Algèbre linéaire avancée et numérique

Antoine Levitt

Université Paris-Saclay, L3DD, 2025-2026

Table des matières

1 Algèbre linéaire élémentaire **7**
- 1.1 Rappels . . . 7
  - 1.1.1 Notions acquises . . . 7
  - 1.1.2 Conventions, vocabulaire, rappels et propriétés élémentaires . . . 7
  - 1.1.3 Matrices 2×2 . . . 10
  - 1.1.4 Formes sesquilinéaires et quadratiques . . . 10
- 1.2 Calculer avec des matrices . . . 11
  - 1.2.1 Changements de base . . . 11
  - 1.2.2 Matrices par blocs . . . 12
  - 1.2.3 Complément de Schur . . . 12
  - 1.2.4 Coût de calcul . . . 13
- 1.3 Modéliser avec des matrices . . . 14
  - 1.3.1 Équations différentielles . . . 14
  - 1.3.2 Graphes . . . 15
  - 1.3.3 Probabilités . . . 15
  - 1.3.4 Électricité . . . 15
  - 1.3.5 Réseaux de neurones . . . 16
- 1.4 Exercices . . . 16

2 Systèmes linéaires et factorisations de matrices **18**
- 2.1 Caractère bien posé ou non . . . 18
- 2.2 Pivot de Gauss . . . 19
- 2.3 Factorisation de matrice : LU . . . 20
- 2.4 Variantes et alternatives au pivot de Gauss . . . 21
  - 2.4.1 Matrices creuses . . . 21
  - 2.4.2 Factorisation de Cholesky . . . 22
  - 2.4.3 Factorisation QR . . . 22
- 2.5 Moindres carrés . . . 23
- 2.6 Exercices . . . 24

3 Le théorème spectral, généralisations et applications **26**
3.1 Le théorème spectral . . . 26
3.2 Décomposition de Schur . . . 27
3.3 Formules min-max . . . 27
3.4 Classification des formes quadratiques . . . 28
3.5 Espaces propres et commutation . . . 29
3.5.1 Matrices normales . . . 29
3.6 La transformée de Fourier discrète . . . 30
3.6.1 Convolution cyclique et matrice circulante . . . 30
3.6.2 Diagonalisation des matrices circulantes . . . 31
3.6.3 La transformée de Fourier discrète . . . 32
3.7 Exercices . . . 33

4 Algèbre linéaire quantitative **35**
4.1 Valeurs singulières . . . 35
4.2 Normes matricielles . . . 36
4.2.1 Norme induite . . . 37
4.2.2 Norme de Frobenius . . . 39
4.3 Meilleure approximation de rang fixé . . . 39
4.4 Séries de Neumann . . . 40
4.5 Conditionnement . . . 41
4.6 Exercices . . . 42

5 Un petit détour par l'analyse complexe **45**
5.1 Formule de Stokes . . . 45
5.2 Fonction holomorphe . . . 46
5.3 Intégrale de contour . . . 46
5.4 Formule de Cauchy . . . 47
5.5 Exercices . . . 47

6 Calcul fonctionnel **49**
6.1 Calcul fonctionnel spectral . . . 49
6.2 Séries entières de matrices . . . 49
6.3 Calcul fonctionnel holomorphe . . . 50
6.4 Exercices . . . 52

7 Spectre et asymptotique **55**
7.1 Puissances d'une matrice . . . 55
7.2 Exponentielle de matrice . . . 56
7.3 Exercices . . . 56

8 Méthodes itératives stationnaires pour $Ax = b$ **57**
8.1 Méthode de Richardson . . . 57
8.2 Préconditionnement . . . 58
8.3 Exercices . . . 59

9 Méthodes de Krylov **60**
9.1 Espace de Krylov 60
9.2 Méthode de Galerkin 60
9.3 Méthodes de Krylov 61
9.4 Procédé d'Arnoldi 61
9.5 Structure de la matrice quand $X = Y$ et procédé de Lanczos 62
9.6 Méthode du gradient conjugué 63
9.7 Vitesse de convergence de la méthode du gradient conjugué 64
9.8 Exercices 65

10 Méthodes numériques pour $Ax = \lambda x$ **67**
10.1 La méthode de la puissance 67
10.2 La méthode d'Arnoldi 68
10.3 Exercices 69

Syllabus prévisionnel

Par séance :

1. Rappels, complément de Schur
2. Formes quadratiques, modélisation avec des matrices
3. Systèmes linéaires et moindres carrés
4. Théorème spectral
5. Transformée de Fourier discrète
6. SVD
7. Normes, séries de Neumann
8. Analyse complexe
9. Calcul fonctionnel, asymptotique
10. Méthode de Richardson
11. Méthodes de Krylov

Les exercices dont les numéros sont **en gras** sont des exercices d'application du cours et doivent être parfaitement maîtrisés.

Préface

Ce cours est un cours d'algèbre linéaire de niveau L3. Il s'adresse à des étudiants ayant déjà reçu un enseignement solide en algèbre linéaire (cf Section 1.1.1 pour les prérequis). L'objectif de ce cours est double. D'abord, compléter la formation en algèbre linéaire théorique (compléments de Schur, matrices normales, valeurs singulières, calcul fonctionnel, ainsi qu'un certain nombre d'applications ou de compléments). Ensuite, donner quelques notions d'algèbre linéaire numérique, c'est-à-dire le calcul effectif d'opérations d'algèbre linéaire (résolution de systèmes linéaires et de systèmes aux valeurs propres), ainsi que les outils théoriques pour les comprendre (notions de conditionnement, analyse de convergence de méthodes itératives...).

L'algèbre linéaire est un sujet central en mathématiques, avec des connections allant des mathématiques les plus pures jusqu'aux plus appliquées. En fonction des buts visés, différentes approches de l'algèbre linéaire peuvent être adoptées. En particulier, vous avez probablement été exposés jusqu'à présent à une vision *algébrique* de l'algèbre linéaire, qui s'attache à des notions de polynômes, de déterminants, de forme de Jordan, etc. Cette vision est importante en vue d'applications en algèbre (typiquement, montrer que la solution d'un problème linéaire à coefficients rationnels est rationnelle, ou faire de l'algèbre linéaire sur des corps finis). Dans ce cours, nous allons développer une vision complémentaire, plus *analytique*, avec des notions d'approximations (contrôlées), d'optimisation, de procédés itératifs, etc.

Cette vision analytique, qui évite en particulier les phénomènes *instables* (par exemple, la forme de Jordan d'une matrice peut changer qualitativement si on modifie un tout petit peu la matrice), est souvent pertinente pour les applications en modélisation (où il ne suffit pas qu'une notion soit bien définie pour être utile, il faut également qu'elle soit robuste aux perturbations), ainsi que pour les généralisations en dimension infinie que vous rencontrerez peut-être plus tard dans vos études (analyse fonctionnelle, théorie spectrale). Beaucoup d'outils algébriques peuvent être remplacés ou complétés par des outils purement analytiques ; à titre d'exemple, nous n'utiliserons pas dans ce cours une seule fois, même implicitement, la notion de déterminant, de polynôme caractéristique ou de forme de Jordan, et prouverons les théorèmes majeurs de l'algèbre linéaire (théorème spectral,

calcul fonctionnel, formule de Gelfand...) sans y faire référence.

Factorisation	Forme	Conditions	Utilisation (en plus de $Ax = b$)	Stable ?	Coût ($N = 5000$)
LU	$A = LU$ L tri. inf. U tri. sup.	Carrée Pivot de Gauss fonctionne sans besoin de pivot		Non	NA
LU pivoté	$PA = LU$ L tri. inf. U tri. sup. P permutation	Carrée		Oui	0.5s
Cholesky	$A = LL^*$ L tri. inf.	HDP	Test si HDP	Oui	0.3s
QR	$A = QR$ Q orthogonale R tri. sup.		Orthogonalisation Moindres carrés	Oui	1.5s
SVD	$A = USV^*$ U, V orthogonales S diagonale ≥ 0		Approx. faible rang Moindres carrés régularisés $\|A\|$ Valeurs sing. Vecteurs sing.	Oui	21s
Diagonalisation	$A = PDP^{-1}$ D diagonale P inversible	Carrée Diagonalisable	$f(A)$ valeurs propres vecteurs propres	Non (sauf normale)	25s
Jordan	$A = PJP^{-1}$ J blocs de Jordan P inversible	Carrée	$f(A)$	Non	NA
Schur	$A = PTP^*$ T tri. sup. P unitaire	Carrée	valeurs propres	Oui	18s

TABLE 1 – Résumé des principales factorisations de matrices utilisées dans ce cours, avec leurs conditions d'application et usages. Abbréviations : tri. = triangulaire, inf. = inférieure, sup. = supérieure, HDP = hermitienne définie positive. NA (non applicable) signifie que l'algorithme n'est pas implémenté dans les bibliothèques standard (parce que trop instable).

1 Algèbre linéaire élémentaire

1.1 Rappels

On insiste sur la distinction conceptuelle entre algèbre linéaire *abstraite* (espaces vectoriel, applications linéaires...) et *concrète* (vecteurs dans $\mathbb{C}^N$, matrices...), le lien étant établi par le choix d'une base. Dans ce document, on va adopter un point de vue concret (on va parler de vecteurs de $\mathbb{C}^N$ et de matrices carrées plutôt que de vecteurs de E et d'endomorphismes), principalement par commodité de vocabulaire, mais le point de vue plus abstrait est très utile pour interpréter les différentes opérations. La plupart des notions étudiées (à l'exception notable de factorisations de matrices comme la décomposition LU et QR) sont indépendantes du choix de base, et s'appliquent donc à des espace abstraits.

On travaillera exclusivement sur le corps de base $\mathbb{R}$ ou $\mathbb{C}$, et la plupart des notions étudiées dans ce cours ne s'étendent pas sur un autre corps de base. On énonce la plupart des résultats de ce cours dans le cas complexe parce qu'ils s'écrivent de la même façon que dans le cas réel, et que la structure complexe est importante pour un certain nombre d'outils (calcul fonctionnel holomorphe notamment). À quelques exceptions près, le lecteur moins intéressé par les nombres complexes peut sans rater grand chose remplacer dans son esprit $\mathbb{C}$ par $\mathbb{R}$ et * par ^{T}.

On se placera exclusivement dans le cas de la dimension finie, mais la plupart des résultats restent valable en dimension infinie, à condition de considérer des *opérateurs bornés* dans des espaces *complets* (espace de Banach et de Hilbert). Les exceptions notables concernent les résultats qui dépendent explicitement de la formulation matricielle du problème (comme le pivot de Gauss) et sont intrinsèquement liés à la dimension finie, ou ceux qui utilisent de façon plus ou moins explicite de la compacité (comme la décomposition en valeurs propres et singulières), et qui nécessitent des hypothèses supplémentaires pour se généraliser correctement.

1.1.1 Notions acquises

Les notions suivantes sont considérées comme acquises et ne feront pas l'objet de rappels :

— Espace vectoriel (sur $\mathbb{R}$ ou $\mathbb{C}$).
— Familles libres, génératrices, bases.
— Sous-espaces, espace engendré, complément orthogonal.
— Normes, produits scalaires. En dimension finie, toutes les normes sont équivalentes.
— Applications linéaires entre espaces vectoriels. Noyau, image, théorème du rang. Composition, inversibilité, transposée, trace. Projection orthogonale, matrices orthogonales.
— Représentation dans une base : vecteurs (colonnes), matrices.
— Valeurs et vecteurs propres, spectre.

On supposera également connues un certain nombre de notions d'analyse et d'optimisation : dérivée, intégration, dérivée sous le signe somme et intégral, théorème de Fubini, séries entières, rayon de convergence, calcul différentiel à plusieurs variables, gradient, hessienne, multiplicateurs de Lagrange...

1.1.2 Conventions, vocabulaire, rappels et propriétés élémentaires

On rappelle, en désordre, quelques notations et propriétés élémentaires. Le lecteur est encouragé à redémontrer les propriétés qu'il n'a pas vues ou dont il ne se souvient pas.

1. En algèbre linéaire, il est d'usage de noter les vecteurs par des lettres minuscules $(x, y, u, v, \dots)$, les matrices par des lettres majuscules $(A, B, P, D, \dots)$, et les scalaires par des lettres grecques $(\alpha, \beta, \lambda, \mu, \dots)$. Aussi, on se dispensera parfois de quantifier les objets ainsi que leurs dimensions, quand leur nature est claire par le contexte (par exemple, si on dit "on cherche à résoudre $Ax = b$", il est sous-entendu que A est une matrice et x, b des vecteurs).
2. Si $x \in \mathbb{C}^N$ est un vecteur, x_i dénote sa i-ième composante (implicitement, dans la base canonique). Il faudra bien prendre garde au risque de confusion lorsque cette notation est utilisée pour des collections de vecteurs : par exemple, si $(p_1, \dots, p_N)$ est une base de $\mathbb{C}^N$, dans une expression de type $y = \sum_{i=1}^N x_i p_i$, la quantité $x_i \in \mathbb{C}$ est un scalaire alors que $p_i \in \mathbb{C}^N$ est un vecteur.
3. On note $\mathbb{C}^{N\times M}$ l'ensemble des matrices de taille $N \times M$, qui s'identifie aux applications linéaires de $\mathbb{C}^M$ dans $\mathbb{C}^N$.
4. On notera $\mathrm{Span}(x_1, \dots, x_n)$ (parfois noté Vect ou Vec) l'espace vectoriel engendré par $x_1, \dots, x_n$. On notera $\mathrm{Im}(A)$ et $\ker(A)$ l'image et le noyau d'une matrice. $A \in \mathbb{C}^{N\times M}$ est injective ssi $\ker(A) = \{0\}$, et surjective ssi $\mathrm{Im}(A) = \mathbb{C}^N$
   On a le théorème du rang, pour une matrice $\mathbb{C}^{N\times M}$:
   $$\dim(\ker(A)) + \dim(\mathrm{Im}(A)) = M$$
   En conséquence, si A est une matrice carrée, A est inversible $\Leftrightarrow \ker(A) = \{0\} \Leftrightarrow \mathrm{Im}(A) = \mathbb{C}^N$.
5. Le produit scalaire, noté
   $$\langle x, y\rangle = x^* y = \sum_{i=1}^N \overline{x_i} y_i,$$
   conjugue à gauche. La convention de conjuguer à droite ou à gauche n'est pas fixée en mathématiques, mais elle l'est en physique, et c'est donc celle-ci que nous adopterons.
   Deux vecteurs sont orthogonaux si leur produit scalaire est nul. L'ensemble des vecteurs orthogonaux à un sous-espace E, noté $E^\perp$, est le *complément orthogonal*, un sous-espace de dimension égale à la dimension de l'espace ambiant moins la dimension de E. On a $(E^\perp)^\perp = E$ (en dimension finie).
6. La norme (euclidienne) d'un vecteur $x \in \mathbb{C}^N$ est $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^N |x_i|^2}$, et c'est celle qui sera utilisée si on ne précise pas qu'on utilise une autre. En dimension finie, toutes les normes sont équivalentes (la preuve utilise la compacité de la sphère unité).
7. La transposée d'une matrice est notée A^T.
8. $A^* = (\overline{A})^T$ (parfois noté $A^\dagger$ en physique) est la transconjuguée, ou adjoint. L'adjoint A^* (à ne pas confondre avec la *comatrice*, parfois aussi appelé adjoint) est l'unique matrice vérifiant
   $$\langle x, Ay\rangle = \langle A^* x, y\rangle$$
   pour tout x, y (prendre $x = e_i, y = e_j$). Cette forme met en évidence les liens très forts entre adjoint et produit scalaire, et c'est souvent elle qui est utile en pratique (plutôt que la définition $(A^*)_{ij} = \overline{A_{ji}}$).
   L'adjoint vérifie
   $$(AB)^* = B^* A^*$$
   pour tout A, B dont les tailles sont compatibles.

9. Un vecteur x est représenté par un *vecteur colonne*, une matrice $N \times 1$ (aussi appelé *ket* $|x\rangle$ en mécanique quantique). L'adjoint x^* est donc interprété comme un *vecteur ligne*, une matrice $1 \times N$ (une forme linéaire, un élément de l'espace dual, un *bra* $\langle x|$ en mécanique quantique). Ainsi, par exemple, x^*x est un scalaire (le produit scalaire entre x et lui-même), xx^* est une matrice (quand x est normalisé, c'est le projecteur orthogonal sur x). L'identité $(AB)^* = B^*A^*$ est encore valable pour des vecteurs compatibles, donc par exemple $x^*A = (A^*x)^*$.
10. Une matrice telle que $A^* = A$ est dite hermitienne : c'est la bonne généralisation des matrices réelles symétriques (les matrices complexes symétriques sont beaucoup moins intéressantes). De même, une matrice telle que $A^* = -A$ est anti-hermitienne ; c'est la bonne généralisation des matrices antisymétriques.
    Si A est hermitienne, la forme quadratique
    $$\langle x, Ax\rangle = \langle Ax, x\rangle = \overline{\langle x, Ax\rangle}$$
    est toujours réelle. En conséquence, les valeurs propres d'une matrice hermitienne sont toujours réelles : si $Ax = \lambda x$, alors $\langle x, Ax\rangle = \lambda\|x\|^2$ et donc $\lambda \in \mathbb{R}$.
    Si A est hermitienne, deux vecteurs propres associés à deux valeurs propres λ et μ différentes sont orthogonaux : si $Ax = \lambda x, Ay = \mu y$, alors $\langle x, Ay\rangle = \mu\langle x, y\rangle = \lambda\langle x, y\rangle$ donc $\langle x, y\rangle = 0$. Ce n'est pas nécessairement le cas pour des valeurs propres confondues : par exemple, n'importe quel vecteur non-nul est vecteur propre de la matrice identité.
11. La matrice identité en dimension N, parfois notée I_N, sera simplement notée 1 quand sa dimension est claire par le contexte. De même, on utilisera des expressions de type $\lambda - A$ pour dénoter $\lambda I_N - A$ si A est une matrice carrée de taille N.
12. On notera parfois $\frac{1}{A} = A^{-1}$, voire même des expressions de type $\frac{A}{\lambda - A} = A(\lambda - A)^{-1} = (\lambda - A)^{-1}A$. Il est en revanche très important de ne pas noter $\frac{A}{B}$ dans le cas où A et B ne commutent pas, car on ne sait pas s'il s'agit de AB^{-1} ou $B^{-1}A$.
13. On écrit $\mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ pour la matrice diagonale avec les éléments $\lambda_1, \ldots, \lambda_N$ sur sa diagonale.
14. Une matrice (non nécessairement carrée) *orthogonale*, ou *isométrie*, est une matrice à colonnes orthonormées, c'est-à-dire que $A^*A = 1$. Les matrices orthogonales sont exactement les matrices qui préservent le produit scalaire ($\langle Ax, Ay\rangle = \langle x, y\rangle$ pour tout x, y), ou, ce qui est équivalent par polarisation, la norme ($\|Ax\| = \|x\|$ pour tout x).
    Une matrice orthogonale carrée est dite *unitaire* (c'est-à-dire que $A^* = A^{-1}$).
15. Une matrice *hermitienne définie positive* (HDP ; on dit aussi symétrique définie positive, SDP, dans le cas réel) est une matrice A hermitienne telle que $x^*Ax \geq 0$ pour tout x, avec $x^*Ax = 0$ seulement si $x = 0$. On montrera au chapitre 3 que A est HDP si et seulement si toutes ses valeurs propres sont strictement positives. Une matrice vérifiant seulement $x^*Ax \geq 0$ (valeurs propres positives ou nulles) est dite semi-définie positive. En particulier, la matrice $A = B^*B$ est toujours semi-définie positive, car $x^*Ax = \|Bx\|^2$; elle est définie positive si et seulement si B est injective.
16. $\sigma(A)$ est le *spectre* de A, l'ensemble de ses valeurs propres.
17. Si une matrice carrée de taille N a N vecteurs propres indépendants, elle est dite *diagonalisable*. Des vecteurs propres associés à des valeurs propres distinctes sont nécessairement indépendants ; ainsi, si une matrice a N valeurs propres distinctes, elle est diagonalisable.

1.1.3 Matrices 2×2

Une matrice 2×2 est de forme

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Sa trace est $a + d$, son déterminant $ad - bc$, son inverse

$$\frac{1}{ad - bc} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}.$$

Ses valeurs propres peuvent souvent se calculer en utilisant les relations $\lambda_1 + \lambda_2 = a + d$, $\lambda_1 \lambda_2 = ad - bc$. Les matrices orthogonales réelles sont soit de forme $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ avec $\theta \in \mathbb{R}$ (matrice de rotation), soit de forme $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ (matrice de réflexion).

1.1.4 Formes sesquilinéaires et quadratiques

Une forme sesquilinéaire sur $\mathbb{C}^N$ est une application $\mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C}$ qui est linéaire par rapport à sa deuxième coordonnée et antilinéaire par rapport à sa première (c'est-à-dire que $f(\alpha x, y) = \overline{\alpha} f(x, y)$). Dans le cas réel, c'est une forme bilinéaire. Une forme sesquilinéaire est dite hermitienne (dans le cas réel, symétrique) si $b(x, y) = \overline{b(y, x)}$. Une forme quadratique (à valeurs réelles) est une application $\mathbb{C}^N \to \mathbb{R}$ de forme $q(x) = b(x, x)$ avec b une forme sesquilinéaire telle que $b(x, x) \in \mathbb{R}$ pour tout $x \in \mathbb{C}^N$. Une forme sesquilinéaire hermitienne fournit évidemment une forme quadratique via $q(x) = b(x, x)$: on peut en fait montrer qu'à toute forme quadratique est associée une forme sesquilinéaire hermitienne, en utilisant l'identité de polarisation

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

(dans le cas réel, $b(x, y) = \frac{1}{2}(q(x + y) - q(x) - q(y))$).

Une forme sesquilinéaire est représentée par une unique matrice : pour toute forme sesquilinéaire b, il existe une matrice A telle que

$$b(x, y) = x^* A y = \sum_{ij} A_{ij} \overline{x_i} x_j$$

avec $A_{ij} = b(e_i, e_j)$. Une forme quadratique (à valeurs réelles) est représentée par une unique matrice *hermitienne* sous la forme

$$q_A(x) = x^* A x = \sum_{ij} A_{ij} \overline{x_i} x_j \in \mathbb{R}.$$

On dit que q_A est définie positive si $q_A(x) \geq 0$ pour tout $x \in \mathbb{C}^N$, et $q_A(x) = 0$ seulement si $x = 0$. q_A est positive définie si et seulement si A est HDP. Si b est une forme sesquilinéaire dont la forme quadratique associée est définie positive, alors b définit un produit scalaire.

Dans le reste de cette section, on va spécialiser la discussion au cas réel. Dans ce cas, une forme quadratique $q_A : \mathbb{R}^N \to \mathbb{R}$ est une application polynomiale en les entrées de x, donc infiniment dérivable. On a

$$\begin{aligned} q_A(x+h) &= q_A(x) + x^T Ah + h^T Ax + q_A(h) \\ &= q_A(x) + 2(Ax)^T h + O(\|h\|^2)) \end{aligned}$$

donc son gradient (l'unique vecteur g_x tel que sa différentielle, l'application linéaire $h \mapsto dq_A(x) \cdot h$, est de forme $g_x^T h$) vaut $\nabla q_A(x) = 2Ax$. Sa Hessienne (la jacobienne de l'application $x \mapsto \nabla q_A(x)$, ou, de façon équivalente, l'unique représentation de la forme quadratique correspondant au terme $O(\|h\|^2)$ sous la forme $\frac{1}{2}h^* Bh + O(\|h\|^3)$ *avec B hermitienne*) est égale à $2A$. Si A est une matrice SDP, q_A est donc strictement convexe.

1.2 Calculer avec des matrices

1.2.1 Changements de base

Strictement parlant, une matrice est un tableau bidimensionnel de nombres et une application linéaire est une application entre deux espaces vectoriels. Les deux objets sont en correspondance bijective *quand une base est choisie* (par défaut, la base canonique). On s'autorisera des abus de langage comme "l'expression de la matrice carrée A dans la base $u_1, \ldots, u_N$" (sous-entendu : l'expression matricielle dans la base $(u_1, \ldots, u_N)$ de l'application linéaire définie par A dans la base canonique de $\mathbb{C}^N$).

Il est utile de reconnaître à vue l'interprétation en termes de changement de base de certains types de produits de matrices :

- Un produit de type $x = Py$ avec P une matrice carrée inversible peut être interprété comme $x = \sum_{i=1}^N y_i p_i$, avec $p_1, \ldots, p_N$ les colonnes de P. Si $(y_1, \ldots, y_N)$ sont les composantes d'un vecteur abstrait dans la base $(p_1, \ldots, p_N)$, $(x_1, \ldots, x_N)$ sont ses composantes dans la base canonique. Une expression $x = Py$ traduit donc un passage de la base $(p_1, \ldots, p_N)$ à la base canonique.
- De même, un produit de type $y = P^{-1}x$ traduit un passage de la base canonique à la base $(p_1, \ldots, p_N)$. Pour ne pas confondre : si je double les vecteurs de la base, $P' = 2P$, et que x ne change pas, les composantes de y dans la base P' sont réduits de moitié : $y' = y/2$, donc c'est bien $y = P^{-1}x$ et pas le contraire.
- En application des points précédents, si une application linéaire a pour matrice B dans la base donnée par les vecteurs de P, elle a pour matrice PBP^{-1} dans la base canonique (cette expression se lit de droite à gauche : on passe dans la base des vecteurs propres, on multiplie chacune des composantes, et on revient dans la base canonique).
- En particulier, si A est diagonalisable, elle agit diagonalement sur ses vecteurs propres, donc si on les range dans une matrice P, alors $A = PDP^{-1}$ avec D une matrice diagonale.
- Une autre façon d'écrire $y = P^{-1}x$ est $y = Q^*x$, avec $Q = (P^{-1})^*$. Les colonnes $q_1, \ldots, q_N$ de Q vérifient $p_i^* q_j = \delta_{ij}$ et $(q_1, \ldots, q_N)$ est appelée la *base duale*.
- En particulier, dans le cas où P est unitaire, $P^* = P^{-1}$ et la base duale est égale à la base elle-même. Le passage de la base canonique à la base $(p_1, \ldots, p_N)$ s'écrit $y = P^*x$, ce qui signifie $y_i = p_i^* x$: les composantes dans une base orthonormée sont extraites par produit scalaire avec les vecteurs de la base.

— Le caractère hermitien est préservé par changement de base *orthogonale* : si $A = PBP^*$ avec P orthogonale et B hermitienne, alors A est hermitienne. Ce n'est pas le cas pour un changement de base quelconque.

1.2.2 Matrices par blocs

Il est souvent utile de décomposer une matrice par *blocs*. Par exemple, une matrice $A = \mathbb{C}^{N\times M}$ peut être vue comme une collection de ses colonnes ou de ses lignes :

$$A = \begin{pmatrix} u_1 & u_2 & \cdots & u_M \end{pmatrix} = \begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_N^T \end{pmatrix}$$

où les vecteurs u et v sont représentés en colonnes. Le produit matriciel *à gauche* est compatible avec la décomposition par *colonnes* au sens suivant :

$$B \begin{pmatrix} u_1 & u_2 & \cdots & u_M \end{pmatrix} = \begin{pmatrix} Bu_1 & Bu_2 & \cdots & Bu_M \end{pmatrix} =$$

De même, le produit matriciel *à droite* est compatible avec la décomposition par *lignes* au sens suivant :

$$\begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_N^T \end{pmatrix} B = \begin{pmatrix} v_1^T B \\ v_2^T B \\ \vdots \\ v_N^T B \end{pmatrix}$$

Cela conduit à l'interprétation suivante du produit de matrices : la multiplication à gauche $A \mapsto BA$ agit sur chaque colonne de A indépendamment, la multiplication à droite $A \mapsto AB$ agit sur chaque ligne de A indépendamment.

On peut également avoir des décompositions par des sous-matrices :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

où l'ensemble $1, \ldots, N$ est partitionné en deux sous-ensembles $(1, \ldots, N_1), (N_1 + 1, \ldots, N)$ (et de même pour M), et plus généralement avec un nombre quelconques de sous-blocs. Cette décomposition est compatible avec la multiplication au sens où le produit $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ se calcule avec les règles habituelles de la multiplication de matrices 2×2, à la différence que les quantités multipliées sont elles-mêmes des matrices.

1.2.3 Complément de Schur

On profite de l'introduction des matrices par blocs pour définir le complément de Schur, qui unifie des techniques souvent utiles dans les applications.

Proposition 1.1. Si la matrice carrée A est partitionnée comme

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

et que A_{11} est inversible, alors A est inversible si et seulement si le *complément de Schur* $A_{22} - A_{21}A_{11}^{-1}A_{12}$ l'est.

Démonstration. On écrit $Ax = b$ sous la forme

$$\begin{cases} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2 \end{cases}$$

Formellement, comme A_{11} est inversible, on peut résoudre x_1 en fonction de x_2 grâce à la première équation : $x_1 = A_{11}^{-1}(b_1 - A_{12}x_2)$, et remplacer dans la deuxième pour obtenir

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 = b_2 - A_{21}A_{11}^{-1}b_1.$$

Grâce à ces manipulations formelles on montre facilement que, si le complément de Schur est inversible, alors $x_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}(b_2 - A_{21}A_{11}^{-1}b_1)$ et $x_1 = A_{11}^{-1}(b_1 - A_{12}x_2)$ résolvent le système $Ax = b$. Réciproquement, on procède par contraposée : supposons que le complément de Schur ne soit pas inversible, de sorte qu'il existe un $x_2 \neq 0$ tel que $(A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 = 0$. Alors $A \begin{pmatrix} -A_{11}^{-1}A_{12}x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} A_{11}(-A_{11}^{-1}A_{12}x_2) + A_{12}x_2 \\ (A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 \end{pmatrix} = 0$, donc A n'est pas inversible. □

Cette méthode consiste simplement à utiliser la première moitié des équations pour résoudre la première moitié des inconnues en fonction de la deuxième moitié, et à ensuite remplacer dans la deuxième moitié des équations pour obtenir un système fermé sur la deuxième moitié des inconnues. En d'autres termes, la deuxième moitié des inconnues obéit à un système linéaire $\widetilde{A}x_2 = \widetilde{b}$, où $\widetilde{A} = A_{22} - A_{21}A_{11}^{-1}A_{12}$ et $\widetilde{b} = b_2 - A_{21}A_{11}^{-1}b_1$ sont modifiés indirectement par l'effet des variables du bloc 1. Les deux blocs peuvent être de taille arbitraire. On peut évidemment procéder dans l'autre sens (résoudre la deuxième moitié en fonction de la première).

1.2.4 Coût de calcul

Les opérations sur des vecteurs, ou les opérations type produit scalaire, ont une complexité en $O(N)$, où N est la dimension des objets manipulés. Les multiplications matrice-vecteur ont une complexité $O(N^2)$. Les multiplications matrice-matrice (pour deux matrices $N \times N$) ont une complexité $O(N^3)$ quand calculées par la formule naive. Il est cependant possible de faire mieux ! Le meilleur algorithme connu à l'heure actuelle a une complexité de $O(N^{2.373})$, et la complexité algorithmique exacte du problème de multiplication de matrices est encore un problème ouvert. Malheureusement, ces algorithmes sont beaucoup plus coûteux (la constante cachée dans le "O" est énorme) et ne sont quasiment pas utilisés en pratique.

Même si la complexité est la même, le temps de calcul effectif peut varier énormément en fonction de l'implémentation. Par exemple, une multiplication de matrice codée simplement par une triple boucle en Python sera environ quatre ordres de grandeur (!) plus lente qu'un appel aux

routines optimisées (`A @ B` en Python). Une raison à cela est que le langage Python est lent : la même triple boucle dans un langage de bas niveau comme le C sera à peu près deux ordres de grandeur plus rapide. Mais il reste une différence de deux ordres de grandeur. Cette différence s'explique par plusieurs facteurs, dont les plus importants sont l'utilisation de la vectorisation et la gestion des caches (sur les ordinateurs modernes, les accès mémoires sont lents par rapport à la vitesse de calcul, et il faut donc les minimiser). Pour avoir un ordre de grandeur, retenons qu'une multiplication de matrice $10,000 \times 10,000$ par des routines optimisées prend une dizaine de secondes sur un ordinateur de bureau classique.

1.3 Modéliser avec des matrices

Vous avez probablement déjà rencontré l'utilisation élémentaire de matrices en mathématiques, notamment via les systèmes dynamiques linéaires discrets

$$x_{n+1} = Ax_n$$

et continus

$$\dot{x}(t) = Ax(t).$$

Par exemple, l'étude de suites récurrentes linéaires de type $x_{n+1} = \alpha x_n + \beta x_{n-1}$ peut se reformuler en $\begin{pmatrix} x_n \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} x_{n-1} \\ x_n \end{pmatrix}$ et l'étude de cette matrice (notamment ses valeurs propres) donne accès au comportement de la suite. De même, l'étude d'un système dynamique $\dot{x} = f(x)$ autour d'un point d'équilibre x_* se fait en étudiant son équation linéarisée, donc les propriétés de la matrice jacobienne $f'(x_*)$. Voyons maintenant quelques autres exemples.

1.3.1 Équations différentielles

Considérons le problème suivant (*problème aux limites*, un modèle simple d'une équation aux dérivées partielles *elliptique*) :

$$-u''(x) = f(x), \quad u(0) = u(1) = 0$$

pour une fonction $f \in C^0([0,1])$ donnée. Ce problème peut par exemple modéliser, dans le régime des petites déformations, l'état d'équilibre d'une corde, avec $u(x)$ la hauteur de la corde, et $f(x)$ le chargement au point x (par exemple, pour une corde soumise à son poids, $f(x) = -1$). L'étude de ce type de problèmes, où les conditions sont données aux limites du domaine, est plus subtile que l'étude de problèmes *de Cauchy* où la donnée initiale est spécifiée (de type $-u'' = f, u(0) = \alpha, u'(0) = \beta$), et il n'est pas du tout évident que ce problème soit bien posé (par exemple, le problème $-u'' - 2\pi u = f$ avec les mêmes conditions aux limites est mal posé puisqu'on peut toujours ajouter $\sin(2\pi x)$ à une solution). Néanmoins, ce problème-ci est bien posé (vous rencontrerez peut-être son étude mathématique dans la suite de vos études).

Pour résoudre ce problème numériquement, on emploie souvent la *méthode des différences finies* : on se donne un maillage uniforme $x_1 = \frac{1}{N}, \dots, x_{N-1} = 1 - \frac{1}{N}$, et on approxime

$$u''(x_n) \approx \frac{u(x_{n-1}) - 2u(x_n) + u(x_{n+1})}{(1/N)^2}$$

avec la convention $u(x_0) = u(x_N) = 0$. On aboutit alors à l'équation

$$-\frac{u(x_{n-1}) - 2u(x_n) + u(x_{n+1})}{(1/N)^2} = f(x_n)$$

qui est une équation *linéaire* pour les inconnues $u(x_n)$. Si on pose $u_n = u(x_n)$, $n = 1, \ldots, N-1$, on aboutit à l'équation

$$N^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{N-1}) \end{pmatrix},$$

un problème de type "$Ax = b$" avec $x = (u_1, \ldots, u_{N-1})$.

1.3.2 Graphes

On rappelle qu'un graphe est une collection de N noeuds, et d'arêtes joignant ces noeuds. Un graphe peut être représenté par sa *matrice d'adjacence*, une matrice $N \times N$ dont l'entrée (i, j) vaut 1 si une arête joint i et j, et 0 sinon. Cette matrice est commode pour formaliser un certain nombre d'algorithmes de graphes. Par ailleurs, ses propriétés en termes d'algèbre linéaire (par exemple ses valeurs propres) sont utiles dans plusieurs contextes. C'est également vrai pour d'autres matrices comme la matrice laplacienne, égale à la matrice d'adjacence moins la matrice diagonale dont chaque élément est la connectivité du noeud associé. On peut également modéliser par des matrices des graphes pondérés (où un coût est associé à chaque noeud).

1.3.3 Probabilités

L'algèbre linéaire intervient de façon naturelle dans plusieurs domaines des probabilités. Les vecteurs gaussiens par exemple sont caractérisés par leur matrice de covariance, une matrice symétrique définie positive. Les chaînes de Markov sont modélisées par des graphes pondérés, où à chaque arête est associée une probabilité de transition ; la matrice associée encode de façon compacte les transitions.

1.3.4 Électricité

Si on branche plusieurs résistances en série ou en parallèle, on obtient un graphe pondéré. Les N noeuds de ce graphe correspondent aux noeuds électriques, où le potentiel électrique est inconnu ; les arêtes correspondent aux résistances, où la loi d'Ohm $U = RI$ lie les différences de potentiels aux courants. Pour un circuit comportant des sources de courant (mais pas de tension), écrire en chaque noeud la loi de Kirchoff (disant que la somme des courants entrants à chaque noeud est nulle) donne un système linéaire en N variables. Si le circuit comporte des éléments inductifs ou capacitifs, la même méthode peut être utilisées via la transformée de Laplace en temps.

1.3.5 Réseaux de neurones

Dans leur version la plus simple, les réseaux de neurones artificiels utilisés en apprentissage automatique sont constitués de "neurones", portant une valeur scalaire, et connectés à d'autres neurones. Chaque neurone i reçoit des neurones d'une couche précédente une valeur x_j, et produit une valeur $\sum_j W_{ij}x_j$, qu'il transforme (de façon non-linéaire) et passe aux neurones de la couche suivante. Les matrices W sont appelés les *poids*, et leur valeur est apprise automatiquement (phase d'*entraînement*) de façon à atteindre un objectif donné (minimiser une fonction coût). Les poids étant fixés, utiliser le réseau de neurones sur une nouvelle donnée (phase d'*inférence*) revient donc à faire des produits de la matrice W donnant les connexions entre une couche et la couche suivante par un vecteur de valeurs qui se propage dans le réseau de neurones. Quand plusieurs entrées sont utilisées à la fois (ce qui est souvent le cas en pratique), ces multiplications matrice-vecteur deviennent des multiplications matrice-matrice.

Les architectures modernes de LLM (Large Language Models) sont des variations sur ce principe, et presque tout le temps de calcul de vos IA préférées est passé dans des routines d'algèbre linéaire : du point de vue computationnel, ChatGPT est une gigantesque machine à multiplier des matrices.

1.4 Exercices

1. (Formes quadratiques) Montrer que la forme quadratique

$$q(x, y) = x^2 + 0.1xy + 2y^2$$

est définie positive, et donner la matrice associée (utiliser plusieurs méthodes : majorer $|xy|$, compléter le carré, diagonaliser la matrice). Que dire de la forme suivante ?

$$q(x, y) = x^2 + 2xy + y^2$$

2. (Calcul différentiel) Calculer le gradient de $x \mapsto \|x\|$ (en développant $\|x + h\|$).

3. (Multiplications de matrices) Si A est une matrice $N \times N$, montrer que la sous-matrice des $n \leq N$ premières lignes et colonnes peut être obtenue par X^*AX, avec X bien choisi. Faire le calcul de plusieurs façons différentes : calcul brutal en utilisant la formule de multiplication de matrices, calcul en utilisant $X = (v_1, \dots, v_n)$, calcul en utilisant $(X^*AX)_{ij} = \langle e_i, (X^*AX)e_j\rangle = \langle Xe_i, AXe_j\rangle$.

4. (Formes quadratiques complexes) Soit q_A une forme quadratique (à valeurs réelle) sur $\mathbb{C}^N$. Soit I l'isomorphisme naturel de $\mathbb{C}$ vers $\mathbb{R}^2$, étendu de $\mathbb{C}^N$ vers $\mathbb{R}^{2N}$. Montrer que, pour tout $x, y \in \mathbb{C}^N$,

$$\mathrm{Re}(x^*y) = I(x)^T I(y).$$

Si $\widetilde{q}_A(\widetilde{x}) = q_A(I^{-1}(\widetilde{x}))$, en déduire que $\nabla\widetilde{q}_A(\widetilde{x}) = I(2AI^{-1}(\widetilde{x}))$. Montrer que la Hessienne de $\widetilde{q}_A$ est l'unique matrice réelle symétrique de taille $2N$ représentant la forme quadratique sur $\mathbb{R}^{2N}$ donnée par $2\widetilde{q}_A$. En déduire que si A est HDP, $\widetilde{q}_A$ est strictement convexe.

5. (Équations différentielles) Vérifier que

$$(u_k)_n = \sin\left(\pi k \frac{n}{N}\right), k = 1, \dots, N-1$$

est vecteur propre de (l'opposé de) la matrice du Laplacien en dimension 1 avec conditions au bord de Dirichlet, associé à la valeur propre

$$\lambda_k = 2N^2 \left(1 - \cos\left(\pi \frac{k}{N}\right)\right)$$

Que se passe-t-il quand $k \ll N$? À quel régime cela correspond-il ?

6. (Équations différentielles) On suppose qu'on sait résoudre le problème $-u'' = f$ avec des conditions homogènes $u(0) = u(1) = 0$. On souhaite maintenant résoudre le problème avec une condition de Neumann à gauche $u'(0) = 0, u(1) = 0$. Écrire le système linéaire correspondant, en rajoutant une inconnue $u(0)$ et en approchant $u'(0)$ par $(u(x_1) - u(x_0))/(1/N)$. Appliquer la méthode du complément de Schur en partitionnant les degrés de liberté 0 et $1, \ldots, N-1$, et en choisissant d'inverser soit le premier bloc soit le deuxième. Montrer que la première approche est équivalente à simplement considérer que $u(x_1) = u(x_0)$ dans la discrétisation.
7. (Électricité) Écrire le système linéaire correspondant à un circuit résistif simple avec une source de courant. Est-il inversible ? Le théorème de Norton affirme que tout circuit ouvert à deux ports est équivalent à une source de courant en parallèle avec une résistance. Montrer ce théorème en utilisant le complément de Schur.
8. (Démographie) Suppons que dans une population, x_i soit le nombre d'individus ayant i ans. Supposons qu'un individu d'âge i ait la probabilité p_i de faire un enfant dans l'année, et la probabilité q_i de mourir dans l'année. Donner la répartition du nombre d'individus l'année $n+1$ en fonction du nombre d'individus l'année n.
9. (Probabilités) Montrer que, sur une chaîne de Markov, si $p = p_1, \ldots, p_N$ sont les probabilités à l'instant n d'être dans les états $1, \ldots, N$, alors les probabilités à l'instant $n+1$ sont données par Ap, où A est une matrice à préciser. Montrer que les entrées de A sont positives.
10. (Probabilités) Supposons qu'on ait N estimateurs non biaisés $X_1, \ldots, X_N$ (variables aléatoires de même espérance) d'une même quantité (sondages, expériences...), et qu'on ait accès à leur matrice de covariance C (une matrice symétrique positive définie). Donner l'espérance et la variance de la variable aléatoire $X = \sum_{i=1}^{N} x_i X_i$. Écrire un problème d'optimisation sous contraintes pour obtenir le X qui estime au mieux la quantité. Utiliser la méthode des multiplicateurs de Lagrange et donner un système linéaire permettant de calculer x. Montrer que la matrice associée est inversible en utilisant un complément de Schur.

2 Systèmes linéaires et factorisations de matrices

2.1 Caractère bien posé ou non

Un système linéaire est une équation de forme $Ax = b$, pour une matrice $A \in \mathbb{C}^{N\times M}$ et un vecteur $b \in \mathbb{C}^N$ donné. C'est un système à N équations et M inconnues, qui ne peut être bien posé (existence et unicité) que si $N = M$ (pourquoi ?), ce qu'on supposera dans la suite.

Il y a ici deux cas. Si A est inversible, le système est bien posé (il existe une solution, et elle est unique). Sinon, il y a deux possibilités : soit il n'y a aucune solution, soit il y en a une infinité (puisqu'on peut toujours ajouter un élément du noyau de A). Cette alternative a une description particulièrement élégante, parfois connue sous le nom d'alternative de Fredholm :

Proposition 2.1. Si $A \in \mathbb{C}^{N\times M}$, alors

$$\mathbb{C}^N = \mathrm{Im}(A) \oplus \ker(A^*)$$
$$\mathbb{C}^M = \mathrm{Im}(A^*) \oplus \ker(A)$$

et ces deux décompositions sont orthogonales.

Démonstration. On va prouver la deuxième égalité $\ker(A) = \mathrm{Im}(A^*)^\perp$; la première égalité $\mathrm{Im}(A) = \ker(A^*)^\perp$ suit en appliquant la deuxième à A^*.

Soit $x \in \ker(A)$. Alors, pour tout $y \in \mathbb{C}^N$, $\langle x, A^*y\rangle = \langle Ax, y\rangle = 0$, donc $x \perp \mathrm{Im}(A^*)$.

Réciproquement, si $x \in \mathbb{C}^M$ est tel que, pour tout $y \in \mathbb{C}^M$, on a $0 = \langle x, A^*y\rangle = \langle Ax, y\rangle$, alors $Ax = 0$. □

Ces égalités peuvent se résumer dans le diagramme de la Figure 1.

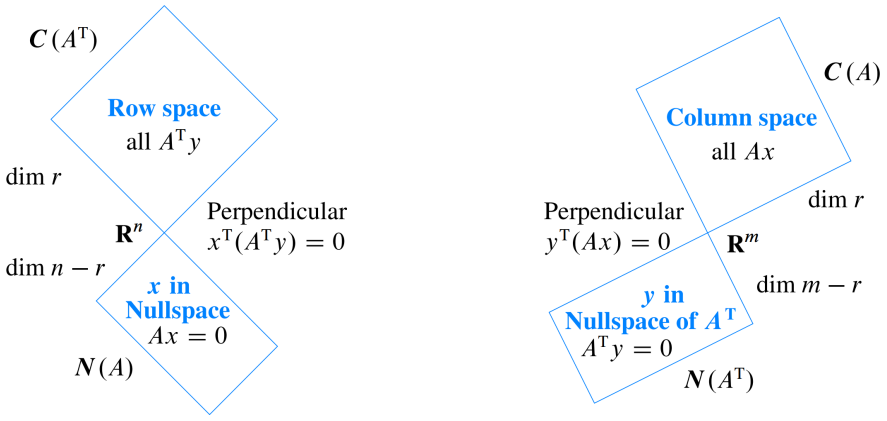


FIGURE 1 – Les quatre sous-espaces fondamentaux, d'après Gilbert Strang, `http://web.mit.edu/18.06/www/Essays/newpaper_ver3.pdf`. Les notations sont $C(A) = \mathrm{Im}(A)$, $N(A) = \ker(A)$.

En particulier, le système $Ax = b$ a une solution si et seulement si $b \in \ker(A^*)^\perp$, c'est-à-dire si b vérifie un nombre fini de conditions de compatibilités. Cela correspond tout simplement à dire que si une certaine combinaison linéaire (de coefficients y_i) de lignes du membre de gauche de l'équation fait zéro ($A^Ty = 0$), alors la même combinaison de lignes du membre de droite doit faire zéro ($y^Tb = 0$).

2.2 Pivot de Gauss

Vous avez tous appris la méthode du pivot de Gauss pour résoudre un système d'équations linéaires. Illustrons sur un exemple :

$$\begin{aligned} x + 2y + 3z &= 1 \\ 4x + 5y + 6z &= 2 \\ 7x + 8y + 10z &= 3 \end{aligned}$$

On commence par éliminer x de toutes les équations sauf une. Pour cela, on choisit un *pivot*, c'est-à-dire une ligne donc le coefficient en face de x est non-nul, mettons la première. On soustrait un certain nombre de fois (ici, 4 et 7) cette équation des deux autres :

$$\begin{aligned} x + 2y + 3z &= 1 \\ -3y - 6z &= -2 \\ -6y - 11z &= -4 \end{aligned}$$

On continue de même pour les lignes 2 et 3 : on choisit un pivot, mettons la deuxième ligne, et on la soustrait un certain nombre de fois (ici, 2) de la troisième :

$$\begin{aligned} x + 2y + 3z &= 1 \\ -3y - 6z &= -2 \\ z &= -3 \end{aligned}$$

On remonte ensuite : on trouve la valeur de z par la dernière équation, on remplace dans la seconde pour trouver y, et finalement dans la première pour trouver x.

Le système linéaire sous sa forme finale est *triangulaire supérieur*[1]

Théorème 2.2. Si A est inversible, la méthode du pivot de Gauss (avec pivot) termine sans erreur (on peut toujours trouver un pivot non-nul).

Démonstration. À l'étape n, le système est de forme

$$\begin{pmatrix} (A_n)_{11} & (A_n)_{12} \\ 0 & (A_n)_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} (b_n)_1 \\ (b_n)_2 \end{pmatrix}$$

où le bloc 1, de taille $n-1$ regroupe les variables déjà éliminées, et le bloc 2, de taille $N-(n-1)$, regroupe les variables encore à éliminer. La matrice $(A_n)_{11}$ est triangulaire supérieure, avec des éléments non-nuls sur la diagonale (les pivots précédents), et est donc inversible. Par ailleurs, la matrice $A_n = \begin{pmatrix} (A_n)_{11} & (A_n)_{12} \\ 0 & (A_n)_{22} \end{pmatrix}$ a été obtenue à partir de A par opérations élémentaires et échange de lignes, qui ne changent pas l'inversibilité (cela peut se voir directement en notant que $(A_n)_{11}$ est obtenue par des multiplications par des matrices élémentaires que nous allons introduire plus tard,

1. On rappelle qu'une matrice triangulaire supérieure (resp. inférieure) est une matrice A telle que $A_{ij} = 0$ pour $i > j$. Une matrice triangulaire a ses valeurs propres sur sa diagonale, et en particulier est inversible si et seulement si ses éléments diagonaux sont tous non-nuls.

ou en notant qu'elles ne changent pas la valeur absolue du déterminant), donc est inversible si et seulement si A l'est. Si l'algorithme du pivot de Gauss ne trouve pas de pivot non-nul, c'est que la première colonne de $(A_n)_{22}$ est nulle, c'est-à-dire que $(A_n)_{22}$ n'est pas inversible. Comme $(A_n)_{11}$ est inversible, par la proposition 1.1 sur le complément de Schur, A_n n'est pas inversible, ce qui est une contradiction. □

Remarques :

— Numériquement, il ne suffit pas que le pivot soit non-nul, il faut encore qu'il ne soit pas trop petit. Dans le cas contraire, on peut se retrouver à diviser par des petits nombres, ce qui pose des problèmes numériques [2]. La stratégie la plus fréquente est simplement de choisir le pivot le plus grand en valeur absolue (*pivot partiel*), c'est-à-dire, à l'étape n, de choisir une ligne telle que $|(A_n)_{nn}| \geq (A_n)_{mn}$ pour tout $m \geq n$ et d'éliminer la variable x_n de toutes les lignes sauf la n-ième. Il existe des cas pathologiques (extrêmement rares en pratique et la plupart du temps ignorés) où cette stratégie aboutit à des erreurs numériques importantes ; il faut dans ce cas utiliser un *pivot complet*, et réordonner les colonnes en plus des lignes.

— Dans le cas où le système linéaire n'est pas inversible, le procédé du pivot de Gauss va finir par aboutir à une ligne dont le membre de gauche est nul. Cela revient à exhiber une combinaison linéaire nulle de lignes de la matrice A, c'est-à-dire un élément de $\mathrm{Ker}(A)$. En poussant le procédé, on peut même en construire une base, et en regardant la nullité du membre de droite cela permet de vérifier la solvabilité du système.

— Le nombre de calculs effectués par la première phase (réduction à une forme triangulaire) de l'algorithme du pivot de Gauss à l'étape n est d'ordre $(N-n+1)^2$: on modifie une fois tous les éléments du sous-système correspondant aux indices $n+1, \ldots, N$. Le coût total est donc $\sum_{n=1}^{N}(N-n+1)^2 = \sum_{m=1}^{N} m^2 = O(N^3)$, donc cubique [3]. Le coût de l'étape de "remontée" est lui $1+2+\cdots+N = O(N^2)$, donc négligeable.

2.3 Factorisation de matrice : LU

Il est instructif de reformuler le pivot de gauss en termes de factorisation de matrice. Considérons pour commencer le cas où le pivot n'est pas nécessaire (à chaque étape, le n-ième élément diagonal est non-nul). L'algorithme du pivot de Gauss part de la matrice A, et la modifie itérativement de sorte à la rendre *triangulaire supérieure*, ce qu'on note U (pour "upper") : dans l'exemple précédent,

$$\underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{pmatrix}}_{A_0=A} \xrightarrow[L_3 \leftarrow L_3 - 7L_1]{L_2 \leftarrow L_2 - 4L_1} \underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & 6 \\ 0 & -6 & -11 \end{pmatrix}}_{A_1} \xrightarrow{L_3 \leftarrow L_3 - 2L_2} \underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & 6 \\ 0 & 0 & 1 \end{pmatrix}}_{A_2=U}$$

Une fois le système d'équations transformé en un système avec une matrice triangulaire supérieure, la résolution est triviale, par remontée.

2. Techniquement, comme les ordinateurs travaillent en virgule flottante, diviser par un petit nombre ne pose pas en soi de problème. Cependant, faire des opérations comme $x = \frac{1}{\varepsilon} + 1.23$, puis $y = x - \frac{1}{\varepsilon}$ introduit une grande erreur sur $y = 1.23$; essayez avec ε de l'ordre de 10^{-13}, la précision machine étant d'ordre 10^{-16}.

3. Visuellement, cette somme correspond à poser N^2 cubes en une couche, puis rajouter une couche de $(N-1)$ cubes, et ainsi de suite. Quand N est grand, on peut oublier le caractère discret des cubes et on obtient un volume égal à $\int_0^N x^2 dx = N^3/6$, qui donne la bonne asymptotique pour $\sum_{m=1}^{N} m^2$. Pour plus de détails, on se rapportera à [1] section 2.5 pour sept (!) preuves de la formule $\sum_{m=1}^{N} m^2 = m(m+1)(2m+1)/6$.

Formalisons cela : pour passer de A_n à A_{n+1}, on effectue des opérations sur les lignes. En se rappelant de comment le produit à gauche de matrice fonctionne (cf. Section 1.2.2), on voit que cela revient à effectuer des multiplications de matrices à gauche : $A_{n+1} = B_n A_n$. Par ailleurs, la structure de B_n n'est pas quelconque : les opérations sont toujours de type $L_m \leftarrow L_m + \alpha_m L_n$, pour $m > n$. Cela implique que B_n est *triangulaire inférieure*, avec des 1 sur la diagonale. On a donc

$$U = B_N B_{N-1} \dots B_1 A. \tag{2.1}$$

On peut alors remarquer la propriété suivante :

Lemma 2.3. L'ensemble des matrices triangulaires inférieures (resp. supérieures) de taille N est stable par multiplication et inversion (pour les matrices qui sont inversibles).

La preuve est immédiate et laissée en exercice. Ce lemme permet de réécrire (2.1) sous la forme

$$A = (B_N B_{N-1} \dots B_1)^{-1} U = LU$$

avec L triangulaire inférieure. On vient en fait de montrer la

Proposition 2.4. Si l'algorithme du pivot de Gauss sans pivotage se déroule sans rencontrer de pivot nul, alors on a

$$A = LU$$

avec L triangulaire inférieure.

Les coefficients de L peuvent être calculés facilement ; on se rapportera par exemple à [2].

L'intérêt de cette écriture factorisée est qu'elle encode de façon compacte toutes les itérations du pivot de Gauss, *sans référence au membre de droite*. Ainsi, si on veut résoudre une séquence de problèmes linéaires $Ax_k = b_k$ pour plusieurs membres de droite $b_k, k = 1, \dots, K$, il suffit de calculer une fois pour toutes la factorisation $A = LU$ (qui était la partie coûteuse de l'algorithme). On peut alors résoudre $Ly_k = b_k$ par remontée (y_k est égal au membre de droite tel que modifié par l'algorithme du pivot de Gauss, mais son calcul par cette méthode, une fois L connu, est plus rapide que l'application du pivot de Gauss), puis $Ux_k = y_k$ par descente, pour un coût total de $O(N^3 + N^2 K)$. En Python, la commande `scipy.linalg.solve` implémente cette méthode.

Dans le cas où le pivot de Gauss sans pivotage ne fonctionne pas (par exemple, la matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$), comme expliqué précédemment, il faut échanger des lignes : on aboutit alors à une factorisation $PA = LU$, avec P une matrice de permutation.

2.4 Variantes et alternatives au pivot de Gauss

2.4.1 Matrices creuses

Une matrice *creuse* est une matrice dont le nombre d'entrées non-nulles est faible devant le nombre total d'entrées. Ce type de matrices apparaît souvent dans les applications, notamment

dans la discrétisation d'équations différentielles ou dans la modélisation par graphes. Quand une matrice est creuse, on peut stocker uniquement les éléments non-nuls et leur position, ce qui donne des algorithmes très efficaces par exemple pour la multiplication d'un vecteur par cette matrice. Cependant, l'application naïve du pivot de Gauss à ces matrices amène souvent à remplir les blocs L et U, ce qui détruit les optimisations possibles (en anglais, phénomène de "fill-in"). On peut parfois s'arranger (notamment par un choix astucieux de pivot) pour minimiser ce phénomène. Un exemple simple est celui d'une matrice tridiagonale (c'est-à-dire telle que $A_{ij} = 0$ si $|i - j| > 1$), laquelle le phénomène de fill-in est absent quand on ne pivote pas.

2.4.2 Factorisation de Cholesky

Si A est hermitienne, sa symétrie n'est pas a priori reflétée dans sa décomposition LU, et la symétrie de A n'est pas utilisée pour accélérer le calul de L et U. En général, il n'est pas évident d'utiliser l'information de symétrie (par exemple, la matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ n'a pas de factorisation qui reflète sa symétrie). Un cas particulier où on peut faire mieux est le cas où A est hermitienne définie positive (HDP) : on a alors la factorisation de Cholesky $A = LL^*$ avec L triangulaire inférieure. Cette factorisation peut être réalisée par un "pivot de Gauss symétrisé" : on ne peut pas se contenter de réaliser des opérations élémentaires $A_{n+1} = B_n A_n$ sur les lignes ; à la place, il faut trouver une opération $A_{n+1} = B_n A_n B_n^*$ qui modifie les lignes et les colonnes de façon symétriques, et il existe plusieurs méthodes standard pour le faire. Toute matrice symétrique définie positive admet une décomposition de Cholesky, et cette décomposition est stable (robuste aux erreurs numériques). On se rapportera à [2] pour plus de détails.

2.4.3 Factorisation QR

Vous avez déjà rencontré la méthode de Gram-Schmidt, qui construit une base orthonormée à partir d'une famille libre de vecteurs. Que se passe-t-il si on applique le procédé de Gram-Schmidt au colonnes $u_1, \dots, u_N$ d'une matrice $N \times M$, avec $N \geq M$, de rang M (donc dont les colonnes forment une famille libre) ? On va produire des vecteurs $q_1, \dots, q_M$ orthonormés, sous la forme

$$q_i = \frac{u_i - \sum_{j<i} \langle q_j, u_i \rangle q_j}{\| u_i - \sum_{j<i} \langle q_j, u_i \rangle q_j \|}$$

On peut montrer que les dénominateurs ne s'annulent jamais (si c'est le cas, cela signifie qu'on a exhibé une combinaison linéaire nulle non-triviale de vecteurs $u_1, \dots, u_N$, ce qui est impossible). Chaque q_i est donc formé comme combinaison linéaire des $u_j, j \leq i$.

D'un point de vue matriciel, si on range les q_i comme colonnes d'une matrice A, on a donc écrit

$$Q = AU$$

avec $U \in \mathbb{C}^{M \times M}$ une matrice triangulaire supérieure et $Q \in \mathbb{C}^{N \times M}$ une matrice orthogonale ($Q^*Q = 1$). En inversant U (ce qui est possible parce que les éléments diagonaux de R sont égaux à l'inverse des dénominateurs de l'équation définissant le procédé de Gram-Schmidt), on obtient

$$A = QR$$

avec $R = U^{-1}$ une matrice triangulaire supérieure. Cette factorisation permet ensuite de résoudre simplement des systèmes linéaires, en résolvant un système linéaire avec Q puis avec R.

Le coût de calcul de cette méthode est cubique, comme pour la méthode LU ; en pratique elle coûte un peu plus cher.

2.5 Moindres carrés

L'exemple le plus élémentaire du problème des moindres carrés est celui de faire passer une droite au plus près d'un nuage de points. On utilise la paramétrisation $y = \alpha x + \beta$, et on cherche les α, β qui minimisent l'erreur

$$\sum_{i=1}^{N} (y_i - (\alpha x_i + \beta))^2$$

La propriété importante de ce problème est que la fonction objectif soit quadratique en *les paramètres* (α, β). Pour mettre cela en évidence, on va poser $c = (\alpha, \beta)$, $d_i = y_i$ et

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix},$$

ce qui permet de reformuler le problème en celui de la minimisation de $\|Ac - d\|^2$.

Plus généralement, la forme générale du problème des moindres carrés est : étant donnée une matrice $A \in \mathbb{C}^{N \times M}$ avec $N \geq M$ et un vecteur $b \in \mathbb{C}^N$, trouver

$$\min_{x \in \mathbb{C}^M} \|Ax - b\|^2.$$

Ce problème de minimisation d'une fonctionnelle quadratique est aisément résolu par les outils du calcul différentiel et de l'optimisation.

Théorème 2.5. Si A est de rang M, alors il existe une unique solution au problème des moindres carrés, donnée par la solution du système

$$A^* A x = A^* b.$$

Démonstration. On va faire pour simplifier la preuve dans le cas réel ; on réfère à l'exercice 4 de la section 1 pour la méthode de preuve dans le cas complexe.

On calcule tout d'abord

$$\begin{aligned} E(x) &= (Ax - b)^*(Ax - b) \\ &= x^* A^* A x - b^* A x - x^* A b + b^* b. \end{aligned}$$

On en déduit (cf section 1.1.4) que le gradient $\nabla E(x)$ est égal à $2(A^* A x - A^* b)$, et sa hessienne à $2A^* A$.

On a par ailleurs

$$x^* A^* A x = \|Ax\|^2.$$

Si A est de rang M, son noyau est vide, donc $x^*A^*Ax = 0 \Rightarrow x = 0$: la matrice A^*A est hermitienne définie positive, donc E est strictement convexe et coercive (tend vers $+\infty$ quand $x \to \infty$).

On peut maintenant conclure : E étant strictement convexe et coercive, elle admet un unique minimiseur. À ce minimiseur, le gradient est nul, ce qui démontre le théorème. □

Pour calculer la solution de ce problème en pratique, on peut résoudre les *équations normales*, mais ce n'est pas une solution très élégante, et elle crée des soucis numériques (on verra qu'elle augmente le *conditionnement* du problème). À la place, on peut utiliser une factorisation $A = QR$ et noter que

$$(A^*A)^{-1}A^*b = (R^*R)^{-1}R^*Q^*b = R^{-1}(R^*)^{-1}R^*Q^*b = R^{-1}Q^*b$$

Comme on l'a vu, la matrice R est inversible si A est de rang plein, ce qui justifie ces opérations.

Si A n'est pas de rang M, son noyau est non-nul et la solution ne peut pas être unique. Dans ces cas-là, on peut essayer de détecter son rang, mais c'est une opération instable ; il vaut mieux *régulariser* le problème, par exemple par le régularisation de Tikhonov (cf exercice).

2.6 Exercices

1. (Électricité) Reprendre l'exemple du système linéaire correspondant à un circuit électrique. Montrer que la matrice correspondante est symétrique. Donner un élément de son noyau. Calculer x^TAx sous la forme d'une somme de carrés et déterminer complètement le noyau dans le cas où le graphe correspondant est connexe. Montrer que le système linéaire a une unique solution à condition d'imposer le potentiel à un noeud particulier.
2. (LU) Montrer que, quand le pivot de Gauss sans pivotage fonctionne, on peut résoudre un système linéaire $Ax = b$ où A est tridiagonale avec une complexité $O(N)$.
3. (LU) Comment calculer un déterminant par une factorisation LU ? Quelle est la complexité de cette méthode ?
4. (Cholesky pour l'orthogonalisation)
   (a) Si $A \in \mathbb{C}^{N\times M}$ avec $N \geq M$, montrer que la matrice de Gram A^*A est symétrique définie positive si et seulement si A est de rang maximal M (c'est-à-dire, si les colonnes de A sont indépendantes).
   (b) Montrer que, si $A^*A = LL^*$ avec les notations précédentes, alors $A(L^*)^{-1}$ est une matrice orthogonale.
   (c) Appliquer cette méthode à $A = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$ avec $v_1, v_2 \in \mathbb{C}^N$ de norme 1. Comparer avec l'orthogonalisation de Gram-Schmidt.
5. (Pivot par blocs) Appliquer la méthode du pivot de Gauss *par blocs* (c'est-à-dire qu'à chaque étape on essaie d'éliminer un groupe d'inconnues) à la matrice par blocs

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

   Quelle méthode retrouvez-vous ?

6. (Givens) On définit la matrice de rotation de Givens par

$$G(i,j,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

avec $c = \cos\theta$ $s = \sin\theta$ et où les lignes et colonnes particularisées sont i et j.

(a) Donner son interprétation géométrique.

(b) Étant donné un vecteur $\begin{pmatrix} a \\ b \end{pmatrix}$, comment peut-on choisir θ de telle sorte que $\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sqrt{a^2+b^2} \\ 0 \end{pmatrix}$?

(c) Utiliser les matrices de Givens pour montrer que, étant donnée une matrice A, on peut amener son coefficient (i,j) à 0 par multiplication à gauche par une matrice unitaire, en ne modifiant que les lignes i et j. En déduire une méthode pour calculer la factorisation QR d'une matrice.

7. (Moindres carrés) Reprendre le problème des moindres carrés avec une parabole au lieu d'une droite.

8. (Moindres carrés) Donner la complexité asymptotique (notation O) du coût de calcul de la méthode des moindres carrés.

9. (Moindres carrés) Refaire le problème des moindres carrés avec une pondération : on minimise $\sum_{i=1}^{N} w_i(y_i - (\alpha x_i + \beta))^2$, avec $w_i > 0$ le poids attaché à chaque observation (x_i, y_i).

10. (Moindres carrés) Avec les notations du problème des moindres carrés, on cherche maintenant à résoudre le problème avec *régularisation de Tikhonov*

$$\|Ax - b\|^2 + \lambda\|x\|^2$$

avec $\lambda > 0$. Montrer que la solution de ce problème est unique, même si A n'est pas de rang M.

3 Le théorème spectral, généralisations et applications

On rappelle que λ est valeur propre d'une matrice carrée A si il existe $x \neq 0$ tel que $Ax = \lambda x$. Une matrice pour laquelle il existe une base de vecteurs propres est dite diagonalisable, et on a alors $A = PDP^{-1}$, où P est la matrice des vecteurs propres (rangés en colonnes) et D est diagonale. Toutes les matrices ne sont pas diagonalisables, l'exemple le plus élémentaire étant la matrice $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, qui a pour seule valeur propre 0 : si elle était diagonalisable, $D = 0$ et $A = PDP^{-1}$ serait donc nulle. Dans ce cas, il n'existe qu'un seul vecteur propre (à facteur près), $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Même si une matrice est diagonalisable, elle peut être "à peine diagonalisable", au sens où une petite perturbation peut la rendre non diagonalisable. Ainsi, la matrice $\begin{pmatrix} 0 & 1 \\ 0 & 0.0001 \end{pmatrix}$ est diagonalisable (car elle a deux valeurs propres distinctes), mais une toute petite perturbation la rend non-diagonalisable. Il est donc utile de trouver des critères permettant d'être sûr que la diagonalisation est robuste. Le plus important est le théorème spectral.

3.1 Le théorème spectral

Théorème 3.1 (Théorème spectral)**.** Une matrice hermitienne est diagonalisable en base orthonormée et a des valeurs propres réelles. Si A est réelle, alors la base orthonormée peut être choisie réelle.

On utilise d'abord le lemme suivant.

Lemma 3.2. Une matrice hermitienne A admet toujours un vecteur propre, associé à une valeur propre réelle. Si A est réelle, ce vecteur peut être choisi réel.

Démonstration du lemme. Une preuve classique, algébrique, utilise le polynôme caractéristique et le fait qu'un polynôme a toujours des racines dans $\mathbb{C}$ (ce qui marche même sur des matrice non-hermitiennes), puis exclut la possibilité que la valeur propre et le vecteur propre soient complexes. On trouvera également une preuve utilisant l'analyse complexe (mais pas le polynôme caractéristique) à l'exercice 6 de la section 6. Nous donnons ici une preuve élémentaire, plus analytique, n'utilisant pas de polynôme, et qui n'utilise pas de nombres complexes si besoin, bien qu'elle soit restreinte au cas hermitien.

On va faire pour simplifier la preuve dans le cas réel ; on réfère à l'exercice 4 de la section 1 pour la méthode de preuve dans le cas complexe. On considère le problème de minimisation du *quotient de Rayleigh* :

$$\min_{x \neq 0} \frac{x^* A x}{x^* x} = \min_{\|x\|=1} x^* A x.$$

Ce problème admet toujours une solution, car il s'agit d'un problème de minimisation d'une fonction continue (ici, polynomiale) sur un compact (on utilise ici la dimension finie). À la solution, on peut utiliser la méthode des multiplicateurs de Lagrange pour assurer l'existence d'un $\lambda \in \mathbb{R}$ tel que le

gradient de x^*Ax soit égal à λ fois le gradient de $\|x\|^2$, ce qui montre que $Ax = \lambda x$. Si A est réelle, il suffit de faire le raisonnement précédent en minimisant sur $\mathbb{R}^N$ plutôt que $\mathbb{C}^N$. □

Démonstration du théorème spectral. On prouve alors le théorème spectral par récurrence. À la première étape, on trouve par le lemme un vecteur propre normalisé v_1 de A. On complète ce v_1 en une base orthonormée $(v_1, u_2, \dots, u_N)$. Dans cette base, on a $\langle u_i, Av_1\rangle = \lambda\langle u_i, v_1\rangle = 0$, et par ailleurs $\langle v_1, Au_i\rangle = \langle Av_1, u_i\rangle = 0$, ce qui montre que, dans cette base, A a l'écriture par blocs

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & \widetilde{A} \end{pmatrix}.$$

On recommence alors l'opération pour $\widetilde{A}$, et ainsi de suite. □

Si A est hermitienne, sa décomposition en valeurs propres s'écrit

$$A = PDP^*.$$

avec D une matrice diagonale, et P une matrice unitaire (contenant une base orthonormée de vecteurs propres rangés par colonnes).

3.2 Décomposition de Schur

Il est intéressant de regarder ce que donne le schéma de preuve du théorème dans le cas où A n'est pas hermitienne. À l'étape de récurrence, on a $\langle u_i, Av_1\rangle = 0$ mais $\langle v_1, Au_i\rangle = \alpha_i$ non-nul a priori. La matrice A s'écrit alors dans la base $(v_1, u_2, \dots, u_N)$

$$A = \begin{pmatrix} \lambda & \alpha^T \\ 0 & \widetilde{A} \end{pmatrix}$$

avec $\alpha \in \mathbb{C}^{N-1}$. En menant la récurrence, on obtient une base orthonormée dans laquelle A est triangulaire supérieure, c'est-à-dire une factorisation

$$A = PTP^*$$

avec P unitaire et T triangulaire supérieure. Cette factorisation, appelée *décomposition de Schur* (à ne pas confondre avec le complément de Schur), permet en particulier de trouver facilement les valeurs propres (mais pas les vecteurs propres !) de A, et est stable numériquement.

3.3 Formules min-max

Proposition 3.3 (Formules min-max)**.** Soit A hermitienne, de valeurs propres $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ (comptées avec leur multiplicité). Alors

$$\lambda_1 = \min_{x \neq 0} \frac{x^*Ax}{x^*x}.$$

Plus généralement,

$$\lambda_n = \min_{S \subset \mathbb{C}^N, \dim S = n} \max_{x \in S \neq 0} \frac{x^*Ax}{x^*x},$$

où le minimum est pris sur tous les sous-espaces vectoriels de $\mathbb{C}^N$ de dimension n.

Démonstration. Par le théorème spectral, quitte à changer de base, on peut sans perte de généralité supposer A diagonale. On a alors

$$\frac{x^*Ax}{x^*x} = \frac{\sum_{n=1}^N \lambda_n |x_n|^2}{\sum_{n=1}^N |x_n|^2}$$

Le quotient de Rayleigh est donc une combinaison barycentrique [4] des λ_n, de poids $|x_n|^2/(\sum_{n=1}^N |x_n|^2)$.

- Pour le premier énoncé, minimiser une combinaison barycentrique par rapport aux poids donne bien le plus petit des éléments, λ_1.
- Dans le cas général, on note qu'en prenant pour S l'espace engendré par les vecteurs propres associés aux n premières valeurs propres, on a bien $\lambda_n = \max_{x \in S \neq 0} \frac{x^*Ax}{x^*x}$ (parce que le quotient de Rayleigh est une combinaison barycentrique des n premières valeurs propres). Il reste donc à montrer que, pour tout S de dimension n, $\max_{x \in S \neq 0} \frac{x^*Ax}{x^*x} \geq \lambda_n$.
- Soit S de dimension n. Par comptage de dimensions, S intersecte de façon non-triviale le sous-espace associé aux valeurs propres $\lambda_n, \ldots, \lambda_N$. Si $x \neq 0$ appartient à cette intersection, alors $\frac{x^*Ax}{x^*x}$ est une combinaison barycentrique de $\lambda_n, \ldots, \lambda_N$, donc est plus grand que λ_n. Il suit que $\max_{x \in S \neq 0} \frac{x^*Ax}{x^*x} \geq \lambda_n$.

□

3.4 Classification des formes quadratiques

On rappelle que la forme quadratique associée à A hermitienne est

$$q_A(x) = x^*Ax.$$

Proposition 3.4. Une matrice A hermitienne est définie positive (HDP) si et seulement si ses valeurs propres sont strictement positives.

Démonstration. Comme A est hermitienne, on a

$$A = PDP^*$$

et donc

$$x^*Ax = (P^*x)^*D(P^*x)$$

On voit alors que A est HDP si et seulement si D l'est. □

En se plaçant dans la base orthonormée de diagonalisation de A, on a donc A diagonale, et

$$q_A(x) = \sum_{i=1}^N \lambda_i |x_i|^2.$$

4. On rappelle qu'une combinaison barycentrique des α_n de poids p_n est une combinaison de type $\sum_{n=1}^N p_n \alpha_n$, avec $p_n \geq 0, \sum_{n=1}^N p_n = 1$. Une telle combinaison varie entre $\min \alpha_n$ et $\max \alpha_n$, et ces minima et maxima sont atteints.

En dimension 2 réelle, les lignes de niveau de q_A sont donnés par l'équation

$$\lambda_1 x^2 + \lambda_2 y^2 = c.$$

Ce sont donc des ellipses (si les deux valeurs propres ont le même signe, c'est-à-dire si A ou $-A$ sont SDP) ou des hyperboles (si les deux valeurs propres ont un signe différent). Dans le cas elliptique, les valeurs propres sont égales à l'inverse des carrés des demi-grands axes de l'ellipse. En revenant dans la base canonique (par une transformation orthogonale, donc une rotation ou une réflexion), on a encore une ellipse ou une hyperbole.

3.5 Espaces propres et commutation

On rappelle que l'espace propre de A associé à la valeur propre λ est $\{x \in \mathbb{C}^N, Ax = \lambda x\}$. Une matrice A a $l \leq N$ sous-espaces propres distincts $X_1, \ldots, X_l$, qui vérifient $\sum_{i=1}^{l} \dim(X_l) \leq N$, avec égalité si et seulement si A est diagonalisable. On dit qu'un sous-espace X est invariant par A si $AX \subset X$.

On dit que A *commute* avec B si $AB = BA$. Deux matrices co-diagonalisées dans la même base (c'est-à-dire $A = PD_AP^{-1}, B = PD_BP^{-1}$ commutent). On va voir que la réciproque est vraie.

La propriété géométrique fondamentale des matrices qui commutent est

Proposition 3.5. Deux matrices qui commutent préservent chacune les espaces propres de l'autre : si $AB = BA$, alors les espaces propres de B sont des espaces invariants de A.

Démonstration. Si $Bx = \lambda x$, alors $BAx = ABx = \lambda Ax$, donc Ax est valeur propre de B associé à la valeur propre λ. □

Corollaire 3.6. Si A et B sont deux matrices hermitiennes qui commutent, alors on peut choisir une base de vecteurs propres communs.

Démonstration. Soit $X_1, \ldots, X_l$ les espaces propres de B. Ce sont des espaces invariants pour A, donc on peut choisir des vecteurs propres orthogonaux pour les matrices $A|_{X_1}, \ldots, A|_{X_l}$, qui forment une base orthogonale de vecteurs propres de A et de B. □

3.5.1 Matrices normales

Définition 3.7. Une matrice carrée est *normale* si elle commute avec son adjoint, i.e. $AA^* = A^*A$.

Des exemples de matrices normales sont les matrices hermitiennes, anti-hermitiennes et unitaires. Plus généralement, toute matrice diagonalisable en base orthonormée (quel que soit leur spectre) est normale, puisque $(PDP^*)^*(PDP^*) = PD^2P^* = (PDP^*)(PDP^*)^*$. C'est en fait une caractérisation :

Théorème 3.8 (Théorème spectral pour les matrices normales)**.** Une matrice est normale si et seulement si elle est diagonalisable en base orthonormée (dans $\mathbb{C}$).

Démonstration. Si A est normale, alors ses parties hermitiennes et anti-hermitiennes $\mathrm{H}(A) = \frac{1}{2}(A + A^*)$ and $\mathrm{AH}(A) = \frac{1}{2i}(A - A^*)$ commutent. Ces deux matrices sont hermitiennes, donc peuvent être co-diagonalisées, ce qui donne une base orthonormale de vecteurs propres pour $A = \mathrm{H}(A) + i\,\mathrm{AH}(A)$. □

Fort de ce théorème, on se rend compte que le corollaire 3.6 s'étend aux matrices normales.

3.6 La transformée de Fourier discrète

Illustrons ces idées avec la notion de transformation de Fourier discrète. Dans cette section seulement, il sera pratique d'indicer les vecteurs à partir de 0, c'est-à-dire $\mathbb{C}^N \ni x = (x_0, \dots, x_{N-1})$. De plus, on utilisera la convention que les vecteurs sont implicitement étendus par périodicité en-dehors de leur domaine de définition via la relation $x_{i+kN} = x_i$ pour tout $k \in \mathbb{Z}$. La donnée d'un vecteur de $\mathbb{C}^N$ doit donc être comprise comme le choix d'une représentation particulière (décidant arbitrairement de démarrer à 0) d'une suite N-périodique de $\mathbb{C}^{\mathbb{Z}}$, ce qu'on notera pour bien expliciter cette interprétation $\mathbb{C}^{\mathbb{Z}/N\mathbb{Z}}$.

3.6.1 Convolution cyclique et matrice circulante

On considère l'opération de convolution cyclique, définie sur $\mathbb{C}^{\mathbb{Z}/N\mathbb{Z}} \times \mathbb{C}^{\mathbb{Z}/N\mathbb{Z}}$ par

$$(a * x)_i = \sum_{j=0}^{N-1} a_j x_{i-j}.$$

Cette opération peut être interprétée comme une moyenne glissante de x pondérée par a. Si par exemple $a_{\mathrm{moy}} = (0, 1/2, 0, \dots, 0, 1/2)$, on a

$$(a_{\mathrm{moy}} * x)_i = \frac{1}{2}(x_{i+1} + x_{i-1}).$$

On voit par un changement de variables $j = i - j'$ que la convolution est également symétrique :

$$\begin{aligned}(a * x)_i &= \sum_{j'=i-(N-1)}^{i-0} a_{i-j'} x_{j'} \\ &= \sum_{j'=0}^{N-1} a_{i-j'} x_{j'} \\ &= (x * a)_i\end{aligned}$$

On a utilisé ici le fait que la suite $j' \mapsto a_{i-j'} x_{j'}$ est N-périodique, et sa somme sur n'importe quelle suite contiguë de N indices est donc indépendante du choix particulier de la suite d'indices.

À a fixé, la convolution $x \mapsto a * x$ est une opération linéaire, qui peut donc se représenter par la matrice $(A_a)_{ij} = a_{i-j}$:

$$A_a = \begin{pmatrix} a_0 & a_{N-1} & a_{N-2} & \dots & a_1 \\ a_1 & a_0 & a_{N-1} & & a_2 \\ a_2 & a_1 & a_0 & & a_{N-3} \\ \vdots & & & \ddots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \dots & a_0 \end{pmatrix}$$

Ce type de matrices est nommée *matrice circulante*.

3.6.2 Diagonalisation des matrices circulantes

La propriété la plus importante des convolutions cyclique est d'être *invariante par translation* (cyclique) : à a fixé, si on translate x, $a * x$ est translaté d'autant. En termes de matrices, cela signifie donc que

$$A_a T x = T(A_a x),$$

où T est la matrice de translation cyclique (shift à gauche) définie par $T(x_0, x_1, \dots, x_{N-1}) = (x_1, x_2, , \dots, x_0)$, de sorte que

$$T = A_{(0,\dots,0,1)} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & 0 & 0 & & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

Cette matrice est plus simple qu'un A_a général. En suivant la philosophie des théorèmes de commutation (sans en suivre exactement la lettre pour l'instant, car nous ne savons pas que A est normale ; cela peut se montrer directement mais c'est un calcul peu élégant que nous allons contourner), on peut commencer par chercher à diagonaliser T.

L'équation $Tx = \lambda x$ s'écrit

$$x_{i+1} = \lambda x_i, \quad i = 0, \dots N-1$$

Satisfaire les équations $0, \dots, N-2$ mène à avoir $x_i = \lambda^i x_0$. Pour que l'équation soit satisfaite pour $i = N-1$, il faut que $x_N = x_0 = \lambda^N x_0$ donc $\lambda^N = 1$ ($x_0 \neq 0$ car $x \neq 0$). Si

$$\omega_N = e^{\frac{2\pi i}{N}}$$

est la racine N-ième de l'unité élementaire, on a donc N valeurs propres distinctes et vecteurs propres associés

$$\lambda_k = \omega_N^k$$
$$(v_k)_j = \frac{1}{\sqrt{N}} \omega_N^{kj}$$

pour $k = 0, \ldots, N-1$. On calcule par ailleurs que

$$\langle v_k, v_{k'} \rangle = \frac{1}{N} \sum_{j=0}^{N-1} \omega_N^{(k-k')j} = \delta_{kk'},$$

et la famille de vecteurs propres $v_0, \ldots, v_{N-1}$ est donc orthonormée (on aurait pu se passer du calcul en notant que $T^* = A_{(1,\ldots,0)} = T^{-1}$, le shift à droite, et donc T est unitaire ; comme les valeurs propres λ_k sont distinctes, les vecteurs v_k sont nécessairement orthogonaux).

On a trouvé une base orthonormée de vecteurs propres de T qui commute avec A_a. De plus, comme les valeurs propres de T sont distinctes, c'est *la seule* base orthonormée à phase près (on peut changer v_k en $v_k e^{i\theta_k}$). En s'inspirant de l'extension du corollaire 3.6 aux matrices normales (qui ne s'applique pas encore ici car on n'a pas encore vérifié que A était normale), on commence à se douter que v_k va être une base orthonormée de vecteurs propres de A_a, et en effet :

$$(A_a v_k)_i = \sum_{j=0}^{N-1} a_j (v_k)_{i-j} = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} a_j \omega_N^{k(i-j)} = \frac{1}{\sqrt{N}} \omega_N^{ki} \left(\sum_{j=0}^{N-1} a_j \omega_N^{-kj} \right) = \left(\sum_{j=0}^{N-1} a_j \omega_N^{-kj} \right) (v_k)_i$$

Une façon plus algébrique de voir cela est de remarquer que $A_a = \sum_{j=0}^{N-1} a_j T^{-j}$ donc $A_a v_k = \sum_{j=0}^{N-1} a_j \omega_N^{-kj}$.

On vient donc de démontrer

Proposition 3.9. Toute matrice circulante est normale, diagonalisée par la base orthonormée de vecteurs propres $(v_k)_j = \omega_N^{kj}$, de valeurs propres associées $\sum_{j=0}^{N-1} a_j \omega_N^{-kj}$.

3.6.3 La transformée de Fourier discrète

Le changement de la base canonique vers la base de vecteurs propres v_k, représenté par la matrice V^* (avec V contenant les vecteurs v_k rangés en colonnes) est, à normalisation près, appelée *transformée de Fourier discrète*. Avec un choix de normalisation plus usuel $F = \sqrt{N} V^*$:

Définition 3.10. La transformation de Fourier discrète est définie par

$$(Fx)_k = \sum_{j=0}^{N-1} e^{-\frac{2\pi i k j}{N}} x_j.$$

Son inverse est donné par

$$(F^{-1} y)_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{-\frac{2\pi i k j}{N}} y_k.$$

et la proposition 3.9 devient alors dans ce nouveau langage le

Théorème 3.11 (de la convolution)**.** La transformation de Fourier discrète transforme les convolutions en produits :

$$(F(x * y))_k = (F(x))_k (F(y))_k$$

La transformée de Fourier discrète est liée à la théorie des séries de Fourier et de la transformée de Fourier ; nous ne détaillerons pas plus ici. La transformée de Fourier discrète est extrêmement utile en pratique, d'autant plus qu'il existe un algorithme de calcul très rapide ($O(N \log N)$ au lieu de $O(N^2)$ naïvement), appelé *Fast Fourier Transform* (FFT), tellement répandu et important qu'on utilise souvent le sigle FFT pour dénoter la transformée de Fourier discrète.

Si l'on souhaite interpréter le résultat de la transformée de Fourier discrète en termes de coefficients de Fourier d'une fonction discrétisée, on prendra bien garde au phénomène d'*aliasing* : comme $(Fx)_{k+lN} = (Fx)_k$ pour tout $l \in \mathbb{Z}$, il n'est pas clair si $(Fx)_k$ doit être interprété comme une approximation du k-ième ou du $(k+lN)$-ième coefficient de Fourier de la fonction $x(t)$ discrétisée aux points $0, \frac{2\pi}{N}, \ldots, \frac{2\pi(N-1)}{N}$. L'approximation du coefficient de Fourier est bonne uniquement quand $e^{-\frac{2\pi i k j}{N}} x_j$ est une représentation fidèle de la fonction sous-jacente (si le pas de grille $(2\pi)/N$ est faible devant l'échelle caractéristique de variation de la fonction $x(t)e^{-kt}$), ce qui n'est pas le cas quand $x(t)$ est régulière et que k commence à approcher de N. Ainsi, bien souvent, on interprète les $N/2$ premières valeurs de la transformée de Fourier discrète comme des approximations des $N/2$ premiers coefficients de Fourier de la fonction discrétisée, et les $N/2$ restants comme les approximations des $N/2$ premiers coefficients de Fourier *négatifs* de la fonction (voir notamment les fonctions `fftfreq` et `fftshift` en Python).

3.7 Exercices

1. (Diagonalisabilité) Montrer que presque toutes les matrices (au sens de la théorie de la mesure) de dimension 2 sont diagonalisables. On notera que le discriminant du polynôme caractéristique dépend polynomialement des entrées de A, et on utilisera le fait que toute variété algébrique de $\mathbb{R}^N$ (ensemble défini par une équation polynomiale) non égale à $\mathbb{R}^N$ est de mesure nulle.
2. (Valeurs propres doubles) Montrer que l'ensemble des matrices symétriques réelles 2×2 ayant une valeur propre double est un sous-espace vectoriel des matrices matrices symétriques réelles 2×2, et donner sa dimension. Que peut-on attendre des valeurs propres d'une matrice réelle symétrique 2×2 $A(t)$ dépendant continûment d'un paramètre t ?
3. (M-adjoint) Montrer que, si M est hermitienne définie positive, alors $\langle x, y \rangle_M = x^* M y$ définit un produit scalaire. On dit que B est l'adjoint de A pour le produit scalaire M si $\langle x, Ay \rangle_M = \langle Bx, y \rangle_M$. Calculer B.
4. (Diagonalisabilité) Montrer que si A est hermitienne définie positive et B est hermitienne, alors AB est diagonalisable à valeurs propres réelles. Indication : on pourra définir le produit scalaire $\langle x, y \rangle_{A^{-1}} = x^* A^{-1} y$.
5. (Matrices M-normales) Montrer que toute matrice diagonalisable est normale pour un certain produit scalaire (commute avec son adjoint pour ce produit scalaire).

6. (Commutation) Montrer que la matrice

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$$

pour $\alpha, \beta \in \mathbb{R}$, étant symétrique par rapport au renommage $e_1 \leftrightarrow e_2$, commute avec la matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. En déduire sans calcul sur A une base de vecteurs propres de A.

7. (Interlacement des valeurs propres) Soit B une sous-matrice de A hermitienne. Montrer que, si $\lambda_i(A)$ est la i-ième première valeur propre (comptée avec multiplicité) de A, alors

$$\lambda_1(A) \leq \lambda_1(B) \leq \lambda_2(A) \leq \dots$$

8. (Théorème de Lax-Milgram discret) Soit $A \in \mathbb{C}^{N \times N}$ HDP, $P \in \mathbb{C}^{N \times n}$ une matrice de rang n avec $n \leq N$. Montrer que P^*AP est inversible. Donner un contre-exemple montrant que ce n'est pas nécessairement le cas si A n'est pas HDP.

9. (Matrices cumulantes) Montrer que si A commute avec T (est invariante par translation), alors A est une matrice cumulante (représentant une convolution).

10. (Transformation de Fourier discrète) Donner la transformée de Fourier discrète de $x_j = \cos\left(\frac{2\pi j}{N}\right)$. Interpréter.

11. (Transformation de Fourier discrète) Donner la matrice représentant la transformation

$$(Ax)_j = \tfrac{x_{j-1}+2x_j+x_{j+1}}{4}$$

Appliquer le théorème de la convolution. Donner l'asymptotique pour $|k| \ll N$. Interpréter. Même question pour

$$(Ax)_j = \frac{x_{j-1} - 2x_j + x_{j+1}}{(2\pi/N)^2}$$

à interpréter à la lumière de la méthode des différences finies.

12. (Transformation de Fourier discrète) Montrer l'égalité de Parseval

$$\sum_{j=0}^{N-1} |x_j|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |(Fx)_k|^2$$

4 Algèbre linéaire quantitative

On va dans cette section examiner la façon dont une matrice modifie la norme d'un vecteur, c'est-à-dire regarder les propriétés du *facteur de croissance*

$$\frac{\|Ax\|}{\|x\|} = \frac{\sqrt{x^*(A^*A)x}}{\|x\|}$$

4.1 Valeurs singulières

On rappelle que les valeurs propres d'une matrice non-normale ne donnent aucune information quantitative sur l'action d'*une* application de la matrice à un vecteur - elles sont utiles pour comprendre les applications répétées. Par exemple, la matrice $A = \begin{pmatrix} 0 & 1000 \\ 0 & 0 \end{pmatrix}$ a des valeurs propres nulles, ce qui ne permet pas de voir par exemple que $\|Ae_2\| = 1000$. Pour ce faire, on introduit la décomposition en valeurs singulières (SVD, pour singular value decomposition) :

Proposition 4.1. Soit $A \in \mathbb{C}^{N\times M}$ (pas forcément carrée) de rang $r \leq \min(N, M)$. Alors il existe deux matrices U et V orthogonales de taille $N \times r$ et $M \times r$, ainsi qu'une matrice diagonale $S = \mathrm{diag}(s_1, \ldots, s_r)$ de taille r à diagonale réelle strictement positive, telles que

$$A = USV^*.$$

Quand A est réelle, U et V peuvent être choisies réelles.

Démonstration. Procédons par analyse-synthèse. On choisit sans perte de généralité $N \geq M$ (sinon, on travaille avec A^*). Si $A = USV^*$, alors

$$A^*A = VS^2V^*.$$

On prend donc la décomposition en valeurs propres de A^*A, qui a r valeurs propres non-nulles (on rappelle que A^*A est hermitienne, à valeurs propres positives ou nulles), et on choisit pour V une famille de vecteurs propres orthonormés associés à toutes les valeurs propres non-nulles, et S la racine de ces valeurs propres. On a

$$AV = US.$$

donc $AV_k = s_kU_k$ avec V_k et U_k les i-ièmes colonnes de V et U. On choisit $U_k = (AV_k)/s_k$ pour $k = 1, \ldots, r$, et on vérifie facilement que les U_k forment une famille orthonormée.

Finalement, on vérifie a posteriori que r est le rang de A. Clairement, $\mathrm{Im}(A) \subset \mathrm{Span}(U_1, \ldots, U_r)$ par propriétés du produit à droite. Mais par ailleurs, $U_k = A\frac{Vk}{s_k} \in \mathrm{Im}(A)$. Finalement, on a $\dim(\mathrm{Span}(U_1, \ldots, U_r)) = r$ puisque les vecteurs sont orthogonaux. □

Commentaires :

— Il est utile d'écrire cette décomposition comme

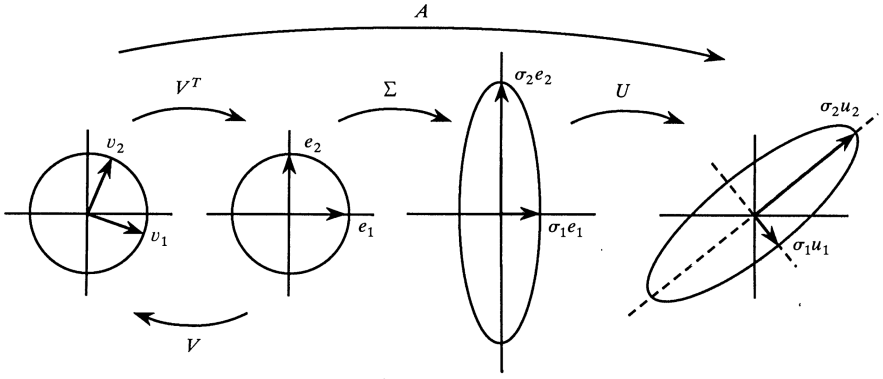
$$A = \sum_{k=1}^{r} s_kU_kV_k^*,$$

où les s_k sont les éléments diagonaux de A, et U_k et V_k sont les colonnes de U et V. C'est une décomposition de A en somme de r matrices de rang 1. Les s_k sont appelées valeurs singulières (ce sont les racines des valeurs propres de A^*A, et elles sont donc déterminées uniquement par A), et les U_k et V_k (qui eux ne sont pas uniques) sont les vecteurs singuliers (à gauche et à droite respectivement).

— Il est parfois utile d'avoir une décomposition "pleine" de type $A = USV^*$ avec $U \in \mathbb{C}^{M\times M}, V \in \mathbb{C}^{N\times N}$ et $S \in \mathbb{C}^{M\times N}$ avec des éléments non-nuls uniquement sur sa diagonale. Pour cela, il suffit de compléter les U et V de la SVD en deux bases orthonormales, et de rajouter des zéros dans S.

— La décomposition en valeurs propres d'une matrice symétrique donne immédiatement une décomposition en valeurs singulières, quitte à prendre $U_k = -V_k$ et $s_k = -\lambda_k$ si $\lambda_k < 0$. Pour une matrice non symétrique, les décompositions en valeurs singulières et propres sont très différentes : les valeurs propres insistent sur le fait que $U = V$, les valeurs singulières sur le fait que U et V soient orthogonales. Pour une matrice carrée, les valeurs propres renseignent sur le comportement de A^n quand $n > 1$ (parce que $UDU^{-1}UDU^{-1} = UD^2U^{-1}$), les valeurs singulières pas du tout (parce que $V^*U \neq 1$).

— Géométriquement, la SVD établit le fait que n'importe quelle transformation linéaire peut s'écrire comme une rotation/réflexion, puis un étirement des axes, puis une autre rotation/réflexion. En particulier, quand $A \in \mathbb{R}^{N\times N}$, il est intéressant de considérer l'image de la sphère unité de $\mathbb{R}^N$ par A. Comme V^* est une isométrie, l'image de la sphère est elle-même. L'image de la sphère par S est une ellipse (faire le calcul pour $N = 2$), dont les longueurs de demi-axes sont les valeurs singulières de A. Finalement, l'image de l'ellipse par une transformation isométrique est également une ellipse, dont les axes principaux sont donnés par les vecteurs de U.



D'après Gilbert Strang, https://www.engineering.iastate.edu/~julied/classes/CE570/Notes/strangpaper.pdf

— La SVD fournit une base orthonormée des quatre sous-espaces fondamentaux de la Figure 1. Par exemple, sur la décomposition pleine, $AV_k = s_kU_k$ montre que les $U_k, k = 1, \ldots, r$ appartiennent à l'image de A et donc en forment une base orthonormée. De même, les $V_k, k = r + 1, \ldots, N$ vérifient $AV_k = 0$ donc forment une base du noyau.

4.2 Normes matricielles

On va doter l'espace vectoriel des matrices d'une structure d'espace vectoriel *normé*, ce qui permet de les mesurer.

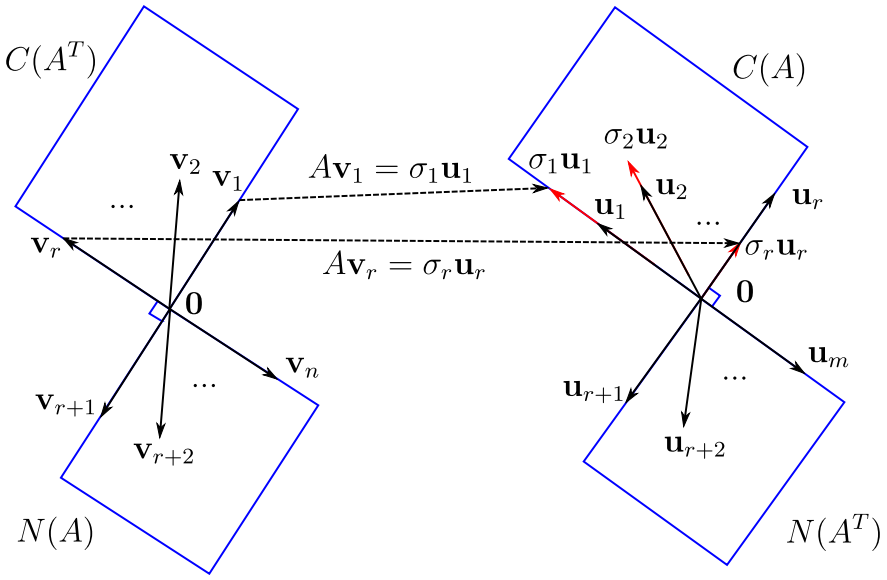


FIGURE 2 – La SVD fournit des bases orthogonales des quatre sous-espaces fondamentaux de la Figure 1. D'après `http://mlwiki.org/index.php/Four_Fundamental_Subspaces`.

4.2.1 Norme induite

La première notion de norme est celle de *norme induite* : pour $A \in \mathbb{C}^{N\times M}$,

$$\|A\| = \max_{x\in\mathbb{R}^M\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x\in\mathbb{R}^M, \|x\|=1} \|Ax\|$$

La raison pour laquelle ce max est atteint est qu'il s'agit de la maximisation d'une fonction continue (exercice) sur un compact. Cette norme (parfois appelée *norme d'opérateur* ou *norme triple*, notée $\|A\|_{\text{op}}$ ou $|||A|||$) dépend des normes choisies sur $\mathbb{R}^M$ et sur $\mathbb{R}^N$. En pratique, on choisira presque toujours la norme euclidienne, de loin la plus utile : si on ne précise pas, c'est celle qui sera choisie.

Proposition 4.2. La norme induite est une norme. Elle est *sous-multiplicative*, c'est-à-dire que $\|AB\| \leq \|A\|\|B\|$.

Démonstration. On vérifie facilement les axiomes d'une norme ; pour l'inégalité triangulaire,

$$\|A+B\| = \max_{x\in\mathbb{R}^M\neq 0} \frac{\|(A+B)x\|}{\|x\|} \leq \max_{x\in\mathbb{R}^M\neq 0} \frac{\|Ax\|+\|Bx\|}{\|x\|} \leq \max_{x\in\mathbb{R}^M\neq 0} \frac{\|Ax\|}{\|x\|} + \max_{x\in\mathbb{R}^M\neq 0} \frac{\|Bx\|}{\|x\|} = \|A\| + \|B\|.$$

Pour la sous-multiplicativité, on utilise

$$\|AB\| = \max_{x\in\mathbb{R}^M\neq 0} \frac{\|ABx\|}{\|x\|} = \max_{x\in\mathbb{R}^M\neq 0} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \leq \|A\|\|B\|$$

□

La norme n'est pas entièrement déterminée par les valeurs propres, comme le montre l'exemple de la matrice $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. On a cependant une majoration dans un sens :

Proposition 4.3. Le *rayon spectral* $\rho(A)$ d'une matrice carrée est défini comme sa plus grande valeur propre en module :

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

On a l'inégalité

$$\|A\| \geq \max_{\lambda \in \sigma(A)} |\lambda|$$

avec égalité dans le cas où A est normale.

Démonstration. Si $Ax = \lambda x$ avec $x \neq 0$, alors $\|Ax\| = |\lambda| \|x\|$ donc $\|A\| \geq |\lambda|$. Dans le cas normal, si $Av_i = \lambda_i v_i$ est une base orthonormée de vecteurs propres, on a par décomposition orthogonale dans la base des v_i la combinaison barycentrique

$$\|Ax\|^2 = \sum_{i=1}^{N} |v_i^* x|^2 |\lambda_i|^2 \leq \rho(A)^2 \sum_{i=1}^{N} |v_i^* x|^2 \leq \rho(A) \|x\|^2.$$

□

On a

$$\|Ax\|^2 = (Ax)^*(Ax) = x^*(A^*A)x.$$

Comme les valeurs propres de A^*A sont les carrés des valeurs singulières de A, cette formule établit une connexion entre la norme et les valeurs singulières :

Proposition 4.4. La norme induite par la norme euclidienne d'une matrice est égale à sa plus grande valeur singulière.

Démonstration. On a

$$\|A\|^2 = \max_{x \in \mathbb{R}^M, \|x\|=1} \|Ax\|^2 = \max_{x \in \mathbb{R}^M, \|x\|=1} x^*(A^*A)x$$

qui est égal par la proposition 3.3 à la plus grande valeur propre de A^*A, c'est-à-dire le carré de la plus grande valeur singulière. □

On vérifie aisément que la norme induite est invariante par multiplication à gauche et à droite par une matrice unitaire, ce qui fournit également une autre preuve de la proposition ci-dessus.

Plus généralement, les valeurs singulières obéissent à des formules de type min-max comme à la Proposition 3.3, en remplaçant le quotient de Rayleigh $\frac{x^*Ax}{x^*x}$ par le *facteur de croissance* $\frac{\|Ax\|}{\|x\|}$.

4.2.2 Norme de Frobenius

Une norme alternative à la norme induite est la *norme de Frobenius* (ou Hilbert-Schmidt)

$$\|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^*A)} = \sqrt{\sum_{i=1,\dots,N, j=1,\dots,M} |A_{ij}|^2}$$

Cette norme provient du produit scalaire

$$\langle A, B\rangle = \mathrm{Tr}(A^*B) = \sum_{i=1,\dots,N, j=1,\dots,M} \overline{A_{ij}} B_{ij}$$

Ces formules montrent que cette norme et ce produit scalaire découlent naturellement de l'identification d'une matrice $A \in \mathbb{C}^{N\times M}$ avec sa forme vectorisée dans $\mathbb{C}^{NM}$.

Proposition 4.5. La norme de Frobenius est égale à la norme euclidienne du vecteur des valeurs singulières.

Démonstration. Pour $A = USV^*$ on calcule

$$\|A\|_{\mathrm{F}}^2 = \mathrm{Tr}(A^*A) = \mathrm{Tr}(V^*S^2V) = \mathrm{Tr}(S^2).$$

□

En particulier, $\|A\|_F \geq \|A\|$.

Proposition 4.6. La norme de Frobenius est sous-multiplicative.

Démonstration. $\|A\|_{\mathrm{F}}^2 = \sum_{j=1}^N \|A_j\|^2$, où A_j est la j-ième colonne de A. Il suit donc par les propriétés de la multiplication à gauche (cf. Section 1.2.2) que

$$\begin{aligned}
\|AB\|_{\mathrm{F}}^2 &= \sum_{j=1}^N \|AB_j\|^2 \\
&\leq \|A\| \sum_{j=1}^N \|B_j\|^2 \\
&= \|A\|\|B\|_{\mathrm{F}} \\
&\leq \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}
\end{aligned}$$

□

L'intérêt principal de la norme de Frobenius est d'être facile à calculer à partir des entrées de A, au contraire de la norme induite par la norme euclidienne.

4.3 Meilleure approximation de rang fixé

La SVD a de nombreuses applications. La principale est celle de l'approximation parcimonieuse d'une matrice, basée sur le théorème suivant :

Théorème 4.7 (Eckart-Young)**.** Soit $A = USV^*$ une matrice $M \times N$ avec ses valeurs singulières rangées par ordre décroissant. Alors, pour $k \leq \min(M, N)$, la troncation de la SVD au rang k

$$A_k = \sum_{n=1}^{k} s_n U_n V_n^*$$

minimise la distance à A en norme d'opérateur parmi toutes les matrices de rang k :

$$\|A - A_k\| = \min_{\text{rang}(B)=k} \|A - B\|.$$

(ce résultat est également vrai en norme de Frobenius)

Démonstration. On utilise une SVD pleine $A = USV^*$ avec $U \in \mathbb{C}^{M\times M}$ et $V \in \mathbb{C}^{N\times N}$, et on va supposer $M \geq N$ sans perte de généralité.

Pour tout B de rang k, on peut faire le changement de variable bijectif $\widetilde{B} = U^*BV$, et voir que $\|A-B\| = \|U(S-\widetilde{B})V^*\| = \|S-\widetilde{B}\|$. On peut donc se restreindre au cas où $A = S$ est une matrice $M \times N$ avec des non-zéros sur la diagonale uniquement. On veut montrer que n'importe quel choix de B de rang k a un écart à A plus grand que le choix $B = \text{diag}(s_1, s_2, \ldots, s_k, 0, \ldots, 0)$, qui réalise un écart s_{k+1}.

Puisque B est de rang k, sa restriction à ses $k+1$ premières colonnes admet un noyau non-nul (l'intersection entre $\ker(B)$ et $\text{Vect}(e_1, \ldots, e_{k+1})$). Soit x un vecteur dans cette intersection, qu'on choisit normalisé. Alors,

$$\|A - B\|^2 \geq \|(A - B)x\|^2 = \|Ax\|^2 = \sum_{i=1}^{k+1} s_i |x_i|^2 \geq s_{k+1}.$$

□

Un exemple classique d'application est l'analyse en composantes principales en statistiques. On prend $A \in \mathbb{R}^{N\times n}$ correspondant à N observations de n variables aléatoires $X_1, \ldots, X_n$. Si on soustrait à chaque colonne sa moyenne (empirique), la matrice A^*A est la matrice de covariance (empirique). Ses vecteurs propres les plus grandes satisfont un principe max-min sur x^*A^*Ax, c'est-à-dire la variance (empirique) de la variable aléatoire $\sum_{i=1}^{n} x_i X_i$: faire une SVD sur A revient donc à chercher les combinaisons linéaires orthogonales de variables aléatoires qui maximisent les variances. C'est très utile par exemple pour visualiser un dataset de grande dimension.

Une autre application possible peut être la compression d'image : si $A \in \mathbb{R}^{M\times N}$ représente une image de $M \times N$ pixels, sa représentation compressée de rang k contient $(M + N)k \ll MN$ nombres. (Cette compression a un intérêt principalement théorique, puisque d'autres méthodes bien plus efficaces existent pour la compression d'image.)

4.4 Séries de Neumann

On rappelle que tout espace vectoriel normé de dimension finie est complet, et que, dans un espace complet, les séries normalement convergentes convergent.

On établit une formule fondamentale :

Proposition 4.8 (Séries de Neumann)**.** Si A et B sont des matrices carrées, que A est inversible et que $\|B\| < \frac{1}{\|A^{-1}\|}$, $A - B$ est inversible, et

$$\begin{aligned}(A-B)^{-1} &= A^{-1}(1-BA^{-1})^{-1} \\ &= A^{-1} + A^{-1}BA^{-1} + A^{-1}BA^{-1}BA^{-1} + \dots\end{aligned}$$

Démonstration. Comme $\|(A^{-1}B)^n\| \leq \|A^{-1}B\|^n$ par sous-multiplicativité, la série $S = \sum_{n\geq 0}(A^{-1}B)^n$ converge normalement dans l'espace complet $M_N(\mathbb{C})$, donc converge. Soit S_N sa somme partielle jusqu'au terme N. Alors on a la série téléscopique

$$(1 - A^{-1}B)S_N = 1 - (A^{-1}B)^{N+1} \xrightarrow{N\to\infty} 1.$$

Par ailleurs, comme la norme d'opérateur est sous-multiplicative, la multiplication est continue : $\|(1 - A^{-1}B)(S_N - S)\| \leq \|(1 - A^{-1}B)\|\|S_N - S\| \to 0$. En passant à la limite, on voit donc que $(1 - A^{-1}B)S = 1$ donc que $S = (1 - A^{-1}B)^{-1}$.

On écrit ensuite

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

donc $A - B$ est inversible, avec

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

On en déduit la formule de la série de Neumann par continuité de la multiplication (à droite cette fois). □

4.5 Conditionnement

Supposons qu'on veuille résoudre le problème

$$Ax = b,$$

bien posé (c'est-à-dire que A est inversible), mais que nos données (A, b) sont entachées d'une petite erreur $(\delta A, \delta b)$. Cette erreur peut provenir de la modélisation (les données sont bruitées, ou le problème n'est pas exactement connu), ou d'erreur numérique (les nombres en double précision sont stockés avec une précision relative de $\approx 10^{-16}$). On résoud alors en pratique

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

À quel point peut-on être sûrs que l'erreur δx est petite ?

Pour δA suffisamment petit, $A + \delta A$ est inversible par série de Neumann et le problème est bien posé. Imaginons que δA et δb soient petits, de sorte que les termes d'ordre $\|\delta A\|\|\delta b\|$ (donc $\|\delta A\|\|\delta x\|$) sont négligeables (cf exercice 10 pour le cas général). On a alors

$$\begin{aligned}Ax + A\delta x + \delta A x &= b + \delta b x \\ \delta x &= A^{-1}(\delta b - \delta A x) \\ \|\delta x\| &\leq \|A^{-1}\|(\|\delta b\| + \|x\|\|\delta A\|)\end{aligned}$$

On est donc tenté de définir par exemple la sensibilité de δx par rapport à δb par $\|A^{-1}\|$. Cette mesure a cependant une *unité*, et n'est donc pas intrinsèque : si on remplace $Ax = b$ par $2Ax = 2b$ (ou qu'on décide de passer des mètres aux kilomètres dans notre mesure d'une quantité), le problème n'est pas intrinsèquement plus sensible à ses données. Par ailleurs, dans de nombreux contextes applicatifs, il est beaucoup plus pratique de raisonner en termes d'erreurs *relatives* plutôt qu'absolues. On va donc diviser cette inégalité par $\|x\| \geq \frac{1}{\|A\|}\|b\|$ pour obtenir

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A^{-1}\|\|A\|}_{\kappa(A)} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right).$$

La quantité $\kappa(A)$ est appelée *conditionnement* (pour la norme ambiante). Comme les valeurs singulières de l'inverse d'une matrice sont les inverses des valeurs singulières de la matrice, $\kappa(A)$ est égal au ratio de la plus grande sur la plus petite valeur singulière de A. Géométriquement, le nombre sans dimension κ mesure à quel point l'ellipse AS, avec S la sphère unité, est *excentrique* (aplatie).

On a raisonné sur la résolution de $Ax = b$, mais on peut également se poser la question : si je fais une petite erreur (relative) $\frac{\|\delta x\|}{\|x\|}$ sur x, quelle erreur (relative) $\frac{\|A\delta x\|}{\|Ax\|}$ fais-je sur Ax ? Il est facile de montrer que la réponse est encore une fois donnée par le conditionnement : ce n'est pas une surprise puisque $\kappa(A) = \|A\|\|A^{-1}\| = \kappa(A^{-1})$.

Une matrice mal conditionnée est souvent une matrice qui possède des échelles très différentes. Par exemple, si on a un problème $Ax = b$ très bien conditionné (disons $A = I$ avec $N = 2$) avec x mesuré en mètres et si on décide (artificiellement) de commencer à mesurer certaines inconnues du problème en kilomètres et d'autres en milimètres, la matrice A va devenir très mal conditionnée (d'un facteur kilomètres/milimètres $= 10^6$). Cet exemple est un peu stupide, mais dans des vrais problèmes on peut souvent chercher à modéliser des phénomènes qui se déroulent à des échelles (de temps ou d'espace) très différentes, ce qui produit un mauvais conditionnement.

4.6 Exercices

1. (SVD) Montrer que les valeurs singulières d'une matrice sont invariantes par multiplication à gauche ou à droite par une matrice unitaire. Montrer que ce n'est pas le cas pour une matrice orthogonale (rectangulaire).
2. (SVD) Montrer que la plus grande (resp. plus petite) valeur singulière de A ne peut que croître (resp. décroître) quand on ajoute une colonne à A.
3. (SVD) Calculer les valeurs propres, singulières, le déterminant et le conditionnement des matrices

$$\begin{pmatrix} 1/100 & 0 \\ 0 & 100 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$$

   Remarquer que le déterminant est un très mauvais indicateur du conditionnement d'une matrice.
4. (SVD) Calculer les valeurs singulières d'une matrice $N \times 2$ constituée de deux vecteurs de taille N normalisés, rangés en colonne. Interpréter géométriquement.
5. (SVD) Si $A = USV^*$ est une matrice $N \times n$ de rang n avec $n \leq N$, on définit $B = UV^*$. Montrer que $B = A(A^*A)^{-1/2}$, où la racine carrée d'une matrice HDP sous forme diagonalisée PDP^* est donnée par $PD^{1/2}P^*$ avec $D^{1/2}$ valant la racine des entrées de D sur la diagonale. Calculer $\|A - B\|$ et $\|A - B\|_F$. Comparer avec la factorisation QR.

6. (Conditionnement) Calculer les vecteurs et valeurs propres de la matrice

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

Calculer le conditionnement de la matrice des vecteurs propres.

7. (SVD) Montrer que toute matrice A peut s'écrire sous *forme polaire* $A = QH$ avec Q orthogonale et H hermitienne définie positive (indice : calculer A^*A)
8. (SVD) Reprendre le problème des moindres carrés en utilisant la SVD. Quel est l'effet de la régularisation de Tikhonov ?
9. (Stabilité des valeurs propres) Montrer que les valeurs propres d'une matrice hermitienne sont stables par perturbation dans le sens suivant : si $\lambda_i(A)$ est la i-ième valeur propre (comptée avec multiplicité) de A, alors

$$|\lambda_i(A+B) - \lambda_i(B)| \leq \|B\|$$

(on utilisera la formule min-max pour obtenir $\lambda_i(A+B) \leq \lambda_i + \|B\|$). Même question pour les valeurs singulières.

10. (Stabilité de $Ax = b$) Refaire le raisonnement de la section 4.5 sans négliger les termes d'ordre $\delta A \delta b$, en utilisant les séries de Neumann.
11. (Série de Neumann) Quel est le lien entre les séries de Neumann et le théorème du point fixe de Banach (aussi appelé Picard), si vous le connaissez ?
12. (Série de Neumann) Montrer par série de Neumann que la matrice

$$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$$

est inversible pour ε suffisamment petit, qu'on précisera, et donner son inverse sous forme de série. Comparer avec l'inverse exact. Même question pour

$$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 2 \end{pmatrix}$$

13. (Série de Neumann et formule de Sherman-Morrison) Soit $A \in \mathbb{C}^{N\times N}$ inversible et $u \in \mathbb{C}^N$. On considère la matrice $M_\varepsilon = A + \varepsilon uu^*$ pour tout $\varepsilon \in \mathbb{C}$. Montrer par série de Neumann que M_ε est inversible pour ε suffisamment petit et montrer que "l'inverse d'une perturbation de rang 1 est une perturbation de rang 1 de l'inverse". En déduire un critère d'inversibilité et un inverse explicite de M_ε pour une valeur quelconque de ε.
14. (Conditionnement) Calculer le conditionnement de la matrice des différences finies en dimension 1. Comment interpréter ce mauvais conditionnement ?
15. (Produit scalaire de Frobenius) Montrer que $M_N(\mathbb{C}) = \mathrm{H}_N \oplus \mathrm{AH}_N$, où H et AH sont les matrices hermitiennes et anti-hermitiennes, et où $\oplus$ est la somme directe orthogonale pour le produit scalaire de Frobenius.
16. (Produit scalaire de Frobenius) Montrer que les matrices de Pauli

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{4.1}$$

forment une famille orthogonale de matrices hermitiennes de dimension 2. Donner une base orthonormée des matrices hermitiennes de dimension 2.

17. (Forme de Schur) Soit A une matrice carrée quelconque. On va montrer que, pour tout $\varepsilon > 0$, il existe une norme $\|\cdot\|_\varepsilon$ sur $\mathbb{C}^N$ telle que $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$, où $\|A\|_\varepsilon$ est la norme induite par $\|\cdot\|_\varepsilon$.
    (a) Si M est une matrice, on pose $\|x\|_M = \|Mx\|$. Calculer la norme induite par cette norme.
    (b) En utilisant la décomposition de Schur ainsi que la matrice $D = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{N-1})$, montrer que, pour tout $\varepsilon > 0$, il existe M_ε telle que $MAM^{-1} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N) + B_\varepsilon$, avec $\|B_\varepsilon\| \leq C\varepsilon$ avec C une constante qui ne dépend pas de ε. En déduire le résultat.

5 Un petit détour par l'analyse complexe

Dans cette section, on développe une introduction très rapide à l'analyse complexe, notre but étant simplement d'arriver à la formule de Cauchy dont nous aurons besoin dans la suite. Dans cette partie, on relâchera notre niveau de rigueur et on se contentera d'énoncés relativement imprécis et de preuves heuristiques, en procédant par analogie avec la formule de Stokes. L'analyse complexe est un sujet extrêmement riche qui mérite un traitement plus détaillé, mais au vu des contraintes de ce cours, le lecteur vraiment intéressé par l'analyse complexe est invité à consulter un cours ou un ouvrage sur le sujet.

5.1 Formule de Stokes

On fait un bref rappel de la formule de Stokes sur $\mathbb{R}^2$. Le théorème de Stokes fait partie d'une grande famille de théorèmes de type "une certaine intégrale d'une fonction sur le bord d'un domaine est égale à une certaine intégrale de sa dérivée sur l'intérieur" dont font partie le théorème fondamental de l'analyse dans $\mathbb{R}$, la formule de Green ainsi que le théorème de la divergence (tous ces théorèmes étant des cas particuliers de la formule de Stokes généralisée aux formes différentielles de dimension arbitraire, que nous n'aborderons pas).

Dans cette sous-section, et dans cette sous-section uniquement, on met des flèches sur les vecteurs. Un *champ de vecteur* $\vec{f}(\vec{x})$ est une application lisse de $\mathbb{R}^2$ dans $\mathbb{R}^2$. On considère une courbe simple fermée lisse C du plan, qui définit un intérieur U. Cette courbe est orientée trigonométriquement (quand on part de l'intérieur du domaine, la courbe "va vers la gauche"). On définit la *circulation*

$$\int_C \vec{f}(\vec{x}) \cdot d\vec{l} = \int_a^b \vec{f}(\vec{\gamma}(t)) \cdot \vec{\gamma}'(t) dt$$

où $\gamma(t) : [a, b] \to C$ est une paramétrisation de la courbe C qui respecte son orientation. On définit également le rotationnel $\vec{\nabla} \wedge \vec{f}$ par

$$\vec{\nabla} \wedge \vec{f}(\vec{x}) = \partial_x f_y - \partial_y f_x.$$

C'est donc une application $\mathbb{R}^2 \to \mathbb{R}$. Le théorème de Stokes convertit la circulation sur C en intégrale du rotationnel sur son intérieur U :

$$\int_C \vec{f}(\vec{x}) \cdot d\vec{l} = \int_U \vec{\nabla} \wedge \vec{f}(\vec{x}) dx$$

Une brève ébauche de preuve est comme suit. On pave le domaine U en un ensemble de petits carrés, de taille $\varepsilon \times \varepsilon$. L'intégrale du rotationnel est égale à la somme sur tous les carrés des intégrales. Sur chaque carré, on peut approcher $\vec{\nabla} \wedge \vec{f}$ par différences finies en fonction de ses valeurs sur chacun des coins du carré, et se rendre compte que la formule de Stokes est vraie sur chacun des carrés, à un reste d'ordre $O(\varepsilon^3)$ près. On somme ensuite les contributions des $O(\varepsilon^{-2})$ carrés : les contributions venant des bords internes se compensent, et il ne reste plus que les contributions venant du bord du domaine pavé en carrés, ainsi qu'un reste $O(\varepsilon^3 \varepsilon^{-2}) = O(\varepsilon)$. On vérifie enfin que les termes de bord approchent bien la circulation.

5.2 Fonction holomorphe

Une fonction de $\mathbb{C} \to \mathbb{C}$ est dite *holomorphe* sur un ouvert $U \subset \mathbb{C}$ si elle est dérivable au sens complexe, c'est-à-dire si $\frac{f(z+h)-f(z)}{h}$ admet une limite, notée $f'(z)$, quand $h \to 0$ pour tout $z \in U$. La subtilité ici est que cette limite ne *dépend pas* de la façon dont $h \to 0$ dans $\mathbb{C}$, c'est-à-dire que la différentielle prend la forme d'une multiplication par un complexe, $df(z) \cdot h = f'(z)h$. Géométriquement, une fonction holomorphe agit localement comme la multiplication de h par un nombre complexe $f'(z)$, donc représente une homothétie fois une rotation. La notion d'holomorphie est donc bien plus forte que celle de simple dérivabilité comme fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$: une fonction régulière $u(x, y) + iv(x, y)$ de $z = x + iy$ n'est généralement pas holomorphe : elle l'est seulement si u et v vérifie certaines relations de compatibilité, appelées équations de Cauchy-Riemann (voir exercice).

Des exemples élémentaires de fonctions holomorphes sont les polynômes, puisque $(z + h)^n = z^n + nz^{n-1}h + O(h^2)$. Il est également facile de vérifier que les fonctions définies par série entière sont holomorphes à l'intérieur de leur rayon de convergence. En particulier, l'exponentielle est holomorphe sur $\mathbb{C}$. Des exemples de fonctions dérivables au sens réel mais non-holomorphes sont les fonctions $z \mapsto \overline{z}$ (qui représente une réflexion, pas une rotation), ainsi que les fonctions qui impliquent une conjugaison, telles que $z \mapsto \mathrm{Re}(z)$, $z \mapsto \mathrm{Im}(z)$, $z \mapsto |z|^2$ (d'ailleurs, par les équations de Cauchy-Riemann, une fonction holomorphe non triviale ne peut pas être à valeurs réelles).

5.3 Intégrale de contour

Un *contour* est une courbe orientée fermée dans $\mathbb{C}$. L'intégrale de contour d'une fonction $f : \mathbb{C} \to \mathbb{C}$, notée

$$\oint_C f(z)dz$$

est une quantité définie en approchant C par des petits segments $[z_i, z_{i+1}]$, en sommant les $f(z_i)(z_{i+1} - z_i)$, et en passant à la limite, de la même façon que l'intégrale de Riemann. En particulier, si $\gamma(t), t \in [a, b]$ est une paramétrisation lisse de C compatible avec son orientation, avec $\gamma(a) = \gamma(b)$, alors

$$\oint_C f(z)dz = \int_a^b f(\gamma(t))\gamma'(t)dt$$

Il est à noter que la quantité $f(\gamma(t))\gamma'(t)$ est une *multiplication complexe*, qui mélange les parties réelles et imaginaires de f et de γ' : l'intégrale de contour n'est *pas* une intégrale séparée des parties réelles et imaginaires de la fonction $f(z)$ le long du contour, ni une circulation du champ de vecteur $(\mathrm{Re}f, \mathrm{Im}f)$ ou une quantité classique de ce genre, mais bien un objet propre à l'analyse complexe qui obéit à ses propres règles. Malgré cela, certaines propriétés usuelles des intégrales fonctionnent toujours : on pourra par exemple sommer ou dériver sous le signe intégrale à condition que l'intégrande soit dominée. On voit également que $|\oint_C f(z)dz| \leq |C| \max_{z \in C} |f(z)|$, avec $|C|$ la longueur de C.

La propriété fondamentale de l'intégrale de contour, qui donne sa richesse à l'analyse complexe, est le théorème de Cauchy : l'intégrale $\oint_C f(z)dz$ sur un contour C entièrement contenu dans un domaine simplement connexe (sans trou) U sur lequel la fonction f est holomorphe vaut 0. C'est une variante des formules de Stokes (voir exercice) : l'intégrale de contour, même si elle n'est pas

une circulation, y est analogue, et l'analogue du "rotationnel" d'une fonction holomorphe est nul, et donc l'intégrale de contour est nulle. Une conséquence est qu'on peut *déformer* un contour sans changer la valeur de l'intégrale de contour, du moment que la déformation reste dans le domaine d'holomorphie. C'est l'analogue des *formes différentielles exactes* en géométrie différentielle, donc l'intégrale ne dépend pas du chemin suivi, ou de l'énoncé en physique que le travail fourni par une force conservative ne dépend pas du chemin suivi.

Un petit calcul simple mais très utile est celui de

$$\oint_{C_r} z^n dz = r^{n+1} \int_0^{2\pi} e^{int} e^{it} i dt = 2\pi i \delta_{n,-1}$$

quand C_r est un cercle de rayon r entourant l'origine (résultat obtenu par la paramétrisation $\gamma(t) = re^{it}$, $\gamma'(t) = ire^{it}$). On note que le résultat de cette intégrale ne dépend pas de r, ce qui est normal parce que z^n est holomorphe dans l'anneau de rayons r et r' et le contour peut donc être déformé d'un cercle en un autre. Quand $n \geq 0$, la fonction est holomorphe sur $\mathbb{C}$ et son intégrale de contour est nécessairement nulle. Dans le cas $n < 0$, z^n n'est pas holomorphe en 0, ce qui rend possible le fait que $\oint_{C_r} z^{-1} dz = 2\pi i$ (au sens des distributions, son "équivalent de rotationnel" est nul presque partout, mais vaut un dirac en 0). Moralement, dans cette intégrale, la phase de $dz = \gamma'(t)dt$ tourne une fois, celle de z^n tourne n fois, donc pour que l'intégrale soit non-nulle, il faut que $n = -1$.

5.4 Formule de Cauchy

Une application extrêmement utile du théorème de Cauchy est la formule de Cauchy

$$\oint_C \frac{f(z)}{z - \lambda} dz = 2\pi i f(\lambda)$$

quand la fonction f est holomorphe dans un ouvert U et que C, contenu dans U, entoure λ une fois dans le sens trigonométrique. Cette formule se démontre de la façon suivante : comme $f(z)/(z - \lambda)$ est holomorphe en dehors de λ, on peut déformer le contour C de façon à former un cercle C_ε de rayon ε autour de λ. Sur ce cercle, on peut approcher $f(z)$ par $f(\lambda)$, et par ailleurs on a déjà calculé que $\int_{C_\varepsilon} \frac{1}{z-\lambda} dz = 2\pi i$.

Cette formule exprime la possibilité, extrêmement puissante, d'obtenir les valeurs de $f(\lambda)$ comme combinaison linéaire continue, paramétrée par z, de fonctions élémentaires $1/(z - \lambda)$, et avec le poids $f(z)$. Cette possibilité de décomposer une fonction compliquée en fonctions plus simples a notamment pour conséquence que $\lambda \mapsto f(\lambda)$ est une fonction analytique (développable en série entière), comme intégrale (dominée) de fonctions analytiques $\lambda \mapsto \frac{1}{z-\lambda}$. Une fonction holomorphe (dérivable au sens complexe) est donc en particulier infiniment dérivable au sens réel.

5.5 Exercices

1. (Cauchy-Riemann) Montrer que, si $f : \mathbb{C} \to \mathbb{C}$ est holomorphe avec $f(x + iy) = u(x, y) + iv(x, y)$, alors sa jacobienne quand elle est vue comme une fonction $(x, y) \mapsto (u, v)$ de $\mathbb{R}^2 \to \mathbb{R}^2$ est proportionnelle à une matrice de rotation. En déduire les équations de Cauchy-Riemann $\partial_x u = \partial_y v, \partial_y u = -\partial_x v$.

2. (Théorème de Cauchy) Dans les conditions d'application du théorème de Cauchy, on a

$$\oint_C f(z)dz = \oint_C (u+iv)(\vec{dl} \cdot (\vec{e_x} + i\vec{e_y}))$$

avec $\vec{dl}$ le vecteur tangent à C. Appliquer la formule de Stokes et utiliser les équations de Cauchy-Riemann pour démontrer le théorème de Cauchy.

3. (Fonction méromorphe) Une fonction *méromorphe* sur un ouvert U est une fonction holomorphe sauf en un nombre fini de pôles, c'est-à-dire de points z_0 tels que $(z-z_0)^n f(z)$ est holomorphe autour de z_0 pour un certain n (l'ordre du pole). Remarquer par déformation de contour que l'intégrale de contour d'une fonction méromorphe ne dépend que du comportement local de f autour de ses pôles.

4. (Intégration de contour) Calculer $\int_{-\infty}^{+\infty} \frac{1}{1+x^2}dx$ en calculant $\oint_C \frac{1}{1+z^2}dz$ pour un contour C formé du segment $[-R, R]$, fermé en un demi-cercle de rayon R dans le plan complexe supérieur.

5. (Formule de Cauchy) Montrer qu'une fonction bornée et analytique sur $\mathbb{C}$ est forcément constante (on pourra dériver la formule de Cauchy). En déduire qu'un polynôme a nécessairement une racine sur $\mathbb{C}$ (on pourra raisonner par l'absurde et montrer que l'inverse est analytique sur $\mathbb{C}$).

6. (Logarithme et racines) Montrer qu'il n'existe pas de logarithme défini globalement, c'est-à-dire une fonction continue satisfaisant $\log(\exp(z)) = z$ pour tout $z \in \mathbb{C}$. (on pourra considérer $z(\theta) = e^{i\theta}$). On définit $\log_p(re^{i\theta}) = \log r + i\theta$ (la *détermination principale* du logarithme). Sur quel ensemble cette fonction est-elle définie ? Est-elle holomorphe ? Comment définir la fonction $\sqrt{z}$? Sur quel ensemble ?

6 Calcul fonctionnel

On va se poser dans cette section la question suivante : étant données une fonction f et une matrice carrée A, peut-on donner un sens à $f(A)$? On appelle *calcul fonctionnel* toute méthode raisonnable permettant de ce faire. Vous avez déjà rencontré le calcul fonctionnel pour des polynômes $p(x) = \sum_{n=0}^{N} a_n x^n$, simplement défini comme $p(A) = \sum_{n=0}^{N} a_n A^n$.

6.1 Calcul fonctionnel spectral

Si $A = PDP^{-1}$ est diagonalisable, il est raisonnable de poser

$$f(A) = Pf(D)P^{-1}$$

où $f(D)$ est la matrice diagonale donc les éléments sont obtenus par application de f aux éléments de D. On vérifie facilement que cette définition coïncide avec la définition usuelle dans le cas où p est un polynôme.

Cette définition ne s'étend pas facilement au cas des matrices non-diagonalisables. Certes, "presque toute matrice est diagonalisable" (cf exercice 1 de la section 3) : une matrice donnée, implémentée sur un ordinateur (donc stockée avec une erreur numérique) a toutes les chances d'être diagonalisable. Cependant (et comme souvent en analyse numérique), vérifier une condition qualitative (être diagonalisable) n'est pas si utile que ça si elle n'est pas accompagnée d'une condition quantitative (être suffisamment loin d'une matrice non-diagonalisable). En effet, une matrice presque non-diagonalisable va avoir des vecteurs propres très mal conditionnés (cf exercice 6 de la section 4) et calculer $f(A)$ en calculant ses vecteurs propres est une mauvaise idée (le calcul subira une erreur arbitrairement grande). On va donc chercher une autre définition de $f(A)$.

6.2 Séries entières de matrices

L'étape suivante naturelle est les séries entières.

Définition 6.1 (calcul fonctionnel par série entière)**.** Si $f(z) = \sum_{n=0}^{\infty} a_n z^n$ est une fonction analytique définie par une série entière de rayon de convergence r, et si A est une matrice carrée avec $\|A\| < r$, alors on définit

$$f(A) = \sum_{n=0}^{\infty} a_n A^n.$$

Cette série converge puisqu'elle converge normalement ($\sum_{n=0}^{\infty} \geq 0 |a_n| \|A^n\| \leq \sum_{n=0}^{\infty} |a_n| \|A\|^n < \infty$).

Cette définition jouit de toutes les bonnes propriétés des séries entières : c'est notamment un *morphisme d'algèbre*, c'est-à-dire que les opérations d'algèbre sur les fonctions (addition, multiplication) accomplissent la même chose sur les matrices : on a par exemple $f(A)g(A) = g(A)f(A) = (fg)(A)$ (exercice).

Cette définition permet par exemple de définir l'exponentielle de matrice

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

pour toute matrice carrée A.

Par des arguments classiques sur les fonctions définies par séries entières qu'on peut copier-coller verbatim dans le cas matriciel, on peut *redévelopper* en séries entière : si $\|B\| < r - \|A\|$, alors $f(A+B)$ est une série entière en B. Cela montre en particulier, par un argument de dérivabilité sous le signe somme des séries entières, que $t \mapsto \exp(tA)$ est dérivable et que sa dérivée est égale à $A\exp(tA)$, et donc que $\exp(tA)x_0$ est solution de $\dot{x} = Ax, x(0) = x_0$.

Il est très important en revanche de noter que

$$\exp(A)\exp(B) \neq \exp(A+B)$$

si A et B ne commutent pas. On peut remarquer par exemple que

$$\exp(A)\exp(B) = 1 + A + B + \frac{1}{2}(A^2 + 2AB + B^2) + \dots$$
$$\exp(A+B) = 1 + A + B + \frac{1}{2}(A^2 + AB + BA + B^2) + \dots$$

6.3 Calcul fonctionnel holomorphe

On a actuellement à notre disposition le calcul fonctionnel analytique, qui fonctionne à condition que la matrice soit dans le rayon de convergence de la série. Ce n'est pas une situation très complète, parce qu'on ne peut par exemple pas définir par ce biais $(1+A)^{-1}$ pour les matrices A de norme > 1 telles que $1+A$ soit inversible, alors qu'on sait que cet objet existe. De même, des expressions comme $\log(1+A)$ peuvent être définies sans que $\|A\|$ soit dans le rayon de convergence du logarithme (par exemple, $A = 2$). On va donc chercher un calcul fonctionnel qui permet de s'affranchir de cette limitation. La solution va provenir de l'analyse complexe, et plus précisément de la formule

$$f(a) = \frac{1}{2\pi i}\oint_C \frac{f(z)}{z-a}dz$$

quand le contour C entoure une fois a dans le sens trigonométrique, qu'on va chercher à étendre au cas où a est une matrice. On commence par généraliser la notion de fonction holomorphe et d'intégrale de contour au cas d'une fonction $\mathbb{C} \to \mathbb{C}^{N\times N}$: une telle fonction est holomorphe si chacune de ses composantes l'est, et l'intégrale de contour est définie composante par composante.

On étudie maintenant la fonction $(z-A)^{-1}$.

Proposition 6.2. Soit A une matrice carrée. La *résolvante*

$$R(z) = (z-A)^{-1}$$

est une fonction holomorphe sur tout ouvert de $\mathbb{C}$ n'intersectant pas le spectre de A.

Démonstration. Soit z_0 un point n'étant pas dans le spectre de A, de sorte que $R(z_0)$ est bien défini. Alors, par série de Neumann on a pour tout z

$$z - A = (z_0 - A) - (z_0 - z)$$

et donc, pour tout z tel que $|z - z_0| < 1/\|R(z_0)\|$, $z - A$ est inversible et

$$R(z) = R(z_0) + (z_0 - z)R(z_0))^{-2} + (z_0 - z)^2 R(z_0))^{-2} + \dots$$

qui est donc analytique (développable en série entière) autour de z_0, donc holomorphe. □

On peut donc poser

Définition 6.3 (calcul fonctionnel holomorphe)**.** Si f est holomorphe à l'intérieur d'un ouvert U simplement connexe et que le spectre d'une matrice carrée A est inclus dans U, on pose

$$f(A) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - A} dz$$

pour tout contour C qui entoure une fois dans le sens trigonométrique le spectre de A.

Cette définition est légitime, car, pour tout $i, j = 1, \ldots, N$, la fonction $z \mapsto ((z - A)^{-1})_{ij}$ est analytique (donc continue) sur C. L'intégrale ne dépend pas du choix de C par déformation de contour.

Dans le cas où f est une série entière de rayon de convergence $r > \|A\|$, $f(A)$ avait été précédemment défini à la section 6.2 est applicable. Ces deux définitions coïncident, heureusement.

Proposition 6.4. Si f est développable en série entière de rayon $r > \|A\|$, et que C est un contour qui entoure une fois le spectre de A dans le sens trigonométrique, alors les définitions du calcul fonctionnel par série entière et holomorphe coïncident.

Démonstration. Par invariance par déformation de contour on peut choisir comme contour $C_{r'}$ un cercle centré en 0 de rayon r', avec $r > r' > \|A\|$. On a alors, avec $f(z) = \sum_{n=0}^{\infty} a_n z^n$,

$$\frac{1}{2\pi i} \oint_C \frac{f(z)}{z - A} dz = \frac{1}{2\pi i} \oint_C \sum_{n=0}^{\infty} a_n \frac{z^n}{z - A} dz = \sum_{n=0}^{\infty} a_n \frac{1}{2\pi i} \oint_C \frac{z^n}{z - A} dz$$

où on a utilisé le fait que la série $\sum_{n=0}^{\infty} a_n \frac{z^n}{z-A}$ converge normalement, uniformément en $z \in C$. On calcule ensuite, pour n fixé,

$$\begin{aligned} \frac{1}{2\pi i} \oint_{C_{r'}} \frac{z^n}{z - A} dz &= \frac{1}{2\pi i} \oint_{C_{r'}} z^{n-1} \sum_{k=0}^{\infty} \left(\frac{A}{z}\right)^k dz \\ &= \sum_{k=0}^{\infty} \frac{1}{2\pi i} \oint_{C_{r'}} z^{n-k-1} A^k dz \\ &= A^n \end{aligned}$$

où on a utilisé le fait que la série de Neumann $\frac{1}{z} + \frac{1}{z^2} A + \ldots$ converge normalement, uniformément en $z \in C$, et la formule $\oint_C z^{n-k-1} dz = 2\pi i \delta_{nk}$. □

Proposition 6.5. Si f et g sont holomorphe à l'intérieur d'un ouvert U simplement connexe et que le spectre d'une matrice carrée A est inclus dans U, $f(A)g(A) = (fg)(A)$.

Démonstration. On calcule, en prenant pour C' un contour inclus dans C et contenant $\sigma(A)$,

$$\begin{aligned} f(A)g(A) &= \frac{1}{(2\pi i)^2}\oint_C\oint_{C'} f(z)g(z')\frac{1}{z-A}\frac{1}{z'-A}dzdz' \\ &= \frac{1}{(2\pi i)^2}\oint_C\oint_{C'} f(z)g(z')\frac{\frac{1}{z'-A}-\frac{1}{z-A}}{z-z'}dzdz'. \end{aligned}$$

Le deuxième terme s'annule, car, quand $z \in C, z' \in C'$, $g(z')/(z-z')$ est holomorphe à l'intérieur de C' et donc son intégrale sur C' est nulle. Il reste donc

$$\begin{aligned} f(A)g(A) &= \frac{1}{(2\pi i)^2}\oint_{C'} g(z')(z'-A)^{-1}\oint_C \frac{f(z)}{z-z'}dzdz' \\ &= \frac{1}{2\pi i}\oint_{C'} g(z')(z'-A)^{-1}f(z') \end{aligned}$$

et le résultat suit. □

Proposition 6.6. Si f est holomorphe dans un disque contenant $\sigma(A)$, alors $\sigma(f(A)) = f(\sigma(A))$.

Démonstration. Pour $\sigma(f(A)) \supset f(\sigma(A))$, il est clair que si $\lambda \in \sigma(A)$, alors il existe un vecteur propre x tel que $Ax = \lambda x$, et donc

$$f(A)x = \frac{1}{2\pi i}\oint_C f(z)(z-A)^{-1}xdz = \frac{1}{2\pi i}\oint_C f(z)(z-A)^{-1}xdz = \frac{1}{2\pi i}\oint_C f(z)(z-\lambda)^{-1}xdz = f(\lambda)x.$$

et donc $f(\lambda) \in \sigma(f(A))$. Il reste donc à montrer l'inclusion inverse, ce que l'on va faire par contraposée : si $A-\lambda$ est inversible, alors $f(A)-f(\lambda)$ l'est également. Cela se fait en notant que $z \mapsto f(z)-\lambda$ est holomorphe et non-nulle sur $\sigma(A)$, et donc que $f(A)-\lambda$ admet un inverse, donné par l'intégrale de contour de $1/(f(z)-\lambda)(z-A)^{-1}$. (techniquement, il faut pour cela montrer que la proposition 6.5 est encore valable pour le cas de fonctions qui ne sont définies que sur un voisinage de $\sigma(A)$, ce que nous ne faisons pas ici). □

6.4 Exercices

1. (Calcul fonctionnel et blocs de Jordan)
   (a) Calculer les puissances de la matrice $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ et en déduire la valeur de $f(A)$ pour une fonction A analytique de rayon de convergence suffisant.
   (b) Même question pour la matrice
   $$\begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$$
   (on pourra développer f en série entière autour de a).
   (c) En déduire qu'on ne peut pas définir de calcul fonctionnel pour les fonctions uniquement continues et les matrices quelconques.

(d) Noter le cas particulier de l'exponentielle, et justifier la méthode d'équations différentielles consistant à chercher une solution de forme $\alpha e^{\lambda t} + \beta t e^{\lambda t}$ pour la solution d'une équation différentielle de type $ay'' + by' + cy = 0$ dans le cas où les racines du polynôme $ax^2 + bx + c$ sont doubles.

2. (Calcul fonctionnel) Montrer que $\ddot{x} = -A^2 x, x(0) = x_0, x'(0) = 0$ a pour unique solution $x(t) = \cos(At)x_0$.
3. (Calcul fonctionnel) Montrer que l'exponentielle d'une matrice anti-hermitienne est unitaire.
4. (Calcul fonctionnel) Calculer

$$\exp\left(t\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\right)$$

de trois façons : par diagonalisation, par série entière, et en résolvant l'équation différentielle associée.

5. (Calcul fonctionnel) Montrer la formule de Lie-Trotter

$$\exp(A+B) = \lim_{n\to\infty} (\exp(A/n)\exp(B/n))^n$$

En déduire un schéma numérique de résolution de $\dot{x} = (A+B)x$ dans le cas où $\exp(tA)$ et $\exp(tB)$ sont faciles à calculer.

6. (Résolvante et valeurs propres) Montrer que, si A est une matrice carrée, il existe $r, C > 0$ tel que, pour $|z| \geq r$, $\|(z-A)^{-1}\| \leq C/|z|$. Déduire de cette inégalité et de l'exercice 5 de la section 5 qu'une matrice carrée admet nécessairement une valeur propre.
7. (Cayley-Hamilton) Montrer le théorème de Cayley-Hamilton : si $p(z) = \det(z-A)$, alors $p(A) = 0$. On pourra utiliser le calcul fonctionnel holomorphe et la formule de Cramer $B^- = \det(B)^{-1}\mathrm{comat}(B)$, avec $\mathrm{comat}(B)$ polynomiale en ses entrées.
8. (Résolvantes) On va montrer que si A et B sont deux matrices carrées, le spectre de AB et celui de BA sont les mêmes (sauf éventuellement 0, qui peut être dans le spectre de l'une mais pas de l'autre).
   (a) Soit $z \in \mathbb{C} \neq 0$. Calculer la série entière de $(z - \varepsilon AB)^{-1}$ et $(z - \varepsilon BA)^{-1}$ en fonction de ε, et préciser son domaine de validité.
   (b) Déduire une formule donnant $(z - \varepsilon AB)^{-1}$ en fonction de $(z - \varepsilon BA)^{-1}$.
   (c) Montrer que cette formule est également valable dès que $z - \varepsilon BA$ est inversible (quelle que soit la valeur de ε).
   (d) Conclure.
9. (Résolvantes) Si λ est une valeur propre de A, et que C est un cercle autour de λ dont l'intérieur ne contient pas d'autres éléments du spectre de A, alors on définit

$$P = \frac{1}{2\pi i}\oint_C (z-A)^{-1} dz.$$

   (a) Montrer que si x est vecteur propre associé à la valeur propre λ, alors $Px = x$.
   (b) Calculer P dans le cas où A est hermitienne.
   (c) Dans le cas général, montrer que $P^2 = P$ (en adaptant les arguments de la proposition 6.5). Montrer que $AP = PA = \lambda P$ en utilisant l'identité

$$A\frac{1}{z-A} = \frac{z}{z-A} - 1$$

et en déformant le contour pour se concentrer autour de λ. Déduire que P est un projecteur sur l'ensemble des vecteurs propres de A associés à la valeur propre λ.

(d) Supposons qu'on perturbe un peu A en $A+\varepsilon B$. Montrer que P est analytique en ε pour ε petit. Dans le cas où λ est une valeur propre simple, montrer que $A+\varepsilon B$ admet une valeur propre simple $\lambda(\varepsilon)$ pour ε petit, et que $\lambda(\varepsilon)$ est analytique en ε.

7 Spectre et asymptotique

7.1 Puissances d'une matrice

On s'intéresse ici aux propriétés de la dynamique linéaire discrète $x_{n+1} = Ax_n$ et on se demande : dans quelle mesure le spectre de A contrôle-t-il le comportement asymptotique (quand n ou t sont grands) de la dynamique ? Dans le cas d'une matrice normale, la réponse est immédiate : on a $\|A^n\| = \rho(A^n) = \rho(A)^n$, avec $\rho(A)$ le rayon spectral (la plus grande valeur propre en module).

Dans le cas non-normal, la situation est plus subtile. Le théorème fondamental est le suivant

Théorème 7.1 (Gelfand)**.** On a

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

où $\rho(A)$ est le rayon spectral de A.

Cela signifie en particulier que, pour tout $\varepsilon > 0$, il existe un rang n_0 à partir duquel

$$(\rho(A) - \varepsilon)^n \leq \|A^n\| \leq (\rho(A) + \varepsilon)^n$$

et donc qu'il existe $c, C > 0$ tel que, pour tout n,

$$c(\rho(A) - \varepsilon)^n \leq \|A^n\| \leq C(\rho(A) + \varepsilon)^n$$

(en fait, on peut même prendre $c = 1$.)

On va commencer par énoncer un corollaire, qui nous servira pour démontrer le théorème

Corollaire 7.2. Si $\rho(A) < 1$, alors $A^n \to 0$.

Démonstration du corollaire. On utilise le calcul fonctionnel holomorphe pour écrire

$$A^n = \frac{1}{2\pi i} \int_C \frac{z^n}{z - A},$$

où C est un cercle de rayon compris strictement entre $\|A\|$ et 1, de sorte que l'intégrale est bornée uniformément sur le contour. On conclut immédiatement puisque $z^n \to 0$ uniformément sur le contour. □

Démonstration du théorème. On peut maintenant montrer le théorème. Comme $\|A^n\| \geq \rho(A^n) = \rho(A)^n$, on a toujours $\|A^n\|^{1/n} \geq \rho(A)$. Pour tout $\varepsilon > 0$, $\rho\left(\frac{A}{\rho(A)+\varepsilon}\right) < 1$ et donc il existe $C > 0$ tel que

$$A^n \leq C(\rho(A) + \varepsilon)^n.$$

Le résultat suit. □

On remarque que, si $A = PDP^{-1}$ est diagonalisable, le corollaire (donc le théorème) est facile à prouver puisque $\|A^n\| = \|PD^nP^{-1}\| \leq \|P\|\|P^{-1}\|\rho(A)^n$. Le problème de cette preuve est que le préfacteur $\|P\|\|P^{-1}\|$ (qui quantifie la non-normalité de A) explose quand A se rapproche d'une matrice non-diagonalisable. Pour obtenir une preuve dans cet esprit sans utiliser le calcul fonctionnel holomorphe comme nous l'avons fait, on peut utiliser la forme de Schur, ou bien l'exercice 17 de la section 4.

7.2 Exponentielle de matrice

Pour la dynamique continue $\dot{x} = Ax$, de solution $x(t) = e^{At}x(0)$, on a de même

Théorème 7.3. On a

$$\lim_{t\to\infty} \frac{1}{t} \log \|e^{tA}\| = \max_{\lambda\in\sigma(A)} \mathrm{Re}(\lambda).$$

L'opération $\frac{1}{t}\log$ est l'analogue de la puissance $1/n$ dans le théorème de Gelfand : il permet d'identifier le λ dans $e^{\lambda t}$. On a en particulier comme avant l'existence pour tout $\varepsilon > 0$ d'une constante C telle que $\|e^{tA}\| \leq Ce^{t(\max_{\lambda\in\sigma(A)} \mathrm{Re}(\lambda)+\varepsilon)}$.

7.3 Exercices

1. (Exponentielle) Donner la preuve du théorème 7.3 en calquant celle du théorème de Gelfand.
2. (Blocs de Jordan) Calculer la solution de $x_{n+1} = Ax_n$ pour le *bloc de Jordan* élémentaire

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

   Rapprocher ce résultat de la méthode générale de résolution des équations récurrentes $x_{n+1} = \alpha x_n + \beta x_{n-1}$. Reprendre cet exercice avec $\dot{x} = Ax$ et $\ddot{x} = \alpha\dot{x} + \beta x$.
3. (Formule de Gelfand et pré-asymptotique) Quelle va être l'allure qualitative de $\|A^n\|$ pour la matrice A suivante ?

$$A = \begin{pmatrix} 0.9 & 1000 \\ 0 & 0.8 \end{pmatrix}$$

   En déduire les limites pratiques de l'application des théorèmes de cette section. Reprendre cette question pour e^{tA} plutôt que A^n ; comment modifier la matrice A pour avoir le même type de comportement ?
4. (Normes matricielles et rayon spectral) Soit A une matrice carrée. On pose pour tout $\varepsilon > 0$

$$\|x\|_\varepsilon = \sum_{n=0}^{\infty} \frac{\|A^n x\|}{(\rho(A)+\varepsilon)^n}.$$

   Montrer que cette série converge, et que $\|A\|_\varepsilon \leq (\rho(A) + \varepsilon)$, avec $\|A\|_\varepsilon$ la norme induite par $\|\cdot\|_\varepsilon$. Comparer avec l'exercice 17 de la section 4.

8 Méthodes itératives stationnaires pour $Ax = b$

On a vu que pour résoudre $Ax = b$, on pouvait utiliser le pivot de Gauss ou des variantes (méthodes *directes*) et le résoudre en un temps $O(N^3)$. Un ordinateur fait de l'ordre de 10^9 opérations par seconde, donc peut travailler sur une matrice de taille ≈ 1000 en $1s$. Un supercalculateur travaillant longtemps (disons, 1000 processeurs pendant un mois, donc $\approx 10^6$ secondes) peut atteindre 10^{18} opérations, ce qui permet de résoudre (une fois...) un problème de taille $\sqrt[3]{\frac{10^{18}}{10^9}} \times 1000 =$ un million d'inconnues. Cela peut paraître beaucoup, mais un modèle de physique comme ceux permettant de prédire le climat implique facilement des milliards d'inconnues.

Comment alors résoudre ce type de problème ? En remarquant que, dans la plupart des problèmes, les inconnues ont une structuration (par exemple spatiale) qui fait que la matrice A contient principalement des zéros (matrice creuse), et en utilisant un stockage adapté. Les méthodes directes peuvent dans une certaine mesure tirer partie de cette structure (cf section 2.4.1) mais se montrent rapidement limitées. On va plutôt utiliser dans ce cas des *méthodes itératives*, reposant uniquement sur le produit matrice-vecteur $x \mapsto Ax$, qui peut être effectué rapidement (en $O(N)$ pour des matrices très creuses).

On supposera dans la suite A carrée inversible.

8.1 Méthode de Richardson

L'objectif d'une méthode itérative est d'annuler itérativement le *résidu* $Ax - b$. On cherche pour commencer une méthode linéaire et *stationnaire* (c'est-à-dire que l'application $x_n \mapsto x_{n+1}$ est la même pour toutes les itérées), donc de forme $x_{n+1} = Bx_n + c$. On doit avoir $Ax = b$ à convergence, donc il est naturel de poser

$$x_{n+1} = x_n - \alpha(Ax_n - b),$$

avec $\alpha \in \mathbb{C}$. C'est la *méthode de Richardson*.

Pour analyser son comportement, comme pour toutes les analyses numériques de méthodes récurrentes (discrétisation d'EDO par exemple), il est naturel d'essayer d'obtenir une équation sur l'*erreur*

$$e_n = x_n - A^{-1}b.$$

On calcule aisément

$$\begin{aligned} e_{n+1} &= x_n - \alpha(Ax_n - b) - A^{-1}b \\ &= e_n - \alpha A(x_n - A^{-1}b) \\ &= (1 - \alpha A)e_n \end{aligned}$$

On est donc amené à calculer le rayon spectral de

$$M_\alpha = 1 - \alpha A,$$

qui est la matrice qui contrôle la dynamique de l'erreur. Son spectre $\sigma(M_\alpha)$ est égal à $1 - \alpha\,\sigma(A)$, et on souhaite choisir α pour que ce spectre passe dans la boule unité. Clairement, ça ne peut pas fonctionner pour toutes les matrices : par exemple, si $A = \mathrm{diag}(-1, 1)$, on aura toujours une valeur

propre de M en-dehors de la boule unité quelle que soit la valeur de α. Il existe cependant une classe importante de matrices où on peut garantir la convergence : les matrices HPD (hermitienne positive définie), ce que nous allons supposer dans la suite. Dans ce cas, on a même

$$\|e_n\| \leq \rho(M_\alpha)^n \|e_0\|.$$

En prenant α réel pour simplifier, on a, avec $0 < \lambda_1 \leq \lambda_2 < \cdots < \lambda_N$ les valeurs propres de A,

$$\rho(M_\alpha) = \max\left(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_N|\right)$$

On a donc convergence quand $\alpha < \frac{2}{\lambda_N}$.

Quel est le α optimal, c'est-à-dire celui qui minimise $\rho(M_\alpha)$? On commence par remarquer que $1-\alpha\lambda_N$ doit être négatif (sinon, on augmente α et on diminue $\rho(M_\alpha)$), et que $|1-\alpha\lambda_N| = |1-\alpha\lambda_1|$ (sinon, on peut augmenter ou diminuer α pour diminuer $\rho(M_\alpha)$). Le α optimal est donc celui pour lequel $-(1 - \alpha\lambda_N) = 1 - \alpha\lambda_1$, c'est-à-dire

$$\alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_N}.$$

Pour ce α, on a

$$\rho(M_{\alpha_{\text{opt}}}) = 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_N} = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} = \frac{\kappa(A) - 1}{\kappa(A) + 1}. \tag{8.1}$$

Pour les matrices mal conditionnées ($\kappa(A) \gg 1$), $\rho(M_{\alpha_{\text{opt}}}) \approx 1 - \frac{1}{\kappa(A)}$. Le nombre d'itérations nécessaires pour que $\rho(M_{\alpha_{\text{opt}}})^n$ soit égal à une précision ε donnée est donc égal à

$$n_{\text{iter}} = \frac{\log(\varepsilon)}{\log(\rho(M_{\alpha_{\text{opt}}}))} \approx \kappa(A)\log(\varepsilon)$$

et est donc proportionnel au conditionnement de la matrice. C'est assez moral : plus un problème est bien posé (bien conditionné), plus il est facile à résoudre.

8.2 Préconditionnement

Analysons les origines de la convergence lente de la méthode quand $\kappa(A)$ est grand. Pour cela, plaçons-nous dans une base orthonormée v_i où A est diagonale. À chaque étape, la i-ième composante de l'erreur est réduite d'un facteur $|1 - \alpha\lambda_i|$. Pour un i donné, le α optimal est $1/\lambda_i$, qui annule l'erreur en un coup. La disparité des λ_i impose cependant que α doit être suffisamment petit pour que $\alpha < 2/\lambda_N$, ce qui rend la convergence lente sur la première composante. Une autre façon de voir ce phénomène est de constater que le résidu

$$Ax - b = Ae = \sum_{i=1}^{N} \lambda_i \langle v_i, e\rangle v_i$$

est une mauvaise représentation de l'erreur, au sens où elle amplifie beaucoup certaines composantes et pas d'autres. Cela fait qu'ajouter $-\alpha(Ax_n - b)$ à x_n est une assez mauvaise façon d'assurer la convergence rapide. Si A est bien conditionnée en revanche, l'erreur et le résidu sont assez proches (du moins, à un facteur multiplicatif scalaire près), et la convergence est rapide.

Il arrive fréquemment que l'on dispose d'une approximation " plus simple" P de A, de sorte que $P^{-1}A$ est une matrice bien conditionnée. Des exemples peuvent être de négliger certains termes dans l'équation (par exemple, couplages entre certaines inconnues, ce qui donne une matrice bloc-diagonale), de remplacer le modèle par un modèle effectif à moins d'inconnues (*coarse-graining*), d'utiliser une précision plus faible, d'utiliser une factorisation inexacte, d'utiliser un inverse d'une matrice calculée à un pas de temps précédent dans le cas d'une évolution en temps... Si c'est le cas et qu'on peut l'inverser facilement, alors on peut espérer "corriger" la relation erreur-résidu, et que $P^{-1}r = P^{-1}Ae$ soit une représentation plus fidèle de e que $r = Ae$. Dans ce cas, on résoud $P^{-1}Ax = P^{-1}b$ par une méthode itérative : c'est la méthode de Richardson préconditionnée.

8.3 Exercices

1. (Optimisation) Montrer que, pour une matrice réelle SDP, la méthode de Richardson s'identifie à la méthode du gradient pour la minimisation de la fonction $\frac{1}{2}x^TAx - b^Tx$. Tracer l'allure des courbes de niveau de cette fonction, et des itérés de la méthode de gradient (en se rappelant que les gradients sont orthogonaux aux lignes de niveau). Comme interpréter la condition $\alpha < 2/\lambda_N$? Et le taux de convergence ?

2. (Optimisation) On rappelle que le gradient de f en x est défini comme l'unique vecteur g tel que $df(x) \cdot h = \langle g, h\rangle$ pour tout h. En particulier, le gradient dépend du produit scalaire. Donner le gradient d'une fonction $f(x)$ pour le produit scalaire $\langle x, y\rangle_M = x^*My$ avec M SDP en fonction du gradient usuel (pour le produit scalaire euclidien). Réinterpréter la méthode de Richardson préconditionnée ainsi que la méthode de Newton à cette lumière.

3. (Équations différentielles) Montrer que la méthode de Richardson appliquée à la résolution de l'équation $-u'' = f$ avec conditions au bord de Dirichlet en utilisant la méthode des différences finies à N points de grille converge, pour un pas suffisamment petit. Quelle est la complexité totale de cette méthode ? Comparer à la méthode LU.

4. (Principe du maximum) Montrer que, si les entrées de b sont positives ou nulles, que la méthode de Richardson avec $\alpha > 0$ converge, et que la matrice $1 - \alpha A$ a des entrées positives ou nulles, alors la solution a des entrées positives ou nulles. Appliquer à la résolution de $-u'' = f$.

5. (Préconditionnement) Comment étendre l'analyse de la méthode de Richardson à la méthode de Richardson préconditionnée dans le cas où A et P sont HDP ?

6. (Préconditionnement) Donner la méthode de Richardson préconditionnée dans le cas où P est la diagonale de A et $\alpha = 1$. Montrer qu'elle s'identifie à la méthode de Jacobi qui consiste à, dans la i-ième ligne de $Ax = b$, "geler" les valeurs des composantes $\neq i$ de x à leur valeur à l'étape n, résoudre le système pour trouver x_{n+1}, et itérer.

7. (Équations différentielles) Un collègue vient vous voir avec une idée brillante pour résoudre $Ax = b$ avec A HDP : il veut résoudre $x' = -(Ax - b)$ par la méthode d'Euler à pas fixe. Que lui répondez-vous ?

8. (Séries de Neumann) Le même collègue vient vous revoir avec une nouvelle idée brillante pour résoudre cette fois $(A + B)x = b$: il veut utiliser la série de Neumann $(A + B)^{-1}x = A^{-1}b - A^{-1}BA^{-1}b + \ldots$. Que lui répondez-vous ? (indice : considérer la méthode de Richardson préconditionnée avec pas $\alpha = 1$.)

9 Méthodes de Krylov

9.1 Espace de Krylov

On a vu que la méthode de Richardson pouvait converger lentement dans le cas où A est mal conditionnée. On peut se demander si cela provient du paramètre α qui est choisi de façon statique : peut-être qu'en le choisissant de façon plus complexe (par exemple, en alternant entre plusieurs pas, ou en les choisissant dynamiquement) on pourrait accélérer la convergence. Ces idées ont été testées historiquement [5] et trouvent leur aboutissement dans la constatation suivante : en prenant $x_0 = 0$ par simplicité, dans la méthode de Richardson avec un pas quelconque α_n, on va avoir

$$x_1 = x_0 - \alpha_1(Ax_0 - b) = -\alpha_1 b$$
$$x_2 = x_1 - \alpha_2(Ax_1 - b) = -(\alpha_1 + \alpha_2)b + \alpha_2\alpha_1 Ab$$

et ainsi de suite. De cette façon, quel que soit le choix de α à l'étape n, on aura toujours

$$x_n \in K_n(A, b) = \mathrm{Span}(b, Ab, \ldots, A^{n-1}b).$$

Cet espace, que nous noterons simplement K_n dans la suite est appelé espace de Krylov. Plutôt que chercher à optimiser une récurrence, on peut se demander directement : parmi les éléments de l'espace de Krylov, lequel résoud le mieux l'équation $Ax = b$? Cette question rentre dans le cadre général de la méthode de Galerkin.

9.2 Méthode de Galerkin

On se place ici dans un cadre plus général. On souhaite résoudre l'équation $Ax = b$ dans $\mathbb{R}^N$, mais N est trop grand pour une résolution directe. On suppose qu'on a à notre disposition un sous-espace $X \subset \mathbb{R}^N$ de dimension $n \ll N$, qu'on espère être suffisamment grand pour contenir une solution approchée de $A^{-1}b$. A priori, la vraie solution $A^{-1}b$ n'a aucune raison de se trouver dans X (on a N équations pour n inconnues), et il n'y a pas de critère naturel pour choisir parmi les éléments de Y lequel est le plus proche de $A^{-1}b$ (on peut imaginer projeter orthogonalement $A^{-1}b$ sur Y, mais cela nécessite de connaître $A^{-1}b$...).

On a donc besoin d'un moyen d'imposer n équations pour n inconnues. Pour cela, on se donne un autre sous-espace $Y \subset \mathbb{R}^N$ de dimension n, et on impose la *condition de Galerkin* : on cherche $x \in X$ tel que

$$Ax - b \perp Y$$

Cela signifie qu'on n'exige pas que la totalité de l'équation soit vérifiée ($Ax - b = 0$), mais simplement qu'une partie (sa projection sur Y) le soit.

Supposons qu'on dispose d'une base $u_1, \ldots, u_n$ de X et d'une base $v_1, \ldots, v_n$ de Y. Alors, en développant

$$x = \sum_{i=1}^{n} c_i u_i$$

5. Notamment dans l'article original de Richardson, "The Approximate Arithmetical Solution by Finite Differences of Physical Problems involving Differential Equations, with an Application to the Stresses in a Masonry Dam", qui en 1918 (!) utilisait ces méthodes pour calculer, sur papier et crayon, des conditions de rupture de barrages hydrauliques.

et en projetant $Ax - b$ sur les $v_1, \ldots, v_n$, on obtient l'équation

$$A_n c = b_n,$$

avec

$$(A_n)_{ij} = \langle v_i, Au_j \rangle$$
$$(b_n)_i = \langle v_i, b \rangle$$

c'est-à-dire un système de n équations pour n inconnues.

9.3 Méthodes de Krylov

Les méthodes de Krylov sont simplement des méthodes de Galerkin appliquées à l'espace $X = K_N$. Il existe deux variantes, selon le choix de Y. Si on prend $Y = K_n$, on obtient la méthode FOM (full orthogonalization method), plus connue, dans le cas particulier où A est HDP, sous le nom de méthode du gradient conjugué (que l'on va voir en plus de détails dans la suite). Si on prend $Y = AK_N$, on obtient la méthode GMRES (generalized minimal residual). En pratique, on utilise souvent le gradient conjugué pour les matrices HDP, et GMRES dans le cas général.

Les méthodes de Krylov trouvent la solution du problème en N étapes au plus :

Théorème 9.1. Si A est une matrice carrée inversible de taille N et $b \in \mathbb{C}^N$, alors

$$A^{-1}b \in K_N(A, b).$$

Démonstration. Si $K_N = \mathbb{R}^N$, alors le résultat est immédiat. Sinon, les N vecteurs $b, Ab, \ldots, A^{N-1}b$ ne sont pas indépendants et il existe donc $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$ non tous nuls tels que

$$\alpha_0 b + \alpha_1 Ab + \cdots + \alpha_{N-1} A^{N-1} b = 0$$

Avec i l'indice du premier α non-nul, on obtient que

$$A^i b = \sum_{j=i+1}^{N-1} \frac{-\alpha_j}{\alpha_i} A^j b.$$

En multipliant à gauche par $A^{-(i+1)}$, on obtient que $A^{-1}b$ est une combinaison linéaire des $A^j b$, $0 \leq j \leq N - i - 2$, d'où le résultat. □

9.4 Procédé d'Arnoldi

Si l'on souhaite implémenter effectivement les méthodes de Krylov, il faut, à l'étape n, résoudre le problème de chercher $x \in K_n$ tel que $Ax - b \perp Y$, avec $Y = K_n$ (méthode du gradient conjugué, quand A est HDP), ou $Y = AK_n$ (méthode GMRES). Pour cela, il faut utiliser une base de K_n. La base naturelle, $b, Ab, \ldots, A^{n-1}b$ est très mal conditionnée ($A^n b$ a tendance à converger, à normalisation près, vers le vecteur propre associé à la plus grande valeur propre de A en module, cf exercice **??** de la section 7, et donc tous les vecteurs $A^i b$ vont avoir tendance à se ressembler, à

normalisation près). On va donc plutôt chercher à construire une base orthonormale de K_n. Comme K_n est construite par addition successive de vecteurs, il est très naturel d'utiliser le procédé de Gram-Schmidt

$$\begin{aligned} q_1 &= \frac{b}{\|b\|} \\ q_2 &= \frac{Aq_1 - \langle q_1, Aq_1\rangle q_1}{\|Aq_1 - \langle q_1, Aq_1\rangle q_1\|} \\ &\dots \\ q_n &= \frac{Aq_{n-1} - \sum_{i=1}^{n-1}\langle q_i, Aq_{n-1}\rangle q_i}{\|Aq_{n-1} - \sum_{i=1}^{n-1}\langle q_i, Aq_{n-1}\rangle q_i\|} \end{aligned}$$

(la convention d'indices est prise telle que $q_n, x_n \in K_n$, et $A^n b \in K_{n+1}$. q commence à $n = 1$, alors que x commence à $n = 0$, avec $x_0 = 0$.)

Ce schéma, appelé *procédé d'Arnoldi*, n'est pas tout à fait la méthode de Gram-Schmidt appliquée aux vecteurs $b, Ab, \dots, A^{n-1}b$: on utilise ici Aq_{n-1} plutôt que $A^{n-1}b$ pour des raisons de stabilité numérique. On vérifie cependant que

Proposition 9.2. Tant que $K_{n-1} \neq K_n$, le procédé d'Arnoldi jusqu'à l'étape n est bien défini (les dénominateurs ne s'annulent jamais), et $\mathrm{Span}(q_1, \dots, q_n) = K_n$.

Démonstration. L'inclusion $\mathrm{Span}(q_1, \dots, q_n) \subset K_n$ est claire. Vérifions par récurrence l'inclusion inverse. Si, à l'étape $n-1$, $K_{n-1} \subset \mathrm{Span}(q_1, \dots, q_{n-1})$ et $K_{n-1} \neq K_n$, il reste à vérifier que $Aq_{n-1} - \sum_{i=1}^{n-1}\langle q_i, Aq_{n-1}\rangle q_i$ ne s'annule pas, et que $A^{n-1}b \in \mathrm{Span}(q_1, \dots, q_n)$.

On voit facilement que

$$Aq_{n-1} = \alpha A^{n-1}b + c,$$

avec $\alpha \neq 0$ (le produit des dénominateurs des étapes $0, \dots, n-1$) et $c \in \mathrm{Span}(q_1, \dots, q_{n-1})$. Si $Aq_{n-1} - \sum_{i=1}^{n-1}\langle q_i, Aq_{n-1}\rangle q_i$ s'annule, c'est donc que $A^{n-1}b$ est combinaison linéaire des $(q_1, \dots, q_{n-1})$, et donc que $K_n = K_{n-1}$, ce qui est impossible par hypothèse. On voit donc que q_n est également de forme $\alpha' A^{n-1}b + c$ (avec comme avant $\alpha' \neq 0$ et $c' \in \mathrm{Span}(q_1, \dots, q_{n-1})$), et donc que $A^{n-1}b \in \mathrm{Span}(q_1, \dots, q_n)$. □

Si $K_{n-1} = K_n$, le raisonnement du théorème 9.1 montre que de toute façon $A^{-1}b \in K_{n-1}$: le procédé d'Arnoldi échoue, mais c'est de toute façon qu'on a déjà trouvé la solution (*happy breakdown*).

9.5 Structure de la matrice quand $X = Y$ et procédé de Lanczos

Jusqu'à présent, on avait considéré le cas général d'un Y quelconque. Explicitons maintenant l'implémentation de la méthode quand $X = Y = K_n$. Cela consiste à trouver $x_n \in K_n$ tel que $Ax_n - b \perp K_n$. On développe

$$x_n = \sum_{i=1}^{n} c_n^i q_i$$

dans la base $q_1, \ldots, q_n$ de K_n. On projette le résidu $Ax_n - b$ sur la base $q_1, \ldots, q_n$ de K_n et on obtient l'équation

$$A_n c_n = b_n,$$

où

$$(A_n)_{ij} = \langle q_i, Aq_j \rangle$$
$$b_i = \langle q_i, b \rangle$$

La matrice A_n possède une structure particulière : en effet, comme $Aq_j \in K_{j+1}$ et que $q_i \perp \mathrm{Span}(q_1, \ldots, q_{i-1}) = K_{i-1}$, la matrice A_n vérifie

$$(A_n)_{ij} = 0 \quad \text{pour } i \geq j+1$$

(elle est dite de type "Hessenberg supérieure"). Cela signifie que tous les éléments en-dessous de sa première sous-diagonale sont nuls.

Quand la matrice A est hermitienne, A_n l'est également (toute expression d'un endomorphisme auto-adjoint *dans une base orthogonale* est hermitienne), et est donc tridiagonale ! Ce petit miracle algébrique signifie en particulier que tous les produits scalaires $\langle q_i, Aq_n \rangle$ utilisés dans le procédé d'Arnoldi sauf les deux derniers ($i = n$ et $i = n-1$) sont nuls : il suffit d'orthogonaliser Aq_n par rapport à q_n et q_{n-1} pour en fait aussi l'orthogonaliser par rapport à $q_1, \ldots, q_{n-2}$. Ainsi, l'algorithme coûte $O(Nn)$ plutôt que $O(Nn^2)$. Cette simplification du procédé d'Arnoldi dans le cas des matrices hermitiennes est appelé procédé de Lanczos.

9.6 Méthode du gradient conjugué

Il reste alors à résoudre l'équation $A_n c = b_n$. Il n'est en général pas évident que la matrice A_n soit inversible. C'est cependant le cas quand A est HDP : on a $A_n = Q^* A Q$ avec Q contenant les vecteurs $q_1, \ldots, q_n$ rangés en colonnes, et A_n est inversible (cf exercice 8 de la section 3).

La structure tridiagonale de la matrice dans le cas hermitien suggère qu'il peut être possible d'obtenir une méthode de résolution rapide (en utilisant par exemple l'exercice 2 de la section 2 sur la résolution de systèmes tridiagonaux, qui résoud le système $A_n c_n = b_n$ en $O(n)$, ramenant la complexité totale de l'algorithme à $O(Nn^2)$). On peut en fait faire mieux et trouver des relations de récurrences qui ramènent cette complexité à $O(Nn)$. En effet, on a

$$A_{n+1} = \begin{pmatrix} A_n & \alpha_n^* e_n \\ \alpha_n e_n^T & \beta_n, \end{pmatrix}$$

où e_n est le n-ième vecteur de la base canonique, et $\alpha_n, \beta_n \in \mathbb{C}$. Pour résoudre le système $A_{n+1} c_{n+1} = b_{n+1}$, on peut donc appliquer un complément de Schur en inversant le bloc A_n (ce qui est légitime parce que A_n était inversible). Cette méthode exprime c_{n+1} en fonction de $A_n^{-1} b_n$ (qui est simplement c_n), de $\langle e_n, A_n^{-1} e_n \rangle$ et de $\langle q_n, b \rangle$. La seule quantité non-triviale à obtenir est $A_n^{-1} e_n$, qu'on peut calculer par récurrence, encore en résolvant le système $A_{n+1} A_{n+1}^{-1} e_{n+1} = e_{n+1}$ par un complément de Schur. L'algèbre résultante est un peu pénible, mais donne au final une récurrence avec un historique court, connue sous le nom de méthode du gradient conjugué (on réfère à [2] pour l'algorithme complet, assez simple). C'est la méthode de choix (avec un préconditionneur adapté) pour la résolution de problèmes hermitiens définis positifs.

9.7 Vitesse de convergence de la méthode du gradient conjugué

On a vu que les méthodes de Krylov convergent exactement en au plus N étapes, ce qui n'est pas très utile en pratique quand N est grand. Ce qui les rend vraiment utiles, c'est qu'on a souvent une convergence approchée en beaucoup moins d'itérations. On va préciser ça pour la méthode du gradient conjugué dans le cas où A est symétrique réelle définie positive. Comme indiqué dans l'exercice 2, cette méthode est équivalent à minimiser la quantité $\frac{1}{2}x^TAx - x^Tb$ dans l'espace de Krylov. Or, on a

$$\frac{1}{2}x^TAx - x^Tb = \frac{1}{2}\|x - A^{-1}b\|_A^2 - \frac{1}{2}\|A^{-1}b\|_A^2,$$

où la A-norme $\|y\|_A$ est définie par $\|y\|_A^2 = y^*Ay$ (cette A-norme n'est pas égale à la norme euclidienne usuelle, mais elle lui est équivalente). On a donc

$$x_n = \text{argmin}_{x\in K_n}\|x - A^{-1}b\|_A \tag{9.1}$$

En ce sens, la méthode du gradient conjugué est *optimale*. On voit ainsi par exemple que la méthode du gradient conjugué est nécessairement meilleure (en A-norme) que la méthode de Richardson, quel que soit le choix du pas (même adaptatif).

En se rappelant que $x_n = p(A)b$ avec p un polynôme de degré $n-1$, on voit que $x_n - A^{-1}b = (p(A)-A^{-1})b$. La méthode du gradient conjugué n'est rien d'autre qu'une méthode d'approximation polynomiale adaptative : elle choisit un polynôme $p(x)$ qui approche la fonction $1/x$, de façon à ce que la A-norme de l'erreur $(p(A) - A^{-1})b$ soit minimale.

On va chercher maintenant à préciser en donnant une borne d'erreur explicite. Pour cela, il suffit de choisir analytiquement un polynôme d'interpolation, ce qui donnera une borne sur $\|x_n - A^{-1}b\|$ par (9.1). Plutôt que de chercher un polynôme d'interpolation de $1/x$ (ce qui est possible mais un peu plus algébriquement compliqué), il est commode de faire la réduction suivante : si $x_n = p(A)$ avec p de degré $n-1$, alors

$$\|x_n - A^{-1}b\|_A^2 = \langle (Ap(A) - 1)b, A^{-1}(Ap(A) - 1)b\rangle \tag{9.2}$$

$$\leq \langle b, A^{-1}b\rangle \max_{\lambda\in\sigma(A)}(1 - \lambda p(\lambda))^2 \tag{9.3}$$

$$\leq \|x_0 - A^{-1}b\|_A^2 \sup_{q\in\mathbb{C}_n[X], q(0)=1} \max_{\lambda\in\sigma(A)} q(\lambda)^2 \tag{9.4}$$

où on a utilisé la décomposition spectrale de A pour obtenir la première inégalité et on a posé $q = 1 - \lambda p$ pour la deuxième (en se rappelant que $x_0 = 0$). On se ramène donc à la question : comment trouver un polynôme q de degré n valant 1 en 0 et prenant des petites valeurs sur $\sigma(A)$? On utilise le lemme d'approximation polynomiale suivant :

Lemma 9.3. Le polynôme de degré n valant 1 en 0 et étant minimal (en valeur absolue) sur $[a, b]$ (avec $0 \notin [a, b]$) est le polynôme de Chebyshev rescalé

$$C_n(x) = \frac{T_n\left(\phi^{-1}(x)\right)}{T_n\left(\phi^{-1}(0)\right)},$$

où T_n est le polynôme de Chebyshev de degré n, l'unique polynôme vérifiant $T_n(y) = \cos(n \arccos(y))$ sur $[-1, 1]$, et ϕ est l'unique application linéaire telle que $\phi(-1) = a, \phi(1) = b$.

On a

$$\max_{x\in[a,b]} |C_n(x)| \leq 2\left(\frac{\sqrt{b/a}-1}{\sqrt{b/a}+1}\right)^n.$$

Démonstration. Le polynôme de Chebyshev a la propriété importante d'*équioscillation* : il vaut ± 1, de façon alternée, à $n+1$ points distincts $y_1, \ldots, y_{n+1}$ de $[-1,1]$ (et est non-nul en dehors). Le polynôme rescalé vaut donc $\pm M$ à $n+1$ points distincts $x_1, \ldots, x_{n+1}$ de $[a,b]$, avec $M = 1/|T_n(\phi^{-1}(0))| > 0$.

Supposons qu'il existe un autre polynôme q valant 1 en 0 et avec $\max_{x\in[a,b]} |q(x)| < 1$. Alors, $C_n - q$ prend des valeurs alternant entre positif et négatif sur les points $x_1, \ldots, x_{n+1}$. Il suit par le théorème des valeurs intermédiaires que $p-q$ a n racines sur $[a,b]$, et donc, en rajoutant 0, $n+1$ racines. Il est donc nul, et $p = q$.

Le calcul de $\max_{x\in[a,b]} |C_n(x)| = 1/|T_n\left(\phi^{-1}(0)\right)|$ est relativement direct, mais pénible [2, Theorem 38.5]. □

En mettant toutes les remarques de cette section bout à bout, on a prouvé

Théorème 9.4. La méthode du gradient conjugué démarrant en $x_0 = 0$ se déroule sans erreur (le procédé d'Arnoldi ne divise pas par zéro, et la matrice A_n est inversible) jusqu'à ce qu'elle trouve la solution exacte, en au plus N itérations. De plus, pour tout n avant sa terminaison, on a

$$\|x_n - A^{-1}b\|_A^2 \leq \|x_0 - A^{-1}b\|_A^2 \, 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n.$$

avec κ le conditionnement de A.

Il est intéressant de comparer ce taux de convergence à celui de la méthode de Richardson, (8.1), qui était le même mais avec κ au lieu de $\sqrt{\kappa}$. Ainsi, la méthode du gradient conjugué obtient une précision donnée en un temps proportionnel à $\sqrt{\kappa}$, plutôt que κ, ce qui est beaucoup mieux quand κ est grand.

9.8 Exercices

1. (Galerkin) Appliquer la méthode de Galerkin au système linéaire de matrice $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ avec $X = Y = \mathrm{Span}\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$.
2. (Galerkin) Dans le cadre d'application de la méthode de Galerkin (résoudre $Ax = b$ sur $\mathbb{R}^N$ avec un sous-espace X donné), dans le cas où A est SDP réelle, on considère le problème d'optimisation suivant :

$$\min_{x\in X} \frac{1}{2}x^T Ax - x^T b$$

Montrer l'existence et l'unicité, donner les conditions d'optimalité au premier ordre, et montrer que la solution s'identifie à la méthode de Galerkin avec $Y = X$. Même question pour

$$\min_{x \in X} \|Ax - b\|$$

avec $Y = AX$, sans hypothèse que A est SDP.
Montrer que la méthode de Galerkin avec $Y = X$ consiste (dans le cas HDP) à projeter la solution exacte $A^{-1}b$ sur X pour le produit scalaire $\langle x, y \rangle_A = x^* Ay$, et que la méthode avec $Y = AX$ consiste (dans le cas général) à projeter pour le produit scalaire $\langle x, y \rangle_{A^*A} = x^* A^* Ay$.

3. (Gradient conjugué) Montrer que si A est une matrice HDP de taille N avec k valeurs propres distinctes, alors le gradient conjugué trouve le résultat exact en k itérations au maximum.

4. (Équation différentielles) Le collègue que vous aviez éconduit au chapitre précédent vient vous revoir avec une idée brillante pour résoudre $Ax = b$ avec A HDP : il veut toujours résoudre $\dot{x} = -(Ax - b)$, mais cette fois avec un schéma d'intégration d'ordre 12 et contrôle adaptatif du pas. Que lui répondez-vous ?
5. (Approximation polynomiale) Montrer que la méthode du gradient conjugué revient à chercher un polynôme p_n de degré n qui minimise

$$\int \lambda (p(\lambda) - 1/\lambda)^2 d\mu(\lambda)$$

avec une mesure μ qu'on précisera. En déduire que p est la projection de la fonction $1/\lambda$ sur l'espace des polynômes de degré n.
6. (GMRES) Refaire l'analyse faite sur la méthode du gradient conjugué pour la méthode GMRES, dans le cas où $A = PDP^{-1}$ est diagonalisable (on pourra faire apparaître le conditionnement de P).
7. (Outliers dans le gradient conjugué) Supposons que A soit HDP et $\sigma(A) = \{x\} \cup [a, b]$, $x, a, b > 0$ et $x \notin [a, b]$. Donner une borne d'erreur sur la méthode du gradient conjugué qui fasse apparaître le ratio $\kappa = b/a$. Que dire des cas $x < a$ et $x > b$?
8. (Complexité) Quel est la complexité asymptotique de la résolution de l'équation $-u'' = f$ en utilisant la méthode des différences finies à N points de grille par la méthode du gradient conjugué ?

10 Méthodes numériques pour $Ax = \lambda x$

Comment calculer numériquement les valeurs propres d'une matrice ? C'est une question intrinsèquement plus délicate que celle de la résolution de systèmes linéaires. En effet, le calcul de valeurs propres de matrices est au moins aussi difficile que celui de racines de polynômes, puisque la construction de la matrice compagnon montre qu'on peut ramener le problème de calcul de racines de polynômes à celui de diagonalisation de matrice. Par ailleurs, la théorie de Galois affirme que les racines de polynômes de degré ≥ 5 ne peuvent pas, en général, se calculer par une suite finie d'opérations élémentaires. Tout calcul de valeurs propres de matrices ne peut donc être qu'*itératif*, au contraire de, disons, la résolution de systèmes linéaires pour laquelle le pivot de Gauss donne la solution exacte en un nombre fini d'étapes.

Néanmoins, on distingue toujours les méthodes *directes* (qui travaillent directement sur les entrées d'une matrice pleine) des méthodes *itératives* (qui n'utilisent que les produits matrice-vecteur), étant entendu que les méthodes directes seront toujours itératives. Les méthodes directes (par exemple, implémentées dans `scipy.linalg.eig` et `scipy.linalg.eigh` dans le cas hermitien) sont simplement des méthodes itératives sophistiquées extrêmement robustes, implémentées de façon optimisée pour les matrices denses ; nous ne détaillerons pas ces méthodes. le lecteur intéressé est invité à se renseigner sur *l'algorithme QR* (à ne pas confondre avec la factorisation QR d'orthogonalisation).

La première méthode qui vient peut-être à l'esprit pour le calcul de valeurs propres est de chercher les racines du polynôme caractéristique. C'est une mauvaise idée pour deux raisons : déjà, le calcul de déterminant numériquement n'est pas trivial (il faut utiliser une factorisation LU), et par ailleurs le calcul de racine de polynôme est assez délicat et mal conditionné. Il vaut bien souvent mieux travailler au niveau des vecteurs directement.

10.1 La méthode de la puissance

En se rappelant de la formule de Gelfand, il est raisonnable d'espérer que $A^n x$ va avoir tendance à s'aligner sur le ou les vecteurs propres associés à la plus grande valeur propre en module. C'est la base de la *méthode de la puissance*

$$x_{n+1} = \frac{Ax_n}{\|Ax_n\|}.$$

Théorème 10.1. Soit A hermitienne de valeurs propres $\lambda_1 \leq \cdots \leq \lambda_N$ (comptées avec multiplicité). Supposons que $\max_{i=1,\ldots,N} |\lambda_i|$ soit atteint en un unique i (égal à 1 ou N). Alors, pour presque toute condition initiale x_0 sur la sphère unité de $\mathbb{C}^N$, on a $x_n \to v_*$ avec v_* un vecteur propre normalisé associé à la plus grande valeur propre en module.

Démonstration. Supposons sans perte de généralité que $\lambda_1 \leq \cdots \leq \lambda_{N-1} < \lambda_N$, avec $|\lambda_N| > |\lambda_1|$, donc $\lambda_N > \max_{i=1,\ldots,N-1} |\lambda_i|$ (l'adaptation à l'autre cas ne pose pas de difficulté). Soit $v_1, \ldots, v_N$ une

base orthonormale de vecteurs propres associés aux valeurs propres $\lambda_1, \ldots, \lambda_N$. Alors on décompose

$$\begin{aligned} x_0 &= \sum_{i=1}^{N} \langle v_i, x_0 \rangle v_i \\ x_n &= \frac{A^n x_0}{\|A^n x_0\|} \\ &= \frac{\sum_{i=1}^{N} \lambda_i^n \langle v_i, x_0 \rangle v_i}{\| \sum_{i=1}^{N} \lambda_i^n \langle v_i, x_0 \rangle v_i \|} \\ &= \frac{\sum_{i=1}^{N} \left(\frac{\lambda_i}{\lambda_N}\right)^n \langle v_i, x_0 \rangle v_i}{\| \sum_{i=1}^{N} \left(\frac{\lambda_i}{\lambda_N}\right)^n \langle v_i, x_0 \rangle v_i \|} \end{aligned}$$

Mais $\left(\frac{\lambda_i}{\lambda_N}\right)^n \to^{n\to\infty} 0$ pour tout $i \neq N$, donc, si $\langle v_N, x_0 \rangle \neq 0$ (ce qui est vrai pour presque tout x_0 dans la sphère unité),

$$\sum_{i=1}^{N} \left(\frac{\lambda_i}{\lambda_N}\right)^n \langle v_i, x_0 \rangle v_i \to \langle v_N, x_0 \rangle v_N$$

$$x_n \to \frac{\langle v_N, x_0 \rangle}{|\langle v_N, x_0 \rangle|} v_N$$

□

Le taux de convergence se lit sur la décomposition ci-dessus : il existe $C > 0$ (dépendant de A et de x_0) tel que

$$\|x_n - v_*\| \leq C \left(\frac{\max_{i=1,\ldots,N-1} |\lambda_i|}{|\lambda_N|} \right)^n$$

C'est assez intuitif : plus une valeur propre est proche d'une autre valeur propre, plus on aura de mal à distinguer les vecteurs propres correspondants.

10.2 La méthode d'Arnoldi

Pour la résolution de très grands problèmes aux valeurs propres avec des matrices creuses, il est également possible d'utiliser une méthode Krylov pour le problème aux valeurs propres, méthode qui prend également le nom de méthode d'Arnoldi (Lanczos dans le cas hermitien). On se donne un sous-espace de Galerkin X, et on va chercher la "meilleure" approximation de certains vecteurs propres de A dans le sous-espace X. Comme pour la résolution de problème linéaire, il faut trouver un moyen de sélectionner ces vecteurs propres. La plus naturelle ici est de chercher $x \in X, \lambda \in \mathbb{C}$ tels que

$$Ax - \lambda x \perp X$$

En écrivant x dans une base orthonormée $(v_1, \ldots, v_n)$ de X, on se ramène comme dans le cas de systèmes linéaires à un problème

$$A_n c = \lambda c,$$

avec $(A_n)_{ij} = \langle v_i, A v_j \rangle$. On résoud ensuite ce problème par une méthode dédiée aux petits problèmes aux valeurs propres denses.

10.3 Exercices

1. (Galerkin) Montrer que l'application de la méthode de Galerkin dans une base non-orthonormale conduit à la résolution d'un système $A_n c = \lambda B_n c$, et préciser A_n et B_n.

2. (Optimisation) Dans le cas hermitien, montrer que la formule

$$\max_{x \in X \neq 0} \frac{x^* A x}{x^* x}$$

conduit à la condition $Ax - \lambda x \perp X$. Même question pour la formule (équivalente) $\max_{x \in X, \|x\|=1} x^* A x$ en utilisant la méthode des multiplicateurs de Lagrange.

3. (Puissance shiftée) Soit A HDP. On va appliquer la méthode de la puissance à la matrice $B = \alpha A + \beta$ (méthode de la puissance *shiftée*). Comment choisir α, β pour que la convergence vers la plus grande valeur propre de A soit la plus rapide possible ? Comment choisir α, β pour faire converger la méthode vers la plus petite valeur propre ? En faisant un parallèle formel avec la convergence de la méthode de Richardson, quel est l'analogue du conditionnement κ pour le problème aux valeurs propres ?
4. (Puissance inverse) Soit A HDP, et on suppose qu'on s'intéresse à la plus petite valeur propre. Montrer que la méthode de la *puissance inverse* $x_{n+1} = \frac{A^{-1} x_n}{\|A^{-1} x_n\|}$ converge pour presque toute condition initiale vers un vecteur propre associé à cette valeur propre. Quels sont les avantages et inconvénients de cette méthode par rapport à la méthode de la puissance shiftée ?
5. (Optimisation) Le collègue persistant des chapitres précédents vient vous revoir avec une idée brillante pour calculer la plus petite valeur propre d'une grande matrice HDP : minimiser le quotient de Rayleigh $x^* A x / x^* x$ par une méthode de descente de gradient. Que lui répondez-vous ?
6. (Puissance inverse) Quand vous expliquez à votre collègue que sa méthode n'est pas aussi révolutionnaire qu'il ne le pensait, il vous expose sa nouvelle idée : appliquer la méthode de la puissance inverse, et calculer $A^{-1} x_n$ par la méthode du gradient conjugué. Quels sont les avantages et inconvénients de la méthode de votre collègue par rapport à celles présentées dans le cours et les exercices ?

Références

[1] Ronald L Graham, Donald E Knuth, and Oren Patashnik. *Concrete mathematics : a foundation for computer science.* Addison-Wesley Longman Publishing Co., Inc., 1989.

[2] Lloyd N Trefethen and David Bau. *Numerical linear algebra.* SIAM, 2022.