Sigma U^*$ avec $U \in \mathbb{U}_n$, $V \in \mathbb{U}_m$,*

$$\Sigma = \begin{pmatrix} D & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

$D = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ et où $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ sont les valeurs singulières de A. Sous ces hypothèses

$$A^\dagger = U\Sigma^\dagger V^* \text{ et } \Sigma^\dagger = \begin{pmatrix} D^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix}.$$

Démonstration. L'égalité $A^\dagger = U\Sigma^\dagger V^*$ est une conséquence de la proposition 9.6 et le calcul de $\Sigma^\dagger$ se mène à partir de la définition de l'inverse généralisé : soit $y \in \mathbb{C}^m$; projeter y sur $\operatorname{Im}\Sigma$ revient à annuler ses $m-r$ dernières coordonnées

$$\Pi_{\operatorname{Im}\Sigma}(y) = \begin{pmatrix} y_1 \\ \vdots \\ y_r \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Les préimages de ce vecteur sont les $x \in \mathbb{C}^n$ qui s'écrivent

$$x = \begin{pmatrix} y_1/\sigma_1 \\ \vdots \\ y_r/\sigma_r \\ x_{r+1} \\ \vdots \\ x_n \end{pmatrix}.$$

De plus

$$\operatorname{Ker}\Sigma = \{x \in \mathbb{C}^n \ : \ x_1 = \ldots = x_r = 0\}$$

et

$$(\operatorname{Ker}\Sigma)^\perp = \{x \in \mathbb{C}^n \ : \ x_{r+1} = \ldots = x_n = 0\}.$$

On voit donc que la seule préimage de $\Pi_{\mathrm{Im}\,\Sigma}(y)$ dans $(\mathrm{Ker}\,\Sigma)^{\perp}$ est

$$\Sigma^{\dagger}(y) = \begin{pmatrix} y_1/\sigma_1 \\ \vdots \\ y_r/\sigma_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} D^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_n \end{pmatrix}.$$

9.2 MOINDRES CARRÉS

Nous considérons ici un système d'équations linéaires $Ax = b$ avec $A \in \mathbb{C}^{m\times n}$, $x \in \mathbb{C}^n$, $b \in \mathbb{C}^m$ où le nombre n d'inconnues et celui m d'équations sont différents. Deux cas sont à considérer :

- Celui des *systèmes surdéterminés* ($m > n$) : le nombre d'équations est plus grand que celui des indéterminées. De tels systèmes se rencontrent dans les problèmes d'identification de paramètres, d'assimilation de données, en géodésie et cetera. En général, un système surdéterminé n'a pas de solution.

- Celui des *systèmes sous-déterminés* ($m < n$) où le nombre d'équations est plus petit que celui des inconnues. En général, un tel système admet une infinité de solutions.

La méthode des *moindres carrés* consiste à rechercher, parmi les $x \in \mathbb{C}^n$, celui ou ceux qui minimisent la quantité

$$f(x) = \|Ax - b\|_2^2$$

appelée *fonction résidu*. La valeur de cet infimum (on verra que c'est un minimum)

$$m = \inf_{x\in\mathbb{C}^n} \|Ax - b\|_2^2$$

est appelé le *résidu minimum* et tout vecteur $x \in \mathbb{R}^n$ qui le réalise, c'est-à-dire pour lequel $f(x) = m$, est appelé *solution au sens des moindres carrés* du système $Ax = b$.

Exemple 9.1 : barycentre. Un exemple simplissime de système surdéterminé est donné par

$$x = b_i,\ 1 \leqslant i \leqslant m.$$

On peut penser à un ensemble de mesures que l'on effectue pour déterminer la valeur numérique d'une grandeur physique. Ici $A = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$ et $x \in \mathbb{R}$. Lorsque les b_i ne sont pas tous égaux, ce qui est toujours le cas pour des mesures faites en précision finie, la solution au sens des moindres carrés

$$\inf_{x \in \mathbb{R}} \sum_{i=1}^{m} (x - b_i)^2$$

est la moyenne arithmétique des b_i : $x = \sum_i b_i / m$.

Exemple 9.2 : régression linéaire. On suppose que des mesures physiques ont été effectuées :

$$(x_i, y_i),\ 1 \leqslant i \leqslant m,\ m > 2,$$

où $x_i \in \mathbb{R}$ est le paramètre de la mesure et $y_i \in \mathbb{R}$ le résultat obtenu. Le modèle linéaire consiste à supposer que $y_i = \alpha x_i + \beta$. En général il est impossible de trouver α et β pour lesquels il y a égalité quel que soit i.On recherche donc une solution au sens des moindres carrés :

$$\inf_{\alpha, \beta \in \mathbb{R}} \sum_{i=1}^{m} (\alpha x_i + \beta - y_i)^2.$$

D'un point de vue matriciel ce problème correspond à

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix},\quad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix},$$

le vecteur des inconnues étant $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. La droite obtenue d'équation $y = \alpha x + \beta$ s'appelle *droite de régression.*

La figure 9.2 montre une droite de régression obtenue à partir d'un nuage de 100 points (x_i, y_i) Pour chaque abscisse x_i la valeur correspondante y_i a été calculée suivant l'égalité

$$y_i = 2x_i + 1 + \varepsilon_i$$

où ε_i est une variable aléatoire gaussienne de moyenne nulle et d'écart-type 0.25. Suivant la théorie statistique, en supposant que les variables aléatoires ε_i sont indépendantes, la droite des moindres carrés est « proche » de la droite d'équation $y = 2x + 1$.

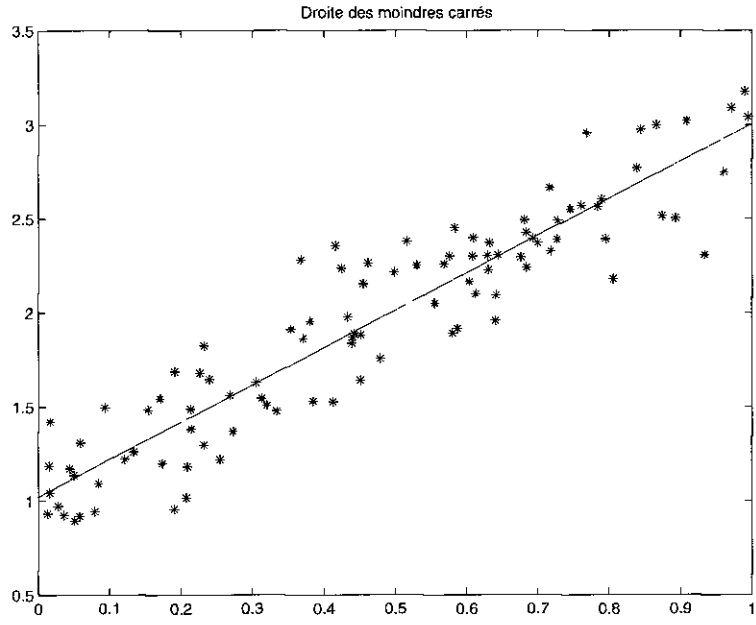


Figure 9.2 Droite de régression. Les points (x_i, y_i) sont notés *

On trouve les valeurs $\alpha = 1.9857$ et $\beta = 1.0209$. Pour 1000 points, on obtient $\alpha = 2.0061$ et $\beta = 1.0067$.

Exemple 9.3 : régression multiple. C'est le même exemple que le précédent mais on suppose que la mesure $y_i \in \mathbb{R}$ dépend de n paramètres $x_i = (x_i^1, \ldots, x_i^n)$, $1 \leqslant i \leqslant m$. Le modèle linéaire consiste maintenant à supposer que

$$y_i = \sum_{j=1}^{n} \alpha_j x_i^j + \beta$$

ce qui conduit au problème des moindres carrés

$$\inf_{(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}} \sum_{i=1}^{m} (\alpha_1 x_i^1 + \ldots \alpha_n x_i^n + \beta - y_i)^2.$$

Exemple 9.4 : erreurs rétrogrades. Nous avons rencontré cet exemple au paragraphe 5.5. Donnons-nous une matrice $A \in \mathbb{GL}_n$, un vecteur $b \in \mathbb{C}^n$, la solution $x = A^{-1}b \in \mathbb{C}^n$ du système $Ax = b$ et une approximation x' de x. L'analyse rétrograde des erreurs consiste à considérer x' comme la solution exacte

d'un système linéaire du type $(A + E)x' = b$. Il s'agit là d'un problème sous-déterminé : il a n équations, n^2 inconnues (les entrées de la matrice E) et il possède une infinité de solutions. Nous en avons, au théorème 5.7, sélectionné une en considérant le problème d'optimisation

$$\min_{\substack{E \in \mathbb{C}^{n\times n} \\ (A+E)x' = b}} \|E\|_2.$$

Cet exemple n'entre pas exactement dans notre cadre d'étude : la norme spectrale $\|E\|_2$ ne se déduisant pas d'un produit scalaire (voir exercice 9.10).

Étudions maintenant l'existence et la caractérisation des solutions de tels problèmes.

Théorème 9.8 *Soient $A \in \mathbb{C}^{m\times n}$ et $b \in \mathbb{C}^m$.*

1. *Le problème des moindres carrés*

$$\inf_{x\in\mathbb{C}^n} \|Ax - b\|_2^2$$

possède au moins une solution.

2. *Ces solutions sont caractérisées par :*

$$A^*Ax = A^*b$$

appelée équation normale.

3. *Si x et x' sont deux solutions alors $Ax = Ax'$. La solution est donc unique si la matrice A est de rang n.*

4. *La solution de norme minimale*

$$\inf_{A^*Ax = A^*b} \|x\|_2^2$$

est donnée par $x = A^\dagger b$.

Démonstration. Le problème d'optimisation

$$\inf_{y\in \operatorname{Im} A} \|y - b\|_2^2$$

possède une et une seule solution $\overline{y} \in \operatorname{Im} A$ qui est la projection orthogonale de b sur $\operatorname{Im} A$. Elle est caractérisée par (voir paragraphe 1.13)

$$\overline{y} \in \operatorname{Im} A \text{ et } \langle \overline{y} - b, y\rangle = 0 \text{ pour tout } y \in \operatorname{Im} A.$$

L'ensemble des solutions du problème des moindres carrés est égal à

$$\mathcal{S} = \{x \in \mathbb{C}^n \ : \ Ax = \overline{y}\}.$$

Ceci prouve l'existence d'une solution et le fait que $Ax = Ax' = \overline{y}$ pour deux telles solutions (assertions 1 et 3). Soit $\overline{x} \in \mathcal{S}$ de sorte que $A\overline{x} = \overline{y}$. L'équation qui caractérise $\overline{y}$ peut aussi s'écrire

$$\langle A\overline{x} - b, Ax\rangle = 0 \text{ pour tout } x \in \mathbb{C}^n$$

c'est-à-dire

$$\langle A^*(A\overline{x} - b), x\rangle = 0 \text{ pour tout } x \in \mathbb{C}^n$$

ou encore

$$A^*(A\overline{x} - b) = 0$$

qui est l'équation normale du problème (assertion 2 prouvée).

L'ensemble $\mathcal{S}$ des solutions est l'image réciproque de la projection orthogonale de b sur Im A. Si $\overline{x}$ est l'une d'entre-elles alors $\mathcal{S} = \overline{x} + \text{Ker } A$. La solution x de norme minimale est la projection orthogonale de 0 sur $\mathcal{S}$ donc $x \in (\text{Ker } A)^{\perp}$ et ceci prouve que $x = A^{\dagger}b$ (assertion 4).

Remarque 9.1. Pour toute fonction $f : \mathbb{R}^n \to \mathbb{R}$ différentiable on définit le gradient de f en x par

$$Df(x)u = \langle \nabla f(x), u\rangle$$

pour tout $u \in \mathbb{R}^n$ où $Df(x) : \mathbb{R}^n \to \mathbb{R}$ est la dérivée de f en x. On sait que $\nabla f(x) = 0$ lorsque x réalise le minimum de f sur $\mathbb{R}^n$.

Prenons pour f la fonction résidu :

$$f : \mathbb{R}^n \to \mathbb{R}, \ f(x) = \|Ax - b\|_2^2.$$

Son gradient est $\nabla f(x) = 2A^*(Ax - b)$: l'équation $\nabla f(x) = 0$ n'est autre que l'équation normale $A^*Ax = A^*b$.

Remarque 9.2. Le problème des moindres carrés

$$\inf_{x \in E} \|Ax - b\|_2^2$$

où E est un sous-espace vectoriel de $\mathbb{C}^n$ est qualifié de *problème des moindres carrés contraints*. En reprenant la démonstration du théorème (9.8) on voit que ce problème possède au moins une solution et que toute solution $\overline{x} \in E$ du problème est caractérisée par

$$\langle A\overline{x} - b, Az\rangle = 0$$

pour tout $z \in E$. Le calcul effectif d'une solution est obtenu à l'aide d'une base $(s_1, \ldots, s_k)$ de E. Notons $S = (s_1 \ldots s_k) \in \mathbb{C}^{n \times k}$ la matrice dont les vecteurs-colonne sont les s_i. Ainsi $x \in E$ si et seulement s'il existe $y \in \mathbb{C}^k$ tel que $x = Sy$. Un vecteur $\overline{x} = S\overline{y}$ est solution du problème si et seulement si

$$\langle AS\overline{y} - b, ASy \rangle = 0$$

pour tout $y \in \mathbb{C}^k$. On obtient ainsi l'équation normale par rapport à y

$$S^*A^*AS\overline{y} = S^*A^*b.$$

9.3 PROBLÈMES SURDÉTERMINÉS

Supposons maintenant que $m > n$ et que A soit de rang n.

9.3.1 L'équation normale

Théorème 9.9 *Soient $A \in \mathbb{C}^{m \times n}$, rang $A = n$ et $b \in \mathbb{C}^m$. Le problème des moindres carrés*

$$\inf_{x \in \mathbb{C}^n} \|Ax - b\|_2^2$$

possède une unique solution

$$\overline{x} = (A^*A)^{-1}A^*b = A^\dagger b.$$

Démonstration. Comme rang $A = n$ la matrice A^*A est inversible et l'équation normale donne $\overline{x} = (A^*A)^{-1}A^*b$. On reconnaît là l'inverse généralisé $A^\dagger$ de A (proposition 9.5).

Exemple 9.5 : Reprenons l'exemple 9.2. L'équation normale s'écrit

$$\begin{pmatrix} x_1 & \ldots & x_m \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_1 & \ldots & x_m \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

c'est-à-dire

$$\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & m \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}.$$

La matrice A^TA est inversible si et seulement si son déterminant $m\sum x_i^2 - \left(\sum x_i\right)^2$ n'est pas nul. Il revient au même de dire que

$$\left(\sum x_i\right)^2 < m\sum x_i^2.$$

Il s'agit là de l'inégalité de Cauchy-Schwarz appliquée aux vecteurs $(x_1, \ldots, x_m)$ et $(1, \ldots, 1)$. Lorsque les x_i ne sont pas tous égaux, ces deux vecteurs ne sont pas proportionnels et l'inégalité de Cauchy-Schwarz est stricte.
Ainsi, lorsque qu'au moins deux des x_i sont distincts, la droite des moindres carrés est unique et donnée par $y = \alpha x + \beta$ avec

$$\alpha = \frac{\Gamma(x,y)}{\sigma^2(x)},\ \beta = \overline{y} - \alpha\overline{x},$$

où

$$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i,\ \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i,$$

$$\sigma^2(x) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})^2,\ \Gamma(x,y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y}).$$

9.3.2 Algorithmique, complexité

a) Via Cholesky

Une méthode classique de résolution du problème des moindres carrés utilise l'équation normale

$$A^*Ax = A^*b.$$

La matrice de ce système est définie positive puisque rang $A = n$ (théorème 7.2). On calcule sa décomposition de Cholesky $CC^* = A^*A$ où $C \in \mathbb{C}^{n\times n}$ est triangulaire inférieure. On résout ensuite les systèmes

$$Cy = A^*b \text{ et } C^*x = y.$$

Il est important de noter qu'en général n est très petit devant m. Par exemple, dans l'exemple 9.2, $n = 2$ alors que m peut être très grand. Ce n'est donc pas la résolution du système donné par les équations normales qui pose problème mais le calcul de ces équations normales c'est-à-dire celui du produit A^*A. Le nombre d'opérations arithmétiques nécessaire est de mn^2 pour le calcul de A^*A parce que A^*A est symétrique (en négligeant dans le compte exact les termes d'ordre inférieur à mn^2) et de $n^3/3$ pour la décomposition de Cholesky. Le calcul de A^*b demande mn opérations et la

résolution des systèmes triangulaires $2n^2$ opérations supplémentaires. On obtient donc le total

$$(m + \frac{n}{3})n^2 \text{ opérations arithmétiques}$$

en négligeant dans le compte exact les termes d'ordre inférieur.

b) Via QR

Lorsque l'on dispose d'une décomposition QR de la matrice A, c'est-à-dire lorsque $m \geqslant n$ et $A = QR$ avec $Q \in \mathbb{S}t_{mn}$ et $R \in \mathbb{C}^{n\times n}$ triangulaire supérieure, l'équation normale $A^*Ax = A^*b$ devient $R^*Q^*QRx = R^*Q^*b$. Comme $Q \in \mathbb{S}t_{mn}$, on a $Q^*Q = I_n$ et comme A est de rang n, R est inversible de sorte que l'équation normale est

$$Rx = Q^*b, \tag{9.1}$$

qui est un système triangulaire $n \times n$. Le projecteur orthogonal $I_m - QQ^*$ fournit en outre le résidu $r = b - Ax$: on a $Ax - b = QRx - QQ^*b - (I_m - QQ^*)b$. Puisque la solution vérifie l'équation (9.1), on a

$$r = b - Ax = (I_m - QQ^*)b.$$

L'interprétation géométrique de cette égalité est claire puisque $(I_m - QQ^*)$ est le projecteur (orthogonal) sur l'orthogonal de Im A.

La résolution de l'équation (9.1) requiert n^2 opérations auxquelles il faut ajouter les $2mn$ opérations nécessaires pour calculer Q^*b.

Par cet algorithme, le calcul de la solution des moindres carrés se ramène essentiellement à celui de la décomposition QR de A. Si l'on utilise la méthode de Householder pour le calcul de QR on obtient

$$2(m - \frac{n}{3})n^2 \text{ opérations arithmétiques}$$

en négligeant dans le compte exact les termes d'ordre inférieur.

9.3.3 Analyse des erreurs

L'analyse des erreurs fait intervenir le concept de conditionnement d'une matrice rectangulaire :

Théorème 9.10 *Étant donné une matrice $A \in \mathbb{C}^{m\times n}$, le conditionnement de A pour la norme spectrale est défini par :*

$$\text{cond}_2(A) = \|A\|_2 \left\|A^\dagger\right\|_2 .$$

Lorsque rang $A = r$, *si l'on note* $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ *les valeurs singulières de* A, *alors*

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_r} \geqslant 1.$$

Soient b *et* $b' \in \mathbb{C}^n$, *notons* $x = A^\dagger b$ *et* $x' = A^\dagger b'$ *les solutions des problèmes de moindres carrés associés aux systèmes* $Ax = b$ *et* $Ax' = b'$. *On a*

$$\frac{\|x' - x\|_2}{\|x\|_2} \leqslant \mathrm{cond}_2(A) \frac{\|b' - b\|_2}{\|b\|_2}.$$

Démonstration. Nous avons vu (remarque 4.1) que $\|A\|_2 = \sigma_1$ la plus grande des valeurs singulières de A. De plus, par le théorème 9.7, les valeurs singulières de $A^\dagger$ sont $\sigma_r^{-1} \geqslant \ldots \geqslant \sigma_1^{-1} > 0$. On a donc $\|A^\dagger\|_2 = \sigma_r^{-1}$ et le théorème est établi. L'inégalité sur les erreurs relatives est une conséquence de

$$\|x' - x\|_2 = \|A^\dagger(b' - b)\|_2 \leqslant \|A^\dagger\|_2 \|b' - b\|_2$$

et de l'inégalité

$$\|b\|_2 = \|Ax\|_2 \leqslant \|A\|_2 \|x\|_2.$$

Corollaire 9.11 *Étant donné une matrice* $A \in \mathbb{C}^{m\times n}$, *le conditionnement de* A^*A *pour la norme spectrale est égal à* $\mathrm{cond}_2(A^*A) = \mathrm{cond}_2(A)^2$.

Remarque 9.3. Le calcul de la solution d'un problème de moindres carrés via l'équation normale $A^*Ax = A^*b$ fait intervenir la matrice A^*A dont le conditionnement est le carré de celui de A (corollaire 9.11). Lorsque cette matrice est mal conditionnée, la matrice A^*A est (mal conditionnée)^2 ce qui rend, dans de tels cas, l'algorithmique fondée sur la décomposition de Cholesky peu attractive.

9.4 ETUDE D'UN EXEMPLE : L'ÉQUATION $AX = B$

Étant données des matrices $A \in \mathbb{C}^{m\times n}$ et $B \in \mathbb{C}^{m\times p}$, existe-t-il une matrice $X \in \mathbb{C}^{n\times p}$ telle que $AX = B$? Cette équation matricielle possède np inconnues (les entrées de X) et mp équations scalaires. Un tel problème n'a pas nécessairement de solution et, s'il en existe, elle n'est pas nécessairement unique. Nous notons

$$\mathcal{L}_A : \mathbb{C}^{n\times p} \to \mathbb{C}^{m\times p},\ \mathcal{L}_A(X) = AX.$$

Proposition 9.12 *Une condition nécessaire et suffisante pour qu'il existe une matrice* $X \in \mathbb{C}^{n\times p}$ *telle que* $AX = B$ *est que* $\mathrm{Im}\, B \subset \mathrm{Im}\, A$. *Dans ce cas* $X_0 = A^\dagger B$ *est solution du problème.*

Démonstration. Si $AX = B$ et si $y \in \operatorname{Im} B$ on a $y = Bx$ pour un certain $x \in \mathbb{C}^p$ d'où $y = A(Xx)$ et donc $y \in \operatorname{Im} A$. Réciproquement, on a : $A(A^\dagger B) = \Pi_{\operatorname{Im} A} B$ (théorème 9.2) et $\Pi_{\operatorname{Im} A} B = B$ puisque $\operatorname{Im} B \subset \operatorname{Im} A$.

Que sont les solutions au sens des moindres carrés ?

Proposition 9.13 *L'équation normale du problème de moindres carrés*

$$\inf_{X \in \mathbb{C}^{n \times p}} \|AX - B\|_F^2$$

est

$$A^* AX = A^* B.$$

La solution de norme minimale est

$$X_0 = A^\dagger B.$$

Démonstration. L'équation normale est $(\mathcal{L}_A)^* \circ \mathcal{L}_A(X) = (\mathcal{L}_A)^*(B)$ c'est-à-dire $A^* AX = A^* B$ parce que $(\mathcal{L}_A)^* = \mathcal{L}_{A^*}$ et que $(\mathcal{L}_A)^* \circ \mathcal{L}_A = \mathcal{L}_{A^* A}$. Passons à la solution de norme minimale. En vertu du théorème 9.8 il suffit de montrer que l'inverse généralisé de $\mathcal{L}_A$ est l'application

$$\mathcal{L}_{A^\dagger} : \mathbb{C}^{m \times p} \to \mathbb{C}^{n \times p},\ \mathcal{L}_{A^\dagger}(Y) = A^\dagger Y.$$

Cela résulte des identités $\mathcal{L}_U \circ \mathcal{L}_V = \mathcal{L}_{UV}$, $(\mathcal{L}_U)^* = \mathcal{L}_{U^*}$ et de la proposition 9.4.

9.5 NOTES ET RÉFÉRENCES

Deux grands noms sont associés à la méthode des moindres carrés : A. Legendre (1752-1833) qui introduisit cette méthode en appendice d'un ouvrage sur la détermination des trajectoires de comètes (1805) et C. F. Gauss (1777-1855) intéressé par la détermination de la trajectoire de l'astéroïde Cérès. Une querelle d'antériorité opposera les deux hommes lorsque Gauss publiera sa méthode.

Les équations normales ont été, jusqu'à une date récente, la seule voie possible pour résoudre les problèmes de moindres carrés. La décomposition de Cholesky a été créée pour résoudre ces équations. On doit à Gene Golub (1932-2007) l'approche fondée sur la triangulation QR via la méthode de Householder (article publié en 1965 [14]). Pour en savoir plus on peut consulter les ouvrages de Björck [5] et de Golub-van Loan [15].

EXERCICES

Exercice 9.1

1. Calculer l'inverse généralisé de la matrice $L = (a, b) \in \mathbb{R}^{1\times 2}$ où a et b sont deux réels donnés tels que $a^2 + b^2 \neq 0$.
2. Soit $M \in \mathbb{R}^{2\times 2}$ définie par $M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Montrer que $M^{-1}L^\dagger \neq (LM)^\dagger$.

Exercice 9.2

Calculer l'inverse généralisé d'une matrice-colonne.

Exercice 9.3

Soient x et $y \in \mathbb{C}^n$, non nuls. Montrer que l'inverse généralisé de la matrice $A = xy^*$ est

$$A^\dagger = \frac{yx^*}{\|x\|_2^2 \|y\|_2^2}.$$

Exercice 9.4

Montrer que pour toute matrice $A \in \mathbb{C}^{m\times n}$ on a $(AA^*)^\dagger = (A^*)^\dagger A^\dagger$.

Exercice 9.5

Soit $A \in \mathbb{C}^{n\times n}$. Montrer qu'en général $(A^k)^\dagger \neq (A^\dagger)^k$, mais que l'égalité a lieu lorsque A est une matrice normale.

Exercice 9.6

Soit $A \in \mathbb{C}^{n\times n}$. Les valeurs propres non nulles de $A^\dagger$ sont-elles les inverses des valeurs propres non nulles de A ?

Exercice 9.7

1. Montrer que pour toute matrice $A \in \mathbb{C}^{m\times n}$ et $t \in \mathbb{C}$, $t \neq 0$, et suffisamment petit, les matrices $tI_n + A^*A$ et $tI_m + AA^*$ sont inversibles.
2. En déduire que

$$A^\dagger = \lim_{t\to 0} (tI_n + A^*A)^{-1}A^* = \lim_{t\to 0} A^*(tI_m + AA^*)^{-1}.$$

Exercice 9.8

Soit $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Résoudre $Ax = b$ par la méthode des moindres carrés et donner la solution de norme minimale lorsque $b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ et $b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Exercice 9.9

Soit le système $Ax = b$ que l'on veut résoudre au sens des moindres carrés, où

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Calculer

1. Im A, Ker A, $(\text{Ker } A)^{\perp}$,
2. La projection orthogonale $\tilde{b}$ de b sur Im A,
3. L'ensemble des solutions du système $Ax = \tilde{b}$,
4. La solution du système $Ax = \tilde{b}$ qui est dans $(\text{Ker } A)^{\perp}$ (solution au sens des moindres carrés).
5. Obtenir ce même résultat via l'inverse généralisé de A et via l'équation normale $A^T Ax = A^T b$.

Exercice 9.10

Donnons nous une matrice $A \in \mathbb{GL}_n$, un vecteur $b \in \mathbb{C}^n$, la solution $x = A^{-1}b$ du système $Ax = b$ et une approximation $x' \neq 0$ de x. Montrer que

$$\min_{\substack{E \in \mathbb{C}^{n \times n} \\ (A+E)x' = b}} \|E\|_F = \frac{\|A(x' - x)\|_2}{\|x'\|_2}$$

et que le minimum est atteint pour

$$E = \frac{A(x - x')x'^*}{\|x'\|_2^2}.$$

Exercice 9.11

Soit $S \in \mathbb{C}^{n \times n}$ une matrice définie positive. On note $\langle ., . \rangle_S$ le produit scalaire sur $\mathbb{C}^n$ défini par

$$\langle x, y \rangle_S = \langle x, Sy \rangle = y^* Sx$$

et par $\|.\|_S$ la norme qui lui est associée.

Étant donnés une matrice $A \in \mathbb{GL}_n$ et un vecteur $b \in \mathbb{C}^n$, montrer que le problème des *moindres carrés pondérés*

$$\inf_{x \in \mathbb{C}^n} \|Ax - b\|_S^2$$

possède au moins une solution et qu'elles sont données par l'équation normale

$$A^* S A x = A^* S b.$$

Exercice 9.12

Soit $A \in \mathbb{C}^{m \times n}$ de rang n. On considère le système

$$\begin{pmatrix} I_m & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{9.2}$$

1. Montrer que la matrice $M = \begin{pmatrix} I_m & A \\ A^* & 0 \end{pmatrix} \in \mathbb{C}^{(n+m)\times(n+m)}$ est inversible.
2. Montrer que le vecteur x obtenu à partir de la solution $\begin{pmatrix} r \\ x \end{pmatrix}$ du système (9.2) est solution du problème des moindres carrés
$$\inf_{x \in \mathbb{C}^n} \|Ax - b\|_2^2.$$
3. Calculer les valeurs propres de M à l'aide des valeurs singulières de A. En déduire le conditionnement $\mathrm{cond}_2 M$ (pour mener le calcul de $\|M^{-1}\|_2$ distinguer les cas $\sigma_{\min} \leqslant \sqrt{2}$ et $\sigma_{\min} > \sqrt{2}$ où $\sigma_{\min}$ est la plus petite valeur singulière non nulle de A).

Exercice 9.13 Moindres carrés régularisés

Soient $A \in \mathbb{C}^{m \times n}$ de rang $p \leqslant n$, $b \in \mathbb{C}^m$ et $\rho > 0$. On considère le problème $\mathcal{P}_\rho$

$$\inf_{x \in \mathbb{C}^n} \|Ax - b\|_2^2 + \rho \|x\|_2^2.$$

1. Écrire $\mathcal{P}_\rho$ sous forme d'un problème des moindres carrés standard, donner l'équation normale du problème et montrer qu'il possède une solution unique x_ρ (montrer que $A^*A + \rho I_n$ est inversible).
2. Soient ρ et ρ' tels que $0 < \rho \leqslant \rho'$. Montrer que $\|x_\rho\|_2 \geqslant \|x_{\rho'}\|_2$ et que $\|Ax_\rho - b\|_2 \leqslant \|Ax_{\rho'} - b\|_2$ (on pourra utiliser les propriétés d'optimalité de x_ρ et $x_{\rho'}$ et montrer que $\left\|(A^*A + \rho' I_n)^{-1}(A^*A + \rho I_n)\right\|_2 \leqslant 1$).

3. Montrer que $\lim_{\rho \to 0} x_\rho = x_0$ la solution de norme minimale du problème de moindres carrés

$$\inf_{x \in \mathbb{C}^m} \|Ax - b\|_2^2$$

(utiliser l'exercice 9.7).

Exercice 9.14

Soit $A \in \mathbb{C}^{m \times n}$ de rang $p \leqslant n$ et $\rho > 0$. On considère le système

$$\begin{pmatrix} I_m & A \\ A^* & -\rho I_n \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{9.3}$$

1. Montrer que la matrice $M_\rho = \begin{pmatrix} I_m & A \\ A^* & -\rho I_n \end{pmatrix}$ est inversible.
2. Montrer que le vecteur x_ρ obtenu à partir de la solution $\begin{pmatrix} r_\rho \\ x_\rho \end{pmatrix}$ du système (9.3) est solution du problème des moindres carrés régularisé $\mathcal{P}_\rho$ (voir exercice 9.13).
3. Calculer les valeurs propres de M_ρ à l'aide des valeurs singulières de A ainsi que le conditionnement $\text{cond}_2 M$.

Chapitre **10**

Méthodes itératives

Dans les chapitres précédents nous avons étudié des décompositions matricielles (LU, Cholesky, QR) qui permettent de ramener la résolution de systèmes linéaires à celle de systèmes triangulaires. Ces méthodes sont qualifiées de *directes* parce que le calcul de la solution est obtenu après un nombre fini d'opérations.

Les méthodes itératives suivent une autre approche : elles calculent une suite d'approximations successives de la solution du problème. En théorie ce processus est infini, en pratique le calcul d'une solution approchée est arrêté dès que l'on estime avoir atteint une précision suffisante. Les méthodes itératives sont utilisées pour résoudre des systèmes de grande taille et à *matrices creuses*[1]. Cette approche est en effet plus naturelle dans ces cas-là puisqu'il n'est pas nécessaire comme dans les méthodes directes de porter à terme le calcul d'une décomposition de la matrice qui serait extrêmement coûteux.

La méthode des approximations successives pour la résolution de l'équation de point fixe $x = f(x)$ est donnée par le schéma itératif $x_{k+1} = f(x_k)$ où le point initial x_0 est donné. Si la suite (x_k) converge vers x et si f est continue en ce point alors $f(x) = x$ et les x_k constituent autant d'approximations du point fixe x. On transforme un système linéaire $Ax = b$ en une équation de point fixe en « cassant » la matrice A en $A = M - N$ où M est inversible et « facile à inverser ». Le système devient

$$x = M^{-1}Nx + M^{-1}b$$

1. On dit qu'une matrice est creuse lorsqu'elle a « beaucoup » de coefficients nuls. C'est le cas notamment des systèmes obtenus par discrétisation d'équations aux dérivées partielles (voir le chapitre 16).

qui conduit au schéma itératif

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b$$

que nous allons étudier ici.

10.1 RÉSULTATS GÉNÉRAUX

Au système linéaire $Ax = b$ avec $A \in \mathbb{GL}_n$ et $b \in \mathbb{C}^n$ nous associons le schéma itératif

$$x_{k+1} = Bx_k + c, \ x_0 \text{ donné},$$

où $B \in \mathbb{C}^{n\times n}$, $c \in \mathbb{C}^n$.

Définition 10.1

1. *On dit que cette méthode itérative est consistante si $I_n - B$ est inversible et si $A^{-1}b = (I_n - B)^{-1}c$.*
2. *On dit qu'elle est convergente si pour tout $x_0 \in \mathbb{C}^n$ la suite (x_k) définie ci-dessus est convergente.*

Remarque 10.1. Lorsqu'une méthode itérative est consistante, le point fixe qu'elle définit est la solution du système $Ax = b$.

Remarque 10.2. Une méthode itérative consistante $x_{k+1} = Bx_k + c$ ne construit pas nécessairement une suite d'approximations de la solution du système $Ax = b$. Un exemple simplissime est donné par le système $2I_n x = b$ et la méthode itérative $x_{k+1} = -x_k + b$. La suite (x_k) vérifie $x_{2k} = b - x_0$ et $x_{2k+1} = b$. Elle ne converge pas vers la solution $x = b/2$ (sauf dans le cas très particulier $b = x_0 = 0$).

La suite (x_k) est donnée par

$$x_k = B^k x_0 + \left(\sum_{i=0}^{k-1} B^i\right) c.$$

Cette identité montre que la convergence de la méthode est liée à celle de la série de matrices $\sum_{k=0}^{\infty} B^k$ que nous avons déjà rencontrée au paragraphe 3.5.

Théorème 10.2 *Pour toute matrice $B \in \mathbb{C}^{n\times n}$ il y a équivalence entre :*

1. *La série $\sum_{k=0}^{\infty} B^k$ est convergente,*
2. $\lim_{k\to\infty} B^k = 0$,

3. *Le rayon spectral de B vérifie $\rho(B) < 1$.*

Sous ces hypothèses, la convergence de la série est absolue, la matrice $I_n - B$ est inversible et $(I_n - B)^{-1} = \sum_{k=0}^{\infty} B^k$.

Démonstration. 1 implique 2 parce que le terme général d'une série convergente a pour limite 0.

2 implique 3 : si $\lambda \in \mathbb{C}$ est une valeur propre de B et si $Bx = \lambda x$ avec $\|x\| = 1$ alors, pour une norme matricielle consistante on a

$$|\lambda^k| = \|\lambda^k x\| = \|B^k x\| \leqslant \|B^k\| \to 0.$$

Si $\lambda^k \to 0$ c'est que $|\lambda| < 1$ et ceci prouve la troisième assertion.

3 implique 1. C'est une conséquence du critère de d'Alembert : la série $\sum_{k=0}^{\infty} B^k$ converge absolument si

$$\lim_{k\to\infty} \|B^k\|^{1/k} < 1.$$

Par le théorème 3.7 cette limite est égale à $\rho(B) < 1$ et donc le critère de d'Alembert est vérifié.

Lorsque ces conditions sont satisfaites, la somme de la série se calcule via l'identité

$$(I_n - B)\sum_{i=0}^{k-1} B^i = I_n - B^k$$

et un passage à la limite.

La conséquence attendue de ce théorème est donnée par :

Théorème 10.3 *Pour toute matrice $B \in \mathbb{C}^{n\times n}$ telle que $I_n - B$ soit inversible et pour tout $c \in \mathbb{C}$, la méthode itérative*

$$x_{k+1} = Bx_k + c$$

est convergente si et seulement si l'une des trois conditions équivalentes du théorème 10.2 est satisfaite.

Démonstration. La condition est nécessaire : soit $x \in \mathbb{C}^n$ tel que $x = Bx+c$. On a

$$x_k - x = B(x_{k-1} - x) = \ldots = B^k(x_0 - x).$$

Comme par hypothèse la méthode itérative est convergente on a $B^k(x_0 - x) \to 0$ pour tout x_0 ; en prenant pour $x_0 - x$ un vecteur propre unitaire de B cela prouve que $\rho(B) < 1$.

La condition est suffisante : cela résulte de l'égalité

$$x_k = B^k x_0 + \left(\sum_{i=0}^{k-1} B^i \right) c.$$

10.2 CHOIX D'UN TEST D'ARRÊT

Dans la pratique, il faut décider d'un test d'arrêt pour savoir quand mettre fin au processus itératif. Deux tests « tombent sous le sens » étant donné un seuil de précision $\varepsilon > 0$ ce sont

$$\|Ax_k - b\| \leqslant \varepsilon$$

pour le premier et

$$\|x_k - x_{k-1}\| \leqslant \varepsilon$$

pour le second.

Il faut noter que le premier test peut n'être jamais satisfait même si la méthode converge. Il se peut, en effet, que les erreurs d'arrondis dues à l'usage d'une arithmétique de précision finie soient du même ordre que le gain de précision obtenu à l'itération en cours.

Le second test est plus réaliste. On arrête l'itération lorsqu'elle ne produit plus de gain significatif de précision. Il se peut qu'alors la quantité $\|Ax_k - b\|$ soit significativement grande.

Le test idéal (mais irréaliste) est bien sûr lié à la distance à la solution :

$$\left\|x_k - A^{-1}b\right\| \leqslant \varepsilon.$$

Ces trois quantités sont reliées par :

Proposition 10.4 *Donnons-nous des normes $\|.\|$ sur $\mathbb{C}^n$, et $\|.\|$ sur $\mathbb{C}^{n\times n}$ consistante avec la précédente. Étant donné une méthode itérative consistante $x_{k+1} = Bx_k + c$ associée au système linéaire $Ax = b$ on a :*

1. *Si $\|Ax_k - b\| \leqslant \varepsilon$ alors*

$$\|x_k - x\| \leqslant \left\|A^{-1}\right\| \varepsilon \text{ et } \frac{\|x_k - x\|}{\|x\|} \leqslant \operatorname{cond}(A)\frac{\varepsilon}{\|b\|}.$$

2. *Si $\|x_k - x_{k-1}\| \leqslant \varepsilon$ alors*

$$\|x_k - x\| \leqslant \left\|(I_n - B)^{-1}\right\| \varepsilon \text{ et } \frac{\|x_k - x\|}{\|x\|} \leqslant \operatorname{cond}(I_n - B)\frac{\varepsilon}{\|c\|}.$$

Démonstration. Dans le premier cas, $x_k - x = A^{-1}(Ax_k - b)$ d'où $\|x_k - x\| \leqslant \|A^{-1}\| \varepsilon$. L'erreur relative est donnée par

$$\|x_k - x\| \leqslant \|A^{-1}\| \varepsilon \|b\| / \|b\| \text{ et } \|b\| = \|Ax\| \leqslant \|A\| \|x\| .$$

Le second cas se traite *mutatis mutandis* de façon similaire.

Un des intérêts des méthodes itératives convergentes est dû à l'absence d'accumulation des erreurs d'arrondis : que l'on utilise l'itéré x_k ou une valeur voisine $\tilde{x}_k$ on a quand même affaire à deux points initiaux pour une méthode convergente. Le résultat suivant précise les propriétés des schémas itératifs approchés :

Proposition 10.5 *Donnons-nous des normes $\|.\|$ sur $\mathbb{C}^n$, et $\|.\|$ sur $\mathbb{C}^{n\times n}$ consistante avec la précédente. Considérons une méthode itérative consistante et convergente $x_{k+1} = Bx_k + c$ associée au système linéaire $Ax = b$ et supposons que $\|B\| \leqslant \lambda < 1$. Soit $\varepsilon > 0$ et soit (x_k) une suite de points de $\mathbb{C}^n$ qui vérifie*

$$\|x_{k+1} - (Bx_k + c)\| \leqslant \varepsilon.$$

On a

$$\|x_k - x\| \leqslant \lambda^k \|x_0 - x\| + \frac{\varepsilon}{1-\lambda}$$

pour tout $k \geqslant 0$.

Démonstration. Nous allons montrer, par récurrence sur k, que

$$\|x_k - x\| \leqslant \lambda^k \|x_0 - x\| + \varepsilon \sum_{i=0}^{k-1} \lambda^i.$$

L'inégalité en résulte puisque $\sum_{i=0}^{k-1} \lambda^i \leqslant 1/(1-\lambda)$. Pour $k = 0$ il n'y a rien à démontrer. Le passage de k à $k+1$ se fait ainsi :

$$\|x_{k+1} - x\| \leqslant \|x_{k+1} - (Bx_k + c)\| + \|(Bx_k + c) - (Bx + c)\| \leqslant \varepsilon + \lambda \|x_k - x\|$$

$$\leqslant \varepsilon + \lambda \left(\lambda^k \|x_0 - x\| + \varepsilon \sum_{i=0}^{k-1} \lambda^i \right) = \lambda^{k+1} \|x_0 - x\| + \varepsilon \sum_{i=0}^{k} \lambda^i.$$

Cette proposition prouve que la suite (x_k) « converge » vers la boule de centre x et de rayon $\frac{\varepsilon}{1-\lambda}$. Ce rayon mesure la précision maximum que l'on peut obtenir.

10.3 EXEMPLES DE MÉTHODES ITÉRATIVES

Nous allons utiliser les notations suivantes : $A \in \mathbb{GL}_n$ est décomposé en

$$A = D - E - F$$

avec

- $d_{ij} = a_{ij}$ si $i = j$ et 0 sinon,
- $e_{ij} = -a_{ij}$ si $i > j$ et 0 sinon,
- $f_{ij} = -a_{ij}$ si $i < j$ et 0 sinon.

Nous supposons aussi que

$$a_{ii} \neq 0 \text{ pour tout } i = 1 \ldots n$$

de sorte que D, $D - E$ et $D - F$ sont inversibles.

10.3.1 Méthode de Jacobi

Cette méthode utilise la décomposition $A = D-(E+F)$. La matrice D étant diagonale elle est bien sûr facile à inverser. On obtient le schéma

$$x_{k+1} = Jx_k + D^{-1}b, \quad J = D^{-1}(E + F)$$

d'où

$$x_{k+1,i} = \frac{1}{a_{ii}} \left(-\sum_{j \neq i} a_{ij} x_{k,j} + b_i \right), \quad 1 \leqslant i \leqslant n.$$

10.3.2 Méthode de Gauss-Seidel

Cette méthode utilise la décomposition $A = (D - E) - F$. La matrice $D - E$ est triangulaire inférieure donc facile à inverser. On obtient

$$x_{k+1} = Gx_k + (D - E)^{-1}b, \quad G = (D - E)^{-1}F,$$

d'où

$$x_{k+1,i} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_{k+1,j} - \sum_{j=i+1}^{n} a_{ij} x_{k,j} + b_i \right), \quad 1 \leqslant i \leqslant n.$$

10.3.3 Méthode de relaxation ou SOR

On se donne un paramètre $\omega \in \mathbb{R}$. Cette méthode est définie par :

$$x_{k+1} = G_\omega x_k + \omega(D - \omega E)^{-1} b, \quad G_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega F)$$

d'où le schéma

$$x_{k+1,i} = \frac{1}{a_{ii}} \left(-\omega \sum_{j=1}^{i-1} a_{ij} x_{k+1,j} + (1-\omega) a_{ii} x_{k,i} - \omega \sum_{j=i+1}^{n} a_{ij} x_{k,j} + \omega b_i \right), \quad 1 \leqslant i \leqslant n$$

Nous verrons que l'on choisit toujours $\omega \in]0,2[$. Le cas $\omega = 1$ correspond à la méthode de Gauss-Seidel. La dénomination SOR vient de l'anglais *Successive Over Relaxation.*

10.3.4 Méthode de relaxation symétrique ou SSOR

Après une étape de type SOR, on effectue une autre étape de même type mais en échangeant les rôles de E et F. On obtient :

$$(D - \omega E) x_{k+1/2} = ((1-\omega)D + \omega F) x_k + \omega b,$$

$$(D - \omega F) x_{k+1} = ((1-\omega)D + \omega E) x_{k+1/2} + \omega b.$$

La dénomination SSOR vient de l'anglais *Symmetric Successive Over Relaxation.*

On obtient l'itération suivante entre x_k et x_{k+1} :

$$x_{k+1} = S_\omega x_k + \omega(2-\omega)(D - \omega F)^{-1} D (D - \omega E)^{-1} b,$$

avec

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

10.3.5 Méthodes par blocs

Supposons que la matrice A s'écrive de la façon suivante :

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ A_{21} & A_{22} & \dots & A_{2p} \\ \vdots & \vdots & & \vdots \\ A_{p1} & A_{p2} & \dots & A_{pp} \end{pmatrix}$$

où les blocs $A_{ii} \in \mathbb{C}^{n_i \times n_i}$ sont inversibles. On peut générer des méthodes de Jacobi, Gauss-Seidel, SOR, SSOR par blocs en utilisant les mêmes formules, la décomposition

$A = D - E - F$ étant ici entendue « par blocs ». Notons aussi $x^T = (X_1^T, \ldots, X_p^T)$ avec

$$X_i = (x_j),\ n_1 + \ldots + n_{i-1} + 1 \leqslant j \leqslant n_1 + \ldots + n_{i-1} + n_i$$

et, de façon similaire, $b^T = (B_1^T, \ldots, B_p^T)$. La méthode de Jacobi par blocs s'écrit :

$$X_{k+1,i} = A_{ii}^{-1} \left(- \sum_{j \neq i} A_{ij} X_{k,j} + B_i \right), \quad 1 \leqslant i \leqslant n.$$

On procède de même pour les autres méthodes. Il faut bien sûr prendre garde à la non commutativité des produits de matrices.

10.4 CONVERGENCE DES MÉTHODES ITÉRATIVES

Le théorème 10.3 fournit un critère général pour déterminer les propriétés de convergence d'une méthode itérative. Dans ce paragraphe nous allons utiliser ce critère pour étudier quelques cas classiques.

10.4.1 Matrices à diagonale strictement dominante

Définition 10.6 *Une matrice $A \in \mathbb{C}^{n \times n}$ est à diagonale strictement dominante lorsque*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

pour tout i.

Nous avons rencontré ces matrices à l'exercice 3.18 où nous avons vu qu'une telle matrice est inversible.

Théorème 10.7 *Si A est à diagonale strictement dominante, les méthodes de Jacobi et de Gauss-Seidel convergent.*

Démonstration. Avec les notations introduites aux paragraphes consacrés à ces méthodes on a $J = D^{-1}(E + F)$, $G = (D - E)^{-1}F$ et l'on doit prouver (théorème 10.3) que $\rho(J)$ et $\rho(G) < 1$.

Pour la méthode de Jacobi, on a : $J_{ij} = 0$ si $i = j$ et $-a_{ij}/a_{ii}$ si $i \neq j$ donc

$$\|J\|_\infty = \max_i \sum_j |J_{ij}| = \max_i \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1$$

(voir l'exemple 3.1 pour la définition de $\|J\|_\infty$) puisque A est à diagonale strictement dominante. On conclut à l'aide de la proposition 3.6 :

$$\rho(J) \leqslant \|J\|_\infty < 1.$$

Passons à la méthode de Gauss-Seidel. Nous allons montrer que $|\lambda| < 1$ pour toute valeur propre λ de $G = (D - E)^{-1}F$. L'inégalité étant évidente si $\lambda = 0$, nous supposons donc que $\lambda \neq 0$. On a $\det((D-E)^{-1}F - \lambda I_n) = 0$ d'où $\det(F - \lambda(D - E)) = 0$ autrement dit 0 est valeur propre de

$$-F + \lambda(D - E) = \begin{pmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nn} \end{pmatrix}.$$

D'après le théorème de Gershgorin (théorème 12.1), il existe i tel que

$$|0 - \lambda a_{ii}| \leqslant \sum_{j<i} |\lambda a_{ij}| + \sum_{j>i} |a_{ij}|.$$

D'après l'hypothèse on a :

$$\sum_{j<i} |\lambda a_{ij}| + \sum_{j>i} |\lambda a_{ij}| < |\lambda a_{ii}| \leqslant \sum_{j<i} |\lambda a_{ij}| + \sum_{j>i} |a_{ij}|$$

ce qui prouve que

$$\sum_{j>i} |\lambda a_{ij}| < \sum_{j>i} |a_{ij}|$$

d'où $|\lambda| < 1$ et $\rho(G) < 1$.

10.4.2 Convergence de la méthode de relaxation

Théorème 10.8 *Le rayon spectral de la matrice G_ω (méthode de relaxation) vérifie :*

$$\rho(G_\omega) \geqslant |\omega - 1|.$$

Une condition nécessaire pour que la méthode de relaxation converge est que $0 < \omega < 2$.

Démonstration. Le résultat est évident si $\omega = 1$. Lorsque $\omega \neq 1$ écrivons :

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

La matrice $(D - \omega E)^{-1}$ est triangulaire inférieure de diagonale D^{-1}, la matrice $((1 - \omega)D + \omega F)$ est triangulaire supérieure de diagonale $(1 - \omega)D$, aussi

$$\det G_\omega = \det D^{-1} \det((1 - \omega)D) = (1 - \omega)^n.$$

Comme $\det G_\omega$ est le produit des valeurs propres de G_ω on a :

$$|\det G_\omega| \leqslant \rho(G_\omega)^n$$

ce qui prouve que

$$|1-\omega| \leqslant \rho(G_\omega).$$

Pour que la méthode converge il faut que $\rho(G_\omega) < 1$ d'où $|1-\omega| < 1$ c'est-à-dire $0 < \omega < 2$.

10.4.3 Le cas des matrices hermitiennes

Le théorème suivant donne un cadre général d'étude de la convergence lorsque A est hermitienne.

Théorème 10.9 *Soit $A \in \mathbb{C}^{n\times n}$ hermitienne et inversible. Écrivons $A = M - N$ avec M inversible et supposons que $M^* + N$ est définie positive. La méthode itérative*

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b$$

converge si et seulement si A est définie positive.

Démonstration. Remarquons tout d'abord que $M^* + N = M^* + M - A$ est hermitienne puisque A est hermitienne.
Supposons que A soit définie positive et notons

$$\|x\|_A^2 = x^*Ax = \langle Ax, x\rangle$$

la norme associée au produit scalaire

$$\langle x, y\rangle_A = \langle Ax, y\rangle .$$

La norme d'endomorphisme associée à cette norme vectorielle est

$$\|X\|_A = \sup_{x\neq 0} \frac{\|Xx\|_A}{\|x\|_A}.$$

Nous allons prouver que $\left\|M^{-1}N\right\|_A < 1$ d'où il résultera que $\rho(M^{-1}N) < 1$ (proposition 3.6) et donc la convergence de la méthode (théorème 10.3).
Compte tenu du fait que le sup définissant $\left\|M^{-1}N\right\|_A$ est un maximum (exercice 3.1) et que $M^{-1}N = I_n - M^{-1}A$, nous devons prouver que

$$x^*(I_n - M^{-1}A)^*A(I_n - M^{-1}A)x < x^*Ax$$

pour tout $x \neq 0$. Partons de l'hypothèse $M^* + N = M^* + M - A$ définie positive. Pour tout $y \neq 0$, on a :

$$y^*(M^* + M - A)y > 0$$

et pour $y = M^{-1}Ax$, $x \neq 0$, on obtient

$$x^*AM^{-*}(M^*+M-A)M^{-1}Ax = x^*A(M^{-1}+M^{-*}-M^{-*}AM^{-1})Ax > 0$$

c'est-à-dire

$$x^*Ax - x^*\left(A - AM^{-1}A - AM^{-*}A + AM^{-*}AM^{-1}A\right)x =$$

$$x^*Ax - x^*(I_n - M^{-1}A)^*A(I_n - M^{-1}A)x > 0$$

ce qui est bien l'inégalité souhaitée.
Supposons maintenant que la méthode soit convergente :

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

Notons $E = I_n - M^{-1}A$ et $F = AM^{-*}(M^* + M - A)M^{-1}A$. Comme $M^* + N = M + M^* - A$ est définie positive, cette matrice peut s'écrire B^*B avec $B \in \mathbb{GL}_n$ (penser par exemple à la décomposition de Cholesky) donc $F = C^*C$ avec $C = BM^{-1}A$ ce qui prouve que F est définie positive. D'autre part

$$A = F + E^*AE$$

d'où, par récurrence,

$$A = E^{*k}AE^k + \sum_{i=0}^{k-1} E^{*i}FE^i.$$

Comme $\rho(E) < 1$ la suite E^k a pour limite 0 (théorème 10.3) et donc

$$A = \sum_{k=0}^{\infty} E^{*k}FE^k.$$

Ceci prouve que A est définie positive.

Corollaire 10.10 *Lorsque les matrices A et $2D - A$ sont définies positives, la méthode de Jacobi est convergente.*

Démonstration. On a : $M = D$ et $N = E+F$. Par hypothèse A et $M^*+N = 2D - A$ sont définies positives et le théorème précédent s'applique.

Remarque 10.3. Ce résultat s'étend à la méthode de Jacobi par blocs mais alors D est la diagonale par blocs de la matrice A.

Corollaire 10.11 *Lorsque la matrice A est définie positive, la méthode de relaxation est convergente si et seulement si $0 < \omega < 2$. En particulier, la méthode de Gauss-Seidel est convergente.*

Démonstration. Le théorème 10.8 montre que cette condition est nécessaire. La méthode de relaxation correspond au découpage $M = \frac{1}{\omega}D - E$ et $N = \frac{1-\omega}{\omega}D + F$ et donc $M^* + N = \frac{2-\omega}{\omega}D$. Notons que les entrées diagonales de D sont positives parce que A est définie positive. Si $0 < \omega < 2$ la matrice $M^* + N$ est définie positive et le théorème précédent s'applique.

Remarque 10.4. Ce résultat s'étend à la méthode de relaxation par blocs. Comme pour la méthode de Jacobi, D est alors la diagonale par blocs de la matrice A.

Corollaire 10.12 *Lorsque la matrice A est définie positive et que $0 < \omega < 2$, la méthode de relaxation symétrique est convergente.*

Démonstration. Ecrivons $x_{k+1/2} = G_E x_k + c_1$ et $x_{k+1} = G_F x_{k+1/2} + c_2$. Soit x la solution du système $Ax = b$. Comme $x = G_E x + c_1$ et $x = G_F x + c_2$ on a :

$$x_{k+1/2} - x = G_E(x_k - x),\ x_{k+1} - x = G_F(x_{k+1/2} - x),\ x_{k+1} - x = G_F G_E(x_k -$$

Puisque $0 < \omega < 2$, les méthodes de relaxation associées aux matrices G_E et G_F convergent et il résulte de la démonstration du théorème 10.9 que $\|G_E\|_A$ et $\|G_F\|_A < 1$. Cette norme étant multiplicative et comme d'autre part $\|G_E\|_A = \|G_F\|_A$ puisque A est hermitienne on a :

$$\|G_F G_E\|_A \leqslant \|G_F\|_A \|G_E\|_A = \|G_E\|_A^2 < 1$$

ce qui prouve que la méthode de relaxation converge.

Remarque 10.5. La démonstration précédente prouve que les suites (x_k^{SOR}) (méthode de relaxation) et (x_k^{SSOR}) (méthode de relaxation symétrique) issues d'un même point initial x_0 vérifient

$$\left\|x_k^{SOR} - x\right\| \leqslant \|G_E\|_A^k \|x_0 - x\|$$

et

$$\left\|x_k^{SSOR} - x\right\| \leqslant \|G_E\|_A^{2k} \|x_0 - x\| .$$

La convergence de SSOR est donc deux fois plus rapide que celle de SOR. Mais une itération SSOR correspond à deux itérations SOR ... ce gain est bien sûr illusoire !

10.5 EXEMPLES

On considère le système $Ax = b$ où A est la matrice tridiagonale définie positive donnée au paragraphe 16.1 :

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

On compare les normes des erreurs $\|e_k\|_2$ où $e_k = x_k - x$, obtenues par les méthodes itératives de Jacobi et de Gauss-Seidel tout au long des itérations. Pour tout schéma itératif $x_{k+1} = Bx_k + c$, la norme de l'erreur vérifie l'inégalité

$$\|e_k\| \leqslant \|B^k\| \, \|e_0\|$$

et donc

$$\|e_k\| \, / \, \|e_0\| \leqslant \|B^k\|.$$

D'après le théorème 3.7 la suite $(\|B^k\|)$ est asymptotiquement équivalente à la suite $(\rho(B)^k)$. Ceci permet de déterminer approximativement le nombre d'itérations nécessaires pour majorer le rapport des erreurs $\|e_k\| \, / \, \|e_0\|$ par une valeur fixée α, $0 < \alpha < 1$. Il suffit pour cela de considérer l'égalité $\rho(B)^r = \alpha$, de laquelle on déduit $r = \log(\alpha)/\log(\rho(B))$, et on prend $k = \lceil r \rceil$ le plus petit entier supérieur où égal à r.

Pour la méthode de Jacobi, le rayon spectral $\rho(J)$ s'obtient facilement à partir du spectre de A (voir paragraphe 16.1) :

$$\rho(J) = \cos(\frac{\pi}{n+1}),$$

où n est la dimension de la matrice. Pour la méthode de Gauss-Seidel, nous avons que $\rho(G_1) = \rho(J)^2$ car la matrice A est tridiagonale à valeurs constantes sur chaque diagonale (voir exercice 10.5).

En prenant $n = 100$ et $\alpha = 1/10$ (c'est-à-dire pour diviser l'erreur par 10) il faut $\ln(0.1) \, / \, \ln(\cos(\frac{\pi}{101})) \approx 4759$ itérations dans le cas de la méthode de Jacobi et deux fois moins dans le cas de la méthode de Gauss-Seidel. C'est ce que l'on observe sur la

figure 10.1. Plus précisemment, on voit que le logarithme de la norme de l'erreur dans le cas de la méthode de Jacobi suit approximativement la droite de pente $\ln(\rho(J))$:

$$\ln(\|e_k\|_2) \approx k \ln(\rho(J)) + \ln(\|e_0\|_2)$$

et dans le cas de Gauss-Seidel la droite de pente double $2 \ln(\rho(J))$:

$$\ln(\|e_k\|_2) \approx 2k \ln(\rho(J)) + \ln(\|e_0\|_2).$$

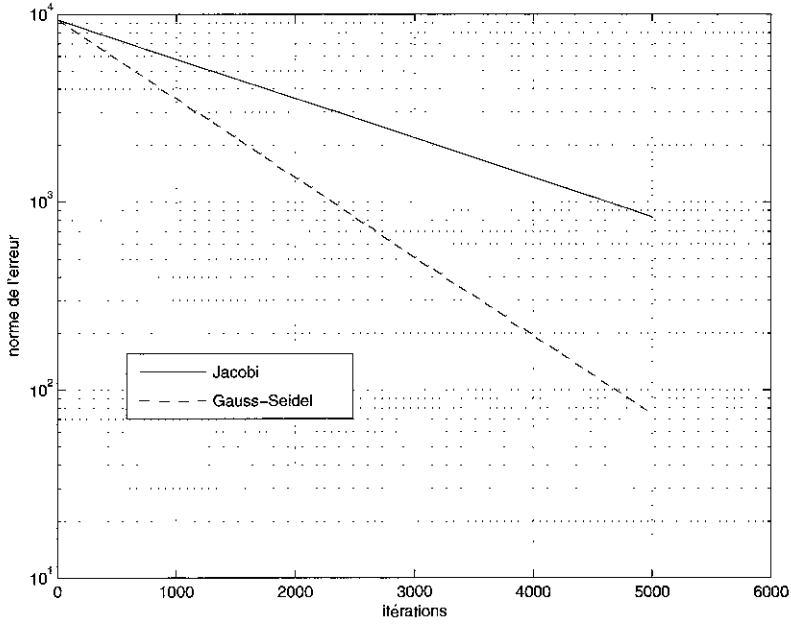


Figure 10.1 Décroissance de la norme de l'erreur $\|e_k\|_2$ en fonction de k.

10.6 MÉTHODES ITÉRATIVES ET PRÉCONDITIONNEMENT

Les méthodes itératives que nous avons considérées jusqu'à présent sont basées sur la recherche d'un point fixe du système

$$x = M^{-1}Nx + M^{-1}b,$$

avec $A = M - N$. On résout donc le système $M^{-1}(M - N)x = M^{-1}b$ c'est-à-dire, en fin de compte, le système initial préconditionné (à gauche) par la matrice M^{-1} :

$$M^{-1}Ax = M^{-1}b.$$

On obtient ainsi les matrices M de préconditionnement suivantes :

- $M = D$ pour la méthode de Jacobi,
- $M = D - E$ pour la méthode de Gauss-Seidel,
- $M = (D - \omega E)/\omega$ pour la méthode de relaxation SOR,
- $M = (D - \omega E)D^{-1}(D - \omega F)/(\omega(2 - \omega))$ pour la méthode de relaxation symétrique SSOR.

Ces matrices de préconditionnement sont utilisées en particulier dans les méthodes de projection sur les sous-espaces de Krylov présentées au chapitre 11.

10.7 NOTES ET RÉFÉRENCES

Les méthodes itératives ont été introduites et utilisées au XIX $^{\text{ième}}$ siècle par Carl Friedrich Gauss (1777-1855), Philipp von Seidel (1821-1896) et Carl Jacobi (1804-1851). Le traitement moderne de cette question est plutôt orienté vers les méthodes de projection présentées au chapitre suivant.

EXERCICES

Les matrices des méthodes de Jacobi, Gauss-Seidel et SOR sont définies respectivement aux paragraphes 10.3.1, 10.3.2 et 10.3.3.

Exercice 10.1

Soient $A \in \mathbb{R}^{2\times 2}$ et $b \in \mathbb{R}^2$. La solution du système $Ax = b$ s'interprète géométriquement comme le point d'intersection des deux droites

$$\begin{array}{ll} (D_1) & a_{11}x_1 + a_{12}x_2 = b_1, \\ (D_2) & a_{21}x_1 + a_{22}x_2 = b_2. \end{array}$$

On suppose que a_{11} et $a_{22} \neq 0$.

1. Calculer les matrices des méthodes de Jacobi et de Gauss-Seidel associées à ce système.
2. Calculer les rayons spectraux de ces matrices. Que remarque-t-on ?
3. Calculer les rayons spectraux des matrices des méthodes de Jacobi et de Gauss-Seidel associées au système obtenu en permutant les deux équations ci-dessus.
4. Interpréter géométriquement les méthodes de Jacobi et de Gauss-Seidel appliquées à $Ax = b$. Pour cette dernière, on notera que le vecteur $x_{k+1} \in \mathbb{R}^2$ est solution du système triangulaire

$$\left\{ \begin{array}{ll} a_{11}x_{k+1,1} + a_{12}x_{k,2} & = b_1 \\ a_{21}x_{k+1,1} + a_{22}x_{k+1,2} & = b_2. \end{array} \right.$$

Exercice 10.2

Calculer les rayons spectraux des matrices des méthodes de Jacobi et de Gauss-Seidel des matrices :

$$\begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ et } \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

Exercice 10.3

Soient

$$A = \begin{pmatrix} 1 & 0 & -1/4 & -1/4 \\ 0 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & 0 \\ -1/4 & -1/4 & 0 & 1 \end{pmatrix} = \begin{pmatrix} I_2 & K \\ K & I_2 \end{pmatrix} \text{ et } b = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

1. Calculer les matrices des méthodes de Jacobi, de Gauss-Seidel et de relaxation associées à A. On les notera J, G et G_ω.
2. Donner l'expression en fonction de k, b et K des itérés de la résolution du système $Ax = b$ par la méthode de Gauss-Seidel et la méthode de Jacobi quand le point initial est l'origine. Il est recommandé d'utiliser la structure bloc de la matrice A.
3. Calculer les rayons spectraux des matrices G et J.
4. a) Montrer que si λ est valeur propre de G_ω, alors $\lambda = 1 - \omega$ ou bien λ est racine de l'équation :

$$\lambda^2 - \left(2(1-\omega) + \frac{\omega^2}{4}\right)\lambda + (1-\omega)^2 = 0.$$

   b) Calculer $\rho(G_\omega)$ en distinguant les cas où les racines de l'équation précédente sont réelles ou non.

   c) Trouver la valeur de ω qui rend $\rho(G_\omega)$ minimum.

Exercice 10.4

1. Soit $B \in \mathbb{C}^{2n\times 2n}$ de la forme :

$$B = \begin{pmatrix} 0 & B_1 \\ B_2 & 0 \end{pmatrix}$$

   où B_1 et $B_2 \in \mathbb{C}^{n\times n}$.

   a) Soit $\lambda \in \text{spec}\,(B)$ et $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ un vecteur propre associé (x_1 et x_2 étant deux vecteurs de $\mathbb{C}^n$). Montrer que $B_1B_2x_1 = \lambda^2 x_1$ et $B_2B_1x_2 = \lambda^2 x_2$.

   b) En déduire que $\rho(B) \leqslant \sqrt{\rho(B_1B_2)}$.

   c) Prouver qu'en fait $\rho(B) = \sqrt{\rho(B_1B_2)}$.

2. Soit $A \in \mathbb{C}^{2n\times 2n}$ de la forme

$$\begin{pmatrix} D_1 & -A_1 \\ -A_2 & D_2 \end{pmatrix}$$

   où D_1 et $D_2 \in \mathbb{C}^{n\times n}$ sont inversibles et où A_1 et $A_2 \in \mathbb{C}^{n\times n}$.

   a) On considère la matrice de la méthode de Jacobi par blocs :

$$J = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & A_1 \\ A_2 & 0 \end{pmatrix}$$

   Calculer $\rho(J)$.

b) On considère la matrice de la méthode SOR par blocs :

$$G_\omega = \begin{pmatrix} D_1 & 0 \\ -\omega A_2 & D_2 \end{pmatrix}^{-1} \begin{pmatrix} (1-\omega)D_1 & \omega A_1 \\ 0 & (1-\omega)D_2 \end{pmatrix}$$

où ω est un réel donné. Lorsque $\omega = 1$, $G_1 = G$ est la matrice de Gauss-Seidel par blocs. Calculer G_ω.

c) Calculer $\rho(G)$ en fonction de $\rho(D_2^{-1}A_2D_1^{-1}A_1)$.

d) On cherche à résoudre le système $Ax = b$ par les méthodes itératives précédentes. Montrer que les méthodes de Jacobi par blocs et de Gauss-Seidel par blocs convergent ou divergent simultanément et écrire la condition nécessaire et suffisante de convergence.

e) Calculer $\det G_\omega$ (le résultat ne dépend que de ω et de n). Montrer qu'une condition nécessaire de convergence pour « SOR par blocs » est que $0 < \omega < 2$.

f) Montrer que cette condition est aussi suffisante lorsque A est aussi définie positive.

Exercice 10.5

a, b et μ désignant des nombres complexes et n étant un entier supérieur ou égal à 2, on considère les matrices tridiagonales $C_n(a,b) = (c_{ij}) \in \mathbb{C}^{n\times n}$ définies par :

$$\begin{array}{lll} c_{ij} = a & \text{pour} & i = j-1 \quad \text{et} \quad 2 \leqslant j \leqslant n, \\ c_{ij} = b & \text{pour} & i = j+1 \quad \text{et} \quad 1 \leqslant j \leqslant n-1, \\ c_{ij} = 0 & \text{sinon.} & \end{array}$$

On définit également les matrices :

$$A_n(\mu, a, b) = \mu I_n + C_n(a,b)$$

où $I_n \in \mathbb{C}^{n\times n}$ est la matrice unité.

Dans cet exercice, a et b sont des nombres complexes non nuls quelconques.

1. a) Soit α un nombre complexe, $\alpha \neq 0$. On considère la matrice diagonale

$$\Delta(\alpha) = \operatorname{diag}(\alpha, \alpha^2, \ldots, \alpha^n).$$

Montrer que pour tout nombre complexe μ on a :

$$A_n(\mu, \alpha^{-1}a, \alpha b) = \Delta(\alpha) A_n(\mu, a, b) \Delta(\alpha^{-1}).$$

b) En déduire que pour tout nombre complexe $\alpha \neq 0$, les matrices $A_n(\mu, a, b)$ et $A_n(\mu, \alpha^{-1}a, \alpha b)$ ont même spectre.

2. Soit maintenant $s > 0$ fixé ; on cherche à résoudre par des méthodes itératives le système :
$$A_n(s, a, b)X = f.$$

   a) Soient $\hat{J}$ et $\hat{G}$ respectivement les matrices des méthodes de Jacobi et de Gauss-Seidel associées à la matrice $A_n(s, a, b)$. Montrer que leurs rayons spectraux respectifs vérifient la relation :
$$\rho(\hat{G}) = \rho(\hat{J})^2.$$

   b) En déduire que ces méthodes convergent ou divergent simultanément. Observer que leur convergence implique l'inversibilité de $A_n(s, a, b)$.

3. Pour quelles valeurs de s les méthodes itératives ci-dessus sont-elles convergentes ?

Exercice 10.6 Méthode de la plus grande pente dite aussi méthode de Richardson

On considère $A \in \mathbb{R}^{n\times n}$ définie positive et $\alpha > 0$ un paramètre fixé. On associe au système $Ax = b$ la méthode itérative
$$x_{k+1} = x_k - \alpha(Ax_k - b).$$

On note $0 < \lambda_1 \leqslant \ldots \leqslant \lambda_n$ les n valeurs propres ordonnées de A.

1. Écrire cette itération sous la forme standard $x_{k+1} = B_\alpha x_k + c$. Pour quelles valeurs de $\alpha > 0$ cette méthode itérative est-elle convergente ?
2. Déterminer la valeur $\alpha > 0$ optimale c'est-à-dire telle que $\rho(B_\alpha)$ soit minimum.

La dénomination « plus grande pente » vient de ce que $Ax_k - b$ est le gradient de la quadrique $q(x) = \frac{1}{2}x^TAx - b^Tx$ (voir paragraphe 7.2). Le minimum unique de la quadrique q correspond à la solution du système : $\nabla q(x) = Ax - b = 0$. À partir de x_k, le nouveau point x_{k+1} est obtenu en « descendant » le long du gradient $\nabla q(x_k) = Ax_k - b$ avec un pas constant α. Cette méthode est aussi appelée méthode du *gradient à pas constant* (voir aussi exercice 11.1 qui traite la méthode du gradient à pas optimal).

Chapitre **11**

Méthodes de projection sur des sous-espaces de Krylov

Les méthodes de projection sur des sous-espaces de Krylov fournissent une autre famille importante de méthodes itératives. Leur développement est plus récent que les méthodes itératives classiques et date des années 1970-80. Ce sont en partie les besoins créés par les applications industrielles et en particulier la nécessité de résoudre des systèmes de grande dimension qui ont motivé leur essor. D'autre part, les progrès constants réalisés en informatique ont permis d'accroître considérablement les capacités de calcul et de stockage des données et ont rendu possible la mise en œuvre effective de ces nouveaux algorithmes. Actuellement, ces méthodes font encore l'objet de recherches mathématiques actives.

Dans ce chapitre, nous nous limiterons à l'étude de deux méthodes parmi les plus représentatives : la *méthode GMRES* (en anglais Generalized Minimum RESsidual) pour des systèmes généraux et la méthode du *gradient conjugué* pour des systèmes à matrice définie positive.

On verra au chapitre 15 que les méthodes de projection sont également utilisées pour calculer les valeurs propres et vecteurs propres de matrices de grande taille. Donnons tout d'abord le cadre général d'une méthode de projection pour le calcul de la solution d'un système linéaire.

11.1 STRUCTURE GÉNÉRALE D'UNE MÉTHODE DE PROJECTION

On considère le système $Ax = b$, où la matrice $A \in \mathbb{C}^{n\times n}$ est inversible, et $x_0 \in \mathbb{C}^n$ un vecteur donné.

Une méthode de projection associée au système consiste à calculer une suite (x_k) d'approximations à l'aide de deux familles de sous-espaces vectoriels $\mathcal{K}_k$ et $\mathcal{L}_k$ de dimension k en imposant les conditions

$$x_k \in x_0 + \mathcal{K}_k$$

et

$$b - Ax_k \perp \mathcal{L}_k.$$

La condition d'orthogonalité est dite condition de *Petrov-Galerkin*. Elle permet de définir de manière unique x_k dans le sous-espace affine $x_0 + \mathcal{K}_k$.

Les sous-espaces $\mathcal{L}_k$ et $\mathcal{K}_k$ ne sont pas nécessairement identiques. On dit que la méthode de projection est *orthogonale* si $\mathcal{L}_k = \mathcal{K}_k$ et *oblique* sinon. Les sous-espaces sont généralement emboîtés : $\mathcal{K}_k \subset \mathcal{K}_{k+1}$ et $\mathcal{L}_k \subset \mathcal{L}_{k+1}$.

Pour toutes les méthodes considérées on a toujours $x_n = x$ la solution du système. Sous certaines hypothèses, le vecteur x_k est une bonne approximation de cette solution. On souhaite bien sûr arrêter l'algorithme pour des valeurs de k petites devant n.

Les espaces d'approximation $\mathcal{K}_k$ que nous allons considérer sont des sous-espaces de Krylov. Le paragraphe suivant présente les propriétés de ces sous-espaces et leur relation avec les algorithmes de réduction de Hessenberg de la matrice A, en particulier l'algorithme d'Arnoldi et l'algorithme de Lanczos.

11.2 ESPACES DE KRYLOV ET RÉDUCTION DE HESSENBERG

Soit $A \in \mathbb{C}^{n\times n}$ une matrice inversible et $v \in \mathbb{C}^n$, $v \neq 0$.

Définition 11.1 *Le sous-espace de Krylov d'ordre k associé à la matrice A et au vecteur v, noté $\mathcal{K}_k(A, v)$, est le sous-espace vectoriel généré par les k vecteurs*

$$v, Av, A^2v, \ldots, A^{k-1}v.$$

On le note $\mathcal{K}_k$ lorsqu'il n'y a pas de risque d'ambiguïté et on convient que $\mathcal{K}_0(A, v) = \{0\}$.

Proposition 11.2 *Pour tout k, les sous-espaces $\mathcal{K}_k(A, v)$ vérifient les propriétés suivantes dont la démonstration est laissée au lecteur :*

1. $\mathcal{K}_k(A, v) \subset \mathcal{K}_{k+1}(A, v)$,

2. $A\,\mathcal{K}_k(A,v) \subset \mathcal{K}_{k+1}(A,v)$,
3. *Le sous-espace* $\mathcal{K}_k(A,v)$ *est invariant par* A *(autrement dit* $A\mathcal{K}_k(A,v) \subset \mathcal{K}_k(A,v)$*. voir chapitre 13) si et seulement si* $\mathcal{K}_l(A,v) = \mathcal{K}_k(A,v)$ *pour tout* $l \geqslant k$.
4. $\mathcal{K}_n(A,v) = \mathcal{K}_{n+1}(A,v)$,
5. $\mathcal{K}_k(A,\alpha v) = \mathcal{K}_k(A,v)$ *pour tout scalaire* $\alpha \neq 0$.

Ayant à utiliser les sous-espaces de Krylov comme sous-espaces d'approximation, il est important d'y disposer d'une base orthonormée. Supposons que les vecteurs $(v, Av, \ldots, A^{k-1}v)$ soient indépendants. Le procédé d'orthonormalisation de Gram-Schmidt permet d'obtenir une base orthonormée de $\mathcal{K}_k(A,v)$. Nous savons cependant qu'en général les directions des itérés successifs $A^j v$, lorsque j augmente, tendent vers la direction d'un vecteur propre associé à la valeur propre de plus grand module de A (voir la méthode de la puissance : théorème 14.3). Cette propriété entraîne que les matrices $K_k = (v\ Av\ \ldots\ A^{k-1}v)$ sont en général mal conditionnées. Nous allons voir qu'une décomposition de Hessenberg de la matrice A définit une base orthonormée d'un espace de Krylov tout en évitant le calcul direct des vecteurs itérés $A^j v$.

Soit $A = QHQ^*$ une décomposition de Hessenberg de A (voir paragraphes 8.6 et suivants). Notons $H_k = H(1:k, 1:k) \in \mathbb{C}^{k\times k}$, $Q_k = Q(1:n, 1:k) = (q_1 \ldots q_k) \in \mathbb{S}t_{nk}$ et $K_k = (q_1\ Aq_1\ \ldots\ A^{k-1}q_1) \in \mathbb{C}^{n\times k}$.

Proposition 11.3 *Si* H_k *est non réduite (définition 8.18), alors les vecteurs* $q_1, \ldots, q_k$ *constituent une base de l'espace de Krylov* $\mathcal{K}_k(A,q_1)$*. En outre,* $K_k = Q_k R_k$ *pour une matrice triangulaire supérieure* $R_k \in \mathbb{C}^{k\times k}$*. Si* $k < n$ *et* $h_{k+1\,k} = 0$*, alors l'espace* $\mathcal{K}_k(A,q_1)$ *est invariant par* A.

Démonstration. Par construction, les vecteurs q_j sont orthonormés. Montrons que $q_j = p_{j-1}(A)q_1$, où p_{j-1} est un polynôme de degré $j-1$. Pour $j = 1$ on a $q_1 = p_0(A)q_1$ avec $p_0 = 1$. Supposons la propriété vraie jusqu'à l'ordre j. De la relation $AQ = QH$ on déduit l'égalité pour la colonne j (voir l'équation (8.3))

$$Aq_j = \sum_{i=1}^{j+1} h_{ij} q_i,$$

et donc

$$h_{j+1\,j}\, q_{j+1} = Aq_j - \sum_{i=1}^{j} h_{ij} q_i. \tag{11.1}$$

L'hypothèse de récurrence implique

$$h_{j+1\,j}\, q_{j+1} = Ap_{j-1}(A)q_1 - \sum_{i=1}^{j} h_{ij}\, p_{i-1}(A)q_1.$$

Puisque $h_{j+1\,j} \neq 0$, il en résulte que $q_{j+1} = p_j(A)q_1$, où p_j est un polynôme de degré égal à j.

Les vecteurs orthonormés q_j appartiennent à l'espace $\mathcal{K}_k(A, q_1)$ et forment donc une base de cet espace. La propriété $q_j = p_{j-1}(A)q_1$, pour $j = 1, \ldots, k$, montre que $Q_k = K_k S_k$, où $K_k = (q_1\, Aq_1 \ldots A^{k-1}q_1)$ et où $S_k \in \mathbb{C}^{k\times k}$ est triangulaire supérieure et inversible puisque Q_k est de rang k. Nous avons donc $K_k = Q_k R_k$ avec $R_k = S_k^{-1}$ triangulaire supérieure.

Si $h_{k+1\,k} = 0$, l'égalité (11.1) pour $j = k$ montre que $Aq_k = \sum_{i=1}^{k} h_{ij} q_i$ et donc que $Aq_k \in \mathcal{K}_k(A, q_1)$. D'autre part, pour $j = 1, \ldots, k-1$, on a $Aq_j = Ap_{j-1}(A)q_1 \in \mathcal{K}_{j+1}(A, q_1) \subset \mathcal{K}_k(A, q_1)$. Ainsi $\mathcal{K}_k$ est invariant par A.

La proposition précédente et la propriété 5 de la proposition 11.2 montrent que les vecteurs-colonne de la matrice $Q_k = (q_1 \ldots q_k)$ avec $q_1 = v/\|v\|_2$ où $v \neq 0$, constituent une base de l'espace de Krylov $\mathcal{K}_k(A, v)$.

Nous avons étudié au chapitre 8 deux méthodes pour obtenir une décomposition de Hessenberg : la méthode d'Arnoldi (modifiée) (paragraphe 8.8) et la méthode de Householder (paragraphe 8.6).

La méthode de Householder jouit d'une meilleure stabilité numérique que la méthode d'Arnoldi. En revanche, la complexité de la méthode d'Arnoldi est plus faible que celle de Householder. Pour les systèmes de grande dimension, on considère que la méthode d'Arnoldi offre actuellement le meilleur compromis entre la complexité des calculs et la fiabilité des résultats.

Remarque 11.1. Il convient de noter que k étapes de la méthode de Householder présentée au paragraphe 8.6 définissent la matrice $Q_k = (q_1 \ldots q_k) \in \mathbb{S}t_{nk}$ où $q_1 = e_1$ et $q_j = H_2 \ldots H_j e_j$ pour $j = 2, \ldots k$. On a ainsi une base orthonormée de l'espace de Krylov $\mathcal{K}_k(A, e_1)$. Afin de pouvoir définir une base orthonormée d'un espace de Krylov général $\mathcal{K}_k(A, v)$, $v \in \mathbb{C}^n$, $v \neq 0$, il est nécessaire de considérer dans cet algorithme une première transformation de Householder H_1 telle que $H_1 v = \alpha e_1$ où $\alpha \in \mathbb{C}$ et $|\alpha| = \|v\|_2$. Cette transformation H_1 est obtenue en vertu du corollaire 8.11. On poursuit ensuite l'algorithme de la même façon en prenant la matrice $H_1 A H_1$.

Le paragraphe suivant est consacré à la méthode GMRES. Le calcul de la base de l'espace de Krylov y est donné par la méthode d'Arnoldi.

11.3 LA MÉTHODE GMRES

11.3.1 Description de la méthode

Soit $x_0 \in \mathbb{C}^n$, notons x la solution du système $Ax = b$ et $r_0 = b - Ax_0$ le résidu en x_0. La méthode GMRES est une méthode de projection de Krylov oblique avec $\mathcal{L}_k = A\mathcal{K}_k(A, r_0)$. L'approximation x_k est donc définie par

$$x_k \in x_0 + \mathcal{K}_k(A, r_0)$$

et

$$r_k = b - Ax_k \perp A\mathcal{K}_k(A, r_0).$$

Ces conditions d'orthogonalité montrent que

Proposition 11.4

1. *Pour tout $k \leqslant n$, x_k est définie par la solution unique du problème des moindres carrés*

$$\min_{z \in x_0 + \mathcal{K}_k(A, r_0)} \|Az - b\|_2^2. \tag{11.2}$$

2. *Pour tout $k < n$, les résidus r_k et r_{k+1} vérifient*

$$\|r_{k+1}\|_2 \leqslant \|r_k\|_2.$$

3. *x_k est solution du système $Ax_k = b$ si et seulement si $\mathcal{K}_k$ est invariant par A.*

Démonstration.

1. Le problème des moindres carrés s'écrit

$$\min_{z \in x_0 + \mathcal{K}_k} \|Az - b\|_2^2 = \min_{u \in \mathcal{K}_k} \|Au - r_0\|_2^2$$

et toute solution optimale u_k vérifie (théorème 9.8) $\langle Au_k - r_0, Au \rangle = 0$ pour tout $u \in \mathcal{K}_k$ c'est-à-dire (avec $x_k = x_0 + u_k$) $\langle b - Ax_k, Au \rangle = 0$ pour tout $u \in \mathcal{K}_k$ autrement dit $b - Ax_k \perp A\mathcal{K}_k(A, r_0)$. La solution du problème des moindres carrés est unique puisque A est inversible (théorème 9.8).

2. Le résultat est évident puisque $\mathcal{K}_k \subset \mathcal{K}_{k+1}$ et que la solution x_{k+1} est définie par

$$\|Ax_{k+1} - b\|_2^2 = \min_{z \in x_0 + \mathcal{K}_{k+1}(A, r_0)} \|Az - b\|_2^2.$$

3. x_k est défini par la condition $x_k - x_0 \in \mathcal{K}_k$ et $b - Ax_k \perp A\mathcal{K}_k$. Si $\mathcal{K}_k$ est invariant par A on a $A\mathcal{K}_k = \mathcal{K}_k$ et donc $b - Ax_k \perp \mathcal{K}_k$. Comme $b - Ax_k \in \mathcal{K}_k$ ($b - Ax_k = b - Ax_0 - A(x_k - x_0) = r_0 - A(x_k - x_0) \in \mathcal{K}_k + A\mathcal{K}_k = \mathcal{K}_k$) on obtient $b - Ax_k = 0$.
Réciproquement, si $Ax_k = b$, puisque $x_k - x_0 \in \mathcal{K}_k$, il existe des scalaires α_i tels que $x_k - x_0 = \alpha_0 r_0 + \ldots + \alpha_{k-1} A^{k-1} r_0$. On obtient $Ax_k = Ax_0 + \alpha_0 Ar_0 + \ldots + \alpha_{k-1} A^k r_0 = b$ de sorte que $\alpha_{k-1} A^k r_0 = r_0 - (\alpha_0 Ar_0 + \ldots + \alpha_{k-2} A^{k-1} r_0)$. Si $\alpha_{k-1} \neq 0$ cette identité prouve que $\mathcal{K}_{k+1} = \mathcal{K}_k$ et donc que $\mathcal{K}_k$ est invariant par A. Si $\alpha_{k-1} = 0$, c'est que $x_k - x_0 \in \mathcal{K}_{k-1}$ et $Ax_k = b$, de sorte que $x_k = x_{k-1}$ par définition de x_{k-1}. On est ramené au même problème à un ordre inférieur. Il reste à traiter le cas $k = 0$. On a alors $Ax_0 = b$ c'est-à-dire $r_0 = 0$. L'espace de Krylov $\mathcal{K}_0 = \{0\}$ est évidemment invariant par A.

11.3.2 Algorithmique

L'algorithme associé à la méthode consiste à déterminer la solution x_k décrite précédemment à l'aide de la base orthonormée $(q_1, \ldots, q_n)$ de $\mathcal{K}_k$ obtenue par l'algorithme d'Arnoldi avec $q_1 = r_0/\|r_0\|_2$.

Proposition 11.5 *L'itéré x_k obtenu par la méthode GMRES est donné par $x_k = x_0 + Q_k y_k$ où $y_k \in \mathbb{C}^k$ est la solution du problème des moindres carrés*

$$\min_{y \in \mathbb{C}^k} \|\tilde{H}_k y - \|r_0\|_2 e_1\|_2^2, \tag{11.3}$$

avec $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{C}^{k+1}$ et où Q_k et $\tilde{H}_k$ sont les matrices déterminées par l'algorithme d'Arnoldi (paragraphe 8.8).

Démonstration. En utilisant la base orthonormée $(q_1, \ldots, q_k)$ on a $z - x_0 \in \mathcal{K}_k$ si et seulement si $z = x_0 + Q_k y$ pour un $y \in \mathbb{C}^k$. Des égalités $AQ_k = Q_{k+1}\tilde{H}_k$ (équation 8.4) et puisque $r_0 = \|r_0\|_2 q_1 = \|r_0\|_2 Q_{k+1} e_1$ on obtient

$$Az - b = AQ_k y - r_0 = Q_{k+1}(\tilde{H}_k y - \|r_0\|_2 e_1)$$

de sorte que, puisque Q_{k+1} est Stiefel,

$$\|Az - b\|_2^2 = \|\tilde{H}_k y - \|r_0\|_2 e_1\|_2^2.$$

L'algorithme GMRES est donc ramené au calcul de la solution d'un problème de moindres carrés associé à une matrice $\tilde{H}_k \in \mathbb{C}^{(k+1)\times k}$ de Hessenberg. Afin de résoudre ce dernier problème nous utilisons une décomposition QR de $\tilde{H}_k$ (paragraphe b).

L'algorithme tire avantage de deux spécificités importantes :

1. la matrice $\tilde{H}_k$ est obtenue en complétant la matrice $\tilde{H}_{k-1}$ par une k-ième colonne,
2. $\tilde{H}_k$ est de Hessenberg.

La décomposition QR de la matrice $\tilde{H}_k$ de Hessenberg peut être calculée par $k-1$ rotations de Givens (paragraphe 8.4).

Supposons avoir calculé $G\tilde{H}_{k-1} = G_{k-2}\dots G_1\tilde{H}_{k-1} = \tilde{U}_{k-1}$ où les G_i sont des rotations de Givens, $\tilde{U}_{k-1} \in \mathbb{C}^{k\times(k-1)}$ est triangulaire supérieure

$$\tilde{U}_{k-1} = \begin{pmatrix} U_{k-1} \\ 0 \end{pmatrix},$$

avec $U_{k-1} \in \mathbb{C}^{(k-1)\times(k-1)}$ triangulaire supérieure. On obtient $\tilde{H}_k$ en ajoutant à $\tilde{H}_{k-1}$ une ligne de 0 et la colonne h_k d'où

$$\tilde{G}\tilde{H}_k = \begin{pmatrix} G & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \tilde{H}_{k-1} & h_k \\ 0 & \end{pmatrix} = \begin{pmatrix} \tilde{U}_{k-1} & \tilde{h}_k \\ 0 & \end{pmatrix}$$

avec $\tilde{h}_k = \tilde{G}h_k$. Par une rotation convenable de Givens G_{k-1} on annule le terme $\tilde{h}_{k+1\,k}$ de $\tilde{H}_k$ (coefficient $k+1$ de h_k) sans changer la structure triangulaire supérieure déjà acquise.

L'étape k de GMRES est donc :

1. Calculer q_{k+1} et h_k par l'algorithme d'Arnoldi,
2. Calculer la rotation de Givens G_{k-1} telle que $(G_{k-1}\dots G_1 h_k)_{k+1} = 0$,
3. Calculer $G_{k-1}\dots G_1 h_k$ et l'adjoindre à $\tilde{U}_{k-1}$ pour obtenir

$$\tilde{U}_k = \begin{pmatrix} U_k \\ 0 \end{pmatrix},$$

4. Calculer $\tilde{g} = (G_1^*\dots G_{k-1}^*)e_1 \in \mathbb{C}^{k+1}$,
5. Résoudre le système triangulaire

$$U_k y_k = \|r_0\|_2 g,$$

où $g \in \mathbb{C}^k$ est tel que $\tilde{g} = \begin{pmatrix} g \\ g_{k+1} \end{pmatrix}$,

6. Calculer la norme du résidu

$$\|Ax_k - b\|_2 = \|g_{k+1}\|_2.$$

On arrête cet algorithme lorsque le résidu $r_k = b - Ax_k$ a une norme suffisamment petite.

La complexité d'une étape k de cet algorithme est dominée par le calcul de q_{k+1} et h_k : on a

- $\approx 2n^2$ opérations pour le produit Aq_k,
- $\approx 2n(k+1) + 2nk + n$ opérations pour le calcul de q_{k+1} et h_k.

On voit que la complexité de la méthode GMRES croît avec la dimension k de l'espace de Krylov. Pour remédier à ce problème on utilise une méthode de réinitialisation cyclique des sous-espaces de Krylov (méthode de redémarrage, en anglais restarting) : à partir de $x_k \in x_0 + \mathcal{K}_k(A, r_0)$, on calcule le résidu $r_k = b - Ax_k$ et on poursuit avec les nouveaux espaces $x_k + \mathcal{K}_j(A, r_k)$ jusqu'à obtenir $x_{2k} \in x_k + \mathcal{K}_k(A, r_k)$. Le processus se poursuit ainsi en réinitialisant les espaces de Krylov au bout de k étapes.

11.4 LA MÉTHODE DU GRADIENT CONJUGUÉ

11.4.1 Description de la méthode

Notons $A \in \mathbb{C}^{n\times n}$ une matrice définie positive, $x_0 \in \mathbb{C}^n$ un vecteur donné et $x \in \mathbb{C}^n$ la solution du système $Ax = b$. Notons aussi

$$\langle u, v\rangle_A = \langle Au, v\rangle = v^* Au$$

le produit scalaire associé à A (paragraphe 7.1).

La méthode du gradient conjugué est une méthode de projection de Krylov oblique avec $\mathcal{L}_k = \mathcal{K}_k(A, r_0)$. L'approximation x_k est donc définie par

$$x_k \in x_0 + \mathcal{K}_k(A, r_0) \tag{11.4}$$

et

$$r_k = b - Ax_k \perp \mathcal{K}_k(A, r_0). \tag{11.5}$$

On a les propriétés suivantes

Proposition 11.6

1. *Pour tout $k \leqslant n$, x_k est définie par la solution unique du problème des moindres carrés*

$$\min_{z\in x_0+\mathcal{K}_k(A,r_0)} \|z - x\|_A^2. \tag{11.6}$$

2. *Pour tout* $k < n$, *les erreurs* $e_k = x - x_k$ *et* $e_{k+1} = x - x_{k+1}$ *vérifient*

$$\|e_{k+1}\|_A \leqslant \|e_k\|_A.$$

3. x_k *est solution du système* $Ax_k = b$ *si et seulement si* $\mathcal{K}_k$ *est invariant par* A.

Démonstration.

1. Le problème de minimisation

$$\min_{z \in x_0 + \mathcal{K}_k(A, r_0)} \|z - x\|_A^2$$

a une solution x_k unique qui est égale à la projection orthogonale de x sur $x_0 + \mathcal{K}_k(A, r_0)$ pour le produit scalaire $\langle\ ,\ \rangle_A$. On a donc $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ et

$$\langle x - x_k, v \rangle_A = 0$$

pour tout $v \in \mathcal{K}_k(A, r_0)$. Cette dernière condition s'écrit

$$\langle A(x - x_k), v \rangle = 0$$

pour tout $v \in \mathcal{K}_k(A, r_0)$ c'est-à-dire

$$r_k = b - Ax_k \perp \mathcal{K}_k(A, r_0).$$

2. Le résultat est évident puisque $\mathcal{K}_k \subset \mathcal{K}_{k+1}$ et que la solution x_{k+1} est définie par

$$\|x_{k+1} - x\|_A^2 = \min_{z \in x_0 \in \mathcal{K}_{k+1}(A, r_0)} \|z - x\|_A^2.$$

3. À partir des conditions $x_k - x_0 \in \mathcal{K}_k$ et $b - Ax_k \perp \mathcal{K}_k$, la démonstration est identique à celle de la proposition 11.4.

Remarque 11.2. Lorsque A est une matrice quelconque, les conditions (11.4) et (11.5) définissent la méthode FOM (en anglais Full Orthogonal Method).

11.4.2 Algorithmique

L'algorithme associé à cette méthode consiste à déterminer la solution x_k à l'aide de la base orthonormée $(q_1, \ldots, q_k)$ de $\mathcal{K}_k$, obtenue par l'algorithme de Lanczos puisque A est hermitienne. La récurrence à trois termes de l'algorithme de Lanczos permet d'obtenir une expression simple de la solution $x_k \in \mathcal{K}_k$ à partir de la solution précédente $x_{k-1} \in \mathcal{K}_{k-1}$.

Soit $(q_1, \ldots, q_k)$ la base orthonormée de $\mathcal{K}_k(A, r_0)$ obtenue par l'algorithme de Lanczos avec $q_1 = r_0/\|r_0\|_2$.

Proposition 11.7 *L'itéré x_k de la méthode du gradient conjugué est donné par $x_k = x_0 + Q_k y_k$ où $y_k \in \mathbb{C}^k$ est l'unique solution du système*

$$T_k y_k = \|r_0\|_2 e_1 \tag{11.7}$$

avec $e_1 = (1.0, \ldots, 0)^T \in \mathbb{C}^k$ et où Q_k et T_k sont les matrices déterminées par l'algorithme de Lanczos (paragraphe 8.9).

Démonstration. En utilisant la base orthonormée $(q_1, \ldots, q_k)$ on a $z - x_0 \in \mathcal{K}_k$ si et seulement si $z = x_0 + Q_k y$ pour un $y \in \mathbb{C}^k$. La condition d'orthogonalité (11.5) s'écrit $\langle b - A(x_0 + Q_k y_k), q_j \rangle = 0$, pour tout $j = 1, \ldots, k$, et donc

$$Q_k^*(r_0 - AQ_k y_k) = 0.$$

Puisque $Q_k^* A Q_k = T_k$ (équation 8.4) et que $r_0 = \|r_0\|_2 q_1 = \|r_0\|_2 Q_k e_1$, on obtient le système

$$T_k y_k = \|r_0\|_2 e_1.$$

Montrons que la matrice $T_k = Q_k^* A Q_k$ est inversible : la matrice définie positive A admet la décomposition de Cholesky $A = LL^*$, donc la condition $Q_k^* A Q_k y = 0$ implique $\|L^* Q_k y\|_2 = 0$ et alors $y = 0$ puisque Q_k est de rang k et L^* inversible.

Le résultat suivant montre que le résidu r_k est colinéaire au vecteur q_{k+1}.

Proposition 11.8 *Le résidu $r_k = b - Ax_k$ obtenu par la méthode du gradient conjugué est donné par*

$$r_k = -\overline{\beta}_k \, e_k^T y_k \, q_{k+1}.$$

Démonstration. D'après les équations (8.4) on a

$$b - Ax_k = r_0 - AQ_k y_k = r_0 - Q_k T_k y_k - \overline{\beta}_k \, q_{k+1} \, e_k^T y_k.$$

Or $T_k y_k = \|r_0\|_2 e_1$, donc $Q_k T_k y_k = \|r_0\|_2 q_1 = r_0$, d'où la conclusion.

La récurrence qui définit x_k à partir x_{k-1} est obtenue grâce à la décomposition LU de la matrice T_k.

La proposition 11.7 donne la solution x_k :

$$x_k = x_0 + Q_k \, T_k^{-1} \, \|r_0\|_2 e_1.$$

Considérons la décomposition LU de T_k : $T_k = L_k U_k$, et posons $P_k = Q_k U_k^{-1}$ et $z_k = L_k^{-1} \|r_0\|_2 e_1$. Nous avons

$$x_k = x_0 + Q_k (U_k^{-1} L_k^{-1}) \|r_0\|_2 e_1 = x_0 + P_k z_k.$$

Nous allons calculer P_k à partir de P_{k-1} et z_k à partir de z_{k-1}. Notons p_j les colonnes de la matrice P_k.

Lemme 11.9 *On a* $P_k = (P_{k-1}\ p_k)$ *où* $P_{k-1} = Q_{k-1} U_{k-1}^{-1}$ *et* $z_k = (z_{k-1}^T, \nu_k)^T$ *où* $z_{k-1} = L_{k-1}^{-1} \|r_0\|_2 e_1$.

Démonstration. La décomposition LU de la matrice T_{k-1} montre que

$$L_k = \begin{pmatrix} L_{k-1} & 0 \\ \times & 1 \end{pmatrix},$$

et

$$U_k = \begin{pmatrix} U_{k-1} & \times \\ 0 & \times \end{pmatrix},$$

où L_{k-1} et U_{k-1} sont les matrices de la décomposition LU de T_{k-1}.
Les matrices inverses de U_k et L_k sont données par

$$L_k^{-1} = \begin{pmatrix} L_{k-1}^{-1} & 0 \\ \times & 1 \end{pmatrix},$$

et

$$U_k^{-1} = \begin{pmatrix} U_{k-1}^{-1} & \times \\ 0 & \times \end{pmatrix}.$$

On a donc

$$P_k = Q_k \begin{pmatrix} U_{k-1}^{-1} & \times \\ 0 & \times \end{pmatrix} = (Q_{k-1} U_{k-1}^{-1}\ p_k)$$

et

$$z_k = \begin{pmatrix} L_{k-1}^{-1} & 0 \\ \times & 1 \end{pmatrix} \|r_0\|_2 e_1 = \begin{pmatrix} L_{k-1}^{-1} \|r_0\|_2 e_1 \\ \nu_k \end{pmatrix}.$$

Il faut noter dans cette dernière égalité que le vecteur de base e_1 a successivement la dimension k et $k-1$.

Nous pouvons déduire de ce lemme une récurrence sur x_k :

$$x_k = x_0 + (P_{k-1}\ p_k) \begin{pmatrix} z_{k-1} \\ \nu_k \end{pmatrix} = x_0 + P_{k-1} z_{k-1} + \nu_k p_k$$

et donc

$$x_k = x_{k-1} + \nu_k p_k. \tag{11.8}$$

En multipliant cette égalité par A et en retranchant b au deux membres de façon à faire apparaître les résidus, nous avons une récurrence sur r_k :

$$r_k = r_{k-1} - \nu_k A p_k. \tag{11.9}$$

Il reste à définir une récurrence sur p_k.

Lemme 11.10 *On a*

$$p_k = r_{k-1} + \mu_k p_{k-1}. \tag{11.10}$$

Démonstration. La relation $P_k = Q_k U_k^{-1}$ implique $P_k U_k = Q_k$. D'autre part, la matrice triangulaire supérieure U_k est bidiagonale puisque T_k est tridiagonale. On déduit que $q_k = \gamma_{k-1} p_{k-1} + \delta_k p_k$, et

$$p_k = \frac{1}{\delta_k} q_k - \frac{\gamma_{k-1}}{\delta_k} p_{k-1}, \tag{11.11}$$

puisque $\delta_k \neq 0$. La proposition 11.8 montre que les vecteurs q_k et r_{k-1} sont colinéaires. Le vecteur p_k est donc une combinaison linéaire des vecteurs r_{k-1} et p_{k-1}. On peut écrire cette relation sous la forme $p_k = r_{k-1} + \mu_k p_{k-1}$ après multiplication des vecteurs p_k par des coefficients appropriés. Nous allons conserver la même notation pour ces nouveaux vecteurs. Il faut noter que la transformation des vecteurs p_k ne modifie pas la forme des récurrences sur x_k et r_k (équations (11.8) et (11.9)).

Nous avons obtenu pour x_k, r_k et p_k trois récurrences (11.8), (11.9) et (11.10). Le calcul des coefficients ν_k et μ_k se fait grâce aux propriétés d'orthogonalité des vecteurs r_k et d'orthogonalité pour le produit scalaire associé à la matrice A des vecteurs p_k.

Proposition 11.11 *On a* $\langle Ap_i, p_j \rangle = 0$, *pour tout* i, j, $i \neq j$.

Démonstration. Il suffit de donner la démonstration pour les vecteurs p_i « d'origine ». On a

$$P_k^* A P_k = U_k^{-*} Q_k^* A Q_k U_k^{-1} = U_k^{-*} T_k U_k^{-1} = U_k^{-*} L_k U_k U_k^{-1} = U_k^{-*} L_k.$$

La matrice $U_k^{-*} L_k$ est triangulaire inférieure. La matrice symétrique $P_k^* A P_k$ égale à $U_k^{-*} L_k$ est donc diagonale.

Les vecteurs r_k colinéaires à q_{k+1} sont orthogonaux entre eux puisque les vecteurs q_k satisfont cette propriété. On est maintenant en mesure de déterminer les coefficients ν_k et μ_k.

Lemme 11.12 *On a*

$$\nu_k = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle Ap_k, p_k\rangle},$$

et

$$\mu_k = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle r_{k-2}, r_{k-2}\rangle}.$$

Démonstration. Grâce à l'orthogonalité des vecteurs r_k on obtient

$$\nu_k = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle Ap_k, r_{k-1}\rangle}$$

à partir de (11.9). L'orthogonalité pour le produit scalaire associé à A et l'égalité (11.10) donnent

$$\langle Ap_k, r_{k-1}\rangle = \langle Ap_k, p_k\rangle.$$

On a donc

$$\nu_k = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle Ap_k, p_k\rangle}.$$

Utilisant à nouveau l'orthogonalité pour le produit scalaire associé à A et (11.10), on obtient

$$\mu_k = -\frac{\langle Ap_{k-1}, r_{k-1}\rangle}{\langle Ap_{k-1}, p_{k-1}\rangle}.$$

De l'égalité (11.9) prise au rang $k-1$ on déduit

$$Ap_{k-1} = \frac{1}{\nu_{k-1}}(r_{k-2} - r_{k-1}).$$

Enfin, utilisant l'orthogonalité des r_k et l'expression de ν_{k-1}, on obtient

$$\mu_k = \frac{1}{\nu_{k-1}} \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle Ap_{k-1}, p_{k-1}\rangle} = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle r_{k-2}, r_{k-2}\rangle}.$$

Nous pouvons ainsi définir l'algorithme du gradient conjugué.

Algorithme du gradient conjugué

$k = 0$, $x_0 \in \mathbb{C}^n$, $\varepsilon > 0$,
$r_0 = b - Ax_0$, $p_1 = r_0$,

tant que $\|r_k\|_2 > \varepsilon$
$k = k + 1$
$z = Ap_k$
$\nu_k = \frac{r_{k-1}^* r_{k-1}}{p_k^* z}$
$x_k = x_{k-1} + \nu_k p_k$
$r_k = r_{k-1} - \nu_k z$
$\mu_{k+1} = \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}}$
$p_{k+1} = r_k + \mu_{k+1} p_k$
fin

À chaque itération k, on effectue

- un produit matrice $\times$ vecteur (Ap_k),
- 3 produits scalaires,
- 3 sommes,
- 3 produits scalaire $\times$ vecteur,
- 2 divisions.

La complexité de chaque itération est dominée par le produit Ap_k, ce qui donne $\approx 2n^2$ opérations.

Remarque 11.3. L'algorithme du gradient conjugué fait partie des outils actuels les plus performants, robustes et simples à mettre en œuvre pour résoudre des systèmes de grande dimension à matrice définie positive.

Remarque 11.4. D'un point de vue pratique, les méthodes de projection dans des sous-espaces de Krylov présentent une particularité importante. La matrice A du système intervient uniquement par son action sur des vecteurs. On dit que A est donnée en évaluation. Contrairement aux méthodes itératives classiques qui requièrent la représentation explicite de tous les coefficients de la matrice A on se contente ici d'utiliser l'information globale donnée par les images Av calculées pour différents vecteurs v. On a ainsi un processus de type « boîte

noire » qui calcule l'image Av d'un vecteur v donné en entrée. Cette situation se rencontre dans la plupart des cas où A est « définie » par un programme informatique ou une suite de programmes informatiques qui s'enchaînent les uns à la suite des autres. Il n'est pas nécessaire alors de calculer et stocker les différents coefficients de la matrice. Ce mode opératoire rend ces méthodes bien adaptées aux grands systèmes pour lesquels le calcul de l'ensemble des coefficients de la matrice est souvent rédhibitoire. De plus, lorsque la structure de la matrice est creuse (voir par exemple le système des éléments finis et le problème de l'assimilation des données aux chapitres 16 et 17), ces méthodes tirent parti de la faible complexité de chaque opération Av ($O(n)$ opérations pour des matrices bande à la place des $O(n^2)$ usuelles).

Remarque 11.5. Il existe de nombreuses approches pour définir l'algorithme du gradient conjugué (voir exercice 11.2). Historiquement, cet algorithme a été introduit dans les années 1950 c'est-à-dire bien avant que ne soient développées les méthodes de projection dans les sous-espaces de Krylov.

Remarque 11.6. Dans le cas des grands systèmes que l'on résout à l'aide des méthodes GMRES ou du gradient conjugué, on cherche à préconditionner la matrice du système afin de limiter le nombre d'itérations. Au paragraphe 10.6 nous avons donné des exemples de matrices de préconditionnement tirées des méthodes itératives classiques. Il existe d'autres approches pour définir des matrices de préconditionnement. On peut citer les méthodes de factorisation LU incomplète ou de Cholesky incomplète lorsque A est définie positive qui sont utilisées dans le cas de matrices creuses. Nous renvoyons le lecteur intéressé aux ouvrages plus spécialisés comme par exemple [29].

11.5 ANALYSE D'ERREUR

En arithmétique exacte les méthodes de projection que nous avons étudiées aboutissent en au plus n itérations. En ce sens, on ne peut pas les qualifier de méthodes itératives puisqu'elles calculent la solution en un nombre fini d'étapes. Cependant, lorsqu'on les utilise pour résoudre de grands systèmes, le nombre d'itérations que l'on effectue est dans la plupart des cas bien inférieur à la dimension du problème. C'est cet usage particulier qui les fait considérer comme des méthodes itératives.

Nous avons vu que la méthode GMRES minimise la norme du résidu $\|b - Ax_k\|_2$ tandis que la méthode du gradient conjugué minimise la norme de l'erreur $\|x - x_k\|_A$. L'analyse des erreurs se fonde sur ces propriétés d'optimalité.

11.5.1 Analyse des erreurs de la méthode du gradient conjugué

Notons $e_k = x - x_k$ l'erreur à l'itération k et $\mathcal{P}_k$ l'espace des polynômes de degré $\leqslant k$. La propriété d'optimalité du gradient conjugué (proposition 11.6) conduit au résultat suivant :

Proposition 11.13 *On a*

$$\|e_k\|_A = \min_{q \in \mathcal{Q}_k} \|q(A)e_0\|_A,$$

où $\mathcal{Q}_k$ est l'ensemble des polynômes q de degré $\leqslant k$ tels que $q(0) = 1$.

Démonstration. D'après la proposition 11.6, x_k est solution du problème

$$\min_{z \in x_0 + \mathcal{K}_k} \|z - x\|_A.$$

Tout $z \in x_0 + \mathcal{K}_k(A, r_0)$ est de la forme $z = x_0 + p_{k-1}(A)r_0$, où $p_{k-1} \in \mathcal{P}_{k-1}$. Sachant que $b = Ax$, on a $z = x_0 + p_{k-1}(A)A(x - x_0)$ et donc

$$x - z = (I_n - p_{k-1}(A)A)(x - x_0) = q_k(A)(x - x_0),$$

où $q_k(X) = 1 - p_{k-1}(X)X$ appartient à $\mathcal{P}_k$ et vérifie $q_k(0) = 1$.

Pour majorer la quantité $\min_{q \in \mathcal{Q}_k} \|q(A)e_0\|_A$ nous allons utiliser un résultat classique de la théorie de l'approximation (voir par exemple P.-J. Laurent [23]). Rappelons que les polynômes de Chebyshev sont définis par $T_k(x) = \cos k\theta$ et $x = \cos\theta$, $k \geqslant 0$. Ils vérifient la relation de récurrence $T_{k+2}(x) + T_k(x) = 2x\, T_{k+1}(x)$.

Proposition 11.14 *Soit $[a, b]$ un intervalle de $\mathbb{R}$ et $c < a$. Alors la solution du problème*

$$\min_{q \in \mathcal{P}_k, q(c)=1} \max_{t \in [a,b]} |q(t)|$$

est obtenue pour le polynôme

$$\tilde{T}_k(t) = \frac{T_k\left(1 + 2\frac{a-t}{b-a}\right)}{T_k\left(1 + 2\frac{a-c}{b-a}\right)}$$

où T_k est le polynôme de Chebyshev de degré k. On a

$$\min_{q \in \mathcal{P}_k, q(c)=1} \max_{t \in [a,b]} |q(t)| = \frac{1}{T_k\left(1 + 2\frac{a-c}{b-a}\right)}.$$

Nous en déduisons la majoration de l'erreur du gradient conjugué :

Proposition 11.15 *Dans la méthode du gradient conjugué la norme de l'erreur* $\|e_k\|_A = \|x_k - x\|_A$ *vérifie*

$$\|e_k\|_A \leqslant 2\left(\frac{\sqrt{\text{cond}_2(A)}-1}{\sqrt{\text{cond}_2(A)}+1}\right)^k \|e_0\|_A.$$

Démonstration. Puisque la matrice A est symétrique définie positive, elle est diagonalisable : $A = Q\Lambda Q^*$ avec Q unitaire, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ et $\lambda_1 \geqslant \ldots \geqslant \lambda_n > 0$. On a

$$\begin{aligned}\|q(A)e_0\|_A^2 &= e_0^* q(A)^* A q(A) e_0 = e_0^* q(Q\Lambda Q^*)^* Q\Lambda Q^* q(Q\Lambda Q^*) e_0 \\ &= e_0^* Q q(\Lambda)^* Q^* Q\Lambda Q^* Q q(\Lambda) Q^* e_0 = y^* q(\Lambda)^* \Lambda q(\Lambda) y\end{aligned}$$

où l'on a posé $y = Q^* e_0$. D'autre part,

$$q(\Lambda)^* \Lambda q(\Lambda) = \text{diag}\left(|q(\lambda_1)|^2\lambda_1, \ldots, |q(\lambda_n)|^2\lambda_n\right).$$

On a donc

$$\|q(A)e_0\|_A^2 = \sum_{i=1}^n |y_i|^2 \lambda_i |q(\lambda_i)|^2 \leqslant \max_{\lambda\in[\lambda_n,\lambda_1]} |q(\lambda)|^2 \, \|e_0\|_A^2 \qquad (11.12)$$

car $\|e_0\|_A^2 = e_0^* A e_0 = e_0^* Q\Lambda Q^* e_0 = \sum_{i=1}^n |y_i|^2 \lambda_i$. D'après la proposition 11.13, on a

$$\|e_k\|_A = \min_{q\in\mathcal{Q}_k} \|q(A)e_0\|_A.$$

De l'inégalité (11.12) et de la proposition précédente en prenant $a = \lambda_n$, $b = \lambda_1$ et $c = 0$, on déduit que

$$\|e_k\|_A \leqslant \frac{\|e_0\|_A}{T_k\left(1 + 2\frac{\lambda_n}{\lambda_1 - \lambda_n}\right)}.$$

Pour tout t tel que $|t| > 1$, on sait que les polynômes de Chebyshev T_k satisfont l'égalité

$$T_k(t) = \frac{1}{2}\{(t + \sqrt{t^2-1})^k + (t + \sqrt{t^2-1})^{-k}\}$$

et donc pour $t > 1$

$$T_k(t) \geqslant \frac{1}{2}\left(t + \sqrt{t^2-1}\right)^k.$$

Posons $\mu := \lambda_n/(\lambda_1 - \lambda_n)$. Nous obtenons

$$T_k(1+2\mu) \geqslant \frac{1}{2}\left(1 + 2\mu + \sqrt{(1+2\mu)^2 - 1}\right)^k.$$

Par un calcul simple, on vérifie que

$$\begin{aligned} 1+2\mu+\sqrt{(1+2\mu)^2-1} &= 1+2\mu+2\sqrt{\mu(1+\mu)} \\ &= \left(\sqrt{\mu}+\sqrt{\mu+1}\right)^2 = \left(\sqrt{\lambda_n}+\sqrt{\lambda_1}\right)^2/(\lambda_1-\lambda_n) \\ &= \frac{\sqrt{\frac{\lambda_1}{\lambda_n}}+1}{\sqrt{\frac{\lambda_1}{\lambda_n}}-1}. \end{aligned}$$

Puisque $\mathrm{cond}_2(A) = \lambda_1/\lambda_n$, on en déduit le résultat.

La fonction $x \mapsto \frac{\sqrt{x}-1}{\sqrt{x}+1}$ est définie et croissante de $[1,+\infty[$ sur $[0,1[$. La vitesse de convergence de la méthode du gradient conjugué est d'autant meilleure que $\mathrm{cond}_2(A)$ est proche de 1.

11.5.2 Erreur de la méthode GMRES

Notons $r_k = b - Ax_k$ le résidu associé à la méthode GMRES.

Proposition 11.16 *L'itéré x_k donné par la méthode GMRES vérifie*

$$\|r_k\|_2 = \min_{q\in\mathcal{Q}_k} \|q(A)r_0\|_2, \tag{11.13}$$

où $\mathcal{Q}_k$ est l'ensemble des polynômes q de degré $\leqslant k$ tels que $q(0) = 1$.

Démonstration. On sait que x_k est solution du problème

$$\min_{z\in x_0+\mathcal{K}_k(A,r_0)} \|Az-b\|_2.$$

Tout $z \in x_0 + \mathcal{K}_k$ est de la forme $z = x_0 + p_{k-1}(A)r_0$, où $p_{k-1} \in \mathcal{P}_{k-1}$. On a donc $b - Az = (I_n - Ap_{k-1}(A))r_0 = q_k(A)b$, où $q_k(X) = 1 - Xp_{k-1}(X)$ est un polynôme de $\mathcal{P}_k$ et tel que $q_k(0) = 1$.

Lorsque la matrice est diagonalisable, nous en déduisons la majoration suivante dont la démonstration est laissée au lecteur.

Proposition 11.17 *Supposons que A soit diagonalisable et soit $A = P\Lambda P^{-1}$ une décomposition avec Λ diagonale. Alors*

$$\|r_k\|_2 \leqslant \|r_0\|_2\,\mathrm{cond}_2(P) \min_{q\in\mathcal{Q}_k} \rho(q(A)),$$

où $\rho(q(A))$ est le rayon spectral de $q(A)$:

$$\rho(q(A)) = \max_{\lambda\in \mathrm{spec}\ A} |q(\lambda)|.$$

Contrairement à la méthode du gradient conjugué où la borne d'erreur ne dépend que du conditionnement de A et de la distance de x_0 à la solution x, la borne d'erreur donnée ici pour GMRES dépend aussi de la norme de la matrice de passage P c'est-à-dire des vecteurs propres de A. Ce dernier terme quantifie d'une certaine façon la distance de A à la « normalité ». On rappelle que toute matrice normale (théorème 1.7) est diagonalisable dans une base de vecteurs orthonormés et que $\text{cond}_2(P) = 1$ si P est une matrice unitaire. Lorsque la matrice A n'est pas normale, on constate que cette borne d'erreur n'est pas optimale contrairement à la borne obtenue pour le gradient conjugué.

La figure 11.1 illustre la décroissance de l'erreur pour le système de Poisson 2D discrétisé par éléments finis (paragraphe 16.2). La dimension du système est égale à 557. L'échelle logarithmique de l'axe des ordonnées montre une vitesse de convergence linéaire dans les premières itérations et superlinéaire [1] lorsque les itérations augmentent.

11.6 NOTES ET RÉFÉRENCES

Alexei Nikolaevich Krylov (1863-1945) est surtout connu pour ses travaux en ingénierie navale. C'est dans un article publié en 1931 [20] qu'apparaît pour la première fois la notion de sous-espace de Krylov.

La méthode du gradient conjugué a été développée indépendamment et à partir d'approches différentes par Cornelius Lanczos d'une part et par Magnus R. Hestenes et Eduard Stiefel d'autre part. Dans un article paru en 1952 [17] ils illustrent les potentialités de l'algorithme en faisant état de la résolution d'un système de 106 équations, ce qui devait représenter une dimension considérable pour l'époque. Actuellement, des systèmes de plusieurs millions de variables sont couramment résolus à l'aide de cet algorithme. On notera que c'est le même auteur E. Stiefel qui a laissé son nom aux matrices définies au paragraphe 8.1.

La méthode du gradient conjugué était conçue à l'origine comme une méthode de résolution directe de systèmes. Pendant de nombreuses années elle est restée dans l'oubli en raison de propriétés numériques moins bonnes que celles obtenues par les méthodes directes classiques. Ce n'est que dans les années 1970 que son intérêt pour la résolution des grands systèmes a matrice creuse est apparu.

1. On dit que la convergence est superlinéaire lorsque $\|e_{n+1}\|_2 / \|e_n\|_2 \to 0$ lorsque $n \to \infty$. Lorsqu'on a simplement $\limsup \|e_{n+1}\|_2 / \|e_n\|_2 = \alpha$ avec $0 < \alpha < 1$ on dit que la convergence est linéaire et α est le taux de convergence linéaire.

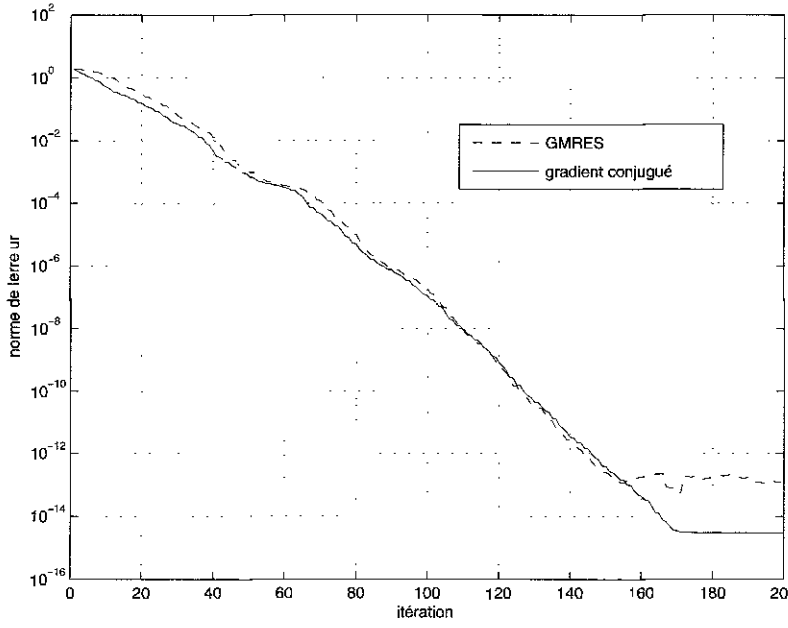


Figure 11.1 Décroissance de la norme de l'erreur $\|e_k\|_2$ en fonction de k

On trouvera dans l'ouvrage de Saad [29] une étude plus approfondie de la convergence de la méthode du gradient conjugué et de la méthode GMRES.

Nous nous sommes limités dans ce chapitre à considérer les deux plus importantes méthodes de projection sur des sous-espaces de Krylov. À partir des années 1980 et 1990 ont été développées de nombreuses autres variantes de cette grande famille permettant de traiter des systèmes non hermitiens. Parmi celles-ci on peut citer la méthode de bi-orthogonalisation BiCG (en anglais BiConjugate Gradient) qui utilise deux espaces de Krylov, l'un associé à A et l'autre à A^* : $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ et $b - Ax_k \perp \mathcal{K}_k(A^*, s_0)$ avec $s_0 \in \mathbb{C}^n$ tel que $\langle u_0, s_0 \rangle \neq 0$. Cette méthode fait intervenir des récurrences « courtes » du même type que le gradient conjugué.

EXERCICES

Exercice 11.1 Méthode du gradient à pas optimal

Soient $A \in \mathbb{R}^{n\times n}$ définie positive et $b \in \mathbb{R}^n$. On considère le système $Ax = b$. Notons $q(x) = \frac{1}{2}x^T Ax - b^T x$ la quadrique qui lui est associée. Nous avons vu au paragraphe 7.2 que la solution x du système est le vecteur qui minimise la fonction $q(x)$ sur $\mathbb{R}^n$. En effet le gradient de $q(x)$ est donné par $\nabla q(x) = Ax - b$.

Pour résoudre le système $Ax = b$, on définit une méthode itérative par

$$x_{k+1} = x_k - \rho_k(Ax_k - b)$$

où $\rho_k > 0$ est un pas de descente le long du gradient de q. La *méthode du gradient à pas optimal* consiste à choisir, à chaque itération k, le pas $\rho_k \in \mathbb{R}$ solution du problème

$$\min_{\rho\in\mathbb{R}} \|x - (x_k - \rho(Ax_k - b))\|_A^2.$$

1. Calculer la valeur du pas optimal ρ_k. Montrer que ρ_k est également solution du problème

$$\min_{\rho\in\mathbb{R}} q(x_k - \rho(Ax_k - b)),$$

et que

$$\langle \nabla q(x_{k+1}), \nabla q(x_k)\rangle = 0.$$

2. Posons $r_k = b - Ax_k$ et $e_k = x - x_k$. Montrer que

$$\|e_{k+1}\|_A^2 = \|e_k\|_A^2 \left(1 - \frac{\langle r_k, r_k\rangle^2}{\langle Ar_k, r_k\rangle \langle A^{-1}r_k, r_k\rangle}\right).$$

Grâce à l'inégalité de Kantorovitch (exercice 5.3), montrer que

$$\|e_{k+1}\|_A^2 \leqslant \left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}\right)^2 \|e_k\|_A^2.$$

En déduire que

$$\|e_k\|_A \leqslant \left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}\right)^k \|e_0\|_A$$

et que la suite (x_k) converge vers la solution x du problème. Comparer la vitesse de convergence avec celle obtenue pour le gradient conjugué (proposition 11.15).

Exercice 11.2

Cet exercice présente l'algorithme du gradient conjugué comme une généralisation de la méthode du gradient à pas optimal. Les notations utilisées sont celles de l'exercice précédent, $A \in \mathbb{R}^{n\times n}$ est définie positive.

On définit une suite de vecteurs (x_k) par la récurrence $x_{k+1} = x_k + g_k$ où g_k est solution du problème

$$\min_{g\in G_k} q(x_k + g)$$

et G_k est l'espace vectoriel engendré par les vecteurs $\nabla q(x_i)$, $0 \leqslant i \leqslant k$:

$$G_k = [Ax_0 - b, Ax_1 - b, \ldots, Ax_k - b]\,.$$

1. Donner les conditions nécessaires et suffisantes d'optimalité du problème d'optimisation précédent. En déduire que $\langle \nabla q(x_{k+1}), \nabla q(x_j)\rangle = 0$, pour tout $j = 0, \ldots, k$.
2. On suppose que $\nabla q(x_j) \neq 0$, pour tout $j = 0, \ldots, k$. Montrer que $G_j = \mathcal{K}_{j+1}(A, r_0)$ pour tout $j = 0, \ldots, k$. En déduire que (x_k) est la suite donnée par l'algorithme du gradient conjugué.

Chapitre **12**

Valeurs propres : sensibilité

Ce chapitre est consacré aux valeurs propres d'une matrice A considérées comme fonctions des entrées de cette matrice. Ce n'est pas une question facile parce que la définition des valeurs propres est donnée implicitement : ce sont les racines du polynôme caractéristique

$$P_A(\lambda) = \det(A - \lambda I_n).$$

Les questions que l'on se pose sont liées à leur localisation dans le plan complexe, à leur continuité et leur dérivabilité. C'est ce que l'on appelle *étude de sensibilité.*

12.1 LE THÉORÈME DE GERSHGORIN

Le théorème qui suit est un résultat de localisation :

Théorème 12.1 *Les valeurs propres d'une matrice $A \in \mathbb{C}^{n\times n}$ sont contenues dans l'union des disques*

$$\mathcal{D}_i(A) = \left\{ \lambda \in \mathbb{C} \ : \ |\lambda - a_{ii}| \leqslant \sum_{j\neq i} |a_{ij}| \right\}, \ 1 \leqslant i \leqslant n,$$

appelés disques de Gershgorin.

Démonstration. Soit $\lambda \in \mathbb{C}$ une valeur propre de A. On a $Ax = \lambda x$ pour un $x \in \mathbb{C}^n$ tel que

$$\|x\|_\infty = \max_i |x_i| = 1.$$

Soit i tel que $|x_i| = 1$. On a :

$$\lambda x_i = (Ax)_i = \sum_{j=1}^{n} a_{ij} x_j$$

de sorte que

$$|\lambda - a_{ii}| = |(\lambda - a_{ii}) x_i| = \left| \sum_{j \neq i} a_{ij} x_j \right| \leqslant \sum_{j \neq i} |a_{ij} x_j| \leqslant \sum_{j \neq i} |a_{ij}| .$$

Ce théorème peut être raffiné de la façon suivante : si un disque de Gershgorin, disons $\mathcal{D}_i(A)$, est isolé des $n-1$ autres disques, c'est-à-dire si $\mathcal{D}_i(A) \cap \mathcal{D}_j(A) = \varnothing$ pour tout $j \neq i$, alors ce disque contient une et une seule valeur propre de A. La démonstration de cette propriété utilise des résultats d'analyse complexe ; nous ne la démontrons pas ici.

12.2 LE THÉORÈME D'ELSNER

Nous allons prouver un résultat de continuité du spectre d'une matrice. Comme c'est un ensemble fini de points, nous devons définir une distance entre de tels ensembles.

Définition 12.2 *Soit $\mathbb{E}$ un espace métrique, d sa distance et $\mathcal{K}(\mathbb{E})$ l'ensemble des parties compactes et non vides de $\mathbb{E}$. La distance de Hausdorff sur $\mathcal{K}(\mathbb{E})$ est donnée par*

$$hd(S,T) = \max \left(\max_{s \in S} \min_{t \in T} d(s,t), \max_{t \in T} \min_{s \in S} d(s,t) \right)$$

pour tout $S, T \in \mathcal{K}(\mathbb{E})$ (exercice 12.1).

Théorème 12.3 *Étant donné deux matrices A et $A' \in \mathbb{C}^{n \times n}$ notons $hd(A, A')$ la distance de Hausdorff entre les spectres de A et de A' :*

$$hd(A, A') = \max \left(\max_{\lambda \in \text{spec } A} \min_{\lambda' \in \text{spec } A'} |\lambda - \lambda'| , \max_{\lambda' \in \text{spec } A'} \min_{\lambda \in \text{spec } A} |\lambda - \lambda'| \right) .$$

On a :

$$hd(A, A') \leqslant \left(\|A\|_2 + \|A'\|_2 \right)^{1 - \frac{1}{n}} \|A - A'\|_2^{\frac{1}{n}} .$$

La démonstration de ce théorème repose sur l'inégalité suivante :

Lemme 12.4 *(Inégalité de Hadamard) Pour toute matrice $A \in \mathbb{C}^{n \times n}$ on a*

$$|\det A| \leqslant \prod_{i=1}^{n} \|a_i\|_2$$

où les a_i sont les vecteurs-colonne de A. Lorsque $\det A \neq 0$, *il y a égalité si et seulement si les colonnes de A sont orthogonales deux à deux.*

Démonstration. (Inégalité de Hadamard) Notons $A = QR$ une décomposition QR de A et r_i les colonnes de R. Comme $a_i = Qr_i$ et que Q est unitaire on a

$$|r_{ii}| \leqslant \|r_i\|_2 = \|a_i\|_2$$

de sorte que, puisque $|\det Q| = 1$,

$$|\det A| = |\det R| = \prod_i |r_{ii}| \leqslant \prod_i \|a_i\|_2 .$$

Lorsque $\det A \neq 0$, il y a égalité si et seulement si $|r_{ii}| = \|r_i\|_2 = \|a_i\|_2$ pour tout i c'est-à-dire si R est diagonale. Cette dernière condition revient à dire que les colonnes de $A = QR$ sont orthogonales deux à deux.

Remarque 12.1. Pour une matrice réelle, $|\det A|$ est le volume du parallélotope de $\mathbb{R}^n$ construit sur les vecteurs a_i. L'inégalité de Hadamard dit que ce volume est majoré par le produit des longueurs de ces vecteurs et qu'il y a égalité si et seulement si ce parallélotope est rectangle.

Démonstration. (Théorème 12.3) A et A' jouant des rôles symétriques dans le résultat que l'on doit prouver, il suffit de le montrer pour

$$\max_{\lambda' \in \operatorname{spec} A'} \min_{\lambda \in \operatorname{spec} A} |\lambda - \lambda'|$$

au lieu de $\operatorname{hd}(A, A')$. Supposons que le maximum soit atteint pour la valeur propre λ'' de A'. Soit $x_1, \ldots, x_n$ une base orthonormée de $\mathbb{C}^n$ telle que $A'x_1 = \lambda'' x_1$ et enfin soit $U \in \mathbb{U}_n$ une matrice unitaire telle que $Ue_i = x_i$ où les e_i désignent les vecteurs de la base canonique de $\mathbb{C}^n$. On a :

$$\max_{\lambda' \in \operatorname{spec} A'} \min_{\lambda \in \operatorname{spec} A} |\lambda - \lambda'|^n \leqslant \prod_{\lambda} |\lambda - \lambda''| = |\det(A - \lambda'' I_n)| =$$

$$|\det((A - \lambda'' I_n)U)| .$$

Par l'inégalité de Hadamard, cette dernière quantité vérifie

$$|\det((A - \lambda'' I_n)U)| \leqslant \prod_i \|((A - \lambda'' I_n)U)_i\|_2 = \prod_i \|(A - \lambda'' I_n)Ue_i\|_2 =$$

$$\prod_i \|(A - \lambda'' I_n)x_i\|_2 = \|(A - A')x_1\|_2 \prod_{i \neq 1} \|(A - \lambda'' I_n)x_i\|_2 .$$

On majore alors

$$\|(A - A')x_1\|_2 \leqslant \|A - A'\|_2 \, \|x_1\|_2 = \|A - A'\|_2$$

et, en utilisant la proposition 3.6,

$$\|(A - \lambda'' I_n)x_i\|_2 \leqslant \|A - \lambda'' I_n\|_2 \, \|x_i\|_2 = \|A - \lambda'' I_n\|_2 \leqslant$$
$$\|A\|_2 + |\lambda''| \leqslant \|A\|_2 + \|A'\|_2 .$$

On a ainsi prouvé l'inégalité

$$\max_{\lambda' \in \text{spec } A'} \min_{\lambda \in \text{spec } A} |\lambda - \lambda'|^n \leqslant \|A - A'\|_2 \left(\|A\|_2 + \|A'\|_2\right)^{n-1} .$$

Il suffit pour conclure de prendre les racines n-ièmes des deux membres.

Remarque 12.2. Le théorème 12.3 montre que l'application « valeur propre » est höldérienne d'exposant $1/n$ ce qui est une propriété un peu décevante. L'exemple de la matrice $n \times n$

$$A_\varepsilon = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \varepsilon & & & 0 \end{pmatrix}$$

avec $\varepsilon > 0$, montre que c'est un mal nécessaire ! A_ε est une perturbation de A_0. Cette dernière a pour valeur propre $\lambda = 0$ de multiplicité n et A_ε a pour valeurs propres $\lambda'_k = \varepsilon^{1/n}\omega_k$, $1 \leqslant k \leqslant n$, où les ω_k sont les racines n-ièmes de l'unité. On a ici

$$|\lambda'_k - \lambda| = \varepsilon^{1/n} = \|A_\varepsilon - A_0\|_2^{1/n} .$$

12.3 SENSIBILITÉ VIA LE THÉORÈME DES FONCTIONS IMPLICITES

Tout au long de ce paragraphe, nous allons supposer que $A \in \mathbb{C}^{n \times n}$ est une matrice hermitienne, de sorte que ses valeurs propres λ_i, $1 \leqslant i \leqslant n$, sont réelles et qu'il existe une base orthonormée $x_i \in \mathbb{C}^n$, $1 \leqslant i \leqslant n$, de vecteurs propres de A.

Le calcul des éléments propres peut se formuler comme la recherche des zéros du système

$$F(A, ., .) : \mathbb{C}^n \times \mathbb{R} \to \mathbb{C}^n \times \mathbb{R}, \quad F(A, x, \lambda) = \begin{pmatrix} (\lambda I_n - A)x \\ \frac{1}{2}\left(\|x\|_2^2 - 1\right) \end{pmatrix} .$$

Lorsque $F(A, x, \lambda) = 0$ c'est que λ est une valeur propre de A et que x est un vecteur propre unitaire associé à cette valeur propre.

Nous allons appliquer le théorème des fonctions implicites dans ce contexte pour prouver l'existence d'une fonction *matrice → (vecteur propre, valeur propre)* qui soit définie et C^∞ dans un voisinage de A puis nous calculerons sa dérivée.

Théorème 12.5 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice hermitienne ; supposons que λ soit une valeur propre simple de A et que x soit un vecteur propre unitaire associé. Alors*

1. *Il existe un voisinage ouvert $\mathcal{V}_A$ de A dans $\mathbb{C}^{n\times n}$ et une unique application C^∞ (et même analytique réelle)*

$$(X, \Lambda) : \mathcal{V}_A \subset \mathbb{C}^{n\times n} \to \mathbb{C}^n \times \mathbb{R}$$

telle que :

a) $X(A) = x$ *et* $\Lambda(A) = \lambda$,

b) $BX(B) = \Lambda(B)X(B)$ *et* $\|X(B)\|_2 = 1$ *pour tout* $B \in \mathcal{V}_A$.

2. *Les dérivées de ces fonctions en A vérifient*

a) $DX(A) : \mathbb{C}^{n\times n} \to \mathbb{C}^n$ *et* $DX(A)\dot{A} = (\lambda I_n - A)^\dagger \dot{A}x$,

b) $D\Lambda(A) : \mathbb{C}^{n\times n} \to \mathbb{R}$ *et* $D\Lambda(A)\dot{A} = x^* \dot{A}x$.

3. *Les normes de ces opérateurs lorsque $\mathbb{C}^{n\times n}$ est muni de la norme spectrale sont*

a) $\|DX(A)\| = \max_{\|\dot{A}\|_2 \leqslant 1} \|DX(A)\dot{A}\|_2 = \max_{\mu\in \text{spec } A,\ \mu\neq\lambda} |\lambda - \mu|^{-1}$,

b) $\|D\Lambda(A)\| = \max_{\|\dot{A}\|_2 \leqslant 1} |D\Lambda(A)\dot{A}| = 1$.

Remarque 12.3.

1. Quoique nous ayons supposé que A soit hermitienne, les fonctions X et Λ sont définies sur un voisinage de A dans $\mathbb{C}^{n\times n}$ et non pas dans le sous-espace des matrices hermitiennes. Autrement dit, nous considérons des perturbations non hermitiennes d'une matrice hermitienne.
2. L'énoncé 3.b. signifie que le calcul des valeurs propres d'une matrice hermitienne est toujours bien conditionné. On a au premier ordre

$$|\Lambda(A) - \Lambda(B)| \leqslant_1\|A - B\|_2 .$$

3. L'énoncé 3.a. donne le conditionnement du calcul des vecteurs propres : un bon conditionnement correspond à des valeurs propres bien séparées, un mauvais conditionnement à des valeurs propres proches.

4. Le théorème 12.5 ne s'étend pas directement au cas non hermitien. En effet, pour une matrice $A \in \mathbb{C}^{n\times n}$ quelconque, le système $Ax = \lambda x$ contient $n+1$ inconnues et n équations complexes ou bien, en séparant parties réelles et parties imaginaires, $2n+2$ inconnues et $2n$ équations réelles. L'équation normalisante pour le vecteur propre $\|x\|_2^2 = 1$ du théorème 12.5 compte pour une équation réelle et ne suffit pas à lever l'indétermination. On se tire de ce mauvais pas en recherchant un vecteur propre dans l'espace affine $x + x^\perp$: on est ainsi amené à étudier des problèmes perturbés du type

$$By = \mu y, \quad \langle y - x, x\rangle = 0.$$

Cette fois-ci le compte y est : $n+1$ équations et $n+1$ inconnues complexes. En termes plus savants, on prend y dans l'espace projectif complexe $\mathbb{P}_{n-1}(\mathbb{C})$ et on utilise la carte locale $x + x^\perp$.

Démonstration. (Théorème 12.5) La première chose à faire est de vérifier les hypothèses du théorème des fonctions implicites (théorème 5.1). Comme F est une application polynomiale (donc C^∞ et même analytique réelle) il suffit de vérifier que sa dérivée en (A, x, λ) par rapport aux variables (x, λ) est un isomorphisme de $\mathbb{C}^n \times \mathbb{R}$. On a : $D_2F(A, x, \lambda) : \mathbb{C}^n \times \mathbb{R} \to \mathbb{C}^n \times \mathbb{R}$,

$$D_2F(A, x, \lambda)(\dot{x}, \dot{\lambda}) = \begin{pmatrix} (\lambda I_n - A)\dot{x} + \dot{\lambda}x \\ \langle \dot{x}, x\rangle \end{pmatrix}.$$

Il suffit de montrer que cette application est injective. Si $D_2F(A, x, \lambda)(\dot{x}, \dot{\lambda}) = 0$ cela signifie que

$$(\lambda I_n - A)\dot{x} + \dot{\lambda}x = 0 \text{ et } \dot{x} \in x^\perp.$$

En multipliant cette équation à gauche par x^* on obtient

$$x^*(\lambda I_n - A)\dot{x} + \dot{\lambda}x^*x = 0.$$

Comme A est hermitienne et que $(\lambda I_n - A)x = 0$ on a aussi $x^*(\lambda I_n - A) = 0$ et donc $\dot{\lambda} = 0$. Ceci prouve que

$$(\lambda I_n - A)\dot{x} = 0, \quad \dot{\lambda} = 0, \quad \dot{x} \in x^\perp.$$

Notons λ_i, $1 \leqslant i \leqslant n$, les valeurs propres de A et $x_i \in \mathbb{C}^n$, $1 \leqslant i \leqslant n$, une base orthonormée de vecteurs propres de A. Supposons que $\lambda = \lambda_1$ et que $x = x_1$. Ainsi $x^\perp$ est le sous-espace engendré par les vecteurs $x_2, \ldots, x_n$ et $\dot{x} \in x^\perp$ s'écrit $\dot{x} = \sum_{i=2}^n \alpha_i x_i$. On a

$$0 = (\lambda I_n - A)\dot{x} = \sum_{i=2}^n \alpha_i(\lambda - \lambda_i)x_i.$$

Comme les x_i sont des vecteurs linéairement indépendants on a $\alpha_i(\lambda - \lambda_i) = 0$ pour tout i ; puisque λ est une valeur propre simple, on a $\lambda - \lambda_i \neq 0$, donc $\alpha_i = 0$ et par conséquent $\dot{x} = 0$. Ainsi $D_2F(A, x, \lambda)$ est un isomorphisme, le théorème des fonctions implicites s'applique ce qui prouve les assertions 1.a et 1.b.

La dérivée de la fonction implicite en A est donnée par

$$D(X, \Lambda)(A) = -D_2F(A, x, \lambda)^{-1} D_1F(A, x, \lambda)$$

où D_1F désigne la dérivée de F par rapport à la variable A et D_2F celle par rapport au couple (x, λ). Soit $\dot{A} \in \mathbb{C}^{n\times n}$. On a :

$$D_2F(A, x, \lambda)(DX(A)\dot{A}, D\Lambda(A)\dot{A}) = -D_1F(A, x, \lambda)\dot{A}$$

ou encore, en posant $\dot{x} = DX(A)\dot{A}$ et $\dot{\lambda} = D\Lambda(A)\dot{A}$,

$$\begin{aligned} &(\lambda I_n - A)\dot{x} + \dot{\lambda}x = \dot{A}x, \\ &\langle \dot{x}, x\rangle = 0. \end{aligned}$$

Notons que

$$\operatorname{Ker}(\lambda I_n - A) = \mathbb{C}x$$

et que

$$\operatorname{Im}(\lambda I_n - A) = x^{\perp}.$$

La première égalité a lieu parce que λ est une valeur propre simple de A et la seconde parce que $x^{\perp}$ est le sous-espace engendré par les vecteurs propres $x_2, \ldots, x_n$. Tout ceci prouve que $(\lambda I_n - A)\dot{x}$ et $\dot{\lambda}x$ sont les projections orthogonales de $\dot{A}x$ sur $x^{\perp}$ et sur $\mathbb{C}x$. Ainsi

$$\begin{aligned} &(\lambda I_n - A)\dot{x} = \Pi_{x^{\perp}}\dot{A}x, \\ &\dot{\lambda}x = \Pi_{\mathbb{C}x}\dot{A}x, \\ &\dot{x} \in x^{\perp} \end{aligned}$$

de sorte que

$$\begin{aligned} &\dot{x} \in \operatorname{Ker}(\lambda I_n - A)^{\perp}, \\ &(\lambda I_n - A)\dot{x} = \Pi_{\operatorname{Im}(\lambda I_n - A)}\dot{A}x, \\ &\dot{\lambda} = x^*\dot{A}x. \end{aligned}$$

donc, par définition de l'inverse généralisé,

$$\dot{x} = (\lambda I_n - A)^{\dagger}\dot{A}x.$$

Ceci prouve la seconde assertion.

La matrice $\lambda I_n - A$ a pour valeurs propres 0 et $\lambda - \lambda_i \neq 0$, $2 \leqslant i \leqslant n$. Puisqu'elle est hermitienne, ses valeurs singulières sont $|\lambda - \lambda_i|$, $2 \leqslant i \leqslant n$, et celles de $(\lambda I_n - A)^\dagger$ sont $|\lambda - \lambda_i|^{-1}$, $2 \leqslant i \leqslant n$, (théorème 9.7).

Passons au calcul de la norme de l'opérateur $DX(A)$. Par définition

$$\|DX(A)\| = \max_{\|\dot{A}\|_2 \leqslant 1} \left\|(\lambda I_n - A)^\dagger \dot{A} x\right\|_2 .$$

Lorsque $\left\|\dot{A}\right\|_2 \leqslant 1$, puisque $\|x\|_2 = 1$, il n'est pas trop difficile de voir que les vecteurs $\dot{A}x$ décrivent la boule unité dans $\mathbb{C}^n$. On a donc

$$\|DX(A)\| = \max_{\|u\|_2 \leqslant 1} \left\|(\lambda I_n - A)^\dagger u\right\|_2 = \left\|(\lambda I_n - A)^\dagger\right\|_2 = \max_{\lambda_i \neq \lambda} |\lambda - \lambda_i|^{-1}$$

par le théorème 3.9.

Pour le calcul de la norme de $D\Lambda(A)$, notons que :

$$\|D\Lambda(A)\| = \max_{\|\dot{A}\|_2 \leqslant 1} \left|x^* \dot{A} x\right| .$$

Lorsque $\left\|\dot{A}\right\|_2 \leqslant 1$, puisque $\|x\|_2 = 1$, les scalaires $x^* \dot{A} x$ décrivent le disque unité dans $\mathbb{C}$ et donc le maximum des modules est égal à 1. Ceci prouve la troisième assertion et le théorème.

Remarque 12.4. Dans le cas de perturbations hermitiennes d'une matrice hermitienne on a le résultat plus précis suivant : soient $A, B \in \mathbb{C}^{n \times n}$ deux matrices hermitiennes dont les valeurs propres sont $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ pour A et $\mu_1 \geqslant \ldots \geqslant \mu_n$ pour B. Alors

$$\max_i |\lambda_i - \mu_i| \leqslant \|A - B\|_2 .$$

Une démonstration de ce résultat est donnée à l'exercice 12.3.

12.4 NOTES ET RÉFÉRENCES

La terminologie *valeur propre, vecteur propre* vient des travaux de Camille Jordan. Il publie en 1870 son *Traité des substitutions et des équations algébriques* sur ce que l'on appelle aujourd'hui *réduction d'endomorphisme*. Les anglophones utilisent les termes *eigenvalue* et *eigenvector* de l'allemand *Eigenwert* dû à David Hilbert (1862-1943).

Trois ouvrages sont à recommander sur ce sujet : Wilkinson [37], Chatelin [7] et Stewart-Sun [34].

EXERCICES

Exercice 12.1

Montrer que la distance de Hausdorff (définition 12.2) est bien définie et vérifie les axiomes des distances sur l'ensemble $\mathcal{K}(\mathbb{E})$ des parties compactes et non vides de $\mathbb{E}$.

Exercice 12.2

Le but de cet exercice est de prouver le théorème suivant dû à E. Fisher (1905) : soient $A \in \mathbb{C}^{n\times n}$ une matrice hermitienne et $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ ses valeurs propres rangées par ordre décroissant. Alors

$$\lambda_i = \max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} x^*Ax$$

où le maximum est pris sur l'ensemble $\mathbb{G}_{n,i}$ des sous-espaces vectoriels de dimension i de $\mathbb{C}^n$ et le minimum sur l'ensemble $\mathbb{S}_X$ des vecteurs de norme 1 dans X (autrement dit la sphère unité dans X). Pour $x \neq 0$ quelconque, le quotient

$$\frac{x^*Ax}{x^*x}$$

est appelé quotient de Rayleigh. Si $x \in \mathbb{S}_X$ alors x^*Ax est un quotient de Rayleigh.

1. Montrer qu'il existe $\overline{x} \in \mathbb{S}_X$ tel que $\overline{x}^*A\overline{x} \leqslant x^*Ax$ pour tout $x \in \mathbb{S}_X$.
2. On pose $A = UDU^*$ avec U unitaire et $D = \mathrm{diag}(\lambda_i)$. Montrer que

$$\max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} x^*Ax = \max_{Y\in\mathbb{G}_{n,i}} \min_{y\in\mathbb{S}_Y} y^*Dy.$$

3. Calculer $\min_{y\in\mathbb{S}_Y} y^*Dy$ lorsque $Y = Y_i = \{y \in \mathbb{C}^n \ : \ y_k = 0,\ i+1 \leqslant k \leqslant n\}$.
4. Soit $Y \in \mathbb{G}_{n,i}$. Montrer qu'il existe $y \in \mathbb{S}_Y$ tel que $y^*Dy \leqslant \lambda_i$. Conclure et donner la valeur de $X \in \mathbb{G}_{n,i}$ qui réalise le maximum.
5. Quelles expressions plus simples obtient-on pour λ_1 et λ_n ?
6. Montrer que λ_1 est une fonction convexe de A et que λ_n est une fonction concave de A.

Exercice 12.3

Dans cet exercice nous allons prouver le théorème suivant (H. Weyl, 1912) : soient $A, B, E \in \mathbb{C}^{n\times n}$ trois matrices hermitiennes avec $B = A + E$. Notons $\lambda_1 \geqslant \ldots \geqslant \lambda_n$

(resp. $\mu_1 \geqslant \ldots \geqslant \mu_n$, $\varepsilon_1 \geqslant \ldots \geqslant \varepsilon_n$) les valeurs propres de A (resp. B, E) rangées par ordre décroissant. Alors, pour tout i,

$$\lambda_i + \varepsilon_n \leqslant \mu_i \leqslant \lambda_i + \varepsilon_1.$$

1. Montrer qu'il existe un sous-espace X de $\mathbb{C}^n$ de dimension i tel que

$$\mu_i \leqslant x^*(A + E)x$$

pour tout $x \in X$, $\|x\|_2 = 1$ (utiliser l'exercice 12.2). En déduire que

$$\mu_i \leqslant \lambda_i + \varepsilon_1.$$

2. Montrer que $\lambda_i + \varepsilon_n \leqslant \mu_i$ (écrire que $A = B - E$).
3. Montrer que

$$\max_i |\lambda_i - \mu_i| \leqslant \|A - B\|_2 .$$

4. Montrer que si E est semi-définie positive alors $\mu_i \geqslant \lambda_i$ pour tout i.

Chapitre **13**

Sous-espaces invariants

13.1 SOUS-ESPACES INVARIANTS, SIMPLES, COMPLÉMENTAIRES

Dans l'étude de la diagonalisation d'une matrice $A \in \mathbb{C}^{n\times n}$ les sous-espaces propres $E_\lambda = \text{Ker}\,(A - \lambda I_n)$ associés aux valeurs propres λ de A ont un rôle essentiel. Une de leurs propriétés est la suivante :

$$AE_\lambda \subset E_\lambda.$$

Ils sont dits *invariants* par A. Les sous-espaces invariants d'une matrice A jouent un rôle important dans le calcul de ses valeurs propres ainsi que, nous l'avons déjà vu, dans l'étude des méthodes de projection sur des sous-espaces de Krylov.

Définition 13.1 *Un sous-espace vectoriel $\mathcal{X} \subset \mathbb{C}^n$ est dit invariant par $A \in \mathbb{C}^{n\times n}$ si*

$$A\mathcal{X} \subset \mathcal{X}.$$

Un autre exemple est fourni par la décomposition de Schur (théorème 1.12) ou la décomposition de Jordan (théorème 1.5) de A. Dans ces deux cas

$$A = PRP^{-1}$$

c'est-à-dire $AP = PR$, où R est triangulaire supérieure ; on en déduit que pour tout $k = 1 \ldots n$

$$AP_k = P_k R_k$$

avec $P_k = P(1:n, 1:k)$ et $R_k = R(1:k, 1:k)$. L'égalité précédente prouve que

$$A \operatorname{Im} P_k \subset \operatorname{Im} P_k \ :$$

les k premières colonnes de P, sans être obligatoirement des vecteurs propres de A, engendrent un sous-espace invariant par A.

La définition précédente montre que, si $\mathcal{X}$ est un sous-espace invariant de A, la restriction de A à $\mathcal{X}$ est un endomorphisme de $\mathcal{X}$:

$$A|_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}.$$

Définition 13.2 *On appelle valeur propre de A dans $\mathcal{X}$ les valeurs propres de cette restriction et on note leur ensemble*

$$\operatorname{spec}(A, \mathcal{X}) = \operatorname{spec}(A|_{\mathcal{X}}).$$

On note enfin $P_{A,\mathcal{X}}(\lambda)$ le polynôme caractéristique de $A|_{\mathcal{X}}$.

Il est clair que toute valeur propre de $A|_{\mathcal{X}}$ est aussi une valeur propre de A. Aussi le polynôme $P_{A,\mathcal{X}}(\lambda)$ divise le polynôme caractéristique $P_A(\lambda)$ de A mais les multiplicités d'une valeur propre dans $P_{A,\mathcal{X}}(\lambda)$ et dans $P_A(\lambda)$ ne sont pas nécessairement identiques.

Définition 13.3 *On dit qu'un sous-espace invariant $\mathcal{X}$ de A est simple lorsque les multiplicités des valeurs propres de $A|_{\mathcal{X}}$ sont égales aux multiplicités des valeurs propres correspondantes de A.*

À titre d'exemple considérons la matrice de Householder H_w avec $w \in \mathbb{C}^n$, $\|w\|_2 = 1$ (définition 8.6). Elle ne possède que deux sous-espaces invariants simples en dehors de $\{0\}$ et $\mathbb{C}^n$: ce sont $\mathbb{C}w$ et $w^\perp$. Ils correspondent aux valeurs propres -1 de multiplicité 1 et 1 de multiplicité $n-1$.

Un sous-espace invariant simple est caractérisé par ses valeurs propres :

Théorème 13.4 *Soit $\mathcal{X}$ un sous-espace invariant simple de A. Alors*

$$\mathcal{X} = \operatorname{Ker} P_{A,\mathcal{X}}(A).$$

De plus, il existe un unique sous-espace invariant simple $\mathcal{Y}$ de A qui soit aussi un supplémentaire de $\mathcal{X}$. On l'appelle le sous-espace invariant complémentaire de $\mathcal{X}$.

Démonstration. Supposons que

$$P_A(\lambda) = \prod_{i=1}^{q} (\lambda_i - \lambda)^{m_i}$$

où les valeurs propres λ_i sont deux à deux distinctes et où les m_i sont leurs multiplicités (donc $m_1 + \ldots + m_q = n$) et que

$$P_{A,\mathcal{X}}(\lambda) = \prod_{i=1}^{p} (\lambda_i - \lambda)^{m_i}$$

où les λ_i, $1 \leqslant i \leqslant p$, sont les valeurs propres de $A\,|_{\mathcal{X}}$. Nous allons utiliser la décomposition en sous-espaces caractérisques (théorème 1.4) :

$$\mathbb{C}^n = E_1 \oplus \ldots \oplus E_q \text{ où } E_i = \operatorname{Ker}(\lambda_i I_n - A)^{m_i} \text{ et } \dim E_i = m_i.$$

Si $x \in \mathcal{X}$ on a $P_{A,\mathcal{X}}(A)x = 0$ par le théorème de Cayley-Hamilton (théorème 1.2) d'où

$$\mathcal{X} \subset \operatorname{Ker} \prod_{i=1}^{p} (\lambda_i I_n - A)^{m_i} = E_1 \oplus \ldots \oplus E_p.$$

Comme $\dim \mathcal{X} = m_1 + \ldots + m_p$ (le degré du polynôme caractéristique $P_{A,\mathcal{X}}$) on a égalité :

$$\mathcal{X} = \operatorname{Ker} \prod_{i=1}^{p} (\lambda_i I_n - A)^{m_i}.$$

L'unique sous-espace invariant simple $\mathcal{Y}$ de A qui soit aussi un supplémentaire de $\mathcal{X}$ est égal à

$$\mathcal{Y} = \operatorname{Ker} \prod_{i=p+1}^{q} (\lambda_i I_n - A)^{m_i}.$$

Le théorème précédent admet la réciproque suivante :

Théorème 13.5 *Soit $S \cup T = \operatorname{spec} A$ une partition du spectre de A ($S \cap T = \varnothing$). Il existe une unique décomposition en sous-espaces invariants de A qui soient simples et complémentaires :*

$$\mathbb{C}^n = \mathcal{X} \oplus \mathcal{Y}$$

avec $S = \operatorname{spec}(A, \mathcal{X})$ et $T = \operatorname{spec}(A, \mathcal{Y})$.

Démonstration. L'unicité est une conséquence du théorème 13.4. Pour l'existence on prend

$$\mathcal{X} = \operatorname{Ker} \prod_{\lambda_i \in S} (\lambda_i I_n - A)^n \text{ et } \mathcal{Y} = \operatorname{Ker} \prod_{\lambda_i \in T} (\lambda_i I_n - A)^n.$$

Remarquons que ces sous-espaces sont invariants par A. Notons

$$P_1(\lambda) = \prod_{\lambda_i \in S} (\lambda_i - \lambda)^n \text{ et } P_2(\lambda) = \prod_{\lambda_i \in T} (\lambda_i - \lambda)^n.$$

Ce sont des polynômes premiers entre-eux donc, par le théorème de Bézout, il existe des polynômes $Q_1(\lambda)$ et $Q_2(\lambda)$ tels que

$$Q_1(\lambda)P_1(\lambda) + Q_2(\lambda)P_2(\lambda) = 1.$$

Les polynômes matriciels correspondants vérifient donc

$$Q_1(A)P_1(A) + Q_2(A)P_2(A) = I_n.$$

On en déduit que

$$\mathcal{X} \cap \mathcal{Y} = \text{Ker } P_1(A) \cap \text{Ker } P_2(A) = \{0\}.$$

Pour prouver que $\mathbb{C}^n = \mathcal{X} \oplus \mathcal{Y}$ on utilise la décomposition

$$x = Q_1(A)P_1(A)x + Q_2(A)P_2(A)x$$

et le fait que $Q_1(A)P_1(A)x \in \mathcal{Y}$ et que $Q_2(A)P_2(A)x \in \mathcal{X}$. En effet, les polynômes matriciels en une même matrice tels que $P_1(A)$, $P_2(A)$ et $Q_1(A)$ commutent entre eux de sorte que

$$P_2(A)(Q_1(A)P_1(A)x) = Q_1(A)P_1(A)P_2(A)x;$$

comme $P_1(\lambda)P_2(\lambda) = \prod_{\lambda_i \in \text{spec } A}(\lambda_i - \lambda)^n$ est un multiple du polynôme caractéristique $P_A(\lambda) = \prod_{\lambda_i \in \text{spec } A}(\lambda_i - \lambda)^{m_i}$ ($m_i \leqslant n$ est la multiplicité de la valeur propre λ_i) on a $P_1(A)P_2(A)x = 0$ par le théorème de Cayley-Hamilton (théorème 1.2). On remarque enfin que les spectres de A dans $\mathcal{X}$ et $\mathcal{Y}$ sont égaux à S et T qui sont deux ensembles d'intersection vide. C'est pourquoi $\mathcal{X}$ et $\mathcal{Y}$ sont simples.

13.2 FORME RÉDUITE

Soit $\mathcal{X}$ un sous-espace invariant de dimension k de A et soit $X \in \mathbb{C}^{n \times k}$ une matrice dont les k vecteurs-colonne constituent une base de $\mathcal{X}$. Notons $X = (x_1 \ldots x_k)$. Puisque $Ax_i \in \mathcal{X}$, ce vecteur s'écrit de façon unique sous la forme d'une combinaison linéaire des vecteurs x_j, $j = 1 \ldots k$. On a donc $Ax_i = Xl_i$ pour un vecteur unique $l_i \in \mathbb{C}^k$ et $AX = XL$ en notant $L = (l_1 \ldots l_k)$. La matrice L est la matrice de l'endomorphisme $A|_{\mathcal{X}}$ dans la base x_j, $j = 1 \ldots k$. Nous venons de prouver que

Proposition 13.6 *Pour toute matrice $X \in \mathbb{C}^{n\times k}$ dont les vecteurs-colonne constituent une base de $\mathcal{X}$, il existe une unique matrice $L \in \mathbb{C}^{k\times k}$ telle que*

$$AX = XL.$$

De plus

$$\text{spec}\,(A, \mathcal{X}) = \text{spec}\,(L).$$

Soit $X \in \mathbb{S}t_{nk}$ une matrice dont les vecteurs-colonne constituent une base orthonormée de $\mathcal{X}$ et soit $L \in \mathbb{C}^{k\times k}$ telle que $AX = XL$. Complétons les vecteurs-colonne de X en une base orthonormée de $\mathbb{C}^n$: on obtient ainsi une matrice unitaire $(X\ Y) \in \mathbb{U}_n$. Dans cette base la matrice A s'écrit sous forme d'une matrice triangulaire supérieure par blocs :

Proposition 13.7 *Soit $X \in \mathbb{S}t_{nk}$ de rang k telle que $AX = XL$. Pour toute matrice $Y \in \mathbb{C}^{n\times(n-k)}$ telle que $(X\ Y) \in \mathbb{C}^{n\times n}$ soit unitaire, on a la forme réduite*

$$(X\ Y)^* A (X\ Y) = \begin{pmatrix} L & H \\ 0 & \tilde{L} \end{pmatrix} \tag{13.1}$$

avec $L = X^*AX$, $H = X^*AY$ et $\tilde{L} = Y^*AY$.

Démonstration. On a $A(X\ Y) = (AX\ AY) = (XL\ AY)$. En multipliant à gauche par $\begin{pmatrix} X^* \\ Y^* \end{pmatrix}$ et en observant que $X^*X = I_k$ et $Y^*X = 0$ on obtient la forme annoncée.

Remarque 13.1. Puisque $(X\ Y)$ est une matrice unitaire, l'espace $\mathcal{Y}$ engendré par les colonnes de Y est égal à $\mathcal{X}^\perp$. Ce n'est pas nécessairement un sous-espace invariant de A sauf si cette matrice est hermitienne. Dans ce cas $H = 0$, $(X\ Y)^* A (X\ Y) = \begin{pmatrix} L & 0 \\ 0 & \tilde{L} \end{pmatrix}$ et $A\ \text{Im}\ Y \subset \text{Im}\ Y$.

Lorsque $\mathcal{X}$ est un sous-espace invariant simple, un changement de base va nous permettre de décomposer A sous-forme d'une matrice diagonale par blocs. Cette construction utilise l'équation de Sylvester que nous étudions ci-dessous.

13.3 ÉQUATION DE SYLVESTER

Définition 13.8 *Soient $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$ et $C \in \mathbb{C}^{n\times m}$. L'équation matricielle*

$$AQ - QB = C \tag{13.2}$$

où $Q \in \mathbb{C}^{n\times m}$ est appelée équation de Sylvester.

Cette équation fait apparaître l'opérateur linéaire

$$\mathcal{S} : \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}, \quad \mathcal{S}(Q) = AQ - QB.$$

Nous souhaitons savoir si l'équation de Sylvester $\mathcal{S}(Q) = C$ admet une solution unique ou, en d'autres termes, si l'opérateur $\mathcal{S}$ est inversible. On a le théorème suivant :

Théorème 13.9 *L'opérateur de Sylvester $\mathcal{S}$ est inversible si et seulement si*

$$\text{spec } A \cap \text{spec } B = \varnothing.$$

Démonstration. Considérons λ une valeur propre de A et $x \in \mathbb{C}^n$, $x \neq 0$, un vecteur propre associé. De même, soit μ une valeur propre de B et soit $y \in \mathbb{C}^m$, $y \neq 0$, tel que $y^*B = \mu y^*$ (y est un vecteur propre de B^* associé à la valeur propre $\overline{\mu}$). La matrice $Q = xy^* \in \mathbb{C}^{n\times m}$ vérifie $Q \neq 0$ et l'on a

$$AQ - QB = Axy^* - xy^*B = (\lambda - \mu)xy^*.$$

Si spec $A \cap$ spec $B \neq \varnothing$ et si l'on prend $\lambda = \mu$ dans cette intersection, alors la matrice $Q = xy^*$ correspondante est un élément non nul du noyau de $\mathcal{S}$ qui n'est donc pas inversible.

Réciproquement, supposons que spec $A \cap$ spec $B = \varnothing$. On va montrer que pour toute matrice C l'équation $\mathcal{S}(Q) = C$ possède une solution et donc $\mathcal{S}$ sera inversible. Considérons une décomposition de Schur de B (théorème 1.12) : $B = VTV^*$. En multipliant à droite le système $\mathcal{S}(Q) = C$ par la matrice V on obtient

$$A(QV) - (QV)T = CV.$$

Posons $P = QV$ et $D = CV$. Il s'agit de résoudre le système

$$AP - PT = D, \tag{13.3}$$

où T est une matrice triangulaire supérieure. Notons p_i (respectivement d_i) les colonnes de P (respectivement de D) et t_{ij} les coefficients de T. La première colonne de ce système est égale à

$$Ap_1 - t_{11}p_1 = d_1$$

et t_{11}, élément de la diagonale de T, est une valeur propre de B. L'hypothèse implique que $A - t_{11}I_n$ est inversible et donc p_1 est bien défini.

Supposons avoir déterminé les colonnes $p_1, \ldots, p_{k-1}$ de P. La $k-$ième colonne de l'équation 13.3 est

$$Ap_k - \sum_{i=1}^{k} t_{ik} p_i = d_k$$

d'où

$$(A - t_{kk} I_n) p_k = \sum_{i=1}^{k-1} t_{ik} p_i + d_k.$$

Comme $t_{kk} \in \text{spec } B$ la matrice $A - t_{kk} I_n$ est inversible ce qui prouve l'existence (et l'unicité) de p_k.

Corollaire 13.10 *Quelles que soient les matrices $A \in \mathbb{C}^{n\times n}$ et $B \in \mathbb{C}^{m\times m}$, les valeurs propres de l'opérateur de Sylvester $\mathcal{S}$ sont $\lambda - \mu$ où $\lambda \in \text{spec } A$ et $\mu \in \text{spec } B$.*

Démonstration. Dans la première partie de la démonstration du théorème précédent on a montré que $\lambda - \mu$ où $\lambda \in \text{spec } A$ et $\mu \in \text{spec } B$ est une valeur propre $\mathcal{S}$.

Réciproquement, soient $Q \neq 0$ et $\sigma \in \mathbb{C}$ tels que $\mathcal{S}Q = \sigma Q$. On a donc $(A - \sigma I_n)Q - QB = 0$ c'est-à-dire que le noyau de ce demier opérateur de Sylvester n'est pas réduit à 0. Le théorème montre que cela n'est possible que s'il existe $\lambda \in \text{spec } A$ et $\mu \in \text{spec } B$ tels que $\lambda - \sigma = \mu$. Ceci prouve que $\sigma = \lambda - \mu$.

Remarque 13.2.

1. La démonstration du théorème précédent donne un procédé de construction de la solution Q de l'équation de Sylvester :
   a) On calcule une décomposition de Schur de B,
   b) On résout les systèmes linéaires

$$(A - t_{kk} I_n) p_k = \sum_{i=1}^{k-1} t_{ik} p_i + d_k, \; k = 1, \ldots, m.$$

   Ce procédé est numériquement coûteux : chaque résolution nécessite $O(n^3)$ opérations arithmétiques.
2. Une méthode plus performante est celle de Bartels et Stewart. [1]

1. Méthode publiée en 1972 dans *Communications of the ACM.*

a) On décompose aussi la matrice A dans la forme de Schur : $A = USU^*$ avec S triangulaire supérieure,

b) En multipliant à gauche l'équation (13.3) par U^* on obtient

$$SR - RT = \tilde{D},$$

où $R = U^*QV$, et $\tilde{D} = U^*CV$.

c) Comme dans la démonstration précédente les colonnes r_i de R sont obtenues en résolvant les systèmes

$$(S - t_{kk}I_n)r_k = \sum_{i=1}^{k-1} t_{ik}r_i + \tilde{d}_k,\ k = 1, \ldots, m,$$

dont les matrices sont triangulaires supérieures ; chaque résolution ne nécessite que $O(n^2)$ opérations arithmétiques.

Remarque 13.3. L'équation de *Lyapunov*

$$AQ + QA^* = D$$

est une forme particulière d'équation de Sylvester. Elle est utilisée en théorie du contrôle et permet de caractériser la stabilité de tels systèmes. Le théorème précédent permet de montrer que, si la matrice A est *stable* (c'est-à-dire si $\Re(\lambda) < 0$ pour toute valeur propre λ de A), il existe une solution unique de cette équation (exercice 13.4).

13.4 DIAGONALISATION PAR BLOCS D'UNE MATRICE

Théorème 13.11 *Soit $\mathcal{X}_1$ un sous-espace invariant simple de A et soit $X_1 \in \mathbb{S}t_{nk}$ une matrice dont les vecteurs-colonne constituent une base orthonormée de $\mathcal{X}_1$. Complétons X_1 en une matrice unitaire $(X_1\ Y_2) \in \mathbb{U}_n$ et considérons la forme réduite*

$$(X_1\ Y_2)^* A (X_1\ Y_2) = \begin{pmatrix} L_1 & H \\ 0 & L_2 \end{pmatrix}.$$

Soit Q la solution de l'équation de Sylvester $L_1Q - QL_2 = -H$ et soient

$$\begin{aligned} X_2 &= Y_2 + X_1Q \\ Y_1 &= X_1 - Y_2Q^*. \end{aligned}$$

Sous ces hypothèses

1. $(X_1\ X_2)$ *est inversible et* $(X_1\ X_2)^{-1} = (Y_1\ Y_2)^*$,
2. $(X_1\ X_2)^* A (Y_1\ Y_2) = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix}$,
3. $\mathcal{X}_2 = \operatorname{Im} X_2$ *est le sous-espace invariant complémentaire de* $\mathcal{X}_1$.

Démonstration. Pour toute matrice $Q \in \mathbb{C}^{k\times(n-k)}$ la matrice $\begin{pmatrix} I_k & Q \\ 0 & I_{n-k} \end{pmatrix}$ est inversible d'inverse $\begin{pmatrix} I_k & -Q \\ 0 & I_{n-k} \end{pmatrix}$ et

$$\begin{pmatrix} I_k & -Q \\ 0 & I_{n-k} \end{pmatrix} \begin{pmatrix} L_1 & H \\ 0 & L_2 \end{pmatrix} \begin{pmatrix} I_k & Q \\ 0 & I_{n-k} \end{pmatrix} = \begin{pmatrix} L_1 & L_1Q - QL_2 + H \\ 0 & L_2 \end{pmatrix}$$

Puisque $\mathcal{X}_1$ est un sous-espace invariant simple, on a spec $L_1 \cap$ spec $L_2 = \varnothing$ et donc, d'après le théorème 13.9, l'équation de Sylvester $L_1Q - QL_2 = -H$ admet une solution unique Q. On obtient X_2 et Y_1 en considérant respectivement les produits

$$(X_1\ Y_2) \begin{pmatrix} I_k & Q \\ 0 & I_{n-k} \end{pmatrix}$$

et

$$\begin{pmatrix} I_k & -Q \\ 0 & I_{n-k} \end{pmatrix} \begin{pmatrix} X_1^* \\ Y_2^* \end{pmatrix}$$

ce qui prouve l'égalité 2. On a $\mathcal{X}_1 \oplus \mathcal{X}_2 = \mathbb{C}^n$ parce que la matrice $(X_1\ X_2)$ est inversible et $A\mathcal{X}_2 \subset \mathcal{X}_2$ parce que $AX_2 = X_2L_2$. Enfin, $\mathcal{X}_2$ est un sous-espace simple parce que les spectres de $A\,|_{\mathcal{X}_1}$ et de $A\,|_{\mathcal{X}_2}$, égaux aux spectres de L_1 et L_2, sont disjoints.

Corollaire 13.12 *Avec les hypothèses et les notations du théorème 13.11,* $\mathcal{X}_1^\perp$ *et* $\mathcal{X}_2^\perp$ *sont deux sous-espaces invariants complémentaires de* A^*.

La démonstration est laissée en exercice au lecteur.

EXERCICES

Exercice 13.1

Déterminer les sous-espaces invariants d'une rotation de Givens. Quels sont ceux qui sont simples ?

Exercice 13.2

Démontrer le corollaire 13.12.

Exercice 13.3

Soient $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$ et $C \in \mathbb{C}^{n\times m}$. On considère l'équation de Sylvester

$$AQ - QB = C.$$

On suppose que spec $A \cap$ spec $B = \varnothing$. Donner l'expression de la solution Q de cette équation lorsque l'on suppose que $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ et $B = \text{diag}(\mu_1, \ldots, \mu_m)$. En déduire l'expression de Q lorsque A et B sont diagonalisables : $A = M\text{diag}(\lambda_i)M^{-1}$ et $B = N\text{diag}(\mu_j)N^{-1}$.

Exercice 13.4

Soient $A, D, Q \in \mathbb{C}^{n\times n}$. On considère l'équation de Lyapunov

$$AQ + QA^* = D$$

et l'on suppose que A est stable c'est-à-dire que $\Re(\lambda) < 0$ pour toute valeur propre λ de A.

1. Montrer que cette équation admet une solution unique Q.
2. Montrer que l'intégrale

$$\int_0^\infty \exp(tA)\, D\, \exp(tA^*)\, dt$$

est convergente et qu'elle est égale à $-Q$.

3. Montrer que si D est hermitienne alors Q l'est aussi.
4. Montrer que si $-D$ est semi-définie positive (resp. définie positive), alors Q l'est aussi.

5. En utilisant l'exercice 13.3 donner l'expression de Q lorsque A est diagonale. Si de plus $D = (d_{ij})$, avec $d_{ij} = 1$ pour tout i, j, montrer que Q est une *matrice de Cauchy* (exercice 7.3).
6. On considère la matrice $H \in \mathbb{C}^{2n \times 2n}$

$$H = \begin{pmatrix} A^* & 0 \\ D & -A \end{pmatrix}.$$

a) Montrer qu'il existe $Y, Z, \Lambda \in \mathbb{C}^{n \times n}$ avec Y inversible et spec $(\Lambda) =$ spec (A^*) telles que

$$H \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} Y \\ Z \end{pmatrix} \Lambda.$$

b) Montrer que $Q = Z Y^{-1}$.

Chapitre **14**

Le calcul des valeurs propres

Puisque les valeurs propres d'une matrice sont les racines du polynôme caractéristique, on doit s'attendre à ce que le calcul des valeurs propres soit aussi compliqué que la recherche des racines d'un polynôme. Ces deux problèmes sont même équivalents puisque tout polynôme est le polynôme caractéristique de sa matrice compagnon (voir l'exercice 1.7). Pourtant, la recherche des valeurs propres ne se fait pas par le biais des zéros des polynômes. On préfère une approche plus géométrique comme nous allons le voir.

14.1 LA MÉTHODE DE LA PUISSANCE

Le principe de la *méthode de la puissance* est contenu dans l'observation suivante : soient $A \in \mathbb{C}^{n\times n}$ et $z \in \mathbb{C}$, $z \neq 0$. En général, la suite des droites vectorielles de direction $A^k z$ converge vers la direction propre associée à la valeur propre de plus grand module. Commençons par préciser ce que l'on entend par « *une suite de droites converge* ».

Définition 14.1 *Étant donnés $u, v \in \mathbb{C}^n$, u et $v \neq 0$, posons*

$$d_P(u, v) = \min_{\lambda\in\mathbb{C}} \frac{\|u - \lambda v\|_2}{\|u\|_2}.$$

Proposition 14.2 *Quels que soient $u, v \in \mathbb{C}^n$, u et $v \neq 0$, $\alpha, \beta \in \mathbb{C}$, α et $\beta \neq 0$ on a :*

1.

$$d_P(u, v) = d_P(\alpha u, v) = d_P(u, \beta v),$$

2.
$$d_P(u,v) = \left(1 - \frac{|\langle u,v\rangle|^2}{\|u\|_2^2\,\|v\|_2^2}\right)^{\frac{1}{2}},$$

3. *d_P est unitairement invariante : $d_P(Qu,Qv) = d_P(u,v)$ pour toute transformation unitaire $Q \in \mathbb{U}_n$,*

4. *d_p est une distance sur l'ensemble des droites vectorielles de $\mathbb{C}^n$ (leur ensemble est noté $\mathbb{P}_{n-1}(\mathbb{C})$ et s'appelle l'espace projectif complexe associé à $\mathbb{C}^n$.)*

Remarque 14.1.

1. On définit de la même manière une distance sur l'ensemble des droites vectorielles de $\mathbb{R}^n$. Dans ce contexte, la distance $d_P(u,v)$ est le sinus de l'angle fait par les droites $\mathbb{R}u$ et $\mathbb{R}v$.
2. $d_P(u,v) \leqslant 1$ et $= 1$ pour des vecteurs u et v orthogonaux.

Démonstration. Le minimum qui définit d_P est atteint lorsque λv est la projection orthogonale de u sur la droite $\mathbb{C}v$. On a donc $\langle u - \lambda v, v\rangle = 0$ d'où
$$\lambda = \frac{\langle u,v\rangle}{\langle v,v\rangle}.$$

Ainsi
$$d_P(u,v) = \frac{\left\|u - \frac{\langle u,v\rangle}{\langle v,v\rangle}v\right\|_2}{\|u\|_2} = \left(1 - \frac{|\langle u,v\rangle|^2}{\|u\|_2^2\,\|v\|_2^2}\right)^{\frac{1}{2}}.$$

La première propriété en découle, l'invariance unitaire aussi. Prouvons que c'est une distance. La symétrie ($d_P(u,v) = d_P(v,u)$) est immédiate ainsi que $d_P(u,u) = 0$. Si $d_P(u,v) = 0$, c'est que u est sa propre projection sur la droite $\mathbb{C}v$. Ceci prouve l'égalité $\mathbb{C}u = \mathbb{C}v$. L'inégalité du triangle est délicate à démontrer. Nous ne le ferons pas ici (exercice 14.3).

Revenons à la méthode de la puissance. Elle est donnée par l'algorithme suivant :

Méthode de la puissance

Entrée : $z \in \mathbb{C}^n$, $z \neq 0$

$$z_0 = \frac{z}{\|z\|_2}$$

pour $k = 1 : \ldots$

$$z_k = \frac{Az_{k-1}}{\|Az_{k-1}\|_2}$$

$$\zeta_k = z_k^* A z_k$$

fin

Théorème 14.3 *Supposons que les valeurs propres de $A \in \mathbb{C}^{n\times n}$ vérifient*

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n| .$$

Soit x_1 un vecteur propre de A associé à λ_1. Il existe un ensemble ouvert et dense $\mathcal{U} \subset \mathbb{C}^n$ tel que, pour tout $z \in \mathcal{U}$,

1. *Les suites (z_k) et (ζ_k) définies ci-dessus sont bien définies,*
2. $\lim_{k\to\infty} d_P(z_k, x_1) = 0$,
3. $\lim_{k\to\infty} \zeta_k = \lambda_1$.

Démonstration. Afin d'établir les propriétés de ces suites nous allons utiliser la décomposition de Jordan de A :

$$A = PJP^{-1}$$

où $J = \operatorname{diag}(J_1, \ldots, J_p)$ est une matrice diagonale par blocs constituée de blocs de Jordan : $J_i = \lambda_{n_i} I_{n_i} \in \mathbb{C}^{n_i \times n_i}$ ou $J_i = \lambda_{n_i} I_{n_i} + N_{n_i}$ (notations du théorème 1.5) et où $n_1 + \ldots + n_p = n$. Notons que cette écriture suppose que les valeurs propres suivantes sont égales

$$\lambda_{n_1+\ldots+n_{i-1}+1} = \ldots = \lambda_{n_1+\ldots+n_{i-1}+n_i},$$

$i = 2, \ldots, p$. Comme λ_1 est une valeur propre simple on peut supposer que le premier bloc de Jordan est $J_1 = (\lambda_1) \in \mathbb{C}^{1\times 1}$. Notons $x_1, \ldots, x_n$ les colonnes de P. Ces vecteurs constituent une base de $\mathbb{C}^n$. Définissons

$$\mathcal{U} = \{\alpha_1 x_1 + \ldots + \alpha_n x_n \ : \ \alpha_1 \neq 0\} ;$$

c'est un ouvert dense de $\mathbb{C}^n$. Soit $z \in \mathcal{U}$, $z = \alpha_1 x_1 + \ldots + \alpha_n x_n$. On a

$$A^k z = \alpha_1 \lambda_1^k x_1 + \ldots + \alpha_n \lambda_n^k x_n \neq 0$$

parce que α_1 et $\lambda_1 \neq 0$. Nous allons voir que

$$z_k = \frac{A^k z}{\|A^k z\|_2}.$$

C'est vrai pour $k = 0$ et si l'égalité a lieu à l'ordre k on a :

$$z_{k+1} = \frac{Az_k}{\|Az_k\|_2} = \frac{A\frac{A^k z}{\|A^k z\|_2}}{\left\|A\frac{A^k z}{\|A^k z\|_2}\right\|_2} = \frac{A^{k+1} z}{\|A^{k+1} z\|_2}.$$

Nous allons montrer que

$$A^k z = \lambda_1^k (\alpha_1 x_1 + e_k)$$

avec $\lim_{k\to\infty} e_k = 0$. On pourra alors en déduire que

$$d_P(z_k, x_1) = d_P(A^k z, x_1) = d_P(x_1 + e_k/\alpha_1, x_1) =$$

$$\left(1 - \frac{|\langle x_1 + e_k/\alpha_1, x_1\rangle|^2}{\|x_1 + e_k/\alpha_1\|_2^2 \, \|x_1\|_2^2}\right)^{\frac{1}{2}} \to 0$$

et que $z_k^* A z_k =$

$$\frac{(A^k z)^*}{\|A^k z\|_2} A \frac{A^k z}{\|A^k z\|_2} = \left(\frac{\overline{\lambda_1}}{|\lambda_1|}\right)^k \frac{\overline{\alpha_1} x_1^* + e_k^*}{\|\alpha_1 x_1 + e_k\|_2} A \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\alpha_1 x_1 + e_k}{\|\alpha_1 x_1 + e_k\|_2}$$

$$= \frac{\left(\overline{\alpha_1} x_1^* + e_k^*\right) A (\alpha_1 x_1 + e_k)}{\|\alpha_1 x_1 + e_k\|_2^2} \to \frac{(\overline{\alpha_1} x_1^*) A(\alpha_1 x_1)}{\|\alpha_1 x_1\|_2^2} = \lambda_1$$

ce qui établit le théorème. Afin de donner l'expression de $A^k z$ notons que

$$z = P \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

et que

$$A^k = P \operatorname{diag}(J_i^k) P^{-1}$$

de sorte que

$$A^k z = \sum_{i=1}^{p} (x_{n_1+\ldots+n_{i-1}+1}, \ldots, x_{n_1+\ldots+n_i}) J_i^k \begin{pmatrix} \alpha_{n_1+\ldots+n_{i-1}+1} \\ \vdots \\ \alpha_{n_1+\ldots+n_i} \end{pmatrix}$$

où n_i est la taille du bloc J_i. Nous avons vu que $n_1 = 1$ de sorte que

$$A^k z = \lambda_1^k \alpha_1 x_1 + \sum_{i=2}^{p} (x_{n_1+\ldots+n_{i-1}+1}, \ldots, x_{n_1+\ldots+n_i}) J_i^k \begin{pmatrix} \alpha_{n_1+\ldots+n_{i-1}+1} \\ \vdots \\ \alpha_{n_1+\ldots+n_i} \end{pmatrix} =$$

$$\lambda_1^k (\alpha_1 x_1 + e_k).$$

De plus, $J_i^k = \lambda_{n_i}^k I_{n_i}$ ou bien, dans le cas nilpotent,

$$J_i^k = \begin{pmatrix} \lambda_{n_i} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_{n_i} \end{pmatrix}^k =$$

$$\begin{pmatrix} \lambda_{n_i}^k & \binom{k}{1}\lambda_{n_i}^{k-1} & \binom{k}{2}\lambda_{n_i}^{k-2} & \cdots & \binom{k}{n_i-1}\lambda_{n_i}^{k-n_i+1} \\ & \lambda_{n_i}^k & \binom{k}{1}\lambda_{n_i}^{k-1} & \cdots & \binom{k}{n_i-2}\lambda_{n_i}^{k-n_i+2} \\ & & \ddots & \cdots & \vdots \\ & & & \lambda_{n_i}^k & \binom{k}{1}\lambda_{n_i}^{k-1} \\ & & & & \lambda_{n_i}^k \end{pmatrix}$$

lorsque $k > n$. Comme $|\lambda_1| > |\lambda_i|$ pour tout $i \geqslant 2$, on a

$$\lim_{k\to\infty} \binom{k}{m} \frac{\lambda_{n_i}^k}{\lambda_1^k} = 0$$

pour tout $m \leqslant n$ de sorte que

$$\lim_{k\to\infty} e_k = \lim_{k\to\infty} \sum_{i=2}^{p} (x_{n_1+\ldots+n_{i-1}+1}, \ldots, x_{n_1+\ldots+n_i}) \frac{J_i^k}{\lambda_1^k} \begin{pmatrix} \alpha_{n_1+\ldots+n_{i-1}+1} \\ \vdots \\ \alpha_{n_1+\ldots+n_i} \end{pmatrix} = 0.$$

Ceci prouve notre assertion et le théorème.

Remarque 14.2.

1. Nous disons qu'une propriété $\mathcal{P}(a)$, qui est vraie ou non suivant la valeur du paramètre $a \in \mathbb{C}^p$, a lieu *de façon générique* lorsque l'ensemble des $a \in \mathbb{C}^p$ pour lesquels la propriété $\mathcal{P}(a)$ est vraie contient un ensemble ouvert et dense de $\mathbb{C}^p$.

2. L'hypothèse du théorème n'est pas vraiment restrictive. On peut montrer qu'une matrice $A \in \mathbb{C}^{n\times n}$ a des valeurs propres de modules distincts *de façon générique*. De plus, un vecteur $z \in \mathbb{C}^n$ a *de façon générique* des coordonnées non nulles dans une base donnée.
3. Le calcul précédent montre que, lorsque A est diagonalisable (c'est-à-dire lorsque les blocs de Jordan sont tous de type $\lambda_{n_i} I_{n_i}$) on a

$$d_P(z_k, x_1) \leqslant C \left|\frac{\lambda_2}{\lambda_1}\right|^k$$

pour une constante $C > 0$ convenable. La convergence de la suite (z_k) est donc linéaire, et la vitesse de convergence donnée par le quotient $|\lambda_2/\lambda_1| < 1$. Dans le cas général, cette inégalité n'a lieu qu'asymptotiquement.
4. Lorsque la matrice A a des valeurs propres de modules distincts, on peut montrer que la méthode de la puissance converge quel que soit le point initial $z \in \mathbb{C}^n$, $z \neq 0$, vers l'une des directions propres de A (exercice 14.2).
5. Il faut noter que cet algorithme n'utilise que des produits matrice-vecteur. On n'a donc pas besoin de stocker la matrice A.
6. La direction propre et la valeur propre de plus petit module peuvent être obtenues, lorsque A est inversible, par la *méthode de la puissance inverse*. Celle-ci est définie par l'itération

$$z_k = \frac{A^{-1} z_{k-1}}{\|A^{-1} z_{k-1}\|_2}.$$

Les valeurs propres de A^{-1} sont les inverses des valeurs propres de A et ces deux matrices se diagonalisent dans la même base de vecteurs propres.

La figure 14.1 illustre la convergence de la suite (z_k) dans le cas d'une matrice symétrique 3×3. Les points sont sur la sphère unité de $\mathbb{R}^3$.

14.2 ITÉRATION DE SOUS-ESPACES

La méthode de la puissance est en défaut pour une matrice A qui possède deux valeurs propres de plus grand module (exercice 14.1). Mais dans un tel cas et sous des hypothèses raisonnables la suite des plans $\mathcal{P}_{k+1} = A\mathcal{P}_k$, où $\mathcal{P}_0$ est un plan donné, « converge » vers le plan $\mathcal{P}$ défini par les vecteurs propres associés à ces deux valeurs propres. C'est pourquoi on étend la méthode de la puissance au cas d'une itération de sous-espaces. On entend par là toute suite définie par $\mathcal{X}_{k+1} = A\mathcal{X}_k$ où $\mathcal{X}_0$ est un sous-espace de dimension p de $\mathbb{C}^n$. Cette suite se stabilise-t-elle vers un sous-espace

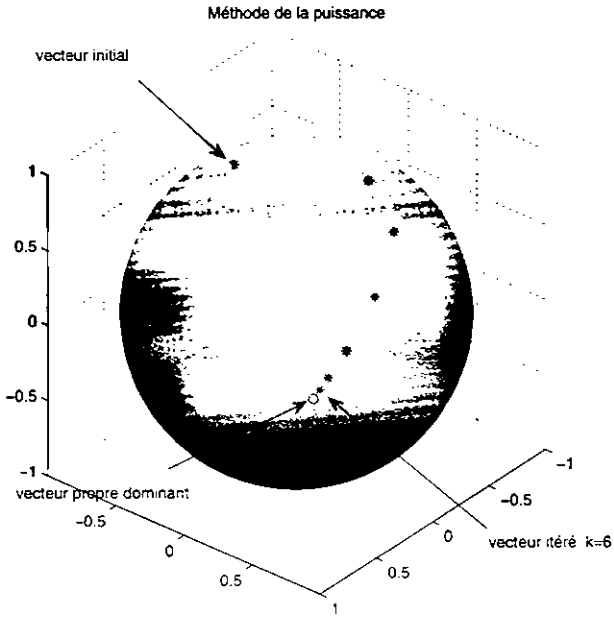


Figure 14.1 Convergence de la suite (z_k).

de $\mathbb{C}^n$? En général la réponse est oui : la suite $(\mathcal{X}_k)$ converge vers le « sous-espace invariant dominant » de dimension p.

Quelle structure de données choisir pour décrire un sous-espace de dimension p de $\mathbb{C}^n$? Et bien tout simplement une matrice $X \in \mathbb{C}^{n\times p}$ de rang p. Ses colonnes engendrent un sous-espace de dimension p : le sous-espace image $\operatorname{Im} X = \{Xu \ : \ u \in \mathbb{C}^p\}$. Un tel choix n'est pas unique :

Lemme 14.4 *Quelles que soient les matrices de rang p : $X, Y \in \mathbb{C}^{n\times p}$ on a* $\operatorname{Im} X = \operatorname{Im} Y$ *si et seulement s'il existe $G \in \mathbb{GL}_p$ telle que $Y = XG$.*

> *Démonstration.* La condition est suffisante : si $Y = XG$ alors $Yv = XGv$ pour tout $v \in \mathbb{C}^p$ de sorte que $\operatorname{Im} Y \subset \operatorname{Im} X$. L'autre inclusion s'obtient via l'égalité $X = YG^{-1}$. La condition est nécessaire : si $\operatorname{Im} X = \operatorname{Im} Y$ les colonnes de Y sont des combinaisons linéaires de celles de X ce qui prouve l'existence de $G \in \mathbb{C}^{p\times p}$ telle que $Y = XG$. Si G n'était pas inversible, il existerait $v \in \mathbb{C}^p$, $v \neq 0$, avec $Gv = 0$. On aurait alors $\dim(\operatorname{Im} XG) \leqslant p-1$ et Y ne serait pas de rang p.

A toute matrice $Z \in \mathbb{C}^{n\times p}$ de rang p on peut associer une unique décomposition $Z = QR$ avec $Q \in \mathbb{St}_{np}$ et $R \in \mathbb{C}^{p\times p}$ triangulaire supérieure à diagonale positive

(proposition 8.4) que nous notons $Q = \mathcal{Q}(Z)$ et $R = \mathcal{R}(Z)$. En vertu du lemme précédent on a :

$$\text{Im } Z = \text{Im } \mathcal{Q}(Z).$$

Comme au paragraphe précédent, où nous avons défini la notion de convergence d'une suite de droites, nous devons définir maintenant un concept de limite pour des suites de sous-espaces dans $\mathbb{C}^n$.

Définition 14.5 *Notons $\mathbb{G}_{np}$ l'ensemble des sous-espaces vectoriels de dimension p contenus dans $\mathbb{C}^n$. Cet ensemble est appelé « grassmannienne ». Etant donnés $\mathcal{U}, \mathcal{V} \in \mathbb{G}_{np}$, posons*

$$d_G(\mathcal{U}, \mathcal{V}) = \sup_{u \in \mathcal{U}, u \neq 0} \inf_{v \in \mathcal{V}} \frac{\|u - v\|_2}{\|u\|_2}.$$

Ce nombre est le supremum du sinus de l'angle fait par un vecteur $u \in \mathcal{U}$ avec sa projection orthogonale sur $\mathcal{V}$. Noter que pour $p = 1$ on retrouve la définition 14.1.

La proposition suivante (difficile) sera prouvée à l'exercice 14.3.

Proposition 14.6

1. $d_G(\mathcal{U}, \mathcal{V}) = \max_{u \in \mathcal{U}, u \neq 0} \min_{v \in \mathcal{V}} \frac{\|u - v\|_2}{\|u\|_2}$,
2. $d_G(\mathcal{U}, \mathcal{V}) = \|\Pi_{\mathcal{V}^\perp} \circ \Pi_{\mathcal{U}}\|_2$ *où Π_X désigne la projection orthogonale dans $\mathbb{C}^n$ sur le sous-espace X,*
3. *d_G est une distance sur $\mathbb{G}_{np}$,*
4. *d_G est unitairement invariante : $d_G(Q(\mathcal{U}), Q(\mathcal{V})) = d_G(\mathcal{U}, \mathcal{V})$ pour toute matrice unitaire Q dans $\mathbb{C}^n$.*

L'itération de sous-espaces est décrite par l'algorithme suivant :

Itération de sous-espaces

Entrée : $Z \in \mathbb{C}^{n \times p}$, de rang p
$Z_0 = \mathcal{Q}(Z)$,
pour $k = 1 : \ldots$
$\quad Z_k = \mathcal{Q}(AZ_{k-1})$
fin

Théorème 14.7 *Soit $A \in \mathbb{C}^{n \times n}$ une matrice diagonalisable dont les valeurs propres vérifient*

$$|\lambda_1| \geqslant \ldots \geqslant |\lambda_p| > |\lambda_{p+1}| \geqslant \ldots \geqslant |\lambda_n|$$

et soit $x_1, \ldots, x_n$ une base de vecteurs propres de A (x_i associé à λ_i). Notons $\mathcal{X}$ et $\mathcal{Y}$ les sous-espaces de $\mathbb{C}^n$ engendrés par les vecteurs $x_1, \ldots, x_p$ pour $\mathcal{X}$ et par $x_{p+1}, \ldots, x_n$ pour $\mathcal{Y}$.

Pour toute matrice $Z \in \mathbb{C}^{n\times p}$ de rang p et telle que

$$(\operatorname{Im} Z) \cap \mathcal{Y} = \{0\},$$

la suite (Z_k) décrite ci-dessus vérifie les propriétés suivantes :

1. $\operatorname{rang} Z_k = p$,
2. $\lim_{k\to\infty} d_G(\operatorname{Im} Z_k, \mathcal{X}) = 0$.

Démonstration. Nous n'allons prouver ce théorème (difficile) que lorsque A est hermitienne. On a alors

$$A = UDU^*$$

avec

$$D = \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix}$$

où $D_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ et $D_2 = \operatorname{diag}(\lambda_{p+1}, \ldots, \lambda_n)$. On peut aussi supposer que les vecteurs x_i sont orthonormés et que

$$U = (x_1 \ldots x_n) = (X \;\; Y)$$

avec $X = (x_1 \ldots x_p)$ et $Y = (x_{p+1} \ldots x_n)$. Ainsi

$$X = U \begin{pmatrix} I_p \\ 0 \end{pmatrix} \text{ et } Y = U \begin{pmatrix} 0 \\ I_{n-p} \end{pmatrix}.$$

On pose aussi

$$Q(Z) = Z_0 = U Q_0 = U \begin{pmatrix} V_0 \\ W_0 \end{pmatrix}$$

avec $V_0 \in \mathbb{C}^{p\times p}$ et $W_0 \in \mathbb{C}^{(n-p)\times p}$. L'hypothèse

$$(\operatorname{Im} Z) \cap \mathcal{Y} = \{0\}$$

qui s'écrit aussi

$$(\operatorname{Im} Z_0) \cap (\operatorname{Im} Y) = \{0\}$$

devient donc

$$\operatorname{Im} \begin{pmatrix} V_0 \\ W_0 \end{pmatrix} \cap \operatorname{Im} \begin{pmatrix} 0 \\ I_{n-p} \end{pmatrix} = \{0\}.$$

Puisque l'on a supposé que Z (et donc Q_0) est de rang p, cette dernière identité signifie que V_0 est inversible (supposons que $V_0 u = 0$. L'hypothèse implique alors $W_0 u = 0$ c'est-à-dire $Q_0 u = 0$. Comme rang $Q_0 = p$ on a $u = 0$ et donc V_0 est inversible.)

Nous allons maintenant montrer que rang $Z_k = p$ pour tout k. Par construction $Z_k = \mathcal{Q}(A Z_{k-1})$ c'est-à-dire que

$$Z_k R_k = A Z_{k-1}$$

pour une matrice $R_k \in \mathbb{C}^{p \times p}$ triangulaire supérieure à diagonale positive. On notera

$$Z_k = U \begin{pmatrix} V_k \\ W_k \end{pmatrix} = U Q_k$$

avec $V_k \in \mathbb{C}^{p \times p}$ et $W_k \in \mathbb{C}^{(n-p) \times p}$ de sorte que

$$Q_k R_k = D Q_{k-1}.$$

Cette égalité implique

$$V_k R_k = D_1 V_{k-1}$$

ce qui prouve, par récurrence, que V_k est inversible pour tout k (D_1 est elle-même inversible) et donc que rang $Z_k = p$ pour tout k.

Nous devons maintenant estimer la distance $d_G(\text{Im } Z_k, \mathcal{X})$. En vertu de l'invariance unitaire de d_G (proposition 14.6-4), elle est égale à

$$d_G(\text{Im } Z_k, \mathcal{X}) = d_G(\text{Im } Z_k, \text{Im } X) = d_G\left(\text{Im } Q_k, \text{Im} \begin{pmatrix} I_p \\ 0 \end{pmatrix}\right).$$

Les projections orthogonales sur Im Q_k et sur $\left(\text{Im} \begin{pmatrix} I_p \\ 0 \end{pmatrix}\right)^{\perp}$ sont données par les matrices

$$\Pi_{\text{Im } Q_k} = Q_k Q_k^* \text{ et } \Pi_{\left(\text{Im} \begin{pmatrix} I_p \\ 0 \end{pmatrix}\right)^{\perp}} = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix}$$

de sorte que, par la proposition 14.6-2,

$$d_G(\text{Im } Z_k, \mathcal{X}) = \left\| \Pi_{\left(\text{Im} \begin{pmatrix} I_p \\ 0 \end{pmatrix}\right)^{\perp}} \circ \Pi_{\text{Im } Q_k} \right\|_2 = \left\| \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix} Q_k Q_k^* \right\|_2$$

$$= \left\| \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix} Q_k Q_k^* Q_k Q_k^* \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix} \right\|_2^{1/2} = \|W_k W_k^*\|_2^{1/2} = \|W_k\|_2 ,$$

quantité que nous allons estimer. L'égalité $Q_k R_k = D Q_{k-1}$ donne, par récurrence,

$$Q_k R_k \dots R_1 = D^k Q_0$$

c'est-à-dire

$$\begin{pmatrix} V_k \\ W_k \end{pmatrix} R_k \dots R_1 = \begin{pmatrix} D_1^k V_0 \\ D_2^k W_0 \end{pmatrix}$$

de sorte que, puisque V_k et D_1 sont inversibles,

$$W_k = D_2^k W_0 V_0^{-1} D_1^{-k} V_k .$$

Notons que $\|V_k\|_2 \leqslant 1$ parce que Q_k est une matrice de Stiefel. Ainsi

$$\|W_k\|_2 \leqslant \|W_0 V_0^{-1}\|_2 \left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k$$

et notre théorème est démontré.

Remarque 14.3.

1. L'hypothèse $(\operatorname{Im} Z) \cap \mathcal{Y} = \{0\}$ du théorème 14.7 est satisfaite de façon générique suivant le sens donné à ce terme dans la remarque 14.2.
2. Le calcul précédent montre que, dans le cas hermitien,

$$d_G(\operatorname{Im} Z_k, \mathcal{X}) \leqslant C \left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k$$

pour une constante $C > 0$ convenable. La convergence de la suite $(\operatorname{Im} Z_k)$ est donc linéaire et la vitesse de convergence donnée par le quotient $|\lambda_{p+1}/\lambda_p| < 1$.

3. L'hypothèse A *diagonalisable* n'est pas vraiment nécessaire. Elle nous permet ici d'éviter de parler de *paire complémentaire de sous-espaces invariants* (les sous-espaces $\mathcal{X}$ et $\mathcal{Y}$ du théorème).
4. La démonstration du cas général (A non diagonalisable) utilise le même argument que dans le cas hermitien mais est techniquement beaucoup plus compliquée. Dans ce cas aussi la vitesse de convergence est linéaire et donnée asymptotiquement par le quotient $|\lambda_{p+1}/\lambda_p| < 1$.
5. Lorsque les valeurs propres de A ont des modules distincts, on peut montrer que l'itération de sous-espaces converge vers un sous-espace invariant quel que soit le sous-espace initial choisi.

6. Cet algorithme utilise à chacune de ses étapes

   - Un produit de deux matrices $n \times n$ et $n \times p$ (calcul de AZ_k). Le coût d'un tel produit est de $2n^2p$ opérations arithmétiques si A est une matrice pleine.

   - La décomposition QR d'une matrice $n \times p$. Le coût de cette décomposition est $\leqslant 2np^2$ opérations arithmétiques par les méthodes de Gram-Schmidt ou de Householder (voir le paragraphe 8.5.3).

7. L'itération de sous-espaces consiste en la suite $(A^k\mathcal{Z})$ qui est représentée par la suite de matrices $(\mathcal{Q}(A^kZ))$ où $\mathcal{Z} = \text{Im } Z$. La décomposition QR sert ici à normaliser les matrices A^kZ qui sans cela auraient des entrées infiniment grandes ou petites (ou les deux à la fois). Il n'est pas utile d'effectuer cette normalisation à chaque étape : une fois de temps en temps suffit !

8. Une autre stratégie utilise non pas la décomposition QR pour normaliser la suite $(A^k\mathcal{Z})$ mais la décomposition LU, c'est la méthode de Treppen. L'algorithme correspondant est le suivant (on note $\mathcal{L}(Z)\mathcal{U}(Z)$ la décomposition LU de la matrice Z) :

Méthode de Treppen

Entrée : $Z \in \mathbb{C}^{n\times p}$, de rang p
$Z_0 = \mathcal{L}(Z)$,
pour $k = 1 : \ldots$
$\quad Z_k = \mathcal{L}(AZ_{k-1})$
fin

14.3 LA MÉTHODE QR

Cette méthode permet le calcul de toutes les valeurs propres d'une matrice. Soit $A \in \mathbb{C}^{n\times n}$ et soit $U \in \mathbb{U}_n$ une matrice unitaire. L'itération QR est donnée par l'algorithme suivant :

Itération QR

Entrée : $A \in \mathbb{C}^{n\times n}$, $U \in \mathbb{U}_n$
$A_0 = U^*AU = Q_0R_0$,
pour $k = 0 : \ldots$
$\quad A_{k+1} = R_kQ_k = Q_{k+1}R_{k+1}$
fin

On constate que, en général, la suite A_k « converge » vers une matrice triangulaire supérieure dont la diagonale contient les valeurs propres de A rangés par modules décroissants. Le mot *converge* a été mis entre guillemets en attendant de lui donner sa signification précise.

La matrice unitaire U qui sert à initialiser cet algorithme a pour seul but de le « rendre générique ». On avait pris, de la même manière, dans la méthode de la puissance un vecteur initial « au hasard ».

L'algorithme QR est à la base de plusieurs méthodes de calcul des valeurs propres, nous allons le décrypter dans le théorème qui suit.

Théorème 14.8 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice dont les valeurs propres sont de modules distincts :*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|\,.$$

Soit $A = QRQ^*$ une décomposition de Schur de A (théorème 1.12) telle que $r_{ii} = \lambda_i$ pour tout i et soit $U \in \mathbb{U}_n$ une matrice unitaire telle que

$$\operatorname{Im} U(1:n, 1:i) \cap \operatorname{Im} Q(1:n, i+1:n) = \{0\}$$

pour tout i (pour la notation $U(1:n, 1:i)$ voir le premier paragraphe du chapitre de rappels).

Sous ces hypothèses :

1. *La suite (A_k) ci-dessus est unitairement semblable à A,*
2. *Il existe des matrices $T_k \in \mathbb{U}_n$ diagonales et $B_k \in \mathbb{C}^{n\times n}$ telles que*

$$A_k = T_kB_kT_k^* \text{ et } B_k \to R.$$

Remarque 14.4.

1. L'hypothèse faite sur U est satisfaite de façon générique suivant le sens donné à ce terme dans la remarque 14.2.
2. Lorsque les valeurs propres sont de modules distincts, pour tout $U \in \mathbb{U}_n$, la suite A_k converge (au sens décrit dans le théorème 14.8) vers une matrice

triangulaire supérieure unitairement semblable à A. Mais les élément diagonaux de cette matrice ne sont pas nécessairement rangés par modules décroissants.

3. Si l'on note $a_{k,ij}$ les entrées de A_k on a, lorsque $k \to \infty$,

$$\begin{array}{lll} a_{k,ij} \to 0 & \text{si} & i > j, \\ a_{k,ij} \to \lambda_i & \text{si} & i = j, \\ |a_{k,ij}| \to |r_{ij}| & \text{si} & i < j. \end{array}$$

Il n'y a donc pas nécessairement convergence de la partie triangulaire supérieure de A_k : les modules de ces entrées convergent mais pas nécessairement leurs arguments.

4. La complexité d'une étape de calcul est celle d'une décomposition QR c'est-à-dire, par exemple, $4n^3/3$ opérations arithmétiques par la méthode de Householder.

Démonstration. (Théorème 14.8) Cette démonstration difficile est partagée en les cinq points suivants.

1. Les matrices A_k sont unitairement semblables à A : $A_0 = U^* A U$ et $A_{k+1} = P_k^* A P_k$ avec $P_k = U Q_0 Q_1 \dots Q_k$. En effet $A_{k+1} = R_k Q_k =$

$$Q_k^* Q_k R_k Q_k = Q_k^* A_k Q_k = \dots = Q_k^* \dots Q_0^* A_0 Q_0 \dots Q_k = P_k^* A P_k.$$

2. On a $P_{k+1} = Q(A P_k)$. En effet

$$A P_k = (U A_0 U^*)(U Q_0 Q_1 \dots Q_k) = U Q_0 R_0 Q_0 Q_1 \dots Q_k =$$

$$U Q_0 Q_1 R_1 \dots Q_k = U Q_0 Q_1 \dots Q_{k+1} R_{k+1} = P_{k+1} R_{k+1}$$

de sorte que $Q(A P_k) = P_{k+1}$ par unicité dans la décomposition QR.

3. L'égalité précédente implique

$$P_{k+1}(1 : n, 1 : i) = Q(A P_k(1 : n, 1 : i)).$$

Autrement dit, $P_k(1 : n, 1 : i)$ est la suite décrite au paragraphe précédent d'itération de sous-espaces associée à A et avec $P_0(1 : n, 1 : i) = U(1 : n, 1 : i)$. De l'hypothèse faite et du théorème 14.7 nous déduisons que

$$\lim_{k \to \infty} d_G(\text{Im } P_k(1 : n, 1 : i), \text{Im } Q(1 : n, 1 : i)) = 0 \text{ pour tout } i = 1 \dots n.$$

4. Un point un peu délicat : les limites précédentes et le fait que les matrices P_k et Q sont unitaires impliquent l'existence d'une suite de matrices T_k unitaires et diagonales telles que

$$\lim_{k \to \infty} P_k T_k = Q.$$

Notons $P_{k,i}$ (resp. Q_i) la $i-$ème colonne de P_k (resp. de Q). Il revient au même de montrer qu'il existe des suites de scalaires $(\theta_{k,i})_k$ telles que $|\theta_{k,i}| = 1$ et $\lim_{k\to\infty} \theta_{k,i} P_{k,i} = Q_i$ pour tout $i = 1 \ldots n$. On prend alors $T_k = \mathrm{diag}(\theta_{k,i})$. Procédons par récurrence sur i.

Pour $i = 1$, partons de la limite suivante

$$\lim_{k\to\infty} d_G(\mathrm{Im}\ P_k(1:n, 1:1), \mathrm{Im}\ Q(1:n, 1:1)) = 0.$$

Comme

$$d_G(\mathrm{Im}\ P_k(1:n, 1:1), \mathrm{Im}\ Q(1:n, 1:1)) = d_P(P_{k,1}, Q_1)$$

(voir les définitions 14.1 et 14.5), par la proposition 14.2 on a

$$\lim_{k\to\infty} |\langle P_{k,1}, Q_1\rangle| = 1.$$

Il existe donc une suite $(\theta_{k,1})_k$ de scalaires de module 1 tels que

$$\lim_{k\to\infty} \langle \theta_{k,1} P_{k,1}, Q_1\rangle = 1$$

(prendre $\theta_{k,1} = \exp\left(-i \arg \langle P_{k,1}, Q_1\rangle\right)$). On en déduit que

$$\lim_{k\to\infty} \|\theta_{k,1} P_{k,1} - Q_1\|^2 = \|\theta_{k,1} P_{k,1}\|^2 + \|Q_1\|^2 - 2\Re\langle \theta_{k,1} P_{k,1}, Q_1\rangle =$$

$$2 - 2\Re\langle \theta_{k,1} P_{k,1}, Q_1\rangle \to 0$$

et ainsi

$$\lim_{k\to\infty} \theta_{k,1} P_{k,1} = Q_1.$$

Supposons maintenant avoir prouvé qu'il existe des suites $(\theta_{k,j})_k$, $1 \leqslant j \leqslant i$, de scalaires de module 1, telles que

$$\lim_{k\to\infty} \theta_{k,j} P_{k,j} = Q_j, \quad 1 \leqslant j \leqslant i.$$

Nous allons utiliser le fait que

$$\lim_{k\to\infty} d_G(\mathrm{Im}\ P_k(1:n, 1:i+1), \mathrm{Im}\ Q(1:n, 1:i+1)) = 0.$$

Cela signifie que

$$\max_{\alpha_j} \min_{\beta_j} \frac{\left\| \sum_{j=1}^{i+1} \alpha_j P_{k,j} - \sum_{j=1}^{i+1} \beta_j Q_j \right\|_2}{\left\| \sum_{j=1}^{i+1} \alpha_j P_{k,j} \right\|_2} \to 0$$

ce qui donne

$$\min_{\beta_j} \left\| P_{k,i+1} - \sum_{j=1}^{i+1} \beta_j Q_j \right\|_2 \to 0.$$

Ce minimum est atteint en la projection orthogonale de $P_{k,i+1}$ sur l'espace Im $Q(1:n, 1:i+1)$ c'est-à-dire pour

$$\beta_j = \langle P_{k,i+1}, Q_j \rangle .$$

On a donc prouvé que

$$\lim_{k\to\infty} P_{k,i+1} - \sum_{j=1}^{i+1} \langle P_{k,i+1}, Q_j \rangle Q_j = 0.$$

Multiplions scalairement cette identité par $\theta_{k,l} P_{k,l}$ avec $1 \leqslant l \leqslant i$. On a

$$\lim_{k\to\infty} \langle P_{k,i+1}, \theta_{k,l} P_{k,l} \rangle - \sum_{j=1}^{i+1} \langle P_{k,i+1}, Q_j \rangle \langle Q_j, \theta_{k,l} P_{k,l} \rangle = 0.$$

Comme $\langle P_{k,i+1}, \theta_{k,l} P_{k,l} \rangle = 0$ et que $\lim_{k\to\infty} \langle Q_j, \theta_{k,l} P_{k,l} \rangle = \langle Q_j, Q_l \rangle$, on obtient

$$\lim_{k\to\infty} \langle P_{k,i+1}, Q_l \rangle = 0$$

pour tout $l \leqslant i$ de sorte que

$$\lim_{k\to\infty} P_{k,i+1} - \langle P_{k,i+1}, Q_{i+1} \rangle Q_{i+1} = 0$$

ce qui est équivalent à

$$\lim_{k\to\infty} d_P \left(P_{k,i+1}, Q_{i+1} \right) = 0$$

par la proposition 14.2. On prouve l'existence d'une suite $(\theta_{k,i+1})_k$ de scalaires de module 1 telle que

$$\lim_{k\to\infty} \theta_{k,i+1} P_{k,i+1} = Q_{i+1}$$

par l'argument développé pour $i = 1$.

5. Conclusion :

$$A_{k+1} = P_k^* A P_k = T_k (P_k T_k)^* Q R Q^* (P_k T_k) T_k^* = T_k B_k T_k^*$$

avec

$$B_k = (P_k T_k)^* Q R Q^* (P_k T_k) \to R$$

lorsque $k \to \infty$.

Remarque 14.5.

1. La démonstration précédente montre que la suite des sous-espaces $A^k(\text{Im } P_0(1:n, 1:i)) = \text{Im } P_k(1:n, 1:i)$ converge vers le sous-espace Im $Q(1:n, 1:i)$ pour tout i, $1 \leqslant i \leqslant n$. Une suite de sous-espaces emboîtés $\mathcal{F} = (\mathcal{F}_1 \subset \mathcal{F}_2 \ldots \subset \mathcal{F}_n)$ avec $\dim \mathcal{F}_i = i$ est appelé *drapeau* de $\mathbb{C}^n$. L'itération QR peut être interprétée comme la méthode des approximations successives associée à l'action de A sur la variété des drapeaux.
2. Un drapeau peut être décrit à l'aide d'une matrice $F \in \mathbb{GL}_n$ en définissant $\mathcal{F}_i$ comme le sous-espace engendré par les i premières colonnes de F ; dans notre cas, $F = P_k$ et $\mathcal{F}_i = \text{Im } P_k(1 : n. 1 : i)$. La quatrième partie de la démonstration précédente donne une interprétation en termes matriciels de la convergence d'une suite de drapeaux.
3. En vertu du théorème 14.7 la suite Im $P_k(1:n, 1:i)$ converge vers Im $Q(1:n, 1:i)$. Cette convergence est linéaire et sa vitesse est donnée par le quotient $|\lambda_{i+1}/\lambda_i|$. On en déduit que la vitesse de convergence de l'algorithme QR est linéaire en $\max_{1 \leqslant i \leqslant n-1} |\lambda_{i+1}/\lambda_i|$. Notons que ce maximum est < 1 puisque les valeurs propres sont rangées par module décroissant.

14.4 LE CAS DES MATRICES RÉELLES

Supposons que A soit réelle et que U soit une matrice orthogonale. L'itération

$$
\begin{aligned}
&A_0 = U^* A U = Q_0 R_0, \\
&\text{pour } k = 0 : \ldots \\
&\qquad A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1} \\
&\text{fin}
\end{aligned}
$$

où $A_k = Q_k R_k$ est une décomposition QR avec Q_k orthogonale et R_k triangulaire supérieure réelle ne saurait « révéler » une décomposition de Schur de A : $A = QRQ^*$ avec Q unitaire et R triangulaire supérieure parce que cette décomposition fait intervenir des matrices complexes dès que A possède des valeurs propres complexes alors que l'algorithme ci-dessus ne calcule que des matrices réelles.

Les propriétés de la méthode de QR réelle sont résumées dans le théorème suivant que nous ne démontrons pas. Il fait appel à la *décomposition de Schur réelle* d'une matrice réelle :

Théorème 14.9 *Soit $A \in \mathbb{R}^{n \times n}$ qui possède des valeurs propres réelles $\lambda_k \in \mathbb{R}$, $1 \leqslant k \leqslant p$, et complexes conjuguées $\alpha_j \pm i\beta_j \in \mathbb{C}$, $1 \leqslant j \leqslant q$, de modules $|\lambda_k|$ et $\sqrt{\alpha_j^2 + \beta_j^2}$ distincts. Pour toute matrice orthogonale $U \in \mathbb{O}_n$, la suite (A_k) définie par*

$A_0 = U^T A U = Q_0 R_0$ et $A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1}$ a la propriété suivante : il existe une décomposition de Schur réelle de A (théorème 1.13)

$$Q^T A Q = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ & R_{22} & \dots & R_{2p} \\ & & \ddots & \vdots \\ & & & R_{pp} \end{pmatrix},$$

telle que les blocs correspondants $A_{k,ij}$ dans A_k vérifient $\lim_{k\to\infty} A_{k,ij} = 0$ si $i > j$ et que les valeurs propres du bloc diagonal $A_{k,ii}$ convergent vers celles de R_{ii}.

14.5 L'UTILISATION DE LA FORME HESSENBERG

Nous avons déjà rencontré les matrices de Hessenberg au paragraphe 8.6 où nous avons montré que toute matrice $A \in \mathbb{C}^{n\times n}$ est unitairement semblable à une matrice de Hessenberg. L'intérêt de ces matrices en ce qui concerne l'itération QR est que :

- Le calcul de la décomposition QR d'une matrice de Hessenberg coûte $O(n^2)$ opérations arithmétiques au lieu de $O(n^3)$ pour une matrice pleine,
- Si H est une matrice de Hessenberg et si $H = QR$ alors RQ est aussi de Hessenberg.

Ceci prouve que la forme Hessenberg est conservée tout au long de l'itération QR d'où un gain en complexité et en erreurs d'arrondis.

Démonstration. Pour obtenir la décomposition QR d'une matrice de Hessenberg H il suffit de la multiplier à gauche par $n-1$ matrices de rotation de Givens (voir le paragraphe 8.4 pour le cas réel et l'exercice 8.8 pour le cas complexe) au lieu de $n(n-1)/2$ pour une matrice pleine (théorème 8.5). Ceci justifie la complexité en $O(n^2)$ opérations arithmétiques dans le cas Hessenberg. Par exemple, pour une matrice 4×4, on a le schéma suivant :

$$\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix} \xrightarrow{G(1,2)} \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix} \xrightarrow{G(2,3)} \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}$$

$$\xrightarrow{G(3,4)} \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \end{pmatrix}.$$

On obtient ainsi la décomposition QR de H : $R = G_{n-1n} \dots G_{23} G_{12} H$ et $Q = G_{12}^* G_{23} \dots G_{n-1n}^*$. La matrice itérée est

$$RQ = R G_{12}^* G_{23}^* \dots G_{n-1n}^*$$

dont il est facile de voir qu'elle est de Hessenberg.

14.6 LA STRATÉGIE DU DÉCALAGE

14.6.1 Principe général

Partons d'une matrice A mise sous forme Hessenberg $A = UHU^*$. L'itération *QR avec décalage* (les franglophones disent *QR avec shift*) est donnée par :

Itération QR avec décalage

Entrée : $H_0 = H \in \mathbb{C}^{n \times n}$ de Hessenberg
pour $k = 0 : \dots$
$\quad H_k - \mu_k I_n = Q_k R_k$
$\quad H_{k+1} = R_k Q_k + \mu_k I_n$
fin

où les μ_k sont des scalaires donnés ; l'itération QR décrite précédemment correspond à $\mu_k = 0$. Les matrices ainsi définies sont unitairement semblables à H (et donc aussi à A) puisque :

$$H_{k+1} = R_k Q_k + \mu_k I_n = Q_k^* (Q_k R_k + \mu_k I_n) Q_k = Q_k^* H_k Q_k$$

de sorte que

$$H_{k+1} = P_k^* H_k P_k \text{ avec } P_k = Q_0 \dots Q_k.$$

Par un argument identique à celui développé au paragraphe 14.3 on obtient :

$$H P_k = P_{k+1} R_{k+1} + \mu_k P_k$$

ou encore

$$(H - \mu_k I_n) P_k = P_{k+1} R_{k+1}$$

et donc

$$P_{k+1} = \mathcal{Q}((H - \mu_k I_n) P_k)$$

(avec les notations du paragraphe 8.10). Si la suite des décalages μ_k est convergente :

$$\lim_{k \to \infty} \mu_k = \mu$$

et si l'on range les valeurs de H de sorte que

$$|\lambda_1 - \mu| > \ldots > |\lambda_n - \mu|$$

alors, sous des hypothèses convenables, la suite (H_k) converge suivant le sens donné au théorème 14.8. La convergence est linéaire en

$$\max_{1 \leqslant i \leqslant n-1} \frac{|\lambda_{i+1} - \mu|}{|\lambda_i - \mu|}.$$

On a donc intérêt à choisir μ de façon à rendre $\max_{1 \leqslant i \leqslant n-1} \frac{|\lambda_{i+1}-\mu|}{|\lambda_i-\mu|}$ aussi petit que possible. Mais comment faire ?

14.6.2 Décalage simple

Cette stratégie consiste à prendre $\mu_k = h_{k,nn}$ (l'entrée nn de la matrice H_k). Elle est fondée sur l'argument suivant : supposons que H soit réelle avec

$$H = \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1\,n-1} & h_{1n} \\ h_{21} & h_{22} & \ldots & h_{2\,n-1} & h_{2n} \\ & \ddots & \ddots & \vdots & \vdots \\ 0 & & \ddots & h_{n-1\,n-1} & h_{n-1\,n} \\ 0 & 0 & & \varepsilon & h_{nn} \end{pmatrix}.$$

Le calcul de la décomposition QR de $H - h_{nn} I_n$ via des rotations de Givens va conduire tout d'abord à une matrice du type

$$G_{n-2\,n-1} \ldots G_{12}(H - h_{nn} I_n) = \begin{pmatrix} \times & \times & \ldots & \times & \times \\ 0 & \times & & \times & \times \\ & \ddots & \ddots & \vdots & \vdots \\ 0 & & \ddots & a & b \\ 0 & 0 & & \varepsilon & 0 \end{pmatrix}.$$

La dernière rotation de Givens appliquée au produit précédent

$$G_{n-1\,n} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & \frac{a}{\sqrt{a^2+\varepsilon^2}} & \frac{\varepsilon}{\sqrt{a^2+\varepsilon^2}} \\ & & & \frac{-\varepsilon}{\sqrt{a^2+\varepsilon^2}} & \frac{a}{\sqrt{a^2+\varepsilon^2}} \end{pmatrix}$$

donne la matrice triangulaire supérieure

$$R = G_{n-1\,n} G_{n-2\,n-1} \dots G_{12}(H - h_{nn} I_n) = \begin{pmatrix} \times & & & & \\ & \ddots & & & \\ & & \times & & \\ & & & \sqrt{a^2+\varepsilon^2} & \frac{ab}{\sqrt{a^2+\varepsilon^2}} \\ & & & 0 & \frac{-b\varepsilon}{\sqrt{a^2+\varepsilon^2}} \end{pmatrix}$$

et la matrice orthogonale

$$Q = G_{12}^T \dots G_{n-2\,n-1}^T G_{n-1\,n}^T = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ 0 & \dots & 0 & \frac{\varepsilon}{\sqrt{a^2+\varepsilon^2}} & \frac{a}{\sqrt{a^2+\varepsilon^2}} \end{pmatrix}.$$

En conséquence, la matrice itérée est du type

$$RQ + h_{nn} I_n = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ 0 & \dots & 0 & \frac{-b\varepsilon^2}{a^2+\varepsilon^2} & * \end{pmatrix}.$$

Cette analyse prouve que, si ε est petit devant a, le terme $h_{k,n\,n-1}$ de la matrice itérée est d'ordre ε^2 d'où une convergence quadratique de ce terme vers 0.

14.6.3 Déflation

L'analyse précédente montre que le coefficient $h_{k,n\,n-1}$ converge vers 0 très rapidement. Lorsque, à l'itération k, le coefficient $h_{k,n\,n-1}$ est jugé suffisamment petit, on supprime la n-ième ligne et la n-ième colonne de la matrice H_k et on continue la méthode QR avec la matrice de Hessenberg réduite $H_k(1 : n-1, 1 : n-1)$ et le décalage $h_{k,n-1\,n-1}$. Ainsi de suite jusqu'à obtenir une matrice réduite 1×1. On appelle *déflation* ce processus de réduction de la dimension du problème.

La figure suivante montre la norme de la sous-diagonale de la matrice de Hessenberg H_k en fonction de k dans le cas des algorithmes QR simple et QR avec décalage simple et déflation. On a choisi une matrice de Hessenberg H de taille 100×100 à coefficients aléatoires.

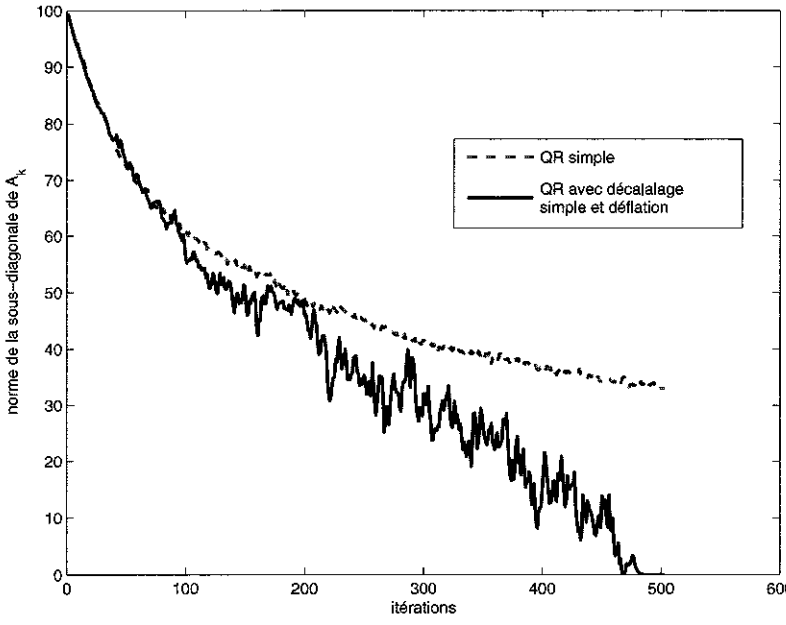


Figure 14.2 Norme de la sous-diagonale de H_k en fonction de k.

14.7 REMARQUES FINALES

La méthode QR est à la base de nombreux algorithmes de calcul des valeurs propres et notre approche ne saurait être exhaustive. Citons, parmi les points qui n'ont pas été abordés :

1. La stratégie du *double décalage* (*double shift* en franglais). Elle est donnée par l'algorithme

Itération QR avec double décalage

Entrée : $H_0 = H \in \mathbb{C}^{n\times n}$ de Hessenberg
pour $k = 0 : \ldots$
$$H_k - \mu_{1,k} I_n = Q_{1,k} R_{1,k}$$
$$H_{1,k} = \mu_{1,k} I_n + R_{1,k} Q_{1,k}$$
$$H_{1,k} - \mu_{2,k} I_n = Q_{2,k} R_{2,k}$$
$$H_{k+1} = R_{2,k} Q_{2,k} + \mu_{2,k} I_n$$
fin

où $\mu_{1,k}$ et $\mu_{2,k}$ sont les valeurs propres de la matrice 2×2

$$\begin{pmatrix} h_{k,n-1\,n-1} & h_{k,n-1\,n} \\ h_{k,n\,n-1} & h_{k,nn} \end{pmatrix}.$$

2. L'itération LU. Elle fonctionne comme l'itération QR mais utilise la décomposition LU (définition 6.5) au lieu de la décomposition QR. Elle est donnée par l'algorithme

Itération LU

Entrée : $A \in \mathbb{C}^{n\times n}$
pour $k = 0 : \ldots$
$\quad A_k = L_k U_k$
$\quad A_{k+1} = U_k L_k$
fin

3. L'itération de Cholesky. On part d'une matrice $A \in \mathbb{C}^{n\times n}$ définie positive dont la décomposition de Cholesky (théorème 7.10) est notée $A = LL^*$. L'itération de Cholesky est donnée par l'algorithme

Itération de Cholesky

Entrée : $A \in \mathbb{C}^{n\times n}$ définie positive
pour $k = 0 : \ldots$
$\quad A_k = L_k L_k^*$
$\quad A_{k+1} = L_k^* L_k$
fin

où $A_k = L_k L_k^*$ est la décomposition de Cholesky de A_k.

4. L'itération QR pour des matrices hermitiennes. Pour une telle matrice, la forme Hessenberg est donnée par une matrice tridiagonale hermitienne (théorème 8.16) et cette forme est conservée au cours de l'algorithme (voir l'exercice 14.4). Les stratégies de décalage utilisent des décalages réels puisque le spectre d'une matrice hermitienne est réel.

14.8 NOTES ET RÉFÉRENCES

La méthode de la puissance pour le calcul de la valeur propre dominante d'une matrice apparaît explicitement en 1913 dans une note de C. Müntz aux Comptes Rendus de l'Académie des Sciences [26]. Cette méthode connaît un regain d'intérêt aujourd'hui puisqu'elle est utilisée par Google pour ses recherches de pages-web.

La méthode QR est due à Kublanovskaya (1961) [21] et Francis (1961) [11] à la suite de travaux de Rutishauser (1955) [28] sur l'algorithme LR. Nombre d'auteurs ont apporté leur brique à l'édifice qui a conduit aux algorithmes actuels, nous renvoyons les lecteurs intéressés et courageux au « LAPACK User's Guide » [1] et au livre de Stewart [33]. Une version plus récente de l'algorithme LR (itération LU) est présentée dans [10] et dans un article de synthèse [27].

EXERCICES

Exercice 14.1

Etudier la méthode de la puissance lorsque A est la matrice 2×2 d'une rotation plane d'angle θ.

Exercice 14.2

On suppose que les valeurs propres de $A \in \mathbb{C}^{n\times n}$ vérifient :

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0.$$

Montrer que pour tout $z \in \mathbb{C}^n$, $z \neq 0$, la suite définie par

$$z_0 = \frac{z}{\|z\|_2}, \quad z_k = \frac{Az_{k-1}}{\|Az_{k-1}\|_2}$$

vérifie :

1. Elle est bien définie,
2. Il existe un vecteur propre x de A tel que $\lim_{k\to\infty} d_P(z_k, x) = 0$.

Exercice 14.3

Le but de cet exercice est d'établir les propriétés principales de la distance d_G (définition 14.5) sur la grassmannienne $\mathbb{G}_{np}$. Nous généralisons ici la définition de cette « distance » en posant

$$d(\mathcal{V}, \mathcal{W}) = \sup_{v\in\mathcal{V}, v\neq 0} \inf_{w\in\mathcal{W}} \frac{\|v - w\|_2}{\|v\|_2}$$

où $\mathcal{V}$ et $\mathcal{W}$ sont des sous-espaces vectoriels de $\mathbb{C}^n$ qui n'ont pas nécessairement la même dimension. Notons $\Pi_{\mathcal{V}}$ la projection orthogonale sur $\mathcal{V}$. Montrer que :

1. $d(\mathcal{V}, \mathcal{W}) = \max_{v\in\mathcal{V}, v\neq 0} \min_{w\in\mathcal{W}} \frac{\|v - w\|_2}{\|v\|_2}$,
2. $d(\mathcal{V}, \mathcal{W}) = \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}}\|_2$,
3. $d(\mathcal{V}, \mathcal{W}) = d(\mathcal{W}^\perp, \mathcal{V}^\perp)$,
4. $d(\mathcal{V}, \mathcal{W}) = d(\mathcal{V} \cap (\mathcal{V} \cap \mathcal{W})^\perp, \mathcal{W} \cap (\mathcal{V} \cap \mathcal{W})^\perp)$,
5. $0 \leqslant d(\mathcal{V}, \mathcal{W}) \leqslant 1$,

6. $d(\mathcal{V}, \mathcal{W}) = 0$ si et seulement si $\mathcal{V} \subset \mathcal{W}$,
7. $d(\mathcal{V}, \mathcal{W}) < 1$ si et seulement si $\mathcal{V} \cap \mathcal{W}^{\perp} = \{0\}$,
8. $d(\mathcal{V}_1, \mathcal{V}_3) \leqslant d(\mathcal{V}_1, \mathcal{V}_2) + d(\mathcal{V}_2, \mathcal{V}_3)$,
9. Si $\mathcal{V}_1 \subset \mathcal{V}_2$ alors $d(\mathcal{V}_1, \mathcal{W}) \leqslant d(\mathcal{V}_2, \mathcal{W})$ et si $\mathcal{W}_1 \subset \mathcal{W}_2$ alors $d(\mathcal{V}, \mathcal{W}_2) \leqslant d(\mathcal{V}, \mathcal{W}_1)$,
10. $d(\mathcal{V}, \mathcal{W}_1 + \mathcal{W}_2) \leqslant \min(d(\mathcal{V}, \mathcal{W}_1), d(\mathcal{V}, \mathcal{W}_2))$,
11. Si $\mathcal{V}_1$ et $\mathcal{V}_2$ sont orthogonaux alors $d(\mathcal{V}_1 \oplus \mathcal{V}_2, \mathcal{W}) \leqslant d(\mathcal{V}_1, \mathcal{W}) + d(\mathcal{V}_2, \mathcal{W})$ et $d(\mathcal{V}_1 \oplus \mathcal{V}_2, \mathcal{W}) \leqslant \sqrt{2}\max(d(\mathcal{V}_1, \mathcal{W}), d(\mathcal{V}_2, \mathcal{W}))$,
12. $d(Q(\mathcal{U}), Q(\mathcal{V})) = d(\mathcal{U}, \mathcal{V})$ pour toute matrice unitaire Q dans $\mathbb{C}^n$,
13. Si $\dim \mathcal{V} = \dim \mathcal{W}$ alors $d(\mathcal{V}, \mathcal{W}) = d(\mathcal{W}, \mathcal{V})$ (on utilisera une transformation unitaire Q dans $\mathbb{C}^n$ qui vérifie $Q^2 = id_n$ et $Q\mathcal{V} = \mathcal{W}$),
14. $d(\mathcal{V}, \mathcal{W})$ est une distance sur l'ensemble $\mathbb{G}_{np}$ des sous-espaces vectoriels de dimension p de $\mathbb{C}^n$.

Exercice 14.4

Soient $T \in \mathbb{C}^{n\times n}$ tridiagonale hermitienne, $\mu \in \mathbb{R}$ et $T - \mu I_n = QR$ la décomposition QR de $T - \mu I_n$. Montrer que $T_+ = RQ + \mu I_n$ est tridiagonale hermitienne et unitairement semblable à T.

Exercice 14.5

Soient A une matrice 2×2, μ une valeur propre de A et $A - \mu I_2 = QR$ la décomposition QR de $A - \mu I_2$. Calculer explicitement Q et R en fonction des entrées de A et de μ et montrer que

$$RQ + \mu I_2 = \begin{pmatrix} \alpha & \beta \\ 0 & \mu \end{pmatrix}$$

pour des scalaires α et β que l'on précisera.

Chapitre **15**

Méthodes de projection pour le problème des valeurs propres

Les méthodes de projection sont également appliquées au calcul des valeurs et vecteurs propres de matrices. Elles sont généralement utilisées pour des matrices creuses de grande taille. On se limite alors à calculer certaines parties du spectre de la matrice et les vecteurs propres associés. Les sous-espaces d'approximation des vecteurs propres que l'on utilise sont des sous-espaces de Krylov.

15.1 PRINCIPE D'UNE MÉTHODE DE PROJECTION POUR LE PROBLÈME DES VALEURS PROPRES

Soit $\mathcal{K}$ un sous-espace vectoriel de $\mathbb{C}^n$ de dimension k. On veut déterminer un vecteur $x \in \mathcal{K}, x \neq 0$ et un scalaire $\lambda \in \mathbb{C}$ vérifiant « au mieux » l'égalité

$$Ax = \lambda x.$$

On impose pour cela des contraintes d'orthogonalité (conditions de Petrov-Galerkin) :

$$Ax - \lambda x \perp \mathcal{L} \tag{15.1}$$

où $\mathcal{L}$ est un sous-espace de dimension k. Ce même principe a déjà été utilisé pour les systèmes linéaires (paragraphe 11.1). Nous allons considérer des méthodes de projection orthogonale pour lesquelles $\mathcal{L} = \mathcal{K}$.

La contrainte d'orthogonalité (15.1) définit un problème de valeurs propres sur une matrice « réduite » de taille $k \times k$. Donnons une forme plus précise aux équations

(15.1). Considérons une base orthonormée $(q_1, \ldots, q_k)$ de $\mathcal{K}$ et notons $Q \in \mathbb{S}t_{nk}$ la matrice ayant les q_i pour colonnes. Pour tout $x \in \mathcal{K}$, nous avons $x = Qy$ pour un unique $y \in \mathbb{C}^k$.

Théorème 15.1 *Pour tout $\lambda \in \mathbb{C}$, $y \in \mathbb{C}^k$, $y \neq 0$ et $x \in \mathbb{C}^n$ tels que $x = Qy$, on a l'équivalence*

$$Q^*AQ\,y = \lambda y \iff Ax - \lambda x \perp \mathcal{K}. \tag{15.2}$$

Démonstration. Il suffit d'observer que $y = Q^*Qy = Q^*x$ puisque $Q^*Q = I_k$. Nous avons $Q^*AQ\,y = \lambda y$ si et seulement si $Q^*Ax - \lambda Q^*x = 0$ c'est-à-dire si et seulement si $Ax - \lambda x \in \text{Ker}\, Q^* = (\text{Im}\, Q)^\perp$.

Le scalaire λ et le vecteur x sont appelés *valeur et vecteur de Ritz*. La procédure consistant à calculer des approximations des valeurs propres de A à partir de la matrice Q^*AQ (équation (15.2)) est appelée *procédure de Rayleigh-Ritz*. Lorsque la matrice A est hermitienne, Q^*AQ est appelée *quotient de Rayleigh* par extension du quotient de Rayleigh défini par $(q^*Aq)/(q^*q)$ (voir l'exercice 12.2 sur le théorème de Fisher).

Notons $P = QQ^*$ le projecteur orthogonal sur l'espace $\mathcal{K}$. Les valeurs et vecteurs de Ritz sont eux-mêmes des valeurs et vecteurs propres associés à la matrice PAP comme le montre la proposition suivante dont la démonstration est laissée au lecteur :

Proposition 15.2 *Soient $\lambda \in \mathbb{C}$, $x \in \mathcal{K}$ et $y \in \mathbb{C}^k$ tels que $x = Qy$. On a l'équivalence*

$$Q^*AQ\,y = \lambda y \iff PAPx = \lambda x.$$

Le résultat suivant montre que, lorsque $\mathcal{K}$ est un sous-espace invariant de A, la procédure de Rayleigh-Ritz donne la solution exacte du problème des valeurs propres.

Proposition 15.3 *Si $\mathcal{K}$ est un sous-espace invariant de A alors les valeurs et vecteurs de Ritz sont égaux aux valeurs et vecteurs propres de la restriction de A au sous-espace $\mathcal{K}$.*

Démonstration. En effet, l'égalité $PAPx = \lambda x$ avec $x \in \mathcal{K}$ est équivalente à $Ax = \lambda x$ puisque $Px = x$ et que $PA = A$.

Lorsque $\mathcal{K}$ est un sous-espace invariant de A, la matrice réduite Q^*AQ est l'unique matrice H telle que $AQ - QH = 0$. Dans le cas général, on a le résultat suivant :

Théorème 15.4 *La matrice $H = Q^*AQ$ est l'unique solution du problème*

$$\min_{H \in \mathbb{C}^{k \times k}} \|AQ - QH\|_F^2. \tag{15.3}$$

Démonstration. C'est une conséquence de la proposition 9.13.

15.2 MÉTHODE DE PROJECTION SUR DES SOUS-ESPACES DE KRYLOV

Prenons pour espace de projection un sous-espace de Krylov : $\mathcal{K} = \mathcal{K}_k(A, v)$. Nous allons étudier quelques propriétés liées à ce choix et en particulier le rôle du vecteur v.

Nous avons vu au paragraphe 11.2 que l'algorithme d'Arnoldi permet d'obtenir une base orthonormée de $\mathcal{K}_k(A, v)$. On suppose que l'algorithme est défini jusqu'à l'étape k, c'est-à-dire $h_{j+1\,j} \neq 0$ pour tout $j \leqslant k-1$. Les vecteurs-colonne de la matrice $Q_k = (q_1 \ldots q_k) \in \mathbb{S}t_{nk}$ forment une base orthonormée de $\mathcal{K}_k(A, v)$ et la matrice réduite $H_k = Q_k^* A Q_k$ est une matrice Hessenberg.

À l'étape k de l'algorithme d'Arnoldi nous avons

$$A Q_k = Q_k H_k + h_{k+1\,k}\, q_{k+1} e_k^T \tag{15.4}$$

(voir équation 8.4) et donc, en considérant la colonne k,

$$A q_k = Q_k h_k + h_{k+1\,k}\, q_{k+1},$$

où h_k est la k-ième colonne de H_k. Nous en déduisons que

$$\|A Q_k - Q_k H_k\|_2 = \|A q_k - Q_k h_k\|_2 = |h_{k+1\,k}|. \tag{15.5}$$

Le résultat suivant complète le théorème 15.4 lorsque l'espace d'approximation est un sous-espace de Krylov.

Proposition 15.5 *On a*

$$|h_{k+1\,k}| = \min_{H \in \mathbb{C}^{k\times k}} \|A Q_k - Q_k H\|_F^2 = \min_{h \in \mathbb{C}^k} \|A q_k - Q_k h\|_2^2 = \min_{p \in \tilde{\mathcal{P}}_k} \|p(A)v\|_2^2$$

où $\tilde{\mathcal{P}}_k$ est l'ensemble des polynômes de degré $\leqslant k$ de même coefficient dominant que le polynôme p_{k-1} tel que $q_k = p_{k-1}(A)v$.

Démonstration. Les équations (15.5) et $H_k = Q_k^* A Q_k$ montrent les deux premières égalités.

Démontrons la dernière égalité. Puisque $\operatorname{Im} Q_k = \mathcal{K}_k(A, v)$, nous avons, pour tout $h \in \mathbb{C}^k$, $Q_k h = p(A)v$ où $p \in \mathcal{P}_{k-1}$ ($\mathcal{P}_{k-1}$ est l'espace des polynômes de degré $\leqslant k-1$). Par ailleurs l'égalité $q_k = p_{k-1}(A)v$ où $p_{k-1} \in \mathcal{P}_{k-1}$ (voir la démonstration de la proposition 11.3) montre que

$Aq_k = Ap_{k-1}(A)v = (\alpha A^k + \tilde{p}(A))v$ en notant α le coefficient dominant du polynôme p_{k-1} et $\tilde{p} \in \mathcal{P}_{k-1}$. On a donc $Aq_k - Q_k h = (\alpha A^k + \tilde{p}(A) - p(A))v$. La minimisation étant réalisée sur l'ensemble des polynômes $p(A) \in \mathcal{P}_{k-1}$, on conclut que

$$\min_{h \in \mathbb{C}^k} \|Aq_k - Q_k h\|_2^2 = \min_{r \in \mathcal{P}_{k-1}} \|(\alpha A^k + r(A))v\|_2^2$$

et donc le résultat.

Les propriétés d'invariance du sous-espace de Krylov $\mathcal{K}_k(A, v)$ sont liées aux propriétés de v par le résultat suivant

Proposition 15.6 *Si v appartient à un sous-espace invariant $\mathcal{V}$ de dimension k alors $\mathcal{K}_k(A, v) = \mathcal{V}$.*

Démonstration.

Puisque v appartient au sous-espace invariant $\mathcal{V}$, alors $Av, A^2v, \ldots, A^{k-1}v$ appartiennent à $\mathcal{V}$. Donc $\mathcal{K}_k(A, v) \subset \mathcal{V}$ et puisque la dimension de $\mathcal{K}_k(A, v)$ est égale à k on déduit l'égalité des ensembles.

Cette proposition implique en particulier que si le vecteur v est une combinaison linéaire de k vecteurs propres de A, alors le sous-espace $\mathcal{K}_k(A, v)$ est invariant par A et les valeurs propres de la restriction de A à $\mathcal{K}_k(A, v)$ sont les valeurs propres associées aux vecteurs propres de la combinaison linéaire.

Il est donc intéressant de prendre un vecteur v « proche » d'une combinaison linéaire de vecteurs propres générant le sous-espace invariant que l'on souhaite obtenir. En réalité ces vecteurs propres ne sont pas connus puisqu'il s'agit justement de déterminer le sous-espace invariant qu'ils génèrent ! La sélection d'un « bon candidat » v se fait plutôt en « éliminant » de celui-ci les composantes correspondant aux parties du spectre que l'on ne souhaite pas approcher. Les méthodes dites de *redémarrage* (*restarting methods* en anglais) utilisent des techniques de filtrage pour annuler les composantes indésirables du vecteur v. Ces méthodes sont actuellement parmi les plus performantes pour calculer des sous-ensembles du spectre et les vecteurs propres correspondants de matrices de très grande dimension ($n \approx 10^6$).

Dans le cas général, si (λ, x) est un couple valeur-vecteur de Ritz, on a $H_k y = \lambda y$, avec $y \neq 0$ et $x = Q_k y$. Le résidu $Ax - \lambda x$ est donné par

$$Ax - \lambda x = (AQ_k - Q_k H_k)y = h_{k+1\,k}\, q_{k+1}\, e_k^T y.$$

Le calcul numérique des valeurs propres de la matrice réduite $H_k = Q_k^* A Q_k$ est obtenu grâce aux méthodes classiques telles que la méthode QR présentée au chapitre

14. L'algorithme de Lanczos est utilisé dans le cas particulier où A est hermitienne et la matrice réduite $T_k = Q_k^* A Q_k$ est alors tridiagonale hermitienne.

La figure 15.1 montre l'évolution des valeurs de Ritz d'une matrice A à coefficients aléatoires de dimension 8. On a utilisé l'algorithme d'Arnoldi pour le calcul de Q_k. Le vecteur v qui définit le sous-espace de Krylov a aussi été choisi aléatoirement.

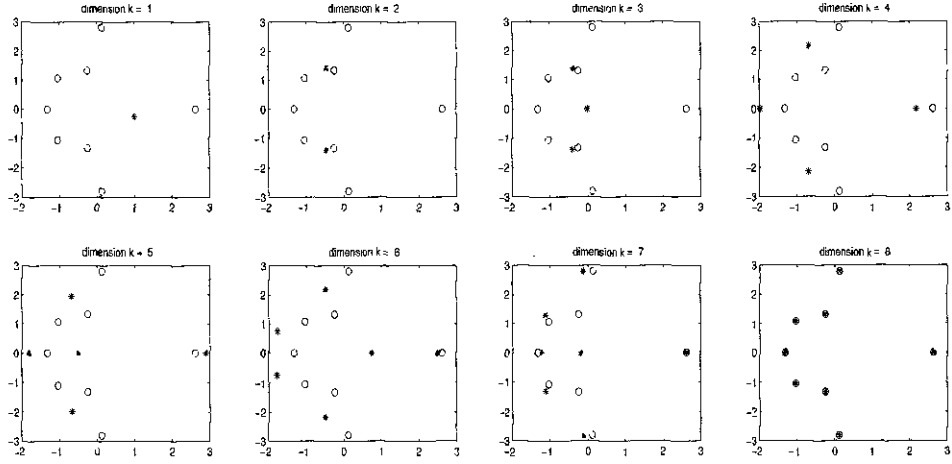


Figure 15.1 Convergence des valeurs de Ritz. Les valeurs propres de A sont notés un rond o et les valeurs de Ritz par une étoile *.

La figure 15.2 montre la convergence des 8 valeurs dominantes de Ritz vers les valeurs propres dominantes de la matrice de raideur K considérée au paragraphe 16.3. La matrice Q_k est obtenue grâce à l'algorithme de Lanczos puisque K est définie positive. L'abscisse représente la dimension k de l'espace de Krylov $\mathcal{K}_k(A, v)$. Le vecteur v est choisi de manière aléatoire et la dimension n de A est égale à 517.

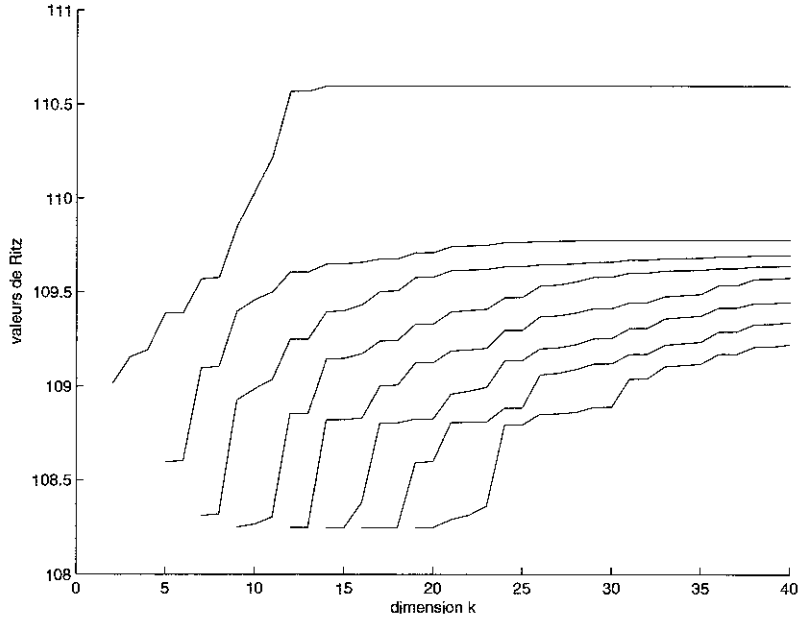


Figure 15.2 Convergence des 8 valeurs de Ritz dominantes de la matrice de raideur K ($n = 517$).

15.3 NOTES ET RÉFÉRENCES

Le physicien John William Strutt (1842-1919) plus connu sous son titre de Lord Rayleigh montra en 1899 comment calculer la fréquence fondamentale d'un système vibrant en minimisant la quantité $x^T Ax \,/\, x^T x$. Suivant la même idée Walther Ritz (1878-1909) proposa en 1909 une méthode générale pour résoudre un problème de minimisation d'une fonctionnelle.

L'analyse de la convergence de ces méthodes reste une question délicate et en partie encore mal comprise. Nous renvoyons le lecteur aux ouvrages spécialisés de Stewart [33] et Saad [30].

On trouve dans [3] une présentation des differentes méthodes qui se rattachent à cette famille ainsi que leurs aspects algorithmiques.

La bibliothèque ARPACK [24] fourni des programmes de calcul qui mettent en œuvre les méthodes de redémarrage.

Enfin on peut citer la méthode de Jacobi-Davidson qui fait partie des méthodes de projection bien qu'elle n'utilise pas pour espace de projection des espaces de Krylov (voir [33]). Cette méthode est aussi conçue pour calculer une partie du spectre d'une matrice hermitienne ou non-hermitienne de grande taille. L'espace de projection est généré par des corrections orthogonales de certains vecteurs de Ritz.

EXERCICES

Exercice 15.1

Soit $Q_k \in \mathbb{S}t_{nk}$ telle que $\text{Im } Q_k = \mathcal{K}_k(A, v)$. Soit $v \in \mathbb{C}^n$, $v \neq 0$, et $P_k = Q_k Q_k^*$ le projecteur orthogonal sur $\mathcal{K}_k(A, v)$.

1. Montrer que pour tout $j \leqslant k$ on a $P_k A^j v = (P_k A P_k)^j v$. En déduire que $P_k p(A)v = p(P_k A P_k)v$ pour tout polynôme de $\mathcal{P}_k$.
2. Soit p_k le polynôme caractéristique de la matrice $Q_k^* A Q_k$. Déduire de la question précédente et du théorème de Cayley-Hamilton que $P_k p_k(A)v = Q_k p_k(Q_k^* A Q_k) Q_k^* v = 0$.
3. Montrer que $\langle p_k(A)v, u\rangle = 0$ pour tout $u \in \mathcal{K}_k(A, v)$. En déduire que $(-1)^k p_k$ est solution du problème

$$\min_{p \in \tilde{\mathcal{P}}_k} \|p(A)v\|_2^2$$

où $\tilde{\mathcal{P}}_k$ est l'espace des polynômes de degré k et de coefficient dominant égal à un.

Chapitre **16**

Exemples de systèmes linéaires

16.1 LE PROBLÈME DE POISSON DISCRÉTISÉ PAR DIFFÉRENCES FINIES

Considérons le problème du fléchissement d'une poutre de longueur unité fixée à ses deux extrémités et soumise à une force transversale de densité f. La déformation transversale, notée u, satisfait à *l'équation de Poisson*

$$\begin{cases} -u''(x) = f(x), \ x \in]0,1[\\ u(0) = u(1) = 0 \end{cases} \tag{16.1}$$

où la fonction f est donnée et où $u : [0,1] \to \mathbb{R}$ est deux fois continûment dérivable. Les valeurs $u(0) = u(1) = 0$ aux extrémités de l'intervalle $[0,1]$ sont fixées.

La solution u de cette équation est unique et donnée par deux quadratures. Elle s'exprime aussi à l'aide du *noyau de Green* G :

$$u(x) = \int_0^1 G(x,y) f(y)\, dy, \tag{16.2}$$

où $G(x,y)$ est défini par

$$G(x,y) = \begin{cases} x(1-y) \text{ si } x \leqslant y, \\ y(1-x) \text{ si } x \geqslant y. \end{cases}$$

Nous allons suivre une autre voie pour calculer une approximation de la solution. Pour cela, nous choisissons la méthode d'approximation par *différences finies*, certainement une des plus anciennes et des plus naturelles.

On considère une subdivision de l'intervalle $[0, 1]$ par des points x_i équidistants : $x_i = ih$, $i = 0, \ldots, n+1$, où $h = 1/n+1$. Il s'agit de calculer une approximation de la valeur de la solution u prise aux différents points x_i. Notons $u_i \approx u(x_i)$ cette approximation. L'inconnue du problème est donc le vecteur $u = (u_1, \ldots, u_n)^T$.

Pour poser l'équation discrète, il faut définir une approximation de la dérivée seconde u'' aux points x_i. Nous commençons par définir une approximation de la dérivée u' :

$$u'(x) \approx \delta u(x) = \frac{1}{h}\left(u\left(x + \frac{h}{2}\right) - u\left(x - \frac{h}{2}\right)\right).$$

On l'appelle *différence finie centrée*. L'approximation de la dérivée seconde $u''(x_i)$ est obtenue en appliquant deux fois la différence finie centrée :

$$u''(x_i) \approx \delta^2 u(x_i) = \delta\, \delta u(x_i) = \frac{1}{h}\left(\delta u\left(x_i + \frac{h}{2}\right) - \delta u\left(x_i - \frac{h}{2}\right)\right)$$

$$= \frac{1}{h^2}\left(u\left(x_{i+1}\right) - 2u\left(x_i\right) + u\left(x_{i-1}\right)\right)$$

(*différence finie d'ordre deux*). On obtient l'équation (16.1) discrétisée

$$-\frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}) = f_i, \text{ pour } i = 1, \ldots, n, \tag{16.3}$$

avec $u_0 = u_{n+1} = 0$, et $f_i = f(x_i)$, $i = 1, \ldots, n$. La forme matricielle de ce système est :

$$A^h u = f, \tag{16.4}$$

en notant $A^h = \frac{1}{h^2} A_2$ avec

$$A_2 = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \tag{16.5}$$

et $f = (f_1, \ldots, f_n)^T$ le vecteur des données. A_2 est une matrice tridiagonale, symétrique définie positive. Ses valeurs propres sont égales à (exercice 1.13)

$$\lambda_k = 2\left(1 + \cos\left(\frac{k\pi}{n+1}\right)\right), \; k = 1, \ldots, n. \tag{16.6}$$

16.2 LE PROBLÈME DE POISSON SUR UN CARRÉ DISCRÉTISÉ PAR DIFFÉRENCES FINIES

Sur le carré $\Omega =]0,1[\times]0,1[$ de $\mathbb{R}^2$, le problème de Poisson (16.1) devient

$$\begin{cases} -\Delta u = f, \text{ sur } \Omega \\ u = 0, \text{ sur la frontière de } \Omega, \end{cases} \tag{16.7}$$

où Δ est le laplacien en dimension 2 : $\Delta = \partial^2_{xx} + \partial^2_{yy}$.

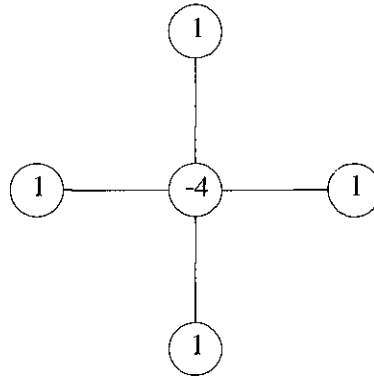
De la même façon que dans l'exemple précédent, cette équation modélise la déformation d'une plaque mince fixée à son bord et soumise à une force tranversable de densité f. De nombreux autres phénomènes physiques sont décrits par cette équation.

On discrétise ce problème en prenant une subdivision régulière du carré Ω par des points $(x_i, y_j) : x_i = ih,\ y_j = jh,\ i,j = 0,\ldots,n+1,\ h = 1/(n+1)$. On note u_{ij} l'approximation de $u(x_i, y_j)$. L'approximation de $\Delta u(x_i, y_j)$ par différences finies est donnée par

$$\Delta u(x_i, y_j) \approx \frac{1}{h^2}\left(u_{i+1\,j} - 2u_{ij} + u_{i-1\,j}\right) + \frac{1}{h^2}\left(u_{i\,j+1} - 2u_{ij} + u_{i\,j-1}\right)$$

$$= \frac{1}{h^2}\left(u_{i+1\,j} + u_{i\,j+1} + u_{i-1\,j} + u_{i\,j-1} - 4u_{ij}\right).$$

Le calcul du laplacien discrétisé au point (x_i, y_j) est obtenu par le schéma en croix suivant



L'équation discrétisée a pour inconnue le vecteur de dimension n^2

$$u = (u_{11},\ldots,u_{n1},u_{12},\ldots,u_{n2},\ldots,u_{1n}\ldots,u_{nn})^T,$$

obtenu en prenant colonne après colonne les valeurs du tableau u_{ij}. De même, en notant $f_{ij} = f(x_i, y_j)$, on a le vecteur des données

$$f = (f_{11},\ldots,f_{n1},f_{12},\ldots,f_{n2},\ldots,f_{1n}\ldots,f_{nn})^T.$$

On obtient le système linéaire

$$B_h u = f,$$

où $B_h = \frac{1}{h^2} B$ et

$$B = \begin{pmatrix} A_4 & -I_n & & & \\ -I_n & A_4 & -I_n & & \\ & \ddots & \ddots & \ddots & \\ & & -I_n & A_4 & -I_n \\ & & & -I_n & A_4 \end{pmatrix} \tag{16.8}$$

avec

$$A_4 = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}. \tag{16.9}$$

La matrice B est définie positive et possède une structure tridiagonale par blocs. Les valeurs propres de B sont données par $\lambda_k + \lambda_l$, où λ_k, λ_l, sont les valeurs propres de A_2 (voir équation (16.6)).

16.3 LE PROBLÈME DE POISSON DISCRÉTISÉ PAR ÉLÉMENTS FINIS

Le problème de Poisson

$$\begin{cases} -\Delta u &= f, \text{ sur } \Omega \\ u &= 0, \text{ sur la frontière de } \Omega \end{cases} \tag{16.10}$$

peut également être discrétisé en utilisant les *éléments finis*. Pour cela, il faut considérer une *formulation variationnelle* du problème : on multiplie les deux membres de l'équation (16.10) par une fonction test v appartenant à un espace V de fonctions régulières qui s'annulent à la frontière du domaine Ω et on intègre sur Ω. Grâce à la formule de Green et à la propriété des fonctions tests, on obtient

$$\int\int_{\Omega} < \nabla u, \nabla v > \, dx\, dy = \int\int_{\Omega} f\, v \, dx\, dy, \tag{16.11}$$

pour tout $v \in V$. On montre que ce problème admet une solution unique $u \in V$ qui est aussi la solution du problème initial lorsque la fonction f est régulière.

Pour discrétiser le problème mis sous forme variationnelle on considère des sous-espaces V_n de dimension finie, $V_n \subset V$, par exemple des espaces de fonctions continues affines par morceaux. Pour cela on subdivise le domaine Ω du plan en triangles,

chacun des triangles admettant soit une intersection vide avec un autre triangle, soit un sommet commun, soit une arête commune. On parle de *maillage* de Ω ; chaque sommet du maillage est appelé *nœud du maillage*. Un exemple de maillage est montré à la figure 16.1. Sur chaque triangle, les fonctions v de V_n sont des fonctions affines $v(x, y) = a + bx + cy$ où les coefficients a, b, c sont définis de manière unique par la valeur de v aux sommets du triangle. Noter que pour une telle fonction ∇v est défini presque partout sur Ω et que l'intégrale (16.11) a un sens. Une base de l'espace V_n est constituée par les fonctions affines par morceaux ϕ_i valant 1 au nœud i du maillage et 0 aux les autres nœuds. La dimension de cet espace est égale au nombre de nœuds internes du maillage.

Le problème variationnel (16.11) est alors formulé dans l'espace de dimension finie V_n : u et les fonctions test v appartiennent à V_n. En exprimant u dans la base $\{\phi_i\}$, $u = \sum_{i=1}^{n} c_i \phi_i$, on obtient un système linéaire vérifié par les coefficients c_i :

$$\sum_{j=1}^{n} c_j \int\int_{\Omega} < \nabla\phi_j, \nabla\phi_i > \, dx\, dy = \int\int_{\Omega} f\, \phi_i \, dx\, dy, \quad \text{pour } i = 1, \ldots, n. \tag{16.12}$$

La matrice du système $K = (k_{ij})$, où $k_{ij} = \int\int_{\Omega} < \nabla\phi_i, \nabla\phi_j > \, dx\, dy$, est appelée *matrice de raideur*.

Dans l'exemple suivant le domaine Ω est en forme de L couché et $f = 1$ sur tout le domaine. La figure 16.1 donne le maillage de Ω produit par la fonction *initmesh* de *Matlab*. La figure 16.2 montre la répartition des coefficients non nuls de la matrice de raideur K obtenue à partir de ce maillage. La matrice K est définie positive et *creuse*. On a $n = 557$ et le nombre de coefficients non nuls de K est égal à 3733. Chaque sommet du maillage a en moyenne six sommets voisins ce qui donne sept coefficients non nuls sur chaque ligne (ou colonne) de K. Le produit $557 \times 7 = 3899$ a une valeur sensiblement plus grande que 3733 du fait que plusieurs sommets de la triangulation ont moins de six voisins. Les figures 16.3 et 16.4 montrent la solution u respectivement sous forme de courbes de niveaux et de surface ombrée.

> ***Remarque 16.1.*** On peut aussi discrétiser par éléments finis le problème de Poisson en dimension 1 défini sur le segment $]0, 1[$. Si l'on prend pour espace de discrétisation V_n l'espace des fonction continues affines par morceaux et un maillage constitué de points équidistants avec un pas $h = 1/(n+1)$, on obtient la base $\{\phi_i\}$ des fonctions-chapeau, chaque chapeau ϕ_i valant 1 au nœud $x_i = ih$ et 0 aux autres nœuds. La matrice de rigidité de ce système est égale à $\frac{1}{h} A_2$ où A_2 est la matrice déjà considérée pour le problème discrétisé par différences finies.

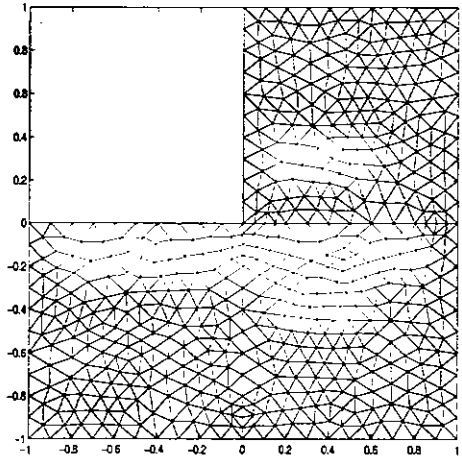


Figure 16.1 Maillage du domaine Ω.

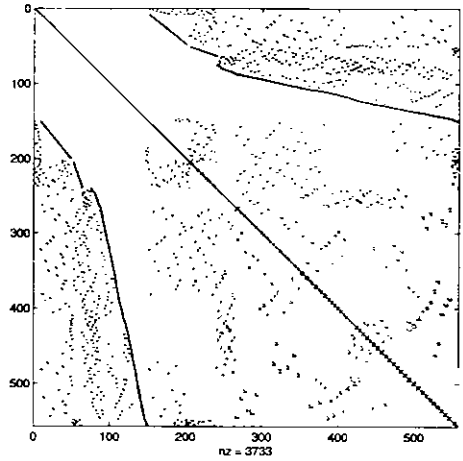


Figure 16.2 Matrice de raideur K. Les points représentent les entrées non nulles de K.

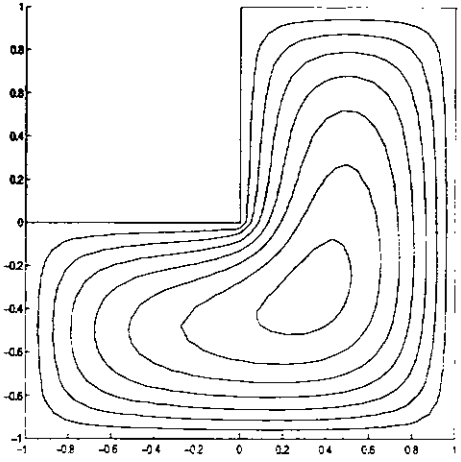


Figure 16.3 Solution u. Isocontours espacés de 0.02 en partant de 0 à la frontière du domaine Ω.

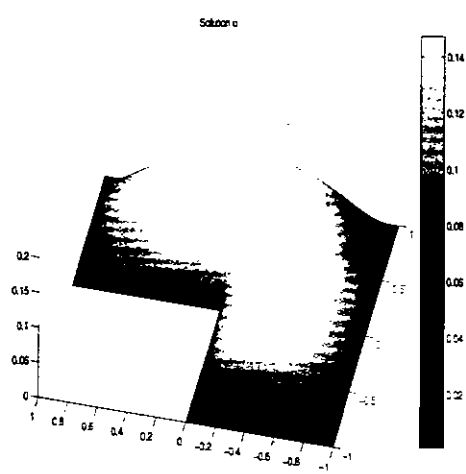


Figure 16.4 Surface $u(x, y)$.

16.4 LA MATRICE DE VANDERMONDE

Cette matrice intervient pour l'évaluation d'un polynôme sur un ensemble fini de points et en interpolation polynomiale.

Soit $p(x) = \sum_{j=0}^{n-1} c_j x^j$ un polynôme de degré inférieur ou égal à $n-1$ et $x_0, \ldots, x_{n-1}$, n points de $\mathbb{C}$. La *matrice de Vandermonde* $V \in \mathbb{C}^{n \times n}$ associée à ces

points est définie par :

$$V = \begin{pmatrix} 1 & x_0^1 & \cdots & x_0^{n-1} \\ 1 & x_1^1 & \cdots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n-1}^1 & \cdots & x_{n-1}^{n-1} \end{pmatrix}. \tag{16.13}$$

On exprime les égalités $p(x_i) = \sum_{j=0}^{n-1} c_j x_i^j$, $i = 0, \ldots, n-1$, sous la forme du produit matrice-vecteur

$$v = Vc, \tag{16.14}$$

où $v = (p(x_0), \ldots, p(x_{n-1}))^T$ et $c = (c_0, \ldots, c_{n-1})^T$ sont respectivement le vecteur des valeurs du polynôme p aux points x_i et le vecteur des coefficients c_i.

Le déterminant de cette matrice est égal à

$$\det V = \prod_{j>i} (x_j - x_i).$$

La matrice V est donc inversible si et seulement si les points x_j sont distincts.

Le problème de l'*interpolation polynomiale* est le problème inverse de l'évaluation : il s'agit de déterminer le polynôme p, c'est-à-dire ses coefficients c_j, à partir des valeurs de p aux points x_i. On doit donc résoudre le système

$$Vc = v.$$

où le vecteur $v = (p(x_0), \ldots, p(x_{n-1}))^T$ est donné. Ce problème a une solution unique si les points x_j sont distincts.

Le conditionnement de la matrice de Vandermonde dépend de la répartition des points x_j.

16.5 LA MATRICE DE FOURIER

La matrice *de Fourier* est une matrice de Vandermonde particulière. Cette matrice joue un rôle important dans le calcul de la transformée de Fourier discrète (TFD) d'une suite finie de nombres complexes.

Soit $\omega = \exp(\frac{-2i\pi}{n})$ une racine primitive n-ième de l'unité. La matrice de Fourier Φ est la matrice de Vandermonde associée au points $\omega^0, \omega^1, \ldots, \omega^{n-1}$. Les coefficients ϕ_{jk} de la matrice de Fourier Φ sont donc définis par

$$\phi_{jk} = \omega^{(j-1)(k-1)}, \quad \text{pour } j, k = 1, \ldots, n.$$

Proposition 16.1 *La matrice* $\Phi \in \mathbb{C}^{n\times n}$ *est symétrique et* $n\Phi^{-1} = \overline{\Phi}$.

Démonstration. Sachant que $\overline{\omega} = \omega^{-1}$, on a

$$(\Phi\overline{\Phi})_{jk} = \sum_{l=1}^{n} \phi_{jl}\overline{\phi}_{lk} = \sum_{l=0}^{n-1} \omega^{(j-1)l}\,\overline{\omega}^{l(k-1)} = \sum_{l=0}^{n-1} \omega^{l(j-k)}.$$

Si $j \neq k$, on a $(\Phi\overline{\Phi})_{jk} = \frac{1-\omega^{n(j-k)}}{1-\omega^{j-k}} = 0$ car $\omega^n = 1$, sinon $(\Phi\overline{\Phi})_{jj} = n$.

La TFD représente la forme discrète de la transformée de Fourier. Soit f une fonction continue périodique de période 1. Les coefficients de Fourier de f sont donnés par

$$\hat{f}(j) = \int_0^1 f(x)\,\exp(-2i\pi jx)\,dx.$$

La discrétisation de cette intégrale par la formule des rectangles donne

$$\hat{f}_j = \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)\,\exp\left(\frac{-2i\pi jk}{n}\right).$$

Cette égalité définit au plus n nombres complexes distincts. En effet $\hat{f}_{j+n} =$

$$\frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)\,\exp\left(\frac{-2i\pi(j+n)k}{n}\right) = \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)\,\exp\left(\frac{-2i\pi jk}{n}\right) = \hat{f}_j.$$

En résumé, notant $f_k = f(\frac{k}{n})$, et les vecteurs $f = (f_0, \ldots, f_{n-1})^T$, $\hat{f} = (\hat{f}_0, \ldots, \hat{f}_{n-1})^T$, le vecteur des coefficients de Fourier discrets $\hat{f}$ est obtenu par le produit

$$\hat{f} = \frac{1}{n}\Phi f.$$

Lorsque n est une puissance de 2, le produit matrice vecteur Φf peut être réalisé en $O(n\log_2(n))$ multiplications au lieu de $O(n^2)$ multiplications normalement requises. L'algorithme qui réalise cette opération, que nous n'allons pas décrire ici, est appelé *transformée de Fourier rapide* (en anglais FFT pour Fast Fourier Transform). Cet algorithme permet d'accélérer les calculs de la transformée de Fourier discrète et en particulier peut être utilisé pour des vecteurs de très grande dimension. On l'utilise dans plusieurs domaines d'application comme par exemple le traitement du signal, l'approximation d'EDP par des méthodes spectrales etc.

16.6 SYSTÈME LINÉAIRE ASSOCIÉ À LA SPLINE CUBIQUE D'INTERPOLATION

Considérons un intervalle $[a, b]$ et une subdivision de celui-ci par n points ordonnés $x_i : a < x_1 < \ldots < x_n < b$. Associé à cette subdivision, nous définissons l'espace $\mathcal{S}_3$ de fonctions $\sigma : [a, b] \to \mathbb{R}$ par

1. La restriction de toute fonction $\sigma \in \mathcal{S}_3$ à chaque intervalle $]x_i, x_{i+1}[$, $i = 1, \ldots, n-1$, est un polynôme de degré inférieur ou égal à 3 ; sur les intervalles extrêmes $[a, x_1[$ et $]x_n, b]$, les restrictions sont des polynômes de degré inférieur ou égal à 1,
2. Les fonctions σ de $\mathcal{S}_3$ ont des dérivées continues jusqu'à l'ordre 2 en chaque point x_i, c'est-à-dire $\sigma^{(k)}(x_i^-) = \sigma^{(k)}(x_i^+)$, pour tout $k = 0, \ldots, 2$, et $i = 1, \ldots, n$.

L'espace $\mathcal{S}_3$ est appelé espace des fonctions *splines cubiques naturelles* sur l'intervalle $[a, b]$ associé à la subdivision $x_1, \ldots, x_n$. La dérivée seconde d'une fonction $\sigma \in \mathcal{S}_3$ est une fonction continue et affine par morceaux sur $[a, b]$. La dimension de $\mathcal{S}_3$ est égale à n ($4(n-1)+4$ coefficients pour décrire ces polynômes et $3n$ conditions de raccordement).

Un problème d'interpolation

Étant données n valeurs y_i, $i = 1, \ldots, n$, on cherche $\sigma \in \mathcal{S}_3$ solution du problème d'interpolation

$$\sigma(x_i) = y_i,\ i = 1, \ldots, n.$$

Notons z_i la valeur (inconnue) de la dérivée seconde aux points x_i : $\sigma''(x_i) = z_i$, $i = 1, \ldots, n$. Puisque σ est un polynôme de degré $\leqslant 1$ dans les intervalles extrêmes, la valeur de la dérivée seconde est nulle aux points x_1 et x_n : $z_1 = z_n = 0$. Nous allons déterminer le système linéaire vérifié par les z_i, $i = 2, \ldots, n-1$. Sur chaque intervalle $[x_i, x_{i+1}]$, $i = 1, \ldots, n-1$, nous avons

$$\sigma''(x) = z_{i+1}\frac{x - x_i}{\Delta x_i} + z_i\frac{x_{i+1} - x}{\Delta x_i}.$$

en notant $\Delta x_i = x_{i+1} - x_i$. Si l'on intègre deux fois cette expression et que l'on utilise les deux conditions d'interpolation $\sigma(x_i) = y_i$, $\sigma(x_{i+1}) = y_{i+1}$, nous obtenons l'expression de σ dans l'intervalle $[x_i, x_{i+1}]$:

$$\sigma(x) = z_{i+1}\frac{(x - x_i)^3}{6\,\Delta x_i} + z_i\frac{(x_{i+1} - x)^3}{6\,\Delta x_i} + B_i(x - x_i) + A_i, \tag{16.15}$$

avec

$$A_i = y_i - z_i\frac{\Delta x_i^2}{6},\ B_i = \frac{\Delta y_i}{\Delta x_i} - \frac{\Delta z_i \Delta x_i}{6},$$

et en notant $\Delta y_i = y_{i+1} - y_i$, $\Delta z_i = z_{i+1} - z_i$.
Le système linéaire qui permet de calculer les z_i (et donc σ d'après la formule (16.15)) est obtenu en imposant la continuité de la dérivée première en chaque point x_i :

$$\sigma'(x_i^+) = \sigma'(x_i^-).$$

En dérivant l'expression (16.15) pour obtenir $\sigma'(x_i^+)$ et en dérivant son analogue sur l'intervalle $[x_{i-1}, x_i]$ pour le calcul de $\sigma'(x_i^-)$, nous avons, après regroupement des facteurs de z_{i-1}, z_i et z_{i+1} :

$$\frac{\Delta x_{i-1}}{6} z_{i-1} + \left(\frac{\Delta x_{i-1}}{3} + \frac{\Delta x_i}{3}\right) z_i + \frac{\Delta x_i}{6} z_{i+1} = \frac{\Delta y_i}{\Delta x_i} - \frac{\Delta y_{i-1}}{\Delta x_{i-1}},$$

pour $i = 3, \ldots, n-2$. Du fait que $z_1 = z_n = 0$, nous obtenons pour $i = 2$

$$\left(\frac{\Delta x_1}{3} + \frac{\Delta x_2}{3}\right) z_2 + \frac{\Delta x_2}{6} z_3 = \frac{\Delta y_2}{\Delta x_2} - \frac{\Delta y_1}{\Delta x_1},$$

et pour $i = n-1$

$$\frac{\Delta x_{n-2}}{6} z_{n-2} + \left(\frac{\Delta x_{n-2}}{3} + \frac{\Delta x_{n-1}}{3}\right) z_{n-1} = \frac{\Delta y_{n-1}}{\Delta x_{n-1}} - \frac{\Delta y_{n-2}}{\Delta x_{n-2}}.$$

La matrice $(n-2) \times (n-2)$ du système est donc donnée par

$$\frac{1}{6} \begin{pmatrix} 2\Delta x_1 + 2\Delta x_2 & \Delta x_2 & & & & \\ \Delta x_2 & 2\Delta x_2 + 2\Delta x_3 & \Delta x_3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \Delta x_{n-3} & 2\Delta x_{n-3} + 2\Delta x_{n-2} & \Delta x_{n-2} & \\ & & & \Delta x_{n-2} & 2\Delta x_{n-2} + 2\Delta x_{n-1} & \end{pmatrix} \tag{16.16}$$

Cette matrice est tridiagonale et définie positive. Le problème d'interpolation dans $\mathcal{S}_3$ admet une solution unique.

La spline cubique naturelle possède une propriété remarquable : elle minimise l'intégrale $\int_a^b \sigma''(x)^2\, dx$ sur l'ensemble des fonctions σ (suffisamment régulières) qui vérifient les conditions d'interpolation $\sigma(x_i) = y_i$.

La figure (16.5) montre une fonction spline cubique naturelle d'interpolation obtenue à partir de dix points d'interpolation. La représentation de courbes et surfaces par fonctions splines est couramment utilisée en modélisation géométrique et en CAO.

16.7 NOTES ET RÉFÉRENCES

L'algorithme de transformation de Fourier rapide FFT est présenté dans un article célèbre de W. Cooley et John W. Tukey paru en 1965 [9] . Il semble cependant que cet algorithme était déjà connu de C. Lanczos dès 1942.

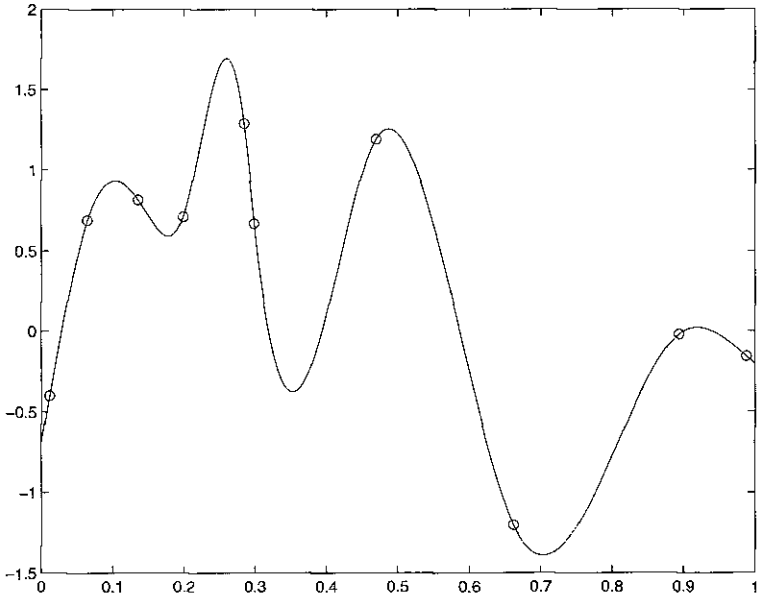


Figure 16.5 Spline cubique naturelle d'interpolation. Les points d'interpolation (x_i, y_i) sont notés avec le symbole o.

Les méthodes de discrétisation par différences finies et par éléments finis sont largement utilisées dans l'industrie.

On trouve de nombreux logiciels mettant en œuvre la méthode des éléments finis. On peut citer un logiciel gratuit FreeFem++ et un gros logiciel professionnel NASTRAN qui représente plusieurs milliers de lignes de code.

Les illustrations numériques ont été faites en utilisant le logiciel MATLAB.

EXERCICES

Exercice 16.1

On considère $u = (u_1, \ldots, u_n)^T$ la solution discrète de l'équation de Poisson (16.4).

1. Déterminer l'expression de u à partir de la matrice A_2^{-1} calculée à l'exercice 7.10.
2. Calculer une approximation de la valeur de la solution exacte du problème de Poisson aux points $x_i = \frac{i}{n+1}$, $i = 1, \ldots, n$, à l'aide de la formule intégrale (formule du noyau de Green (16.2)) que l'on discrétise par la méthode des trapèzes et en prenant pour subdivision de l'intervalle $[0, 1]$ les points équidistants $y_j = \frac{j}{n+1}$, $j = 0, \ldots, n+1$. Comparer avec la solution discrète obtenue à la question précédente.

Exercice 16.2 Diagonalisation d'une matrice circulante

Une matrice circulante $C \in \mathbb{C}^{n\times n}$ est définie par une suite de n scalaires $c_0, \ldots, c_{n-1}$ tels que

$$C = \begin{pmatrix} c_0 & c_1 & & c_{n-2} & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \ddots & c_{n-2} \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ c_2 & \ddots & c_{n-1} & c_0 & c_1 \\ c_1 & c_2 & & c_{n-1} & c_0 \end{pmatrix}.$$

Montrer qu'une matrice circulante C est diagonalisable par la matrice de Fourier $\Phi \in \mathbb{C}^{n\times n}$.

Chapitre **17**

Gauss-Newton et l'assimilation des données

L'algorithme de Gauss-Newton est une des méthodes les plus performantes pour résoudre les problèmes des moindres carrés non-linéaires. Ce chapitre illustre une application de cette méthode pour résoudre le problème d'*assimilation des données* considéré en météorologie et en océanographie.

17.1 LA MÉTHODE DE NEWTON

Présentons brièvement la *méthode de Newton* qui permet de calculer les zéros d'une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ non-linéaire.

Considérons $x \in \mathbb{R}^n$ une approximation de la solution cherchée et linéarisons la fonction f autour de x : on a

$$f(y) \approx f(x) + Df(x)(y - x),$$

où $Df(x)$ est la dérivée de f au point x. L'itération de Newton consiste à remplacer la fonction $f(y)$ par son approximation affine $f(x) + Df(x)(y - x)$ et à calculer les zéros de cette fonction. Si l'on suppose que $Df(x)$ est inversible, on obtient alors

$$y = x - Df(x)^{-1} f(x).$$

L'algorithme de Newton est défini par l'itération

$$x_{k+1} = x_k - Df(x_k)^{-1} f(x_k).$$

La convergence quadratique de cette suite fait de l'algorithme de Newton un outil puissant de calcul des zéros de fonctions. La méthode de Newton permet aussi de calculer la solution d'un problème d'optimisation tel que

$$\min_{x\in\mathbb{R}^n} G(x)$$

où G est une fonction scalaire $G : \mathbb{R}^n \to \mathbb{R}$. En effet, une solution x du problème est obtenue comme zéro du gradient de G : $\nabla G(x) = 0$. L'itération de Newton est dans ce cas donnée par

$$x_{k+1} = x_k - \nabla^2 G(x_k)^{-1} \nabla G(x_k),$$

où $\nabla^2 G(x_k)$ est la matrice hessienne de G en x_k.

17.2 GAUSS-NEWTON ET MOINDRES CARRÉS

Considérons une fonction $g : \mathbb{R}^n \to \mathbb{R}^m$ régulière et le problème d'optimisation

$$\min_{x\in\mathbb{R}^n} \|g(x)\|_2^2. \tag{17.1}$$

Lorsque la fonction g est affine, $g(x) = Ax - b$, on obtient le problème des moindres carrés classique. Lorsque g n'est pas une fonction affine ce problème est qualifié de *moindres carrés non-linéaires.*

Pour résoudre (17.1), on peut utiliser l'algorithme de Newton. Dans ce cas, il est nécessaire, à chaque itération, de calculer la matrice hessienne de la fonction $G(x) = \|g(x)\|_2^2$. Ce calcul peut s'avérer difficile, en particulier lorsque l'expression de $g(x)$ est complexe et la dimension n importante. La méthode proposée par Gauss consiste à linéariser la fonction g autour du point courant x et à calculer la solution du problème des moindres carrés correspondant. L'approximation au premier ordre de g autour de x est

$$g(y) \approx g(x) + Dg(x)(y - x),$$

et le problème des moindres carrés associé

$$\min_{y\in\mathbb{R}^n} \|g(x) + Dg(x)(y - x)\|_2^2. \tag{17.2}$$

Si $m \geqslant n$ (problème surdéterminé) et rang $Dg(x) = n$, c'est-à-dire $Dg(x)$ injective, on sait que la solution du problème (17.2) est unique et donnée par (théorème 9.9)

$$y = x - (Dg(x)^T\, Dg(x))^{-1}\, Dg(x)^T\, g(x). \tag{17.3}$$

L'algorithme se poursuit ainsi à partir du nouveau point y. On évite donc le calcul de la matrice hessienne de la fonction $G(x) = \|g(x)\|^2$. De plus, cette méthode est peu sensible au choix du point initial de l'itération contrairement à la méthode de Newton.

17.3 LE PROBLÈME DE L'ASSIMILATION DES DONNÉES

Un des objectifs des sciences de la terre et en particulier de la météorologie est de prévoir l'évolution des phénomènes physiques que l'on observe à notre échelle sensible. Grâce au développement de l'informatique et des capacités de calcul, il est devenu possible de réaliser des simulations numériques à l'échelle de la planète et de prévoir l'évolution météorologique du temps. La *prévision du temps*, conçue à l'origine pour les besoins de l'aviation, est maintenant largement utilisée dans plusieurs secteurs de la vie socio-économique. D'autre part, les préoccupations actuelles autour des problèmes d'environnement et de changements climatiques montrent l'importance de mieux comprendre et prévoir les phénomènes atmosphériques.

Les prévisions météorologiques sont réalisées à partir de deux composantes essentielles : d'une part un modèle d'évolution de l'atmosphère basé sur les équations générales de la physique, essentiellement les équations d'Euler de la mécanique des fluides, d'autre part des données provenant de mesures réalisées en différents lieux et à différents instants (mesures au sol, radiosondages, données satellitaires ...). La prévision du temps est définie mathématiquement par un *problème d'évolution* : il s'agit d'intégrer un système d'équations aux dérivées partielles d'évolution, non-linéaires, à partir de conditions initiales connues (problème dit de Cauchy). D'un point de vue pratique, l'intégration est obtenue à l'aide d'équations discrétisées.

Une des difficultés majeures de la prévision météorologique réside dans l'instabilité des équations. Elle est due, en particulier, à leur non-linéarité et aux multiples échelles des phénomènes qu'elles représentent. Le résultat d'une prévision dépend de manière cruciale de la précision observée sur les valeurs initiales. Un des objectifs de l'*assimilation des données* est précisément de proposer une solution de ce problème.

Considérons les équations de la météorologie après discrétisation de la variable d'espace. Nous avons un système différentiel

$$x'(t) = f(x(t)), \tag{17.4}$$

que l'on doit intégrer à partir d'une valeur initiale $x(t_0) = x_0$. On note $t \to x(t; x_0)$ la trajectoire (appelée aussi *vecteur d'état* dans le langage du *contrôle optimal*) issue de la valeur initiale x_0. Le vecteur d'état $x(t; x_0)$ décrit les variables météorologiques fondamentales sur l'ensemble des points de discrétisation du modèle. Il s'agit des trois composantes de la vitesse du vent (u, v, w), de la pression p, de la température T et de l'humidité q. Avec les besoins actuels de précision des modèles, on est amené à

considérer des vecteurs d'état x de dimension très importante, de l'ordre de 10^7 et plus.

Les données quant à elles permettent en principe de déterminer la valeur initiale x_0 du système différentiel. En réalité, les données seules ne sont pas suffisantes pour définir avec la précision voulue la valeur initiale x_0. Le problème est largement sous-déterminé. Pour le définir correctement et également pour obtenir une solution plus régulière, on ajoute aux observations une information supplémentaire : on considère également le vecteur d'état à l'instant t_0 issu d'une prévision antérieure. Ce vecteur, noté x_b, est appelé l'*ébauche* (l'indice b vient de l'anglais background). Le meilleur compromis entre ces deux sources d'informations (les mesures et l'ébauche) est obtenu comme solution d'un problème des moindres carrés.

On note z_0 le vecteur des observations disponibles à l'instant t_0 et x_b l'ébauche à ce même instant. Pour passer de l'espace du vecteur d'état x du modèle à l'espace des observations, on utilise un opérateur H que l'on suppose linéaire. Cet opérateur permet par exemple de calculer par interpolation linéaire les valeurs des variables du modèle sur les points où sont réalisées des observations : il y a en effet peu de chances que les observations soient réalisées précisément aux points de grille du modèle. La fonction G des moindres carrés que l'on utilise est pondérée par deux matrices symétriques définies positives B et R, respectivement matrice de covariance d'erreur d'ébauche et matrice de covariance d'erreur d'observation.

En adoptant la notation du paragraphe précédent, on a

$$g(x_0) = (x_0 - x_b, Hx_0 - z_0)$$

et la norme euclidienne dans l'expression $\|g(x_0)\|_2^2$ est remplacée par la norme associée à la matrice définie positive (voir exercice 9.11)

$$S = \frac{1}{2}\begin{pmatrix} B^{-1} & 0 \\ 0 & R^{-1} \end{pmatrix}.$$

La fonction quadratique G s'écrit donc

$$G(x_0) = \frac{1}{2}(x_0 - x_b)^T B^{-1}(x_0 - x_b) + \frac{1}{2}(Hx_0 - z_0)^T R^{-1}(Hx_0 - z_0).$$

La valeur optimale $\tilde{x}_0$ est celle qui minimise G. Le calcul du gradient de G au point x_0 donne

$$\nabla G(x_0) = B^{-1}(x_0 - x_b) + H^T R^{-1}(Hx_0 - z_0).$$

On cherche donc $\tilde{x}_0$ solution de $\nabla G(\tilde{x}_0) = 0$. Le système à résoudre est donc

$$(B^{-1} + H^T R^{-1} H)\tilde{x}_0 = (B^{-1}x_b + H^T R^{-1} z_0), \tag{17.5}$$

où la matrice $M := (B^{-1} + H^T R^{-1} H)$ est définie positive. Comme la dimension du système est très importante et que les différents opérateurs B, R, H sont connus en évaluation, la méthode du gradient conjugué est toute désignée pour résoudre ce système. [1]

Cette méthode d'assimilation est appelée *3D-Var*, le suffixe Var pour Variationnel et le préfixe 3D pour exprimer qu'il s'agit d'une analyse qui ne prend en compte que l'information présente à un seul instant, donc uniquement spatiale, par opposition à une analyse plus complète qui considère également la dimension temporelle (le *4D-Var*) et que l'on va considérer dans la suite.

Remarque 17.1. Dans la théorie de l'estimation, en particulier la théorie du *filtrage de Kalman*, la matrice $K := (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}$, est appelée matrice de *gain*. Un calcul facile montre que la solution $\tilde{x}_0$ de (17.5) s'écrit aussi

$$\tilde{x}_0 = x_b + K(z_0 - H x_b).$$

L'égalité matricielle $(B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} = B H^T (R + H B H^T)^{-1}$, [2] donne une expression de K numériquement plus intéressante puisqu'en général on a moins d'observations (vecteur z) que de variables d'état (vecteur x) et donc la dimension de la matrice $(R + H B H^T)$ que l'on doit inverser est plus petite que celle de $(B^{-1} + H^T R^{-1} H)$.

Le principe de l'assimilation 4D-Var est de considérer non plus une image instantanée de l'atmosphère à l'instant t_0, mais un ensemble d'observations $z_i, i = 0, \ldots, m$, obtenues à différents instants t_i d'une fenêtre temporelle $[t_0, t_m]$ fixée. Ces valeurs z_i sont comparées avec le vecteur d'état $x(t_i; x_0)$ solution du modèle de prévision (17.4) aux différents instants t_i.

Remarque 17.2. Il faut considérer que les instants $t_0, \ldots, t_{m-1}$ sont des instants passés et t_m l'instant présent, à partir duquel une nouvelle prévision sera effectuée au terme de la phase d'assimilation.

Le problème de l'assimilation des données s'exprime à nouveau sous forme d'un problème de moindres carrés. La fonction g est donnée par

$$g(x_0) = (x_0 - x_b, H_0 x(t_0; x_0) - z_0, \ldots, H_m x(t_m; x_0) - z_m)$$

1. En pratique, sachant qu'il n'est possible d'effectuer qu'un nombre très limité d'itérations vue la taille du problème, on utilise pour accélérer la convergence différents types de préconditionnements de ce système.

2. Démontrer cette égalité en exercice.

et la norme est définie par la matrice définie positive

$$S = \frac{1}{2}\begin{pmatrix} B^{-1} & & & \\ & R_0^{-1} & & \\ & & \ddots & \\ & & & R_m^{-1} \end{pmatrix}.$$

Il est naturel de considérer que les opérateurs H_i et R_i dépendent des instants t_i.

La fonction des moindres carrés G s'écrit donc

$$G(x_0) = \frac{1}{2}(x_0 - x_b)^T\, B^{-1}\,(x_0 - x_b) + \frac{1}{2}\sum_{i=0}^{m}(H_i x(t_i; x_0) - z_i)^T\, R_i^{-1}\,(H_i x(t_i; x_0) - z_i),$$

et le problème des moindres carrés

$$\min_{x_0} G(x_0), \tag{17.6}$$

est non-linéaire puisque les fonctions $x(t_i; x_0)$ dépendent de manière non-linéaire de la valeur initiale x_0. En théorie du contrôle optimal la variable x_0 joue le rôle de variable de contrôle du problème. Nous allons utiliser la méthode de Gauss-Newton pour calculer la solution optimale du problème.

> ***Remarque 17.3.*** À partir de la valeur optimale obtenue $\tilde{x}_0$, la solution $x(t_m; \tilde{x}_0)$ calculée à l'instant t_m fournit la nouvelle condition initiale pour une prévision effective initiée à l'instant t_m.

Calcul du gradient du problème linéarisé

La méthode de Gauss-Newton requiert la linéarisation des fonctions $x(t_i; x_0), i = 0, \ldots, m$, par rapport à x_0. Ce calcul est réalisé grâce aux équations (17.4) linéarisées autour de la trajectoire $t \mapsto x(t; x_0)$ issue de x_0.

Pour cela, nous définissons le système linéaire

$$\delta x'(t) = Df(x(t; x_0))\delta x(t), \tag{17.7}$$

vérifié par la variable δx et où $Df(x(t; x_0))$ est la dérivée de f au point $x(t; x_0)$. La variable δx est définie par la condition initiale $\delta x(t_0) = \delta x_0$. Notons également $\delta x(t; \delta x_0)$ la trajectoire de δx issue de la valeur initiale δx_0. Au premier ordre, nous avons

$$x(t; x_0 + \delta x_0) \approx x(t; x_0) + \delta x(t; \delta x_0).$$

Soit $\mathcal{R}(t,t')$ la résolvante [3] associée au système différentiel linéaire (17.7). On a

$$\delta x(t;\delta x_0) = \mathcal{R}(t,t_0)\delta x_0.$$

On peut donc écrire

$$x(t;x_0+\delta x_0) \approx x(t;x_0) + \mathcal{R}(t,t_0)\delta x_0.$$

À partir du problème linéarisé nous considérons la fonction quadratique $\tilde{G}$ de la variable δx_0 :

$$\begin{aligned}\tilde{G}(\delta x_0) &= \frac{1}{2}(x_0+\delta x_0 - x_b)^T\, B^{-1}\,(x_0+\delta x_0 - x_b)\\ &+ \frac{1}{2}\sum_{i=0}^{m}(H_i x(t_i;x_0) + H_i\mathcal{R}(t_i,t_0)\delta x_0 - z_i)^T\, R_i^{-1}(H_i x(t_i;x_0) + H_i\mathcal{R}(t_i,t_0)\delta x_0 - z_i),\end{aligned} \tag{17.8}$$

et le problème des moindres carrés classique

$$\min_{\delta x_0} \tilde{G}(\delta x_0).$$

Le gradient de $\tilde{G}$ est donné par

$$\begin{aligned}\nabla\tilde{G}(\delta x_0) &= B^{-1}\,(\delta x_0 + x_0 - x_b)\\ &+ \sum_{i=0}^{m}\mathcal{R}(t_i,t_0)^T\, H_i^T\, R_i^{-1}\,(H_i\mathcal{R}(t_i,t_0)\delta x_0 + H_i x(t_i;x_0) - z_i),\end{aligned} \tag{17.9}$$

et l'on cherche δx_0 solution du système

$$\nabla\tilde{G}(\delta x_0) = 0.$$

La matrice M du système

$$M = \left(B^{-1} + \sum_{i=0}^{m}\mathcal{R}(t_i,t_0)^T\, H_i^T\, R_i^{-1}\, H_i\mathcal{R}(t_i,t_0)\right)$$

est définie positive. Dans ce cas également, la méthode la plus appropriée pour résoudre numériquement ce système est la méthode du gradient conjugué.

Afin de pouvoir effectuer le produit matrice vecteur $M\delta x_0$, pour tout vecteur δx_0, il est donc nécessaire d'interpréter les opérateurs qui définissent M. L'opérateur résolvant $\mathcal{R}(t_i,t_0)$ correspond à l'intégration de l'équation linéaire entre les instants t_0 et

3. La résolvante associée à un système différentiel linéaire est l'opérateur linéaire $\mathcal{R}(t,t')$ tel que $y(t) = \mathcal{R}(t,t')y(t')$ où y est solution du système différentiel considéré.

t_i. Qu'en est-il de l'opérateur $\mathcal{R}(t_i, t_0)^T$? On introduit pour cela le système adjoint associé au système linéaire (17.7) :

$$p'(t) = -Df(x(t; x_0))^T p(t). \tag{17.10}$$

Il s'agit d'un système différentiel linéaire en p. Ces équations classiques interviennent dans les conditions d'optimalité des problèmes de contrôle optimal. Notons $\mathcal{S}(t', t)$ la résolvante de ce système entre les instants t et t'. L'opérateur $\mathcal{R}(t_i, t_0)^T$ est lié au système adjoint par la propriété suivante.

Proposition 17.1 *Pour tout t et t', on a l'égalité*

$$\mathcal{R}(t', t)^T = \mathcal{S}(t, t').$$

Démonstration. Considérons δx et p respectivement solutions du système linéaire (17.7) et du système adjoint (17.10). Calculons la dérivée du produit scalaire $\langle \delta x(t), p(t) \rangle$:

$$\begin{aligned} \frac{d}{dt}\langle \delta x(t), p(t) \rangle &= \langle \delta x'(t), p(t) \rangle + \langle \delta x(t), p'(t) \rangle \\ &= \langle Df(x(t; x_0))\delta x(t), p(t) \rangle + \langle \delta x(t), -Df(x(t; x_0))^T p(t) \rangle \end{aligned} \tag{17.11}$$

Par la propriété de la transposée, on obtient $\frac{d}{dt}\langle \delta x(t), p(t) \rangle = 0$. Le produit scalaire $\langle x(t), p(t) \rangle$ est donc constant par rapport à t. Considérons deux instants t et t'. On a donc $\langle \delta x(t), p(t) \rangle = \langle \delta x(t'), p(t') \rangle$. En utilisant les résolvantes $\mathcal{R}(t', t)$ et $\mathcal{S}(t, t')$, on déduit que

$$\langle \delta x(t), \mathcal{S}(t, t')p(t') \rangle = \langle \mathcal{R}(t', t)\delta x(t), p(t') \rangle = \langle \delta x(t), \mathcal{R}(t', t)^T p(t') \rangle.$$

Comme cette égalité est vérifiée pour tout $\delta x(t)$ et $p(t')$, on a donc $\mathcal{R}(t', t)^T = \mathcal{S}(t, t')$.

L'opération $\mathcal{R}(t_i, t_0)^T$ correspond ainsi à une intégration rétrograde, de l'instant t_i à l'instant t_0, de l'équation adjointe (17.10).

Le calcul $\mathcal{R}(t_i, t_0)^T H_i^T R_i^{-1} H_i \mathcal{R}(t_i, t_0)\delta x_0$ est donc obtenu par la succession d'une intégration directe entre les instants t_0 et t_i de l'équation linéaire (17.7) à partir de la valeur initiale δx_0 (produit $\mathcal{R}(t_i, t_0)\delta x_0$) suivie du produit $H_i^T R_i^{-1} H_i$ et enfin d'une intégration rétrograde entre les instants t_i et t_0 de l'équation adjointe à partir du vecteur $H_i^T R_i^{-1} H_i \mathcal{R}(t_i, t_0)\delta x_0$ (opérateur $\mathcal{R}(t_i, t_0)^T$ appliqué à $H_i^T R_i^{-1} H_i \mathcal{R}(t_i, t_0)\delta x_0$). En ajoutant les différentes contributions venant de chaque indice i on voit que la somme $\sum_{i=0}^{m} \mathcal{R}(t_i, t_0)^T H_i^T R_i^{-1} H_i \mathcal{R}(t_i, t_0)\delta x_0$ est donnée par une seule intégration

du système direct entre les instants t_0 et t_m suivie d'une intégration rétrograde du système adjoint augmentée à chaque instant d'observation t_i de la valeur $H_i^T R_i^{-1} H_i \mathcal{R}(t_i, t_0)\delta x_0$ obtenue au cours de l'intégration directe.

Le calcul du gradient que nous avons présenté est un outil classique de la théorie du contrôle optimal. Il est utilisé ici pour résoudre numériquement ce problème des moindres carrés non-linéaires de très grande dimension. Il est clair que dans la pratique le système différentiel (17.4) est aussi discrétisé suivant la variable t. Les étapes de calcul du gradient de la fonctionnelle $G(x_0)$ sont analogues à celles présentées ici.

Actuellement, plusieurs centres météorologiques utilisent cette approche pour initialiser les modèles de prévision numérique.

Corrigés des exercices

Exercice 1.1. **1.** Si A et B sont triangulaires inférieures on a $\sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=j}^{i} a_{ik}b_{kj} = 0$ si $i < j$. **2.** Récurrence sur n, développer $\det A$ par rapport à la première ligne de A. **3.** Calculer $P_A(\lambda)$ à l'aide de la question précédente. **4.** Conséquence de 2. **5.** Effectuer le produit $AA^{-1} = I_n$. **6.** Effectuer le produit $A^{-1}A = I_n$. **7.** On remplace a_i par $-a_i$ dans A pour obtenir A^{-1}.

Exercice 1.2. **1.** Les colonnes de uv^* sont $\overline{v_i}u$, $1 \leqslant i \leqslant n$, elles sont donc proportionelles. L'une d'elles est $\neq 0$ donc rang $(uv^*) = 1$. Réciproquement, si l'espace image $\operatorname{Im} A$ est de dimension 1, les colonnes de A qui en forment une base sont proportionnelles et l'une d'elles est non nulle. **2.** $(uv^*)x = 0$ pour tout $x \in u^{\perp}$ qui est de dimension $n-1$ et $(uv^*)u = \langle u, v\rangle u$. **3.** Lorsque $\langle u, v\rangle \neq 0$ une base de $u^{\perp}$ et u constituent une base de vecteurs propres de uv^* qui est donc diagonalisable. Lorsque $\langle u, v\rangle = 0$, la seule valeur propre est 0 et comme $uv^* \neq 0$ cette matrice n'est pas diagonalisable.

Exercice 1.3. $\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} a+ib & 0 \\ 0 & a-ib \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}$.

Exercice 1.4. $\begin{pmatrix} 1+\alpha^2 & \alpha\beta \\ \alpha\beta & 1+\beta^2 \end{pmatrix} =$

$$\frac{1}{\sqrt{\alpha^2+\beta^2}} \begin{pmatrix} \alpha & \beta \\ \beta & -\alpha \end{pmatrix} \begin{pmatrix} 1+\alpha^2+\beta^2 & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{\alpha^2+\beta^2}} \begin{pmatrix} \alpha & \beta \\ \beta & -\alpha \end{pmatrix}.$$

Exercice 1.5. Les valeurs propres et vecteurs propres associés sont $1+\|b\|_2^2$ et $\begin{pmatrix} 0 \\ x \end{pmatrix}$, $x \in a^\perp$, $1+\|a\|_2^2$ et $\begin{pmatrix} y \\ 0 \end{pmatrix}$, $y \in b^\perp$, $1+\|a\|_2^2+\|b\|_2^2$ et $\begin{pmatrix} b\,\|a\|_2^2 \\ a\,\|b\|_2^2 \end{pmatrix}$, 1 et $\begin{pmatrix} -b \\ a \end{pmatrix}$.

Exercice 1.6. 1. Cela résulte des égalités

$$Bu = A\left(u - \frac{xy^*A}{y^*Ax}u\right) = Au - \frac{y^*Au}{y^*Ax}Ax.$$

4. Par 1., rang $A-1 \leqslant$ rang $B \leqslant$ rang A. Si rang $B =$ rang A alors dim Ker $B =$ dim Ker A, impossible par 3.

Exercice 1.7. Notons $P(a_0, \ldots, a_{n-1}, \lambda)$ le polynôme caractéristique de A. On obtient une formule de récurrence en développant ce déterminant par rapport à la première ligne.

Exercice 1.8. Écrivons $A = PSP^{-1}$ et $B = QTQ^{-1}$ avec S et T triangulaires supérieures. On obtient $M = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}\begin{pmatrix} S & P^{-1}BQ \\ 0 & T \end{pmatrix}\begin{pmatrix} P^{-1} & 0 \\ 0 & Q^{-1} \end{pmatrix}$ et on peut utiliser l'exercice 1.1. $M^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}BD^{-1} \\ 0 & D^{-1} \end{pmatrix}$.

Exercice 1.9. 5. $A+xy^* = A(I_n + A^{-1}xy^*)$. Les valeurs propres de $I_n + A^{-1}xy^*$ sont 1 et $1+y^*A^{-1}x$ (exercice 1.2). Aussi $A+xy^*$ est inversible si et seulement si $y^*A^{-1}x \neq -1$. Prenons $B=-x$, $C=y^*$ et $D=1$. La formule précédente donne $(A+xy^*)^{-1} = A^{-1} + A^{-1}(-x)(1-y^*A^{-1}(-x))^{-1}y^*A^{-1} = A^{-1} - \frac{A^{-1}xy^*A^{-1}}{1+y^*A^{-1}x}$.

Exercice 1.10. Par addition de lignes et de colonnes $\det\begin{pmatrix} A & B \\ B & A \end{pmatrix} =$

$$\det\begin{pmatrix} A-B & B \\ B-A & A \end{pmatrix} = \det\begin{pmatrix} A-B & B \\ 0 & A+B \end{pmatrix} = \det(A-B)\det(A+B).$$

Exercice 1.11. Facile. Noter que $\det(A-iB) = \det(\overline{A+iB}) = \overline{\det(A+iB)}$ parce que A et B sont réelles.

Exercice 1.12. Écrire cette matrice $\begin{pmatrix} I_3 & B \\ C & D \end{pmatrix}$.

Exercice 1.13. 1. Noter que $v^{(p)} \neq 0$ et que les relations $bv_{k+1}^{(p)} - \lambda_p v_k^{(p)} + cv_{k-1}^{(p)} = 0$ sont satisfaites avec $\lambda_p = 2\sqrt{bc}\cos\frac{p\pi}{n+1}$. Noter aussi que $v_0^{(p)} = v_{n+1}^{(p)} = 0$. **2.** $\lambda_p = a + 2\sqrt{bc}\cos\frac{p\pi}{n+1}$, vecteurs propres identiques à ceux de $A(0,b,c)$.

Exercice 1.14. 1. Soit λ valeur propre de A et $x \neq 0$ vecteur propre associé. De l'égalité $Ax = \lambda x$ on déduit $x^*Ax = \lambda x^*x$ et donc $x^*A^*x = \overline{\lambda} x^*x$ par adjonction. L'hypothèse $A^* = -A$ implique $-x^*Ax = \overline{\lambda} x^*x$ et donc $-\lambda x^*x = \overline{\lambda} x^*x$ et $-\lambda = \overline{\lambda}$. λ est donc un nombre complexe imaginaire pur. **2.** $I_n - A$ est inversible parce que 1 n'est pas valeur propre de A. **3.** A est normale et peut se diagonaliser en $A = U \operatorname{diag}(i\beta_k) U^*$ avec U unitaire et $\beta_k \in \mathbb{R}$. Ainsi $Q = U \operatorname{diag}\left(\frac{1+i\beta_k}{1-i\beta_k}\right) U^*$. Il est alors évident que $QQ^* = I_n$ et que $(1 + i\beta_k)/(1 - i\beta_k) \neq -1$.

Exercice 1.15. 1. Pour tout $u \in \mathbb{C}^n$, $Au = (y^*u)x + (x^*u)y$ donc Im A est de dimension 2 engendré par x et y. **2.** Si $u \in (\text{Im } A)^\perp$ c'est-à-dire si $x^*u = y^*u = 0$ on a $Au = 0$ donc 0 est valeur propre et $(\text{Im } A)^\perp$ est le sous-espace propre associé. Si $u \in \text{Im } A$, $u = \alpha x + \beta y$, le système $Au = \lambda u$ s'écrit

$$x(\alpha(y^*x - \lambda) + \beta y^*y) + y(\alpha x^*x + \beta(x^*y - \lambda)) = 0$$

ce qui est équivalent à

$$\begin{pmatrix} \langle x, y\rangle - \lambda & \|y\|_2^2 \\ \|x\|_2^2 & \overline{\langle x, y\rangle} - \lambda \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Les valeurs propres λ sont données par l'équation du second degré $\lambda^2 - 2\Re\langle x, y\rangle \lambda + |\langle x, y\rangle|^2 - \|x\|_2^2 \|y\|_2^2 = 0$ qui possède deux racines réelles d'où λ puis α et β. Lorsque $x, y \in \mathbb{R}^n$ cette équation s'écrit $\lambda^2 - 2\langle x, y\rangle \lambda + |\langle x, y\rangle|^2 - \|x\|_2^2 \|y\|_2^2 = 0$ et les valeurs propres sont $\lambda = \langle x, y\rangle \pm \|x\| \|y\|$. Des vecteurs propres correspondant sont $u = \|y\| x \pm \|x\| y$. **3.** Im B est le sous-espace engendré par x et y, sa dimension est 2, c'est un sous-espace invariant de $\mathbb{C}^n$ par B. Les valeurs propres λ de $B|_{\text{Im } B}$ sont données par l'équation caractéristique $\lambda^2 - 2i\Im\langle x, y\rangle \lambda - |\langle x, y\rangle|^2 + \|x\|_2^2 \|y\|_2^2 = 0$. Les autres valeurs propres sont $\lambda = 0$ associées au sous-espace propre $(\text{Im } B)^\perp$. Lorsque x et y sont réels on obtient $\lambda = \pm i(\|x\|_2^2 \|y\|_2^2 - |\langle x, y\rangle|^2)^{1/2}$.

Exercice 2.1. Les flottants positifs sont :

$$\left(\frac{d_1}{10} + \frac{d_2}{100}\right) 10^e$$

avec $e = -1, 0$ ou 1. $1 \leqslant d_1 \leqslant 9$ et $0 \leqslant d_2 \leqslant 9$. On obtient 270 nombres qui sont, en écriture décimale,

.010, .011, ... , .019, .020, .021, ... , .029, ... , .090, .091, ... , .099,

.10, .11, ... , .19, .20, .21, ... , .29, ... , .90, .91, ... , .99,

1.0, 1.1, ... , 1.9, 2.0, 2.1, ... , 2.9, ... , 9.0, 9.1, ... , 9.9.

Noter que ces nombres ne sont pas régulièrement espacés mais que leur espacement est constant entre deux puissances consécutives de $\beta = 10$.

Exercice 2.3. On trouve 0. Pour Maple, l'instruction $> evalf(3.*(4./3.-1.)-1.)$; donne $-.110^{-18}$. Noter la syntaxe : on a écrit 3. et non pas 3 de façon à ce que ces nombres soient traités avec l'arithmétique flottante et non pas avec l'arithmétique des entiers.

Exercice 2.4. Récurrence sur n.

Exercice 3.1. Notons S_i et λ_i les ensembles et leurs bornes supérieures respectives considérées aux questions $i = 1, \ldots, 6$. Pour tout $x \neq 0$ et $\alpha > 0$, on a $\frac{\|Lx\|}{\|x\|} = \frac{\|L\,\alpha x\|}{\|\alpha x\|}$. On en déduit facilement l'égalité des ensembles S_1, S_2, S_5 et donc l'égalité de leurs bornes supérieures. L'égalité $\frac{\|Lx\|}{\|x\|} = \|L\,\frac{x}{\|x\|}\|$ pour tout $x \neq 0$, montre que $S_3 = S_1$ et donc $\lambda_3 = \lambda_1$. On a $\|Lx\| \leqslant \frac{\|Lx\|}{\|x\|}$ pour tout x tel que $\|x\| \leqslant 1$. On en déduit que $\lambda_4 \leqslant \lambda_2$. Or $S_3 \subset S_4$ implique que $\lambda_3 \leqslant \lambda_4$. Sachant que $\lambda_2 = \lambda_3$, on a donc $\lambda_4 = \lambda_2$. Montrons enfin que $\frac{\|Lx\|}{\|x\|} \leqslant \lambda_6$, pour tout $x \neq 0$. Supposons qu'il existe $x \neq 0$ tel que $\frac{\|Lx\|}{\|x\|} > \lambda_6$. Posons $a = \frac{\|Lx\|}{\|x\|}$ et prenons α tel que $\frac{\lambda_6}{a} < \alpha < 1$. On a alors $\|L\ \alpha\,\frac{x}{\|x\|}\| = \alpha\,\frac{\|Lx\|}{\|x\|} = \alpha\, a > \lambda_6$. D'autre part $\|\alpha\,\frac{x}{\|x\|}\| < 1$ montre qu'il y a contradiction puisque λ_6 est un majorant de S_6. On a donc $\frac{\|Lx\|}{\|x\|} \leqslant \lambda_6$ et $\lambda_1 \leqslant \lambda_6$. Comme $\lambda_4 = \lambda_1$ et $\lambda_6 \leqslant \lambda_4$ puisque $S_6 \subset S_4$, on en déduit que $\lambda_6 = \lambda_4$. On a montré l'égalité des bornes supérieures $\lambda_i, i = 1, \ldots, 6$. **7.** λ_3 est la borne supérieure de l'image par la fonction continue $x \mapsto \|Lx\|$ de la sphère unité $S_{n-1} = \{x \in \mathbb{C}^n,\ \|x\| = 1\}$ qui est un ensemble compact. Elle appartient donc à l'ensemble S_3. Les bornes $\lambda_1, \lambda_2, \lambda_5$ appartiennent aussi aux ensembles respectifs S_1, S_2, S_5 puisque ces ensembles sont tous égaux à S_3. λ_4 appartient à S_4 puisqu'il s'agit de la borne supérieure de l'image de la boule unité compacte $B_n(0,1) = \{x \in \mathbb{C}^n,\ \|x\| \leqslant 1\}$ par la fonction continue $x \mapsto \|Lx\|$.

Exercice 3.2. 1. Pour tout vecteur x on a $\|Ax\|_1 = \sum_{i=1}^n |(Ax)_i| = \sum_{i=1}^n |\sum_{j=1}^n a_{ij}x_j|$. On a donc

$$\|Ax\|_1 \leqslant \sum_{j=1}^n |x_j| \sum_{i=1}^n |a_{ij}| \leqslant \|x\|_1 \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$$

et $\|A\|_1 \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$. Soit k un indice tel que $\sum_{i=1}^n |a_{ik}| = \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$. Définissons le vecteur $x = (x_1, \ldots, x_n)^T$ tel que $x_i = \delta_{ik}$. On a $\|x\|_1 = 1$ et $\|Ax\|_1 = \sum_{i=1}^n |\sum_{j=1}^n a_{ij}\delta_{jk}| = \sum_{i=1}^n |a_{ik}| = \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$. Ceci montre que $\|A\|_1 \geqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$ et donc l'égalité. **2.** Pour tout vecteur x on a $\|Ax\|_\infty = \max_{1\leqslant i\leqslant n} |\sum_{j=1}^n a_{ij}x_j|$. On a donc $\|Ax\|_\infty \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}| \max_{1\leqslant j\leqslant n} |x_j| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}|\, \|x\|_\infty$.

Ceci montre que $\|A\|_\infty \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}|$. Soit k un indice tel que $\sum_{j=1}^n |a_{kj}| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}|$. Définissons le vecteur $x = (x_1, \ldots, x_n)^T$ tel que $x_j = 1$ si $a_{kj} = 0$ et $x_j = |a_{kj}|/a_{kj}$ si $a_{kj} \neq 0$. On a $\|x\|_\infty = 1$ et $\sum_{j=1}^n a_{kj}x_j = \sum_{j=1}^n |a_{kj}| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}|$. Ceci montre que $\|A\|_\infty \geqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^n |a_{ij}|$ et donc l'égalité.

Exercice 3.3. Pour toute valeur propre λ de A et tout vecteur propre associé x, $x \neq 0$, on a $Ax = \lambda x$ et donc $\|Ax\|_1 = |\lambda|\,\|x\|_1$. On en déduit que et $|\lambda| = \|Ax\|_1 / \|x\|_1 \leqslant \|A\|_1$. On obtient le résultat grâce à l'égalité $\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^n |a_{ij}|$ (voir exercice 3.2).

Exercice 3.4. Le spectre de $I_n - A$ est égal à $\{1 - \lambda,\ \lambda \in \text{spec } A\}$. On a $|1 - \lambda| \geqslant 1 - |\lambda| > 0$ puisque $|\lambda| < 1$ pour toute valeur propre de A. Les valeurs propres de $I_n - A$ sont donc toutes distinctes de zéro et la matrice est inversible.

Exercice 3.5. La matrice J donnée par la décomposition de Jordan (voir théorème 1.5) a une structure diagonale par blocs. Chaque bloc J_k est soit de la forme $J_k = \lambda_k I_{n_k}$, soit de la forme $J_k = \lambda_k I_{n_k} + N_{n_k}$ où $N_{n_k} \in \mathbb{C}^{n_k \times n_k}$ est la matrice nilpotente

$$N_{n_k} = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{pmatrix}.$$

Nous avons $A^p = P J^p P^{-1}$ et la matrice J^p a également une structure diagonale par blocs avec des blocs de la forme J_k^p. Lorsque $J_k = \lambda_k I_{n_k}$ on a $J_k^p = \lambda_k^p I_{n_k}$ et lorsque $J_k = \lambda_k I_{n_k} + N_{n_k}$, par la formule du binôme on a

$$J_k^p = (\lambda_k I_{n_k} + N_{n_k})^p = \sum_{l=0}^{n_k-1} \begin{pmatrix} p \\ l \end{pmatrix} \lambda_k^{p-l} N_{n_k}^l.$$

La sommation est effectuée jusqu'à l'indice $l = n_k - 1$ puisque la matrice N_{n_k} est nilpotente d'ordre n_k (c'est-à-dire $N_{n_k}^{n_k} = 0$). La matrice J_k^p est triangulaire supérieure. Ses coefficients sont de la forme $\begin{pmatrix} p \\ l \end{pmatrix} \lambda_k^{p-l}$. Pour la norme considérée dans la démonstration du théorème 3.7 nous avons

$$\|A^p\| = \max_k \max_{1\leqslant i,j\leqslant n_k} |(J_k^p)_{ij}|.$$

Considérons la limite $|(J_k^p)_{ij}|^{1/p}$ lorsque $p \to \infty$. Pour les coefficients diagonaux on a $|(J_k^p)_{ii}|^{1/p} = |\lambda_k|$ et leur limite est égale à $|\lambda_k|$. Les autres coefficients non nuls sont de la forme $\left|\begin{pmatrix} p \\ l \end{pmatrix} \lambda_k^{p-l}\right|^{1/p}$. Leur limite est égale à $|\lambda_k|$ puisque $\lim_{p\to\infty} \begin{pmatrix} p \\ l \end{pmatrix}^{1/p} = 1$ et que $\lim_{p\to\infty} |\lambda_k|^{(p-l)/p} = |\lambda_k|$. On conclut en utilisant la propriété $\max_{k,i,j} \lim_{p\to\infty} = \lim_{p\to\infty} \max_{k,i,j}$.

Exercice 3.6. On a

$$\|A\|_F = \sqrt{a^2 + d^2 + |b + ic|^2 + |b - ic|^2} = \sqrt{a^2 + d^2 + 2(b^2 + c^2)}.$$

La matrice A est hermitienne, donc $\|A\|_2 = \rho(A)$. Calculons les valeurs propres de A. On a

$$p_A(\lambda) = \det(A - \lambda I_2) = \begin{vmatrix} a - \lambda & b + ic \\ b - ic & d - \lambda \end{vmatrix}$$

donc $p_A(\lambda) = (a - \lambda)(d - \lambda) - (b^2 + c^2) = \lambda^2 - (a + d)\lambda + ad - (b^2 + c^2)$. Les deux racines du polynôme p_A sont données par

$$\frac{a + d \pm \sqrt{(a - d)^2 + 4(b^2 + c^2)}}{2}$$

et

$$\rho(A) = \max \left| \frac{a + d \pm \sqrt{(a - d)^2 + 4(b^2 + c^2)}}{2} \right|.$$

Exercice 3.7. Les normes $\|.\|_1$, $\|.\|_2$, $\|.\|_\infty$ sont des normes d'opérateur. Pour celles-ci on a

$$\|I_n\| = \sup_{x \in \mathbb{C}^n,\, x \neq 0} \frac{\|I_n(x)\|}{\|x\|} = 1.$$

Pour la norme de Frobenius on a $\|I_n\|_F = \sqrt{n}$.

Exercice 3.8. La proposition 3.10 donne $\|U\|_2 = \|UI_n\|_2 = \|I_n\|_2 = 1$. Pour la norme de Frobenius on a $\|U\|_F = \sqrt{\text{trace}\,(U^*U)} = \sqrt{\text{trace}\,(I_n)} = \sqrt{n}$.

Exercice 3.9. La matrice A est hermitienne. Les valeurs de cette matrice tridiagonale sont égales à

$$4 + 2\cos(p\pi/4)$$

avec $p = 1, 2, 3$ (voir exercice 1.13). Donc $\|A\|_2 = \rho(A) = 4 + 2\cos(\pi/4) = 4 + \sqrt{2}$. Pour la norme de Frobenius on obtient $\|U\|_F = \sqrt{4 + 3\ 16} = \sqrt{52}$.

Exercice 3.10. Si A est diagonale il est évident que les coefficients de sa diagonale sont ses valeurs propres. Réciproquement : supposons que A symétrique ait ses

coefficients diagonaux égaux à ses valeurs propres $\lambda_i, i = 1, \ldots, n$. On a $\|A\|_F = \sqrt{\text{trace}\,(A^2)}$. Les valeurs propres de la matrice A^2 sont égales à λ_i^2. Sachant que la trace d'une matrice est égale à la somme de ses valeurs propres on a donc $\|A\|_F = \sqrt{\sum_{i=1}^n \lambda_i^2}$. Par ailleurs, le calcul direct des coefficients diagonaux de A^2 donne $\|A\|_F = \sqrt{\sum_{i=1}^n \lambda_i^2 + \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2}$. On déduit donc que $\sum_{j \neq i} a_{ij}^2 = 0$ pour tout $i = 1, \ldots, n$, et donc A est diagonale.

Exercice 3.11. 1. On a $N(x) = \alpha^{p-1}\|x\|_2 + \alpha^{p-2}\|Ax\|_2 + \ldots + \|A^{p-1}x\|_2$. Il est évident que $N(x)$ est une norme. **2. et 3.** Par définition $N(A) = \sup_{x \neq 0} \frac{N(Ax)}{N(x)}$. On vérifie que $N(Ax) = \alpha N(x) - \alpha^p \|x\|_2 + \|A^p x\|_2$. D'autre part, pour tout $x \neq 0$, on a

$$\frac{\|A^p x\|_2}{\|x\|_2} \leqslant \|A^p\|_2 < \alpha^p$$

et donc $\|A^p x\|_2 - \alpha^p \|x\|_2 \leqslant 0$. Cette inégalité et l'expression de $N(Ax)$ permet de conclure que $N(Ax) \leqslant \alpha\, N(x)$. On a donc $N(Ax)/N(x) \leqslant \alpha$ et $N(A) \leqslant \alpha$.

Exercice 3.12. On a $\|xy^*\|_2 = \sqrt{\rho(yx^*\, xy^*)} = \|x\|_2 \sqrt{\rho(yy^*)}$. Par l'exercice 1.2 on sait que l'unique valeur propre non nulle de la matrice de rang un yy^* est égale à $\langle y, y\rangle = \|y\|_2^2$. On conclut que $\|xy^*\|_2 = \|x\|_2 \|y\|_2$. Pour la norme de Frobenius on a $\|xy^*\|_F = \sqrt{\sum_i \sum_j |x_i\, \overline{y_j}|^2} = \sqrt{\sum_i \sum_j |x_i|^2 |y_j|^2} = \sqrt{\sum_i |x_i|^2} \sqrt{\sum_j |y_j|^2} = \|x\|_2 \|y\|_2$. L'exercice 3.2 montre que

$$\|xy^*\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^n |x_i\, \overline{y_j}| = \max_{1 \leqslant j \leqslant n} |y_j| \sum_{i=1}^n |x_i| = \max_{1 \leqslant j \leqslant n} |y_j|\, \|x\|_1 = \|y\|_\infty \|x\|_1$$

et que

$$\|xy^*\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |x_i\, \overline{y_j}| = \max_{1 \leqslant i \leqslant n} |x_i| \sum_{j=1}^n |y_j| = \max_{1 \leqslant i \leqslant n} |x_i|\, \|y\|_1 = \|x\|_\infty \|y\|_1.$$

Exercice 3.13. Prenons $H \in \mathbb{C}^{n \times n}$ tel que $\|H\| < 1/\|A^{-1}\|$. On a alors $\|A^{-1}H\| \leqslant \|A^{-1}\|\, \|H\| < 1$. Le théorème de perturbation de Neumann (proposition 3.14) montre que $(I_n + A^{-1}H)$ est inversible et donc aussi $(A + H) = A\,(I_n + A^{-1}H)$. De $(I_n + A^{-1}H)^{-1} = \sum_{k=0}^\infty (-1)^k (A^{-1}H)^k = I_n - A^{-1}H + (A^{-1}H)^2 + \ldots$ on déduit $(I_n + A^{-1}H)^{-1} A^{-1} = A^{-1} - A^{-1}H\, A^{-1} + (A^{-1}H)^2 A^{-1} + \ldots$ et $(A + H)^{-1} - A^{-1} + A^{-1}H\, A^{-1} = (A^{-1}H)^2 (I_n - (A^{-1}H) + (A^{-1}H)^2 + \ldots)\, A^{-1}$ $= (A^{-1}H)^2\, (I_n + A^{-1}H)^{-1}\, A^{-1}$.

Ainsi $\|(A+H)^{-1} - A^{-1} + A^{-1}HA^{-1}\| \leqslant \|H\|^2 \|A^{-1}\|^3 \|(I_n + A^{-1}H)^{-1}\|$ et $\|(A+H)^{-1} - A^{-1} + A^{-1}HA^{-1}\|/\|H\| \to 0$ lorsque $H \to 0$.

Exercice 3.14. 1. Cette série est absolument convergente du fait que pour toute norme multiplicative on a $\|A^k/k!\| \leqslant \|A\|^k/k!$ qui est le terme général d'une série convergente. On a ainsi

$$\left\| \sum_{k=0}^{\infty} \frac{A^k}{k!} \right\| \leqslant \sum_{k=0}^{\infty} \frac{\|A\|^k}{k!} = \exp(\|A\|).$$

2. Évident. **3.** Si $AB = BA$ on peut appliquer la formule du binôme $(A+B)^k = \sum_{l=0}^{k} \binom{k}{l} A^{k-l} B^l$ et $(A+B)^k/k! = \sum_{i+j=k} A^i/i!\, B^j/j!$. Pour toute matrice M notons $S_k(M)$ la somme partielle $S_k(M) = \sum_{i=0}^{k} \frac{M^i}{i!}$. On a

$$S_k(A+B) = \sum_{i+j \leqslant k} \frac{A^i}{i!} \frac{B^j}{j!}$$

et

$$S_k(A)\, S_k(B) - S_k(A+B) = \sum_{\substack{k+1 \leqslant i+j \leqslant 2k \\ 1 \leqslant i,j \leqslant k}} \frac{A^i}{i!} \frac{B^j}{j!}.$$

On obtient la majoration

$$\|S_k(A)\, S_k(B) - S_k(A+B)\| \leqslant \sum_{\substack{k+1 \leqslant i+j \leqslant 2k \\ 1 \leqslant i,j \leqslant k}} \frac{\|A\|^i}{i!} \frac{\|B\|^j}{j!}.$$

Le majorant est égal

$$\sum_{i=0}^{k} \frac{\|A\|^i}{i!} \sum_{i=0}^{k} \frac{\|B\|^i}{i!} - \sum_{l=0}^{k} \frac{\left(\|A\| + \|B\|\right)^l}{l!}$$

et il converge vers zéro lorsque $k \to \infty$ en vertu de l'égalité $\exp(\|A\|)\exp(\|B\|) = \exp(\|A\| + \|B\|)$. **4.** La matrice $-A$ commute avec la matrice A. On a d'une part $\exp(A + (-A)) = \exp(A)\exp(-A)$ et d'autre part $\exp(0) = I_n$ d'où le résultat. **5.** De l'égalité $A^k = (PDP^{-1})^k = PD^kP^{-1}$ on déduit $S_k(A) = PS_k(D)P^{-1}$ et le résultat par passage à la limite. **6.** La décomposition de Jordan de la matrice A, $A = PJP^{-1}$, montre que $\exp(A) = P\exp(J)P^{-1}$. La matrice $\exp(J)$ est triangulaire supérieure et a pour coefficients diagonaux $\sum_{i=0}^{\infty} \lambda^i/i! = \exp(\lambda)$, où λ est une valeur propre de A. **7.** La décomposition de Jordan de la matrice A donne

$\det(\exp(A)) = \det(P\exp(J)P^{-1}) = \det(P)\det(\exp(J))\det(P^{-1}) = \det(\exp(J)) = \prod_i \exp(\lambda_i) = \exp(\sum_i \lambda_i) = \exp(\text{trace}(A))$ où λ_i sont les valeurs propres de A. **8.** La question 3 montre que $\exp((t+h)A) = \exp(tA)exp(hA)$ donc $\exp((t+h)A) - \exp(tA) = (\exp(hA) - I_n)exp(tA)$. On a $\exp(hA) - I_n = \sum_{k=1}^{\infty}(hA)^k/k! = hA\,(\sum_{k=0}^{\infty}(hA)^k/(k+1)!)$ et

$$\lim_{h\to 0}\frac{\exp(hA)-I_n}{h} = A.$$

On en déduit le résultat. **9.** Pour tout entier $k > 0$, on a $(xy^*)^k = (y^*x)^{k-1}xy^* = \langle x,y\rangle^{k-1}xy^*$. On a $\exp(xy^*) = I_n + xy^* + (xy^*)^2/2! + \ldots$. Si $\langle x,y\rangle = 0$ alors $\exp(xy^*) = I_n + xy^*$, sinon $\exp(xy^*) = I_n + xy^*\ (\exp(\langle x,y\rangle) - 1)/\langle x,y\rangle)$. **10.** Notons u_i les colonnes de la matrice unitaire U. La décomposition $A = U\Lambda U^*$ montre que A peut s'écrire sous la forme d'une somme de matrices de rang un : $A = \sum_{i=1}^n \lambda_i u_i u_i^*$. Pour $i \neq j$, on a $(u_iu_i^*)(u_ju_j^*) = \langle u_j,u_i\rangle u_iu_j^* = 0$ puisque les vecteurs u_i et u_j sont orthogonaux. D'après la question 3 on a donc $\exp(A) = \exp(\sum_{i=1}^n \lambda_i u_i u_i^*) = \prod_{i=1}^n \exp(\lambda_i u_i u_i^*)$. La question précédente montre que $\exp(A) = \prod_{i=1}^n (I_n + u_iu_i^*(\exp(\lambda_i) - 1))$.

Exercice 3.15. Posons $M = \begin{pmatrix} I_m & A \\ 0 & I_n \end{pmatrix}$. On a $M^*M = \begin{pmatrix} I_m & A \\ A^* & A^*A + I_n \end{pmatrix}$. Considérons λ une valeur propre de M^*M. On a

$$\begin{cases} x + Ay &= \lambda x \\ A^*x + (A^*A + I_n)y &= \lambda y \end{cases}$$

où $\begin{pmatrix} x \\ y \end{pmatrix} \neq 0$ est un vecteur propre associé à λ. La première équation donne $A^*x + A^*Ay = \lambda A^*x$. Supposons $\lambda \neq 1$. On a donc $A^*x = 1/(\lambda - 1)\,A^*Ay$. En remplaçant A^*x dans la deuxième équation, on obtient $\lambda A^*Ay = ((\lambda-1)\lambda + 1 - \lambda)y$. On constate facilement que $\lambda \neq 0$ puisque la matrice M est injective et donc M^*M inversible. Ainsi, l'égalité précédente donne $A^*Ay = ((\lambda^2 - 2\lambda + 1)/\lambda)\ y$. On observe également que $y \neq 0$ car sinon on aurait $\lambda = 1$. On a ainsi $(\lambda^2 - 2\lambda + 1)/\lambda = \sigma^2$, où σ est une valeur singulière de A. λ est donc la solution positive de l'équation $\lambda^2 - (2+\sigma^2)\lambda + 1 = 0 : \lambda = (2 + \sigma^2 + \sigma\sqrt{\sigma^2+4})/2$. La fonction $s \mapsto (2 + x^2 + x\sqrt{x^2+4})/2$ est croissante sur $\mathbb{R}_+$. En considérant $\sigma_{\max}$ la plus grande valeur singulière de A et sachant que $(2 + \sigma_{\max}^2 + \sigma_{\max}\sqrt{\sigma_{\max}^2+4})/2 \geqslant 1$, on a dans tous les cas $(2 + \sigma_{\max}^2 + \sigma_{\max}\sqrt{\sigma_{\max}^2+4})/2 \geqslant \lambda$ pour toute valeur propre de M^*M d'où le résultat puisque $\|A\|_2 = \sigma_{\max}$.

Exercice 3.16. Des égalités (voir exercice 3.2) $\|A\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^m |a_{ij}|$ et $\|A\|_\infty = \max_{1\leqslant i\leqslant m}\sum_{j=1}^n |a_{ij}|$ on déduit $\|A^*\|_\infty = \max_{1\leqslant i\leqslant m}\sum_{j=1}^n |\overline{a_{ji}}| = \|A\|_1$.

L'inégalité est démontrée grâce à $\|A\|_2 = \sqrt{\rho(A^*A)}$ (voir théorème 3.9) et à $\rho(A^*A) \leqslant \|A^*A\|_1$ (voir proposition 3.6).

Exercice 3.17. La somme des coefficients d'un carré magique d'ordre n vaut $\sum_{i=1}^{n^2} i = n^2(n^2+1)/2$. La somme S_n des termes d'une même ligne (ou d'une même colonne) est donc égale à $(n^2(n^2+1)/2)/n = n(n^2+1)/2$. Le calcul pour les normes $\|A\|_1$ et $\|A\|_\infty$ est évident en utilisant l'exercice 3.2. L'inégalité $\|A\|_2 \leqslant \sqrt{\|A\|_1 \|A\|_\infty}$ (voir l'exercice 3.16) montre que $\|A\|_2 \leqslant S_n$. D'autre part, en prenant le vecteur $u = (1, \ldots, 1)^T$, on a $\|Au\|_2 / \|u\|_2 = S_n$. On en déduit donc que $\|A\|_2 = S_n$.

Exercice 3.18. Considérons D la matrice diagonale obtenue à partir des termes diagonaux de A : $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$. On décompose $A = D - M$ où $-M$ est la matrice des termes extra-diagonaux de A. On a $A = D(I_n - D^{-1}M)$ et $\|D^{-1}M\|_\infty < 1$. Le lemme de Neumann (proposition 3.14) montre alors que $I_n - D^{-1}M$ est inversible et donc la matrice A également.

Exercice 3.19. La matrice A est symétrique et

$$A^2 = \begin{pmatrix} n & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 1 \end{pmatrix}$$

est diagonale par blocs. Ses blocs sont le scalaire n et la matrice $n \times n$ de rang 1 uu^T où $u = (1, \ldots, 1)^T$. Les valeurs propres de cette matrice sont 0 et $u^T u = n$ (voir exercice 1.2). On obtient donc $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = n$.

Exercice 3.20. Les valeurs propres de $I_n + M^*M$ sont égales à $1 + \lambda_i$ où λ_i sont les valeurs propres (réelles et positives) de M^*M. On a $\det(I_n + M^*M) = \prod_{i=1}^n (1+\lambda_i)$ et $\|M\|_F^2 = \mathrm{trace}\,(M^*M) = \sum_{i=1}^n \lambda_i$. Considérons les réels positifs $a_i = 1 + \lambda_i$, $i = 1, \ldots, n$. L'inégalité entre les moyennes géométriques et arithmétiques

$$\left(\prod_{i=1}^n a_i\right)^{1/n} \leqslant \frac{\sum_{i=1}^n a_i}{n}$$

(qui se démontre facilement à partir de la concavité de la fonction logarithme) donne le résultat.

Exercice 3.21. Soit $A = URU^*$ la décomposition de Schur de la matrice A où R est triangulaire supérieure et U est unitaire. Pour tout entier $p > 0$ posons $A_p = UR_pU^*$ où R_p est la matrice obtenue à partir de R en remplaçant les entrées nulles de sa diagonale par $1/p$. On a $\|A - A_p\|_F = \|R - R_p\|_F = \sqrt{k}/p$, où k est le

nombre d'entrées nulles de la diagonale de R. Les matrices A_p sont inversibles et $\lim_{p\to\infty} \|A - A_p\|_F = 0$.

Exercice 4.1. Soit λ une valeur propre non nulle de AB. Il existe donc un vecteur x non nul tel que $ABx = \lambda x$. On note que le vecteur Bx est distinct de zéro car sinon on aurait $\lambda = 0$. En multipliant par B les deux membres de l'égalité on a $BAB\,x = \lambda Bx$. Cette égalité montre que λ est une valeur propre de BA et que Bx est un vecteur propre associé. L'inclusion opposée s'obtient de manière identique. Lorsque $m = n$ on utilise l'égalité $\det(AB) = \det(BA) = \det(A)\det(B)$ qui montre que si zéro est valeur propre de AB alors c'est aussi une valeur propre de BA.

Exercice 4.2. On a $A^TA = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$. Les valeurs propres de cette matrice sont 2 et 4. Les valeurs singulières de A sont donc égales à $\sqrt{2}$ et 2.

Exercice 4.3. Soit $A = x$ où $x \in \mathbb{C}^n$ est un vecteur colonne. La seule valeur singulière de A est égale à $\sqrt{x^*x} = \|x\|_2$. Considérons une décomposition en valeurs singulières de A : $A = V\Sigma U^*$. La matrice V est obtenue en complétant en une base orthonormée de $\mathbb{C}^n$ le vecteur unitaire $v_1 = zx/\|x\|_2$ où $z \in \mathbb{C}$ et $|z| = 1$. La matrice U est donnée par $U = z$ et $\Sigma = \|x\|_2$.

Exercice 4.4. Soit A une matrice hermitienne. Elle se diagonalise en $A = UDU^*$ avec $U = (u_1 \ldots u_n)$ unitaire et $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, où $\lambda_i \in \mathbb{R}^*$ sont les valeurs propres non nulles de A. Les valeurs singulières de A sont égales à $|\lambda_i|$ pour $i = 1, \ldots, r$. On obtient une décomposition en valeurs singulières de A : $A = U\Sigma V^*$ avec $\Sigma = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_r|, 0, \ldots, 0)$ et $V = (\pm u_1 \ldots \pm u_r\, u_{r+1} \ldots u_n)$ où le signe $\pm$ dépend du signe de λ_i.

Exercice 4.5. 1. On a $Ay = \lambda x$ et $A^*x = \lambda y$, avec $\lambda \neq 0$. La première égalité montre que si $y = 0$ alors $x = 0$ et la seconde que si $x = 0$ alors $y = 0$. Donc $x \neq 0$ et $y \neq 0$ puisque x ou y sont par hypothèse non nuls. La première égalité donne $A^*Ay = \lambda A^*x = \lambda^2 y$. Donc $\lambda = \pm\sigma$ où σ est une valeur singulière de A. **2.** Réciproquement, soit σ une valeur singulière de A. On a donc $A^*Ay = \sigma^2 y$ avec $y \neq 0$. En posant $x = (1/\sigma)\,Ay$ on a $A^*x = (1/\sigma)\,A^*Ay = \sigma y$ et $(x, y)^T$ est un vecteur propre associé à la valeur propre σ du système augmenté. En posant $x = (-1/\sigma)\,Ay$ on a $A^*x = (-1/\sigma)\,A^*Ay = -\sigma y$ et $(x, y)^T$ est un vecteur propre associé à la valeur propre $-\sigma$ du système augmenté.

Exercice 4.6. Par le théorème 3.9 on a $\|A\|_2 = \sqrt{\rho(A^*A)}$. Sachant que les valeurs singulières sont les racines carrées des valeurs propres non nulles de A^*A on a donc

$\|A\|_2 = \sigma_{\max}$ où $\sigma_{\max}$ est la plus grande valeur singulière de A. La proposition 3.6 montre que $\rho(A) \leqslant \|A\|_2$.

Exercice 4.7. De l'égalité $Ax = \lambda x$ avec $\lambda \neq 0$ et $x \neq 0$ on déduit que $x = \lambda A^{-1}x$ et donc le premier résultat. De même, l'égalité $A^*Ax = \sigma^2 x$ avec $\sigma \neq 0$ et $x \neq 0$ montre que $x = \sigma^2 A^{-1}A^{-*}x$. $1/\sigma^2$ est donc valeur propre de $A^{-1}A^{-*}$ qui a même spectre que $A^{-*}A^{-1}$.

Exercice 4.8. Le polynôme caractéristique de Z est égal à $(1-x)^{n+1}$. Z a pour seule valeur propre 1. On vérifie que

$$Z^*Z = \begin{pmatrix} 1+\sum_{i=1}^n |z_i|^2 & \overline{z_1} & \dots & \overline{z_n} \\ z_1 & 1 & & \\ \vdots & & \ddots & \\ z_n & & & 1 \end{pmatrix}.$$

En développant le déterminant $\det(Z^*Z - xI_{n+1})$ par rapport à la dernière colonne (ou ligne) on montre facilement, par récurrence sur n, que le polynôme caractéristique de Z^*Z est égal à $(1-x)^{n-1}((1-x)^2 - x\sum_{i=1}^n |z_i|^2)$ dont les racines sont 1 et $r_1, r_2 = (2+\sum_{i=1}^n |z_i|^2 = \sqrt{\sum_{i=1}^n |z_i|^2}\sqrt{4+\sum_{i=1}^n |z_i|^2})/2$. Les valeurs singulière de Z sont donc 1 et $\sqrt{r_1}, \sqrt{r_2}$.

Exercice 4.9. On vérifie que

$$Z^*Z = \begin{pmatrix} 1+\beta^2 & 0 \\ 0 & \alpha^2+\beta^2 \end{pmatrix}.$$

En supposant que $\alpha, \beta \neq 0$, les valeurs singulières de Z sont donc $\sqrt{1+\beta^2}$ et $\sqrt{\alpha^2+\beta^2}$. Une décomposition en valeurs singulières de Z est donnée par $Z = V\Sigma U^*$ où V est la matrice obtenue à partir de Z en normalisant ses colonnes $V = (v_1\ v_2)$ avec $v_1 = (1/\sqrt{1+\beta^2})(1,0,\beta,0)^T$, $v_2 = (1/\sqrt{\alpha^2+\beta^2})(0,\alpha,0,\beta)^T$, et $\Sigma = \mathrm{diag}(\sqrt{1+\beta^2}, \sqrt{\alpha^2+\beta^2})$, $U = I_2$.

Exercice 5.1. On trouve $X = (.9999341067, -.9999087092)^T$ pour le premier système, $X = (.3410001318, -.08700018258)^T$ pour le second, $\mathrm{cond}_2 A = .2193219001\ 10^7$ et $\det A = 10^{-6}$ (calculs flottants).

Exercice 5.2. $\|AB - I_n\|_2 = \left\|A(BA - I_n)A^{-1}\right\|_2 \leqslant \|A\|_2 \|BA - I_n\|_2 \left\|A^{-1}\right\|_2$.

Exercice 5.3. 2. La première inégalité provient de la convexité de $x > 0 \to 1/x$, la seconde se ramène à $1/\lambda_i \leqslant (\lambda_1 + \lambda_n - \lambda_i)/(\lambda_1\lambda_n)$ pour tout i. **3.** Le max est égal à $(\lambda_1+\lambda_2)^2/(4\lambda_1\lambda_2)$.

Exercice 5.4. On obtient $\text{cond}_2(A) \geqslant 141.99295$.

Exercice 5.5. Les valeurs singulières de cette matrice hermitienne sont égales à $\sigma_p = \left|a + 2|b| \cos \frac{p\pi}{n+1}\right|$, $1 \leqslant p \leqslant n$, en vertu de l'exercice 1.13 et $\text{cond}_2 A(a, b, \overline{b}) = (\max_p \sigma_p)/(\min_p \sigma_p)$.

Exercice 6.1. Posons $A = LU$. On calcule alternativement les lignes de U et les colonnes de L de 1 à n. L'égalité $a_{ij} = \sum_{k=1}^{i} l_{ik}u_{kj}$ pour $j = i, \ldots, n$ donne la ligne i de U : $u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}$. L'égalité $a_{ji} = \sum_{k=1}^{i} l_{jk}u_{ki}$ pour $j = i+1, \ldots, n$ donne la colonne i de L : $l_{ji} = (a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki})/u_{ii}$.

Exercice 6.2. Les matrices triangulaires obtenues à partir de A sont

$$U_1 = \begin{pmatrix} 0.0001 & 1 \\ 0 & 9999 \end{pmatrix} \text{ et } U_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0.9999 \end{pmatrix}$$

après permutation des lignes. Leur conditionnement est $\text{cond}_\infty A = 4.000400040$, $\text{cond}_\infty U_1 = 10^8$ et $\text{cond}_\infty U_2 = 4.000200020$. Cet exemple illustre bien l'intérêt d'une stratégie de type « pivot partiel ».

Exercice 6.3. 1. L'inverse d'une matrice d'élimination est donné par $E(i, \lambda_{i+1}, \ldots, \lambda_n)^{-1} = E(i, -\lambda_{i+1}, \ldots, -\lambda_n)$. Le produit de k telles matrices est la matrice triangulaire inférieure obtenue à partir de I_n en y substituant aux colonnes $i_1 \ldots i_k$ la colonne i_1 de $E(i_1)$ ... la colonne i_k de $E(i_k)$. **2.** Si $E(i_k) \ldots E(i_1)A = U$ alors $A = LU$ avec $L = E(i_1)^{-1} \ldots E(i_k)^{-1}$ qui s'obtient sans calcul.

Exercice 6.4. 5. Notons $A = L_A U_A$ la décomposition LU de A. On a $B = A + uv^* = L_A U_A + uv^* = L_A(I_n + (L_A^{-1}u)(U_A^{-*}v)^*)U_A$ ce qui donne $L_B = L_A(E + \Delta)\Delta^{-1}$ et $U_B = \Delta_+^{-1}(F + \Delta)U_A$ avec les notations des questions précédentes.

Exercice 6.5. Il faut considérer $a_{ij} = \sum_{k=1}^{min(i,j)} l_{ik}u_{kj}$ avec $i = 1$, $j > p$: on obtient $u_{1j} = 0$ puis, avec $j = 1$ et $i > p$, on obtient : $l_{i1} = 0$ et ainsi de suite.

Exercice 7.1. 3. $c_{ij} = 1$ si $j \leqslant i$, 0 sinon. La matrice $C^{-1} = (d_{ij})$ est donnée par $d_{ii} = 1$, $d_{i\,i-1} = -1$ et 0 sinon. On obtient la matrice tridiagonale symétrique $A^{-1} = (b_{ij})$: $b_{ii} = 2$, $b_{i+1\,i} = b_{i\,i+1} = -1$ pour $i = 1, \ldots, n-1$ et $b_{nn} = 1$.

Exercice 7.2. 1. Par récurrence : $u_{11} = a_{11} > 0$. Si $A = LU$ alors $A_{n-1} = L_{n-1}U_{n-1}$ où les matrices A_{n-1}, L_{n-1} et U_{n-1} sont obtenues à partir de A, L et U en supprimant la dernière ligne et la dernière colonne. Comme $u_{ii} > 0$ pour $i = 1 \ldots n-1$ par l'hypothèse de recurrence, on a $u_{nn} > 0$ parce que $\det A = u_{11} \ldots u_{nn} > 0$. On a $A = LU = A^* = U^*L^* = (U^*D^{-1})(DL^*)$ d'où $L = U^*D^{-1}$ par unicité de la

décomposition LU et donc $\left(LD^{1/2}\right)^* = D^{-1/2}U$. La décomposition de Cholesky de A est donnée par

$$A = (U^*D^{-1/2})(D^{1/2}L^*) = (LD^{1/2})(D^{1/2}L^*) = (LD^{1/2})(LD^{1/2})^*.$$

2. Avec les notations de l'exercice 6.4, la décomposition LU de B est donnée par $L_B = L(E+\Delta)\Delta^{-1}$ et $U_B = \Delta_+^{-1}(F+\Delta)U$. On a $D_B = \Delta_+^{-1}\Delta D$ et donc la décomposition de Cholesky de B est donnée par

$$C_B = L(E+\Delta)\Delta^{-1}(\Delta_+^{-1}\Delta D)^{1/2} = C_A(D^{-1/2}(E+\Delta)\Delta^{-1/2}\Delta_+^{-1/2}D^{1/2}).$$

Exercice 7.3. 2. Si $\sum_{k=1}^n \alpha_k\varphi_k(x) = 0$ pour tout $x \in]0,1[$ alors $\sum_{k=1}^n \alpha_k x^{a_k+p} = 0$, $p = 1/2+n$. Par dérivations successives et passage à la limite pour $x \longrightarrow 1$ on obtient $\sum_{k=1}^n \alpha_k = \sum_{k=1}^n \alpha_k(a_k+p) = \ldots = \sum_{k=1}^n \alpha_k(a_k+p)\ldots(a_k+p-n+1) = 0$ d'où $\sum_{k=1}^n \alpha_k a_k^l = 0$, $l = 0\ldots n-1$. Le déterminant de la matrice de ce système en les α_k est un déterminant de Vandermonde qui est non nul parce que les a_k sont distincts. Donc $\alpha_k = 0$ pour tout k et les fonctions φ_k sont indépendantes.

Exercice 7.4. Noter que si $A = UDU^*$ avec D diagonale réelle et U unitaire alors $\exp A = U(\exp D)U^*$ est définie positive. Toute matrice définie positive $M = U\Delta U^*$, Δ diagonale positive, est de ce type puisque $M = \exp(U(\log\Delta)U^*)$ avec $\log\Delta = \mathrm{diag}(\log\delta_i)$ lorsque $\Delta = \mathrm{diag}(\delta_i)$. Ainsi l'application exponentielle est surjective. On prouve son injectivité en suivant les indications données.

Exercice 7.5. La décomposition de Cholesky de la matrice de ce système est donnée par $C = \begin{pmatrix} 3 & 0 & 0 \\ 2 & 4 & 0 \\ 1 & 1 & 1 \end{pmatrix}$ et sa solution est $x = 2$, $y = 4$, $z = -1$.

Exercice 7.6. $A = CC^*$ avec $C = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 1 & 2 & 1 \end{pmatrix}$, $t = 1$, $x = 0$, $y = 1$, $z = 0$.

Exercice 7.7. 1.a. $\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ax_1^2 + (b+c)x_1x_2 + dx_2^2$ est strictement positif pour tout $(x_1,x_2) \neq (0,0)$ si et seulement si $a > 0$ et $at^2 + (b+c)t + d > 0$ pour tout $t \in \mathbb{R}$ ce qui équivaut à $a > 0$ et $\Delta = (b+c)^2 - 4ad = (b-c)^2 - 4(ad-bc) < 0$. **1.b.** On a donc $a > 0$ et $\det A = ad - bc > 0$ mais ces conditions ne suffisent pas comme le montre $A = \begin{pmatrix} 1 & -4 \\ 0 & 1 \end{pmatrix}$, $x_1 = x_2 = 1$. **3.** Faire $x = e_i$ dans $x^TAx > 0$, où e_i est le i–ème vecteur de la base canonique. **4.** La première

étape de la décomposition LU donne (puisque $a_{11} \neq 0$) $L_1 A = \begin{pmatrix} a_{11} & w^T \\ 0 & C \end{pmatrix}$ d'où $L_1 A L_1^T = \begin{pmatrix} a_{11} & v^T \\ 0 & C \end{pmatrix}$ puisque $L_1 \in \mathcal{L}_1$. Prenons $x = \begin{pmatrix} 0 \\ y \end{pmatrix}$, $y \in \mathbb{R}^{n-1}$, $y \neq 0$. On a $x^T L_1 A L_1^T x = y^T C y > 0$ donc $C \in \mathcal{P}$ et l'on peut continuer par récurrence. On obtient ainsi $LAL^T = U$ avec $L \in \mathcal{L}_1$, U triangulaire supérieure et $U \in \mathcal{P}$. Noter que A est inversible et admet une décomposition LU. **5.** On obtient $A = L^{-1} U L^{-T} = M U M^T$ avec $M \in \mathcal{L}_1$ et UM^T triangulaire supérieure. C'est la décomposition LU de A qui est unique. Lorsque A est symétrique on a $U = U^T$ donc U est diagonale positive et $A = (MU^{1/2})(U^{1/2}M)$ est la décomposition de Cholesky de A.

Exercice 7.8. 1. Tracer le graphe de $y = \frac{1}{2}\left(x + \frac{a}{x}\right)$. Noter que $x_1 \geqslant \sqrt{a}$ et que la suite (x_p), $p \geqslant 1$, est décroissante et minorée par $\sqrt{a}$. **2.** On se ramène au cas précédent par diagonalisation : $A = UDU^*$ avec U unitaire et D diagonale positive. Les matrices X_p vérifient $X_p = U D_p U^*$. $D_p = \text{diag}(x_{pi})$. $x_{0i} = 1$ et $x_{p+1i} = \frac{1}{2}\left(x_{pi} + \frac{d_i}{x_{pi}}\right)$, $p \geqslant 0, i = 1 \ldots n$.

Exercice 7.9. 1. Il faut montrer que $\int_0^\infty \|\exp(-tB) H \exp(-tB)\|_2 \, dt$ converge. Cette norme est majorée par $\|H\|_2 \|\exp(-tB)\|_2^2 = \|H\|_2 \exp(-2t\lambda_n)$ où $\lambda_1 \geqslant \ldots \geqslant \lambda_n > 0$ sont les valeurs propres de B et l'intégrale $\int_0^\infty \exp(-2t\lambda_n) dt$ converge. **2.** Noter que $\mathcal{PH}_n$ est un ouvert de $\mathcal{H}_n$. C est une application polynomiale donc C^∞ et $DC(B)H = \lim \left((B + \varepsilon H)^2 - B^2\right)/\varepsilon = BH + HB$. **3.** C'est une conséquence du théorème des fonctions inverses. Il faut prouver que $DC(B)$ est un isomorphisme. Si $DC(B)H = BH + HB = 0$, par diagonalisation de $B = UDU^*$ (D diagonale, U unitaire) on se ramène à $DK + KD = 0$ avec $K = U^* H U$. On a $k_{ij}(\lambda_i + \lambda_j) = 0$ pour tout i et j d'où $K = 0$. Ainsi $DC(B) : \mathcal{H}_n \to \mathcal{H}_n$ est injective et c'est un isomorphisme. **4.** Pour tout $K \in \mathcal{H}_n$ on a $D\left(\sqrt{A}\right) K = (DC(B))^{-1} K$ d'où $DC(B)H = BH + HB = K$. Pour prouver que $H = \int_0^\infty \exp(-tB) K \exp(-tB) dt$ il faut montrer que cette intégrale satisfait à l'équation $BH + HB = K$. Comme

$$B \exp(-tB) K \exp(-tB) + \exp(-tB) K \exp(-tB) B = -\frac{d}{dt}\left(\exp(-tB) K \exp(-tB)\right)$$

on obtient

$$BH + HB = \int_0^\infty B \exp(-tB) K \exp(-tB) + \exp(-tB) K \exp(-tB) B =$$

$$- \exp(-tB) K \exp(-tB)\Big|_{t=0}^{t=\infty} = K.$$

Comme nous l'avons déjà vu,

$$\left\| D\sqrt{A} K \right\|_2 \leqslant \|K\|_2 \int_0^\infty \exp(-2t\lambda_n) dt = \frac{\|K\|_2}{2\lambda_n} = \frac{\|K\|_2 \left\|B^{-1}\right\|_2}{2}.$$

Exercice 7.10. 1. Par développement par rapport aux dernières lignes et colonnes on obtient $\det B_{2,n} = 2\det B_{2,n-1} - \det B_{2,n-2}$, $\det B_{2,1} = 1$ et $\det B_{2,2} = 1$ d'où $\det B_{2,n} = 1$ pour tout $n \geqslant 1$.

$$C = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & 0 & \\ & \ddots & \ddots & & \\ & & 0 & \ddots & \ddots \\ & & & -1 & 1 \end{pmatrix}, \quad C^{-1} = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & 0 & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix},$$

$B^{-1} = C^{-T}C^{-1}$ donc $\left(B^{-1}\right)_{ij} = n - \max(i,j) + 1 = \min(n-i+1, n-j+1)$. La formule de S-M-W (exercice 1.9) donne ici $A^{-1} = (B + e_1 e_1^*)^{-1} = B^{-1} - \frac{B^{-1}e_1 e_1^* B^{-1}}{1+e_1^* B^{-1} e_1} = B^{-1} - \frac{B^{-1}e_1 e_1^* B^{-1}}{n+1}$ d'où $(A^{-1})_{ij} = \min(n-i+1, n-j+1) - \frac{(n-i+1)(n-j+1)}{n+1} = \min(i,j) - \frac{ij}{n+1}$.

Exercice 7.11. Si $A = LU$ (décomposition LU) et si D est la diagonale de U alors D est définie positive et la décomposition de Cholesky de A est $A = CC^*$ avec $C = LD^{1/2}$ (exercice 7.2). Il est donc clair que, comme pour la décomposition LU (exercice 6.5), la structure bande de A se transfère sur C.

Exercice 8.1. $A = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix} \frac{1}{\sqrt{5}} \begin{pmatrix} 5 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix}$.

Exercice 8.4. Les valeurs propres de $H = I_n - 2xx^*$, $\|x\|_2 = 1$, sont 1 de multiplicité $n-1$, associée au sous-espace $x^\perp$ et -1 de multiplicité 1 associée au vecteur propre x. Les valeurs propres de $G(r,s,\theta)$ sont 1 de multiplicité $n-2$, associée au sous-espace propre engendré par les vecteurs de la base canonique e_k, $k \neq r$ et s, et $\exp(\pm i\theta)$ associées aux vecteurs propres $(1,i)^T$ et $(1,-i)^T$.

Exercice 8.5. Par récurrence sur n. Soit λ une valeur propre de A et z, $\|z\|_2 = 1$, un vecteur propre associé. **Premier cas :** $\lambda = \pm 1$ et $z \in \mathbb{R}^n$. Notons $Q_1 = \begin{pmatrix} z & Q_2 \end{pmatrix}$ une matrice orthogonale dont la première colonne est z de sorte que $Q_2^T z = 0$. On a $Q_1^T A Q_1 = \begin{pmatrix} z^T Az & z^T AQ_2 \\ Q_2^T Az & Q_2^T AQ_2 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & A_{n-1} \end{pmatrix}$ où A_{n-1} est $n-1 \times n-1$ orthogonale. En effet : $z^T Az = z^T \lambda z = \lambda \|z\|_2^2 = \lambda$, $Q_2^T Az = Q_2^T \lambda z = 0$, $(z^T AQ_2)^T = Q_2^T A^T z = Q_2^T A^{-1} z = Q_2^T \lambda z = 0$. Il suffit d'utiliser l'hypothèse de récurrence pour conclure. **Deuxième cas :** $\lambda = \exp(i\theta)$, $\theta \neq 0$ et π, $z \in \mathbb{C}^n$. Posons $z = x + iy$; on a $Ax = x\cos\theta - y\sin\theta$ et $Ay = y\cos\theta + x\sin\theta$. Noter que x et $y \neq 0$ sont linéairement indépendants parce que $\sin\theta \neq 0$. Dans une base

orthonormée du plan Oxy on a $A \begin{pmatrix} u & v \end{pmatrix} = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$. Soit $Q_1 = \begin{pmatrix} u & v & Q_2 \end{pmatrix}$ une matrice orthogonale obtenue en complétant la base (u, v). On a : $Q_1^T A Q_1 =$

$$\begin{pmatrix} \begin{pmatrix} u & v \end{pmatrix}^T A \begin{pmatrix} u & v \end{pmatrix} & \begin{pmatrix} u & v \end{pmatrix}^T A Q_2 \\ Q_2^T A \begin{pmatrix} u & v \end{pmatrix} & Q_2^T A Q_2 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} & 0 \\ 0 & A_{n-1} \end{pmatrix}$$

et l'on raisonne comme dans le premier cas.

Exercice 8.6. Soit A orthogonale et soit $A = QR$ la décomposition QR obtenue par la méthode de Householder. Q est le produit d'au plus $n-1$ symétries orthogonales. R est triangulaire supérieure et orthogonale donc diagonale : $R = \mathrm{diag}(\pm 1)$. Comme la méthode de Householder permet de contrôler les signes des entrées r_{ii}, $1 \leqslant i \leqslant n-1$, on peut supposer que $R = \mathrm{diag}(1, \ldots, 1, \pm 1)$ c'est donc l'identité ou une symétrie orthogonale. On obtient ainsi A produit d'au plus $n-1+1$ symétries orthogonales.

Exercice 8.7. On procède comme dans la méthode de Householder mais en utilisant cette fois des rotations (pas nécessairement des rotations de Givens). Le premier vecteur-colonne a_1 de A est « rabattu » sur $\|a_1\|_2 e_1$ par une rotation R_1 et cetera. On obtient ainsi $R_{n-1} \ldots R_1 A = R$ triangulaire supérieure et orthogonale et donc, comme dans l'exercice précédent, $R = \mathrm{diag}(1, \ldots, 1, \pm 1)$ est l'identité ou une symétrie orthogonale. On obtient ainsi A produit d'au plus $n-1$ rotations et au plus 1 symétrie orthogonale.

Exercice 8.8. 1. $U = \begin{pmatrix} \rho_1 e^{i\theta_1} & \rho_3 e^{i\theta_3} \\ \rho_2 e^{i\theta_2} & \rho_4 e^{i\theta_4} \end{pmatrix}$ avec $\rho_i \geqslant 0$, $\rho_1^2 + \rho_2^2 = \rho_3^2 + \rho_4^2 = 1$ et $\rho_1\rho_3 e^{i(\theta_1-\theta_3)} + \rho_2\rho_4 e^{i(\theta_2-\theta_4)} = 0$. Cette dernière équation impose $\rho_1\rho_3 = \rho_2\rho_4$ et $\theta_2 - \theta_4 = \theta_3 - \theta_1$ modulo (2π). Prenons $\rho_1 = \cos\alpha$, $\rho_2 = \sin\alpha$, $\rho_3 = \sin\beta$, $\rho_4 = \cos\beta$ avec $0 \leqslant \alpha, \beta \leqslant \pi/2$. Comme $\rho_1\rho_3 = \rho_2\rho_4$ on a $\alpha = \beta$. La condition sur les θ_i permet d'obtenir $\theta_4 = \theta_2 + \theta_3 - \theta_1 - \pi$ (modulo 2π). On prend donc $\theta_1 = \sigma$, $\theta_2 = \tau + \pi$, $\theta_3 = \nu$, $\theta_4 = \tau + \nu - \sigma$. **2.** $U \in \mathbb{SU}_n$ lorsque $\det U = e^{i(\tau+\nu)} = 1$ c'est-à-dire lorsque $\nu = -\tau$.

Exercice 9.1. 1. $L : \mathbb{R}^2 \to \mathbb{R}$ est surjective donc $L^\dagger = L^*(LL^*)^{-1} = \begin{pmatrix} a \\ b \end{pmatrix} \frac{1}{a^2+b^2}$.

2. $LM = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & a+b \end{pmatrix}$, $(LM)^\dagger = \begin{pmatrix} a \\ a+b \end{pmatrix} \frac{1}{a^2+(a+b)^2}$, $M^{-1}L^\dagger = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \frac{1}{a^2+b^2} = \begin{pmatrix} a-b \\ b \end{pmatrix} \frac{1}{a^2+b^2}$.

Exercice 9.2. Si $L = x = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}^T \neq 0$ alors (cas injectif) $L^\dagger = (L^*L)^{-1}L^* = x^*/\|x\|_2^2$.

Exercice 9.3. A est un opérateur de rang 1, $(\mathrm{Ker}\, A)^\perp = \mathbb{C}y$ et $\mathrm{Im}\, A = \mathbb{C}x$. Soit $z \in \mathbb{C}^n$; sa projection orthogonale sur $\mathrm{Im}\, A$ est $\Pi_{\mathrm{Im}\, A} z = x\langle z, x\rangle / \|x\|_2^2$. On recherche $u \in (\mathrm{Ker}\, A)^\perp$ tel que $Au = \Pi_{\mathrm{Im}\, A} z$. On obtient $u = \frac{\langle z,x\rangle y}{\|x\|_2^2\|y\|_2^2} = \frac{yx^*z}{\|x\|_2^2\|y\|_2^2}$.

Exercice 9.4. Utiliser la proposition 9.4 avec $L = AA^*$ et $P = A^{*\dagger}A^\dagger$.

Exercice 9.5. Considérer une matrice nilpotente comme $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

Exercice 9.6. Non. Considérer la matrice A de l'exercice 9.3.

Exercice 9.7. 1. Puisque les valeurs propres de A^*A sont en nombre fini, il existe un disque centré en 0 qui ne contient pas de valeurs propres sauf éventuellement 0. Ainsi $tI_n + A^*A$ est inversible pour tout $t \neq 0$ dans ce disque. Idem pour AA^*. **2.** Par une décomposition en valeurs singulières $A = V\Sigma U^*$ on obtient $(tI_n + A^*A)^{-1}A^* = U(tI_n + \Sigma^*\Sigma)^{-1}\Sigma^*V^*$. Un calcul direct montre que $\lim_{t\to 0}(tI_n + \Sigma^*\Sigma)^{-1}\Sigma^* = \Sigma^\dagger$.

Exercice 9.8. $\mathrm{Im}\, A = \{y \in \mathbb{R}^2 : y_1 = y_2\}$. $\mathrm{Ker}\, A = \{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$, $(\mathrm{Ker}\, A)^\perp = \{x \in \mathbb{R}^2 : x_1 = x_2\}$, la projection de $b \in \mathbb{R}^2$ sur $\mathrm{Im}\, A$ est $\frac{b_1+b_2}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ dont l'antécédent dans $(\mathrm{Ker}\, A)^\perp$ est $\frac{b_1+b_2}{4}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Exercice 9.9. $\mathrm{Im}\, A = \left\{x_1 \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}^T + x_2 \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^T : x_1, x_2 \in \mathbb{R}\right\}$, $\mathrm{Ker}\, A = \{x \in \mathbb{R}^3 : x_1 = x_2 = 0\}$, $(\mathrm{Ker}\, A)^\perp = \{x \in \mathbb{R}^3 : x_3 = 0\}$, $\tilde{b} = \frac{1}{3}\begin{pmatrix} 2b_1 + b_2 + b_3 \\ b_1 + 2b_2 - b_3 \\ b_1 - b_2 + 2b_3 \end{pmatrix}$, les solutions du sytème $Ax = \tilde{b}$ sont $x = \frac{1}{3}\begin{pmatrix} b_1 + 2b_2 - b_3 \\ b_1 - b_2 + 2b_3 \\ x_3 \end{pmatrix}$ et la solution de norme minimale correspond à $x_3 = 0$.

Exercice 9.10. L'identité $(A+E)x' = b$ donne $A(x - x') = Ex'$ d'où $\|A(x - x')\|_2 \leqslant \|E\|_F \|x'\|_2$. On prend ensuite $E = A(x - x')x'^*/\|x'\|_2^2$ et l'on vérifie facilement que $(A + E)x' = b$ et que $\|E\|_F = \|A(x' - x)\|_2 / \|x'\|_2$.

Exercice 9.11. Les solutions $\tilde{x}$ de ce problème sont les antécédents par A de la projection orthogonale $\tilde{b}$ de b sur $\mathrm{Im}\, A$ pour le produit scalaire $\langle\ ,\ \rangle_S$. On a donc

$\tilde{b} = A\tilde{x}$ et $\langle \tilde{b} - b, Ax \rangle_S = 0$ pour tout $x \in \mathbb{R}^n$ c'est à dire $\langle A\tilde{x} - b, SAx \rangle = \langle A^*S(A\tilde{x} - b), x \rangle = 0$ pour tout $x \in \mathbb{R}^n$.

Exercice 9.12. 1. Si $M \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ alors $r + Ax = 0$ et $A^*r = 0$. On obtient $A^*(r + Ax) = A^*Ax = 0$ et donc $x = 0$ puisque rang $A = n$ (dans ce cas A^*A est inversible). Mais alors $r = r + Ax = 0$ et ceci prouve que M est inversible. **2.** $M \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$ s'écrit $r + Ax = b$ et $A^*r = 0$ d'où $A^*r + A^*Ax = A^*Ax = A^*b$ qui est l'équation normale du problème des moindres carrés $\inf_x \|Ax - b\|_2$. Réciproquement, si $A^*Ax = A^*b$, posant $r = b - Ax$, on a $A^*r = A^*b - A^*Ax = 0$ et donc $M \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$. **3.** Notons $\sigma_1 \geqslant \ldots \geqslant \sigma_n > 0$ les valeurs singulières de A. Si $M \begin{pmatrix} r \\ x \end{pmatrix} = \lambda \begin{pmatrix} r \\ x \end{pmatrix}$ on a $r + Ax = \lambda r$ et $A^*r = \lambda x$. Une solution de ce système est $\lambda = 1$, $x = 0$, $r \in \operatorname{Ker} A^* = (\operatorname{Im} A)^\perp$ qui est un espace de dimension $m - n$; $\lambda \neq 1$ conduit à $x \neq 0$ d'où $A^*(r + Ax) = \lambda A^*r$ et donc à $\lambda x + A^*Ax = \lambda^2 x$. Ceci prouve que $\lambda^2 - \lambda = \sigma_k^2$ donc que $\lambda = \left(1 \pm \sqrt{1 + 4\sigma_k^2}\right)/2$ ($2n$ valeurs propres). Comme M est hermitienne, ses valeurs singulières sont les valeurs absolues de ses valeurs propres : 1 et $|\lambda| = \left(\sqrt{1 + 4\sigma_k^2} \pm 1\right)/2$. La plus grande d'entre-elles est $\|M\|_2 = \left(\sqrt{1 + 4\sigma_1^2} + 1\right)/2$ et la plus petite est $\|M^{-1}\|_2 = 1$ si $\sigma_n \geqslant \sqrt{2}$ et $\left(\sqrt{1 + 4\sigma_n^2} - 1\right)/2$ sinon, d'où le conditionnement de M : $\operatorname{cond}_2(M) = \left(\sqrt{1 + 4\sigma_1^2} + 1\right)/2$ si $\sigma_n \geqslant \sqrt{2}$ et $\left(\sqrt{1 + 4\sigma_1^2} + 1\right) / \left(\sqrt{1 + 4\sigma_n^2} - 1\right)$ sinon.

Exercice 9.13. 1. $\inf_{x \in \mathbb{C}^n} \|Ax - b\|_2^2 + \rho \|x\|_2^2 = \inf_{x \in \mathbb{C}^n} \left\| \begin{pmatrix} A \\ \sqrt{\rho} I_n \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2$. L'équation normale de ce dernier problème est $(A^*A + \rho I_n)x = A^*b$ d'où $x_\rho = (A^*A + \rho I_n)^{-1} A^*b$. Noter que les valeurs propres de $A^*A + \rho I_n$ sont $\sigma_k^2 + \rho > 0$ où $\sigma_1^2 \geqslant \ldots \geqslant \sigma_n^2 \geqslant 0$ sont les valeurs propres de A^*A. **2.** $\|Ax_\rho - b\|_2^2 + \rho \|x_\rho\|_2^2 \leqslant \|Ax_{\rho'} - b\|_2^2 + \rho \|x_{\rho'}\|_2^2 \leqslant \|Ax_{\rho'} - b\|_2^2 + \rho' \|x_{\rho'}\|_2^2$. D'autre part $x_{\rho'} = (A^*A + \rho' I_n)^{-1}(A^*A + \rho I_n)x_\rho$. Les valeurs singulières de cette matrice sont $(\sigma_k^2 + \rho)/(\sigma_k^2 + \rho') \leqslant 1$, $1 \leqslant k \leqslant n$, donc $\|x_{\rho'}\|_2 \leqslant \|x_\rho\|_2$ et $\|Ax_\rho - b\|_2 \leqslant \|Ax_{\rho'} - b\|_2$. **3.** $\lim_{\rho \to 0} x_\rho = \lim_{\rho \to 0} (A^*A + \rho I_n)^{-1} A^*b = A^\dagger b = x_0$ par l'exercice 9.7.

Exercice 9.14. Raisonner comme dans l'exercice 9.12.

Exercice 10.1. 1. La matrice J de la méthode de Jacobi est égale à

$$J = \begin{pmatrix} 0 & -a_{12}/a_{11} \\ -a_{21}/a_{22} & 0 \end{pmatrix}$$

et celle de la méthode de Gauss-Seidel

$$G = \begin{pmatrix} 0 & -a_{12}/a_{11} \\ 0 & (a_{12}a_{21})/(a_{11}a_{22}) \end{pmatrix}.$$

2. Le rayon spectral de J est égal à $\sqrt{|(a_{12}a_{21})/(a_{11}a_{22})|}$. Pour la matrice G on a $\rho(G) = |(a_{12}a_{21})/(a_{11}a_{22})|$ et donc $\rho(J)^2 = \rho(G)$. Lorsque les deux méthodes sont convergentes, alors la méthode de Gauss-Seidel converge plus rapidement que la méthode de Jacobi. **3.** Après permutation des lignes et en supposant $a_{12}a_{21} \neq 0$, on a $\rho(J) = \sqrt{|(a_{11}a_{22})/(a_{12}a_{21})|}$ et $\rho(G) = |(a_{11}a_{22})/(a_{12}a_{21})|$. Les rayons spectraux sont les inverses des précédents. **4.** Faire un dessin. On remarque que pour Gauss-Seidel, les itérés successifs se situent sur la droite (D_2).

Exercice 10.2. Pour la première matrice, on a $\rho(J) = 0$ et $\rho(G) = 2$. La méthode de Jacobi est convergente et la méthode de Gauss-Seidel ne l'est pas. Pour la seconde matrice, on a $\rho(J) = \sqrt{5}/2$ et $\rho(G) = 1/2$. C'est le contraire qui se passe.

Exercice 10.3. 1. On a

$$J = \begin{pmatrix} 0 & -K \\ -K & 0 \end{pmatrix},$$

$$G = \begin{pmatrix} I_2 & 0 \\ -K & I_2 \end{pmatrix} \begin{pmatrix} 0 & -K \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -K \\ 0 & K^2 \end{pmatrix}$$

et

$$G_\omega = \begin{pmatrix} I_2 & 0 \\ -\omega K & I_2 \end{pmatrix} \begin{pmatrix} (1-\omega)I_2 & -\omega K \\ 0 & (1-\omega)I_2 \end{pmatrix} = \begin{pmatrix} (1-\omega)I_2 & -\omega K \\ \omega(\omega-1)K & \omega^2K^2 + (1- \end{pmatrix}$$

2. Pour la méthode de Jacobi, on a

$$x_{k+1,i} = \frac{x_{k,3} + x_{k,4}}{4} + \frac{1}{2}$$

pour $i = 1, 2$, et

$$x_{k+1,i} = \frac{x_{k,1} + x_{k,2}}{4} + \frac{1}{2}$$

pour $i = 3, 4$. Sachant que $x_{0,i} = 0$, on obtient $x_{k,i} = \sum_{l=1}^{k} (1/2)^l = (1-(1/2)^k)$ pour toutes les coordonnées i. Pour la méthode de Gauss-Seidel, les coordonnées $i = 1, 2$ suivent la récurrence

$$x_{k+1,i} = \frac{x_{k,3} + x_{k,4}}{4} + \frac{1}{2}$$

et les coordonnées $i = 3, 4$ la récurrence

$$x_{k+1,i} = \frac{x_{k,3} + x_{k,4}}{8} + \frac{3}{4}.$$

Les conditions initiales et cette récurrence montrent que $x_{k,i} = (1 - (1/2)^k)$ pour $i = 1, 2$, et $x_{k,i} = (1 - (1/4)^k)$ pour $i = 3, 4$. **3.** Les valeurs propres de J sont égales à $0, \pm 1/2$ et donc $\rho(J) = 1/2$. Pour G on obtient les valeurs propres 0, $1/4$ et donc $\rho(G) = 1/4$. **4.a** En utilisant les blocs de la matrice G_ω qui commutent entre eux on obtient le polynôme caractéristique $p_{G_\omega}(x)$ de G_ω comme déterminant de la matrice 2×2

$$(1 - \omega - x)^2 I_2 + (1 - \omega - x)\omega^2 K^2 + \omega^2(\omega - 2)K^2 = (1 - \omega - x)^2 I_2 - \omega^2(1 + x)K^2$$

(voir exercice 1.9). On a $p_{G_\omega}(x) = (1 - \omega - x)^2\,((1 - \omega - x)^2 - \omega^2(1 + x)/4)$ qui admet pour racines $1 - \omega$ (racine double) et les deux racines du polynôme $q(x) = x^2 - (2(1 - \omega) + \omega^2/4)x + (1 - \omega)^2$. **4.b** On étudie le signe du déterminant Δ du polynôme $q(x)$. On a $\Delta = (1 - \omega)\omega^2 + \omega^4/16$. Ses racines sont 0 (racine double) et $4(2 \pm \sqrt{3})$. La racine $\overline{\omega} = 4(2 - \sqrt{3})$ appartient à l'intervalle $]0, 2[$. Pour $\omega \in]0, \overline{\omega}[$ Δ est positif et les deux racines réelles du polynôme $q(x)$ sont $1 - \omega + \omega^2/8 \pm (\omega/2)\sqrt{1 - \omega + \omega^2/16}$. Soit $r = 1 - \omega + \omega^2/8 + (\omega/2)\sqrt{1 - \omega + \omega^2/16}$ la plus grande des deux racines. Pour $\omega \in]0, 1]$ on a $r \geqslant 1 - \omega$ et donc $\rho(G_\omega) = r$. Pour $\omega \in]1, \overline{\omega}]$ on vérifie que $r \geqslant \omega - 1$ et donc aussi $\rho(G_\omega) = r$. Pour $\omega \in]\overline{\omega}, 2[$ la racine complexe $s = 1 - \omega + \omega^2/8 + (\omega/2)\,i\,\sqrt{\omega - 1 - \omega^2/16}$ satisfait $|s| = \omega - 1$. On a donc $\rho(G_\omega) = \omega - 1$. **4.c** On vérifie que la fonction $\omega \mapsto r$ est décroissante sur l'intervalle $]0, \overline{\omega}[$. La valeur minimale de $\rho(G_\omega)$ est donc $1 - \overline{\omega} + \overline{\omega}^2/8 = \overline{\omega} - 1$.

Exercice 10.4. 1.a Évident à partir de l'égalité $Bx = \lambda x$. **1.b** La question précédente montre que $\{\lambda^2,\ \lambda \in \text{spec}\,(B)\} \subset \text{spec}\,(B_1 B_2)$, et ainsi $\rho(B)^2 \leqslant \rho(B_1 B_2)$. **1.c** Réciproquement, soit $\mu \in \text{spec}\,(B_1 B_2)$, $\mu \neq 0$ et λ tel que $\lambda^2 = \mu$. Il existe $x \neq 0$ tel que $B_1 B_2 x = \mu x$. Définissons $y = (1/\lambda)\,B_2 x$. On a $y \neq 0$ et $B_1 y = (1/\lambda)\,B_1 B_2 x = \lambda x$. λ vérifie donc $B_2 x = \lambda y$ et $B_1 y = \lambda x$. On voit que λ est valeur propre de B. On a donc $\rho(B_1 B_2) \subseteq \rho(B)^2$. **2.a** On a

$$J = \begin{pmatrix} 0 & D_1^{-1} A_1 \\ D_2^{-1} A_2 & 0 \end{pmatrix}$$

et donc $\rho(J) = \sqrt{\rho(D_1^{-1} A_1 D_2^{-1} A_2)}$ d'après la question précédente. **2.b** Par identification on obtient facilement l'expression de l'inverse :

$$\begin{pmatrix} D_1 & 0 \\ -\omega A_2 & D_2 \end{pmatrix}^{-1} = \begin{pmatrix} D_1^{-1} & 0 \\ \omega D_2^{-1} A_2 D_1^{-1} & D_2^{-1} \end{pmatrix}.$$

On obtient

$$G_\omega = \begin{pmatrix} (1 - \omega) I_n & \omega D_1^{-1} A_1 \\ \omega(1 - \omega) D_2^{-1} A_2 & \omega^2 D_2^{-1} A_2 D_1^{-1} A_1 + (1 - \omega) I_n \end{pmatrix}.$$

2.c Pour $\omega = 1$ on a

$$G_1 = G = \begin{pmatrix} 0 & D_1^{-1}A_1 \\ 0 & D_2^{-1}A_2D_1^{-1}A_1 \end{pmatrix}$$

et donc $\rho(G) = \rho(D_2^{-1}A_2D_1^{-1}A_1)$. **2.d** La méthode de Jacobi converge si et seulement si $\rho(J) = \sqrt{\rho(D_1^{-1}A_1D_2^{-1}A_2)} < 1$. Pour Gauss-Seidel la condition est $\rho(G) = \rho(D_2^{-1}A_2D_1^{-1}A_1) < 1$. On sait que pour toutes matrices A et B de $\mathbb{C}^{n\times n}$, on a $\rho(AB) = \rho(BA)$ (voir exercice 4.1). On a donc $\rho(D_2^{-1}A_2D_1^{-1}A_1) = \rho(D_1^{-1}A_1D_2^{-1}A_2)$ et les deux conditions sont équivalents. **2.e** Le calcul de $\det(G_\omega)$ est obtenu grâce au produit des déterminants des deux matrices blocs triangulaires qui constituent G_ω. On obtient $\det(G_\omega) = \det(D_1^{-1})\det(D_2^{-1})\det((1-\omega)D_1)\det((1-\omega)D_2) = (1-\omega)^{2n}$. Sachant que $|\det(G_\omega)| \leqslant \rho(G_\omega)^{2n}$ puisque le déterminant est le produit des valeurs propres, une condition nécessaire pour avoir la convergence est que $|1-\omega| \leqslant \rho(G_\omega) < 1$ c'est-à-dire $0 < \omega < 2$. **2.f** La condition donnée par le théorème 10.9 est que $(2-\omega)D_1/\omega$ et $(2-\omega)D_2/\omega$ soient définies positives ce qui est le cas puisque D_1 et D_2 sont des blocs diagonaux d'une matrice définie positive.

Exercice 10.5. 2.a Utilisons la décomposition $A = D - E - F$. On a $\hat{J} = D^{-1}(E+F)$ pour la matrice de Jacobi et $\hat{G} = (D-E)^{-1}F$ pour Gauss-Seidel. Le polynôme caractéristique de $\hat{J}$ est donné par $p_{\hat{J}}(x) = \det(D^{-1}(E+F) - xI_n) = \det(-D^{-1})\det(xD - E - F)$. Pour Gauss-Seidel, on a le polynôme $p_{\hat{G}}(x) = \det((D-E)^{-1}F - xI_n) = \det((E-D)^{-1})\det(xD - xE - F)$. Pour $\lambda \neq 0$ on a d'après la question 1.a précédente $\det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n \det(\lambda D - E - F)$. On voit donc que si $\mu \neq 0$ est une valeur propre de $\hat{G}$ alors ses deux racines carrées complexes $\pm\lambda$ sont valeurs propres de $\hat{J}$ et réciproquement, si $\lambda \neq 0$ est valeur propre de $\hat{J}$, alors λ^2 est valeur propre de $\hat{G}$. On a ainsi $\rho(\hat{G}) = \rho(\hat{J})^2$. **2.b** La convergence (ou divergence) simultanée des deux méthodes est évidente. Pour montrer que leur convergence implique que $A_n(s,a,b)$ est inversible il suffit d'observer que $A_n(s,a,b) = D(I_n - D^{-1}(E+F))$ et que $\rho(D^{-1}(E+F)) < 1$ implique $(I_n - D^{-1}(E+F))$ inversible (voir exercice 3.4). **3.** D'après l'exercice 1.13 on sait que les n valeurs propres de la matrice $C_n(a,b)$ sont données par $2\sqrt{ab}\cos(k\pi/(n+1))$ avec $k = 1,\ldots,n$ et $\sqrt{ab}$ une racine de ab. Les valeurs propres de la matrice $\hat{J}$ sont donc égales à $2(\sqrt{ab}/s)\cos(k\pi/(n+1))$ et $\rho(\hat{J}) = 2|\sqrt{ab}/s|\cos(\pi/(n+1))$. Les deux méthodes convergent si et seulement si $2|\sqrt{ab}/s|\cos(\pi/(n+1)) < 1$.

Exercice 10.6. 1. L'itération s'écrit $x_{k+1} = (I_n - \alpha A)x_k + \alpha b$. Notons $\lambda_1 \geqslant \ldots \geqslant \lambda_n > 0$ les valeurs propres de A. Les valeurs propres de la matrice $B_\alpha = (I_n - \alpha A)$ sont de la forme $1 - \alpha\lambda_i$ pour tout $i = 1,\ldots,n$. Le graphe des différentes fonctions $\alpha \mapsto |1 - \alpha\lambda_i|$ montre que pour $\alpha \in]0, 2/\lambda_1[$ la méthode est convergente. **2.** Le graphe

des fonctions montre que la valeur optimale de $\rho(B_\alpha)$ est obtenue pour α solution de $1-\alpha\lambda_n = \alpha\lambda_1 - 1$ c'est-à-dire $\alpha = 2/(\lambda_1+\lambda_n)$.

Exercice 11.1. 1. La dérivée de la fonction $\rho \mapsto \|x-(x_k-\rho(Ax_k-b))\|_A^2$ vaut $\langle A(x-x_k+\rho(Ax_k-b)), Ax_k-b\rangle$. À l'optimum on a donc $\rho_k \langle A(Ax_k-b), Ax_k-b\rangle = \langle Ax_k-b, Ax_k-b\rangle$ et $\rho_k = \langle r_k, r_k\rangle / \langle Ar_k, r_k\rangle$. Par ailleurs, la dérivée de la fonction $\rho \mapsto q(x_k-\rho(Ax_k-b))$ est égale à $-\langle A(x_k-\rho(Ax_k-b)), Ax_k-b\rangle + \langle b, Ax_k-b\rangle$ et on retrouve la valeur ρ_k à l'optimum. De l'égalité $x_{k+1} = x_k - \rho_k(Ax_k-b)$ on déduit $Ax_{k+1}-b = Ax_k-b-\rho_k A(Ax_k-b)$ et donc $\langle Ax_{k+1}-b, Ax_k-b\rangle = \langle Ax_k-b, Ax_k-b\rangle - \rho_k\langle A(Ax_k-b), Ax_k-b\rangle$. La valeur de ρ_k montre que $\langle Ax_{k+1}-b, Ax_k-b\rangle = 0$. **2.** À partir de $\|x-x_{k+1}\|_A^2 = \|x-x_k+\rho_k(Ax_k-b)\|_A^2$ et de la valeur de ρ_k on tire $\|x-x_{k+1}\|_A^2 = \|x-x_k\|_A^2+\rho_k^2\|Ax_k-b\|_A^2+2\rho_k\langle A(x-x_k), Ax_k-b\rangle$ et $\|x-x_{k+1}\|_A^2 = \|x-x_k\|_A^2 - \langle r_k, r_k\rangle^2/\langle Ar_k, r_k\rangle$. On remarque que $\|x-x_k\|_A^2 = \langle A(x-x_k), x-x_k\rangle = \langle r_k, A^{-1}r_k\rangle = \langle A^{-1}r_k, r_k\rangle$. On obtient donc l'égalité $\|e_{k+1}\|_A^2 = \|e_k\|_A^2\left(1-\langle r_k, r_k\rangle^2/(\langle Ar_k, r_k\rangle\langle A^{-1}r_k, r_k\rangle)\right)$. L'inégalité de Kantorovitch donne $\left(1-\langle r_k, r_k\rangle^2/(\langle Ar_k, r_k\rangle\langle A^{-1}r_k, r_k\rangle)\right) \leqslant \left(1-4/\left((\mathrm{cond}_2 A)\right)^{1/2}+(\mathrm{cond}_2 A)^{-1/2}\right)^2 = (\mathrm{cond}_2 A-1)^2/(\mathrm{cond}_2 A+1)^2$, qui implique le résultat. La majoration de l'erreur obtenue pour le gradient conjugué est de la forme $(\sqrt{\mathrm{cond}_2 A}-1)/(\sqrt{\mathrm{cond}_2 A}+1)$. On a $(\sqrt{x}-1)/(\sqrt{x}+1) \leqslant (x-1)/(x+1)$ pour tout $x \geqslant 1$. La vitesse de convergence méthode du gradient conjugué est donc supérieure à la vitesse de la méthode du gradient à pas optimal.

Exercice 11.2. 1. Les conditions d'optimalité sont $x_{k+1} = x_k + g_k$ avec $g_k \in G_k$ et $\nabla q(x_{k+1}) \perp G_k$ c'est-à-dire $Ax_{k+1}-b \perp G_k$. On a $Ax_k - b \in G_k$ donc $\nabla q(x_{k+1}) \perp \nabla q(x_k)$. **2.** Par récurrence sur j. Pour $j=0$, on a $G_0 = \mathcal{K}_1(A, r_0)$. Supposons que $G_j = \mathcal{K}_{j+1}(A, r_0)$ pour $j \leqslant k-1$. De l'égalité $x_{j+1} = x_j + g_j$ avec $g_j \in G_j$ et de $A\mathcal{K}_{j+1}(A, r_0) \subset \mathcal{K}_{j+2}(A, r_0)$ on déduit $Ax_{j+1}-b = Ax_j - b + Ag_j$ et donc $Ax_{j+1}-b \in \mathcal{K}_{j+2}(A, r_0)$. Puisque par hypothèse les vecteurs $Ax_j - b$ sont non nuls et orthogonaux entre eux d'après la question 1, on a $\dim G_{j+1} = j+2$. De $\dim \mathcal{K}_{j+2}(A, r_0) \leqslant j+2$ on déduit donc que $G_{j+1} = \mathcal{K}_{j+2}(A, r_0)$. On montre facilement par récurrence sur $j \geqslant 1$ que $x_j \in x_0 + \mathcal{K}_j(A, r_0)$. On retrouve donc les conditions (11.4) et (11.5) définissant de manière unique la solution du gradient conjugué.

Exercice 12.1. Posons $\delta(A, B) = \max_{x\in A}\min_{y\in B} d(x, y)$ de sorte que $hd(A, B) = \max(\delta(A, B), \delta(B, A))$. $\delta(A, B)$ et $hd(A, B)$ sont bien définis parce que A et B sont compacts et non vides. Prouvons l'inégalité triangulaire, les autres axiomes des distances sont faciles à vérifier. Prenons $x \in A$, $y \in B$ et $z \in C$ où A, B et C sont compacts et non vides. On a successivement : $d(x, z) \leqslant d(x, y)+d(y, z)$, $\min_{z\in C} d(x, z) \leqslant d(x, y) + \min_{z\in C} d(y, z)$, $\min_{z\in C} d(x, z) \leqslant d(x, y) + \max_{y\in B}\min_{z\in C} d(y, z)$, $\min_{z\in C} d(x, z) \leqslant \min_{y\in B} d(x, y) + \delta(B, C)$, $\delta(A, C) \leqslant \delta(A, B) + \delta(B, C)$. On montre de même que $\delta(C, A) \leqslant \delta(B, A) + \delta(C, B)$ d'où $hd(A, C) \leqslant hd(A, B) + hd(B, C)$.

Exercice 12.2. 1. La sphère $\mathbb{S}_X$ est compacte, non vide et x^*Ax est une fonction continue. **2.** $\max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} x^*Ax = \max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} (U^*x)^*D(U^*x) = \max_{Y\in\mathbb{G}_{n,i}} \min_{y\in\mathbb{S}_Y} y^*Dy$ en posant $y = U^*x$ et $Y = U^*X$. On a $U^*\mathbb{S}_X = \mathbb{S}_Y$ parce que U est unitaire. **3.** Lorsque $y_k = 0, i+1 \leqslant k \leqslant n$, et $\|y\|_2 = 1$ on a $y^*Dy = \sum_{k=1}^{i} \lambda_k |y_k|^2 \geqslant \lambda_i$. **4.** Notons $Z_i = \{y \in \mathbb{C}^n : y_k = 0,\ 1 \leqslant k \leqslant i-1\}$ qui est de dimension $n-i+1$ et soit $Y \in \mathbb{G}_{n,i}$. Puisque $\dim Y = i$ on a $Y \cap Z_i \neq \{0\}$ et il existe y de norme 1 dans cette intersection. On a $y^*Dy = \sum_{k=i}^{n} \lambda_k |y_k|^2 \leqslant \lambda_i$. On obtient ainsi, pour tout $Y \in \mathbb{G}_{n,i}$, $\min_{y\in\mathbb{S}_Y} y^*Dy \leqslant \lambda_i$ et il y a égalité pour le sous-espace $Y = Y_i$ décrit à la question 3. Ceci prouve que

$$\max_{Y\in\mathbb{G}_{n,i}} \min_{y\in\mathbb{S}_Y} y^*Dy = \lambda_i = \max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} x^*Ax$$

obtenu lorsque $U^*X = Y_i$ c'est-à-dire pour X engendré par des vecteurs propres $u_1, \ldots, u_i$ associés aux valeurs propres $\lambda_1, \ldots, \lambda_i$. **5.** $\lambda_1 = \max_{\|x\|_2=1} x^*Ax$, $\lambda_n = \min_{\|x\|_2=1} x^*Ax$. **6.** $\lambda_1(A)$ (resp. $\lambda_n(A)$) est convexe (resp. concave) parce que c'est l'enveloppe supérieure (resp. l'enveloppe inférieure) de la famille de formes linéaires $A \to x^*Ax$.

Exercice 12.3. 1. C'est une conséquence de l'exercice 12.2 : $\mu_i = \max_{X\in\mathbb{G}_{n,i}} \min_{x\in\mathbb{S}_X} x^*Bx$. Si $\tilde{X} \in \mathbb{G}_{n,i}$ réalise ce maximum on obtient $\mu_i = \min_{x\in\mathbb{S}_{\tilde{X}}} x^*Bx \leqslant x^*Bx$ pour tout $x \in \tilde{X}$, $\|x\|_2 = 1$, d'où $\mu_i \leqslant \min_{x\in\mathbb{S}_{\tilde{X}}} x^*Ax + \max_{\|x\|_2=1} x^*Ex \leqslant \lambda_i + \varepsilon_1$. **3.** Puisque $\lambda_i + \varepsilon_n \leqslant \mu_i \leqslant \lambda_i + \varepsilon_1$ on obtient $|\mu_i - \lambda_i| \leqslant \max_k |\varepsilon_k| = \|A - B\|_2$. **4.** Lorsque E est semi-définie positive on a $\varepsilon_n \geqslant 0$ et l'inégalité $\lambda_i + \varepsilon_n \leqslant \mu_i$ donne $\lambda_i \leqslant \mu_i$.

Exercice 13.1. Les sous-espaces de $\mathbb{R}^n$ invariants par la rotation de Givens $G(i, j, \theta)$ sont le plan de rotation $[e_i, e_j]$, les axes $[e_k]$, $k \neq i$ et j, et toute somme directe d'iceux. Les sous-espaces simples sont $[e_i, e_j]$ et son complémentaire orthogonal.

Exercice 13.3. Lorsque A et B sont diagonales $AQ - QB = C$ devient $\lambda_i q_{ij} - q_{ij}\mu_j = c_{ij}$ d'où $q_{ij} = c_{ij}/(\lambda_i - \mu_j)$. Dans le cas diagonalisable $AQ - QB = C$ devient $MD_\lambda M^{-1}Q - QND_\mu N^{-1} = C$ et donc $D_\lambda(M^{-1}QN) - (M^{-1}QN)D_\mu = M^{-1}CN$. Ainsi $Q = MRN^{-1}$ avec $r_{ij} = (M^{-1}CN)_{ij}/(\lambda_i - \mu_j)$.

Exercice 13.4. 1. La condition $\operatorname{spec} A \cap \operatorname{spec}(-A^*) = \varnothing$ est réalisée parce que A est stable. On utilise alors le théorème 13.9. **2.** Notons que $\frac{d}{dt}\exp(tA) = A\exp(tA) = \exp(tA)A$. On obtient : $\frac{d}{dt}\exp(tA)Q\exp(tA^*) = \exp(tA)(AQ + QA^*)\exp(tA^*) = \exp(tA)D\exp(tA^*)$ de sorte que $\int_0^T \exp(tA)D\exp(tA^*)dt = \int_0^T \frac{d}{dt}\exp(tA)Q\exp(tA^*)dt = \exp(TA)Q\exp(TA^*) - Q$. Nous allons montrer que $\lim_{T\to\infty}\exp(TA) = \lim_{T\to\infty}\exp(TA^*) = 0$ ce qui permettra de conclure. Pour ce faire on utilise la décomposition de Jordan : $A = PJP^{-1}$, $J = \operatorname{diag}(J_1, \ldots, J_p)$, $J_k = \lambda_k I_{n_k} + N_{n_k}$ (théorème 1.5). On a : $\exp(TA) = P\exp(TJ)P^{-1}$ et

$\|\exp(TA)\|_2 \leqslant \|P\|_2 \|P^{-1}\|_2 \max_k \|\exp(TJ_k)\|_2 = \|P\|_2 \|P^{-1}\|_2 \max_k |\exp(T\lambda_k)|$ $\|\exp(TN_{n_k})\|_2$. Comme $N_{n_k}^n = 0$ et que $\|N_{n_k}\|_2 \leqslant 1$ on a $\|\exp(TN_{n_k})\|_2 \leqslant \sum_{l=0}^{n-1} \frac{T^l}{l!}$ et donc, en résumé, $\|\exp(TA)\|_2 \leqslant \|P\|_2 \|P^{-1}\|_2 \max_k \exp(T\Re(\lambda_k)) \sum_{l=0}^{n-1} \frac{T^l}{l!}$. Il est clair que cette expression a pour limite 0 lorsque T tend vers l'infini parce que $\Re(\lambda_k) < 0$ pour tout k. Même chose pour $\exp(TA^*)$ puisque les parties réelles des valeurs propres de A et de A^* sont les mêmes. **6.** L'identité $H \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} Y \\ Z \end{pmatrix} \Lambda$ s'écrit $A^* = Y\Lambda Y^{-1}$ et $AZ + Z\Lambda = DY$. On peut prendre Y et Λ donnés par une décomposition de Schur de A^* et Z solution de l'équation de Sylvester $AZ + Z\Lambda = DY$. Sa solution est unique puisque spec $A \cap$ spec $(-\Lambda) = \varnothing$. On a alors $A(ZY^{-1}) + (ZY^{-1})A^* = (AZ + Z\Lambda)Y^{-1} = D$ et donc $Q = ZY^{-1}$ par unicité de la solution de l'équation de Lyapunov.

Exercice 14.1. Il revient au même d'étudier le comportement de la suite $e^{in\theta}$, $n \geqslant 0$. Cette suite est périodique si $\theta = \alpha\pi$ avec $\alpha \in \mathbb{Q}$. Lorsque $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, la suite est dense dans le cercle unité.

Exercice 14.2. Soit (x_k) une base de vecteurs propres de A. On écrit que $z = \alpha_p x_p + \ldots + \alpha_n x_n$ avec $1 \leqslant p \leqslant n$ et $\alpha_p \neq 0$. On a alors $z_k = A^k z / \|A^k z\|_2$, $A^k z = \alpha_p \lambda_p^k x_p + \ldots + \alpha_n \lambda_n^k x_n$ et l'on montre que $\lim_{k\to\infty} d_p(z_k, x_p) = 0$.

Exercice 14.3. **1.** $d(\mathcal{V}, \mathcal{W}) = \sup_{v\in\mathcal{V}, \|v\|_2=1} \inf_{w\in\mathcal{W}} \|v - w\|_2$. L'infimum est atteint lorsque $w = \Pi_{\mathcal{W}} v$ est la projection orthogonale de v sur $\mathcal{W}$. Ainsi $d(\mathcal{V}, \mathcal{W}) = \sup_{v\in\mathcal{V}, \|v\|_2=1} \|v - \Pi_{\mathcal{W}} v\|_2$ et c'est un maximum par compacité de la sphère unité. **2.** Comme $v - \Pi_{\mathcal{W}} v = \Pi_{\mathcal{W}^\perp} v$ on obtient $d(\mathcal{V}, \mathcal{W}) = \sup_{v\in\mathcal{V}, \|v\|_2=1} \|\Pi_{\mathcal{W}^\perp} v\|_2 = \sup_{\|z\|_2=1} \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}} z\|_2 = \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}}\|_2$. **3.** $d(\mathcal{W}^\perp, \mathcal{V}^\perp) = \|\Pi_{\mathcal{V}} \circ \Pi_{\mathcal{W}^\perp}\|_2 = \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}}\|_2$ parce que les normes d'un opérateur et de son adjoint sont les mêmes. **4.** $\mathcal{V}$ se décompose orthogonalement en $\mathcal{V} = (\mathcal{V} \cap \mathcal{W}) \oplus (\mathcal{V} \cap (\mathcal{V} \cap \mathcal{W})^\perp)$ et de même $\mathcal{W} = (\mathcal{V} \cap \mathcal{W}) \oplus (\mathcal{W} \cap (\mathcal{V} \cap \mathcal{W})^\perp)$. Soit $v \in \mathcal{V}$, $v = v_1 + v_2$ avec $v_1 \in (\mathcal{V} \cap \mathcal{W})$ et $v_2 \in (\mathcal{V} \cap (\mathcal{V} \cap \mathcal{W})^\perp)$. Il se projette sur $\mathcal{W}$ en $\Pi_{\mathcal{W}} v = v_1 + \Pi_{(\mathcal{W}\cap(\mathcal{V}\cap\mathcal{W})^\perp)} v_2$ et donc $v - \Pi_{\mathcal{V}} v = v_2 - \Pi_{(\mathcal{W}\cap(\mathcal{V}\cap\mathcal{W})^\perp)} v_2$. **5.** $0 \leqslant d(\mathcal{V}, \mathcal{W}) = \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}}\|_2 \leqslant \|\Pi_{\mathcal{W}^\perp}\|_2 \|\Pi_{\mathcal{V}}\|_2 \leqslant 1$. **6.** Par 1., $d(\mathcal{V}, \mathcal{W}) = 0$ si et seulement si pour tout $v \in \mathcal{V}$ il existe $w \in \mathcal{W}$ tel que $\|v - w\|_2 = 0$ c'est-à-dire si $\mathcal{V} \subset \mathcal{W}$. **7.** $d(\mathcal{V}, \mathcal{W}) = \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}} z\|_2$ pour un z de norme 1. Si $z \notin \mathcal{V}$ alors $\|\Pi_{\mathcal{V}} z\|_2 < \|z\|_2 = 1$ et donc $d(\mathcal{V}, \mathcal{W}) < 1$. De même, si $\Pi_{\mathcal{V}} z \notin \mathcal{W}^\perp$ on aura $d(\mathcal{V}, \mathcal{W}) < 1$. Ceci prouve que $d(\mathcal{V}, \mathcal{W}) = 1$ implique $\mathcal{W}^\perp \cap \mathcal{V} \neq \{0\}$. Réciproquement si $\mathcal{W}^\perp \cap \mathcal{V} \neq \{0\}$, pour un $z \in \mathcal{W}^\perp \cap \mathcal{V}$, $\|z\|_2 = 1$, on a $\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}} z = z$ donc $d(\mathcal{V}, \mathcal{W}) \geqslant \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}} z\|_2 = \|z\|_2 = 1$ d'où $d(\mathcal{V}, \mathcal{W}) = 1$. **8.** $d(\mathcal{V}_1, \mathcal{V}_3) = \left\|\Pi_{\mathcal{V}_3^\perp} \circ \Pi_{\mathcal{V}_1}\right\|_2 = \left\|\Pi_{\mathcal{V}_3^\perp} \circ (\Pi_{\mathcal{V}_2} + \Pi_{\mathcal{V}_2^\perp}) \circ \Pi_{\mathcal{V}_1}\right\|_2 \leqslant \left\|\Pi_{\mathcal{V}_3^\perp} \circ \Pi_{\mathcal{V}_2} \circ \Pi_{\mathcal{V}_1}\right\|_2 + \left\|\Pi_{\mathcal{V}_3^\perp} \circ \Pi_{\mathcal{V}_2^\perp} \circ \Pi_{\mathcal{V}_1}\right\|_2 \leqslant \left\|\Pi_{\mathcal{V}_3^\perp} \circ \Pi_{\mathcal{V}_2}\right\|_2 \|\Pi_{\mathcal{V}_1}\|_2 +$

$\left\|\Pi_{\mathcal{V}_3^\perp}\right\|_2 \left\|\Pi_{\mathcal{V}_2^\perp} \circ \Pi_{\mathcal{V}_1}\right\|_2 \leqslant \left\|\Pi_{\mathcal{V}_3^\perp} \circ \Pi_{\mathcal{V}_2}\right\|_2 + \left\|\Pi_{\mathcal{V}_2^\perp} \circ \Pi_{\mathcal{V}_1}\right\|_2 = d(\mathcal{V}_2, \mathcal{V}_3) + d(\mathcal{V}_1, \mathcal{V}_2)$.
9. Facile. **10.** Provient de 9. **11.** $d(\mathcal{V}_1 \oplus \mathcal{V}_2, \mathcal{W}) =$

$\|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}_1 \oplus \mathcal{V}_2}\|_2 = \|\Pi_{\mathcal{W}^\perp} \circ (\Pi_{\mathcal{V}_1} + \Pi_{\mathcal{V}_2})\|_2 \leqslant \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}_1}\|_2 + \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}_2}\|_2 =$

$d(\mathcal{V}_1, \mathcal{W}) + d(\mathcal{V}_2, \mathcal{W})$. De plus si $v_1 \in \mathcal{V}_1$ et $v_2 \in \mathcal{V}_2$ on a $\|\Pi_{\mathcal{W}^\perp}(v_1 + v_2)\|_2 \leqslant$

$\|\Pi_{\mathcal{W}^\perp} v_1\|_2 + \|\Pi_{\mathcal{W}^\perp} v_2\|_2 \leqslant \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}_1}\|_2 \|v_1\|_2 + \|\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}_2}\|_2 \|v_2\|_2 \leqslant$

$\max(d(\mathcal{V}_1, \mathcal{W}), d(\mathcal{V}_2, \mathcal{W}))(\|v_1\|_2 + \|v_2\|_2) \leqslant \max(d(\mathcal{V}_1, \mathcal{W}), d(\mathcal{V}_2, \mathcal{W}))\sqrt{2}\,\|v_1 + v_2\|_2$.

13. $d(\mathcal{V}, \mathcal{W})$ est la plus grande valeur singulière de $\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}} = \Pi_{\mathcal{V}} - \Pi_{\mathcal{W}} \circ \Pi_{\mathcal{V}}$ et, de la même manière, $d(\mathcal{W}, \mathcal{V})$ est la plus grande valeur singulière de $\Pi_{\mathcal{V}^\perp} \circ \Pi_{\mathcal{W}} = \Pi_{\mathcal{W}} - \Pi_{\mathcal{V}} \circ \Pi_{\mathcal{W}}$. Soit Q une transformation unitaire, involutive ($Q^2 = id_n$) et telle que $Q\mathcal{V} = \mathcal{W}$. On a $\Pi_{\mathcal{W}} = Q \circ \Pi_{\mathcal{V}} \circ Q$ donc $\Pi_{\mathcal{V}} - \Pi_{\mathcal{W}} \circ \Pi_{\mathcal{V}} = \Pi_{\mathcal{V}} - Q \circ \Pi_{\mathcal{V}} \circ Q \circ \Pi_{\mathcal{V}}$ et $\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}} \circ \Pi_{\mathcal{W}} = Q \circ (\Pi_{\mathcal{V}} - Q \circ \Pi_{\mathcal{V}} \circ Q \circ \Pi_{\mathcal{V}}) \circ Q$ de sorte que $\Pi_{\mathcal{W}^\perp} \circ \Pi_{\mathcal{V}}$ et $\Pi_{\mathcal{V}^\perp} \circ \Pi_{\mathcal{W}}$ ont les mêmes valeurs singulières. L'existence d'une involution qui « rabat » $\mathcal{V}$ sur $\mathcal{W}$ est admise. Dans le cas de droites on peut prendre une symétrie orthogonale. **14.** C'est une conséquence de 5, 6, 8 et 13.

Exercice 14.4. $T_+ = RQ + \mu I_n = Q^*(QR + \mu I_n)Q = Q^* T Q$ donc T_+ est hermitienne et unitairement semblable à T. Nous devons maintenant prouver que QR tridiagonale et hermitienne implique RQ tridiagonale. Un argument de continuité permet aussi de supposer que QR est inversible. Comme T est tridiagonale et R^{-1} triangulaire supérieure, $Q = TR^{-1}$ est une matrice de Hessenberg. Mais alors RQ est elle-aussi de Hessenberg et, puisqu'elle est hermitienne, elle est tridiagonale.

Exercice 14.5. Puisque $A - \mu I_n$ n'est pas inversible, on peut prendre $A - \mu I_2 = QR$ avec $R = \begin{pmatrix} u & v \\ 0 & 0 \end{pmatrix}$. Mais alors $RQ + \mu I_2 = \begin{pmatrix} u' & v' \\ 0 & \mu \end{pmatrix}$.

Exercice 15.1. 1. L'égalité est évidente pour $j = 1, \ldots, k-1$ puisque $\mathcal{K}_k(A, v)$ est le sous-espace vectoriel généré par les vecteurs $v, Av, A^2v, \ldots, A^{k-1}v$. Pour $j = k$ on a $P_k A^k v = P_k A A^{k-1} v = (P_k A P_k)\, P_k A^{k-1} v = (P_k A P_k)(P_k A P_k)^{k-1} v = (P_k A P_k)^k v$. Pour tout polynôme $p \in \mathcal{P}_k$, $p(x) = \sum_{j=0}^k a_j x^j$, on a $P_k p(A) v = \sum_{j=0}^k a_j P_k A^j v = \sum_{j=0}^k a_j (P_k A P_k)^j v = p(P_k A P_k) v$. **2.** Sachant que $P_k = Q_k Q_k^*$, on vérifie facilement que $(P_k A P_k)^j = Q_k (Q_k^* A Q_k)^j Q_k^*$ pour tout j. On a donc $P_k p_k(A) v = p_k(P_k A P_k) v = Q_k\, p_k(Q_k^* A Q_k)\, Q_k^* v$. D'après le théorème de Cayley-Hamilton 1.2 on a $p_k(Q_k^* A Q_k) = 0$ et donc $P_k p_k(A) v = 0$. **3.** On a $\langle p_k(A) v, P_k w \rangle = \langle P_k p_k(A) v, w \rangle$ pour tout $w \in \mathbb{C}^n$. Sachant que $P_k p_k(A) v = 0$ par la question précédente et que $P_k u = u$ pour tout $u \in \mathcal{K}_k(A, v)$, on a donc $\langle p_k(A) v, u \rangle = 0$ pour tout $u \in \mathcal{K}_k(A, v)$. L'espace affine $\tilde{\mathcal{P}}_k$ est égal à $\mathcal{P}_{k-1} + x^k$ où $\mathcal{P}_{k-1}$ est l'espace vectoriel des polynômes

de degré inférieur ou égal à $k-1$. La solution unique p du problème

$$\min_{p\in\tilde{\mathcal{P}}_k} \|p(A)v\|_2^2$$

est caractérisée par $p \in \tilde{\mathcal{P}}_k$ et $\langle p(A)v, q(A)v\rangle = 0$ pour tout $q \in \mathcal{P}_{k-1}$ autrement dit $\langle p(A)v, u\rangle = 0$ pour tout $u \in \mathcal{K}_k(A, v)$. Le polynôme $(-1)^k p_k$ satisfait ces conditions et est donc la solution de ce problème.

Exercice 16.1. 1. À partir de l'expression de A_2^{-1} on a donc $u_i = h^2\left(\sum_{j=1}^{i} j\,(1-ih)\,f_j + \sum_{j=i+1}^{n} i\,(1-jh)\,f_j\right)$ pour tout $i = 1,\ldots,n$. **2.** La solution u exprimée à partir du noyau de Green est donnée par $u(x) = \int_0^1 G(x,y)\,f(y)\,dy$. Si l'on discrétise cette intégrale par la méthode des trapèzes en utilisant la subdivision de $[0,1]$ par les points $y_j = jh$, $j = 0,\ldots,n+1$ et sachant que $G(x,y_0) = G(x,y_{n+1}) = 0$, on obtient, aux points $x_i = ih$, $u(x_i) = \int_0^1 G(x_i,y)\,f(y)\,dy \approx h\sum_{j=1}^{n} G(x_i,y_j)\,f(y_j)$. On a $G(x_i,y_j) = jh\,(1-ih)$ si $j \leqslant i$ et $G(x_i,y_j) = ih\,(1-jh)$ si $j > i$. En prenant $f_j = f(y_j)$, on obtient ainsi l'expression obtenue à la question 1.

Exercice 16.2. On voit que le coefficient jl de C est donné par $c_{[l-j]}$ où $[l-j]$ est le représentant compris entre 0 et $n-1$ de $l-j$ dans le groupe additif $\mathbb{Z}/n\mathbb{Z}$. Calculons le coefficient jk du produit $C\Phi$. On a $(C\Phi)_{jk} = \sum_{l=1}^{n} c_{[l-j]}\omega^{(l-1)(k-1)}$. Pour que la colonne k de Φ soit un vecteur propre de C il faut qu'il existe un scalaire λ_k tel que $(C\Phi)_{jk} = \lambda_k\,\omega^{(j-1)(k-1)}$, pour tout $j = 1,\ldots,n$. En multipliant les deux membres de cette égalité par $\omega^{-(j-1)(k-1)}$ on obtient $\lambda_k = \sum_{l=1}^{n} c_{[l-j]}\omega^{(l-j)(k-1)}$. On va montrer que cette somme ne dépend pas de j. Pour cela on partage la somme en deux : $\sum_{l=1}^{j-1}$ et $\sum_{l=j}^{n}$. On a d'une part $\sum_{l=j}^{n} c_{[l-j]}\omega^{(l-j)(k-1)} = \sum_{l=0}^{n-j} c_l\omega^{l\,(k-1)}$, d'autre part $\sum_{l=1}^{j-1} c_{[l-j]}\omega^{(l-j)(k-1)} = \sum_{l=1}^{j-1} c_{[n+l-j]}\omega^{(n+l-j)(k-1)} = \sum_{l=n-j+1}^{n-1} c_l\omega^{l\,(k-1)}$. En définitive $\lambda_k = \sum_{l=0}^{n-1} c_l\omega^{l\,(k-1)}$. Les valeurs propres λ_i de C sont données par

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \Phi \begin{pmatrix} c_0 \\ \vdots \\ c_{n-1} \end{pmatrix}.$$

Bibliographie

[1] ANDERSON E. AND AL., *LAPACK User's Guide*, second edition, SIAM, 1995.

[2] ARNOLDI W. E., *The principle of minimized iteration in the matrix eigenproblem*, Quart. Appl. Math., 9, pp. 17-29, 1951.

[3] BAI Z. AND AL., *Templates for the Solution of Algebraic Eigenvalue Problems - A Practical Guide*, SIAM, 2000.

[4] BAJARD J.-C., J.-M. MULLER, *Calcul et arithmétique des ordinateurs*, Lavoisier, 2004.

[5] BJÖRCK, A., *Numerical Methods for Least Squares Problem*, SIAM, 1996.

[6] BLUM L., F. CUCKER, M. SHUB, S. SMALE, *Complexity and Real Computation*, Springer Verlag, 1997.

[7] CHATELIN F., *Valeurs Propres de Matrices,* Masson, 1988.

[8] CHAITIN-CHATELIN F, V. FRAYSSÉ., *Lectures on Finite Precision Computations,* SIAM, Philadelphia, 1996.

[9] COOLEY J. W. AND J. W. TUKEY, *An Algorithm for the Machine Calculation of Complex Fourier Series*, Mathematics of Computation, Vol. 19, No. 90, pp. 297-301, 1965.

[10] FERNANDO K. V. AND B.N. PARLETT, *Accurate singular values and differential qd algorithms*, Numer. Math. 67, no. 2, pp. 191-229, 1994.

[11] FRANCIS J. G. F., *The QR transformation, part I and II*, Computer Journal, 4, pp. 265-271, pp. 332-345, 1961, 1962.

[12] GODEMENT R., *Cours d'algèbre,* troisième édition, Hermann, 1997.

[13] GOLDSTINE H. H., J. VON NEUMANN, *Inverting of Matrices of High Order.* In : J. Von Neumann, Collected Works, Vol. 5, Pergamon Press, 1963.

[14] GOLUB G., *Numerical methods for solving least squares problems*, Num. Math., 7, pp. 206-216, 1965.

[15] GOLUB G., C. VAN LOAN, *Matrix Computations*, third edition, The Johns Hopkins University Press, 1996.

[16] GRIFONE J., *Algèbre linéaire,* deuxième édition, Cepadues, 2002.

[17] HESTENES M. R. AND E. STIEFEL, *Methods of Conjugate Gradients for Solving Linear Systems,* J. Res. Natl. Bur. Stand. 49, pp. 409-436, 1952.

[18] HIGHAM N., *Accuracy and Stability of Numerical Algorithms*, SIAM, Second Edition, 2002.

[19] HOUSEHOLDER A., *The Theory of Matrices in Numerical Analysis*, Dover, 1964.

[20] KRYLOV A. N., *On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined*, Otdel. mat. i estest. nauk, VII, Nr.4, pp. 491-539, 1931 (in Russian).

[21] KUBLANOVSKAYA V. N., *On some algorithms for the solution of the complete eigenvalue problem*, USSR Comput. Math. Math. Phys., pp. 637-657, 1961.

[22] LANCZOS C., *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Natl. Bur. Stand. 45, pp. 225-280, 1950.

[23] LAURENT P.-J., *Approximation et optimisation*, Hermann, 1972.

[24] LEHOUCQ R. B. AND AL., *ARPACK Users' Guide : Solution of Large Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*, SIAM, 1998.

[25] MULLER J.-M., *Arithmétique des ordinateurs*, Masson, 1989.

[26] MÜNTZ C., *Solution directe de l'équation séculaire et de quelques problèmes analogues transcendants*, Comptes Rendus Acad. Sci. Paris, pp. 43-46, 1913.

[27] PARLETT B. N., *The new qd algorithms*, Acta Numerica, Cambridge Univ. Press, Vol. 4, pp. 459-491, 1995.

[28] RUTISHAUSER H., *Une méthode pour la détermination des valeurs propres d'une matrice*, Comptes Rendus Acad. Sci. Paris, 240, pp. 34-36, 1955.

[29] SAAD Y., *Iterative Methods for Sparse Linear Systems*, Second Edition, SIAM, 2003.

[30] SAAD Y., *Numerical Methods for Large Eigenvalues Problems : Theory and Algorithms*, John Wiley, 1992.

[31] SCHREIBER R., L. TREFETHEN, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl. 11, pp. 335-360, 1990.

[32] STEWART G. W., *Matrix Algorithms 1 : Basic Decompositions*, Cambridge University Press, 1998.

[33] STEWART G. W., *Matrix Algorithms 2 : Eigensystems*, Cambridge University Press, 2003.

[34] STEWART G. W., J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, 1990.

[35] TURING A., *Rounding-off Errors in Matrix Processes*, Quart. J. Mech. Appl. Math. 1, pp 287-308, 1948.

[36] WILKINSON J. H., *Rounding Errors in Algebraic Processes,* Prentice-Hall, 1964.

[37] WILKINSON J. H., *The Algebraic Eigenvalue Problem,* Clarendon Press, 1965.

Index

Symboles

$A(p:q,r:s)$ 2
$\mathcal{L}(E,F)$ 33
$\mathbb{C}^{m\times n}$ 1
$\mathbb{GL}_n$ 2
$\mathbb{GL}_n(\mathbb{C})$ 2
$\mathbb{GL}_n(\mathbb{R})$ 2
$\mathbb{G}_{np}$ 232
$\mathbb{O}_n$ 10
$\mathrm{P}_{n-1}(\mathbb{C})$ 226
$\mathbb{R}^{m\times n}$ 1
$\mathbb{SO}_n$ 10
$\mathbb{SU}_n$ 9
$\mathbb{St}_{mn}$ 107
$\mathbb{U}_n$ 9
$\rho(A)$ 35
3D-Var 275
4D-Var 275

A

arrondi 26
assimilation des données 271, 273

C

col 94
complément de Schur 15
conditionnement
    d'un problème 56
    d'une matrice 61
    d'une matrice rectangulaire 153
    inverse d'un problème 57
conditions de Petrov-Galerkin 251
contrôle optimal 273
creuse 263

D

décomposition
    Cholesky 96
    LU 73
    polaire 96
    QR 108
    Schur 15
    Schur ordonnée 16
    Schur réelle 16
    valeurs singulières 48
    valeurs singulières réduite 49
déflation 245
déterminant 3
différence finie
    centrée 260
    d'ordre deux 260
disques de Gershgorin 203
distance de Hausdorff 204
drapeau 241

E

ébauche 274
éléments finis 262
epsilon machine 27

équation
Lyapunov 220
normale 149
Poisson 259
Sylvester 217, 218
erreur inverse 56
espace
euclidien 8
hermitien 7
préhilbertien 8
préhilbertien complexe 7
projectif complexe 226
exponentielle de matrice 43

F

fonction résidu 146
formulation variationnelle 262
formule de Sherman-Morrison-Woodbury 15

G

Gershgorin 203
grassmannienne 232
groupe
linéaire 2
orthogonal 10
rotations 10
spécial orthogonal 10
spécial unitaire 9
unitaire 9

I

image 2
inégalité de Cauchy-Schwarz 7
interpolation polynomiale 265
inverse de Moore-Penrose 142
inverse généralisé 142

M

maillage 263
matrice
échelonnée 79
élémentaire 71
élimination 72
antisymétrique 11
bande 85
Cauchy 101, 223
circulante 270
compagnon 20
creuse 161, 195
définie négative 94
définie positive 87
diagonale par blocs 12
diagonale strictement dominante 44, 168
diagonalisable 4
Fourier 265
hermitienne 8
Hessenberg 121
Hessenberg non réduite 129
Hilbert 89
Householder 116
normale 10
orthogonale 10
par blocs 12
permutation 76
racine carrée 95
raideur 263
semblables 4
semi-définie positive 87
sous-matrice 2
stable 220
Stiefel 107
symétrique 9
transposition 76
triangulaire inférieure par blocs 12
triangulaire supérieure par blocs 12
triangularisable 5
tridiagonale 124
unitaire 9
Vandermonde 264
méthode
différences finies 260
directe 161
Gauss-Newton 272
GMRES 181

gradient à pas constant 179
gradient à pas optimal 201
gradient conjugué 181
Newton 271
plus grande pente 179
puissance 225
QR 236
QR avec stratégie du décalage 243
redémarrage 188, 254
Richardson 179
méthode itérative
consistante 162
convergente 162
Gauss-Seidel 166
Jacobi 166
par blocs 167
relaxation 167
relaxation symétrique 167
SOR 167
SSOR 167
moindres carrés 146
contraints 150
non-linéaires 272
régularisés 158
multiplicité algébrique 3
multiplicité géométrique 4

N

nœud du maillage 263
nombre flottant
base 25
double précision 27
exposant 25
mantisse 25
nombre 25
normalisé 26
précision 25
norme
1−norme 34
2−norme 36
∞−norme 35
consistante 33
d'opérateur 33
F-norme 38
Frobenius 38
multiplicative 33
spectrale 36
noyau 3
noyau de Green 259

O

opérateur
adjoint 8
hermitien 8
symétrique 9
orthonormalisation de Gram-Schmidt 109
overflow 26

P

Petrov-Galerkin 182
pivot
Gauss 74
partiel 77
total 77
point-selle 94
polynôme
annulateur 5
caractéristique 3
matriciel 5
préconditionnement
à droite 65
à gauche 64
à gauche et à droite 65
d'un système linéaire 64
problème d'évolution 273
procédé de réorthogonalisation 112
procédure de Rayleigh-Ritz 252
produit
hermitien 7
scalaire 8
produit par blocs 12
projecteur 11
projection orthogonale 11

Q

quadrique 91
quotient de Rayleigh 211, 252

R

rang 2
rayon spectral 35
résidu minimum 146
rotation de Givens 113

S

sous-espace caractéristique 5
sous-espace invariant 213
    complémentaire 214
    simple 214
sous-espace propre 4
spectre 3
splines cubiques naturelles 267
système
    sous-déterminé 146
    surdéterminé 146

T

trace 6

U

underflow 26
unité d'arrondi 26

V

valeur de Ritz 252
valeur propre 3
valeur singulière 47
vecteur d'état 273
vecteur de Ritz 252
vecteur propre 4

052085 - (I) - (1,2) - OSB 80° - PUB - MPN

Achevé d'imprimer sur les presses de
Snel
Z.I. des Hauts-Sarts - Zone 3
Rue Fond des Fourches 21 – B-4041 Vottem (Herstal)
Tél +32(0)4 344 65 60 - Fax +32(0)4 286 99 61
Août 2008 – 45772

Dépôt légal : septembre 2008

Imprimé en Belgique

SCIENCES SUP

Luca Amodei
Jean-Pierre Dedieu

ANALYSE NUMÉRIQUE MATRICIELLE

Cet ouvrage est destiné aux étudiants en Master de mathématiques appliquées, aux élèves ingénieurs, ainsi qu'aux candidats au CAPES ou à l'Agrégation.

Il propose un panorama des problèmes abordés en analyse numérique matricielle : normes sur les espaces de matrices, décompositions matricielles, méthodes directes ou itératives de résolution des systèmes linéaires, problèmes des valeurs propres. On y aborde les aspects théoriques de ces questions, l'algorithmique qui y est associée ainsi que les problèmes de complexité, de sensibilité aux erreurs et de stabilité. Le cours est illustré par des exercices corrigés qui mettent en œuvre les techniques introduites dans chaque chapitre.

JEAN-PIERRE DEDIEU

est professeur à l'Institut de Mathématiques de Toulouse. Ses travaux de recherche portent sur l'algèbre linéaire, l'optimisation, la résolution des système d'équations polynomiales et les problèmes de complexité sur les nombres réels.

LUCA AMODEI

est maître de conférences à l'Institut de Mathématiques de Toulouse. Ses centres d'intérêt sont l'algèbre numérique matricielle, la théorie de l'approximation, l'optimisation et le contrôle optimal.

MATHÉMATIQUES
PHYSIQUE
CHIMIE
SCIENCES DE L'INGÉNIEUR
INFORMATIQUE
SCIENCES DE LA VIE
SCIENCES DE LA TERRE

SMAI

LICENCE 1 2 3 | MASTER 4 5 | DOCTORAT 6 7 8

9 782100 520855

6656615
ISBN 978-2-10-052085-5

www.dunod.com

DUNOD