ÉNS de Lyon
M2 FEADéP

Cours d'algèbre
2019–2020

Matrices et Groupes linéaires
Normes, Décompositions matricielles et Conditionnement

Table des matières

Leçons concernées (2020)

(103) Conjugaison dans un groupe. Exemples de sous-groupes distingués et de groupes quotients. Applications.

(106) Groupe linéaire d'un espace vectoriel de dimension finie E, sous-groupes de $\mathrm{GL}(E)$. Applications.

(108) Exemples de parties génératrices d'un groupe. Applications.

(150) Exemples d'actions de groupes sur les espaces de matrices.

(151) Dimension d'un espace vectoriel (on se limitera au cas de la dimension finie). Rang. Exemples et applications.

(152) Déterminant. Exemples et applications.

(158) Matrices symétriques réelles, matrices hermitiennes.

(159) Formes linéaires et dualité en dimension finie. Exemples et applications.

(162) Systèmes d'équations linéaires ; opérations élémentaires, aspects algorithmiques et conséquences théoriques.

(226) Suites vectorielles et réelles définies par une relation de récurrence $u_{n+1} = f(u_n)$. Exemples. Applications à la résolution approchée d'équations.

(233) Analyse numérique matricielle. Résolution approchée de systèmes linéaires, recherche d'éléments propres, exemples.

Ce qui est dans le programme (2020)

1.2(b) Applications multilinéaires. Déterminant d'un système de vecteurs, d'un endomorphisme. Groupe spécial linéaire $\mathrm{SL}(E)$. Orientation d'un $\mathbb{R}$-espace vectoriel.

1.2(c) Matrices à coefficients dans un anneau commutatif. Opérations élémentaires sur les lignes et les colonnes, déterminant, inversibilité.
Matrices à coefficients dans un corps. Rang d'une matrice. Représentations matricielles d'une application linéaire. Changement de base.
Méthode du pivot de Gauss. Notion de matrices échelonnées. Applications à la résolution de systèmes d'équations linéaires, au calcul de déterminants, à l'inversion des matrices carrées, à la détermination du rang d'une matrice, à la détermination d'équations définissant un sous-espace vectoriel.

13.1 Notion de conditionnement. Théorème de Gershgorin-Hadamard. Pivot de Gauss, décomposition LU. Méthodes itératives (par exemple méthode de Jacobi, méthode de Gauss-Seidel) ; analyse de convergence : normes subordonnées, rayon spectral.
Décomposition en valeurs singulières.
Exemple de la matrice de discrétisation par différences finies du laplacien en dimension un.

Option B (a) Systèmes linéaires : mise en œuvre de méthodes directes (pivot de Gauss, LU, Cholesky), coût de calcul de ces méthodes ;

Bibliographie

Pour le cours

- Ciarlet, Introduction à l'analyse numérique matricielle et à l'optimisation.
- Gourdon, Algèbre.
- Goblot, Algèbre linéaire.
- Paugam, Questions délicates en algèbre et géométrie.
- Perrin, Cours d'algèbre.
- Serre, Matrices.
- À suivre...

1 Méthodes de pivot

Pour la simplicité, vous pouvez penser que, partout $A = K$ est un corps. Lorsque ce n'est pas le cas, il y a de bonnes chances que A soit un anneau intègre et que vous puissiez voir $A \subset K = \mathrm{Frac}(A)$.

Motivations

Lorsqu'on est amené à étudier un problème en algèbre ou en analyse, il est souvent commode de se ramener à un cas qu'on sait déjà traiter. L'algèbre linéaire est un ensemble d'outils efficaces pour résoudre des problèmes (calculs de dimension, inversion de systèmes, invariants, etc...).

Ce chapitre met en avant les techniques de pivot qu'on est amené à rencontrer dans plusieurs situations :

— calcul efficace du déterminant (mise sous forme triangulaire) ;

— déterminer le rang d'une matrice (orbites : matrices J_r) ;

— résoudre un système linéaire sur un corps (mise sous forme normale de Gauss-Jordan) ;

— résoudre un système d'équations linéaires diophantien (mise sous forme normale de Hermite) : prendre $A = \mathbb{Z}$;

— déterminer la classe d'isomorphisme d'un groupe abélien (de type) fini (mise sous forme normale de Smith) : prendre $A = \mathbb{Z}$;

— déterminer les invariants de similitude d'un endomorphisme (mise sous forme normale de Smith) : prendre $A = k[X]$.

Dans toute la suite, on se place dans le A-module libre de type fini $\mathcal{M}_{m,n}(A)$ des matrices rectangulaires, à m lignes et n colonnes, à coefficients dans A.

Le but est d'exhiber une traduction de l'action de certains éléments du groupe $\mathrm{GL}_r(A)$ pour $r \in \{m, n\}$ sur l'espace $\mathcal{M}_{m,n}(A)$. On pourra donc chercher des invariants, calculer des orbites, etc...Typiquement, l'algorithme du pivot de Gauss consiste à échelonner des matrices rectangulaires dans un anneau principal (on pourra se limiter à A euclidien, voire A corps, pour l'agrégation).

Le principe du pivot de Gauss est de regarder simultanément l'action par multiplication à gauche de $\mathrm{GL}_m(K)$ sur $\mathcal{M}_{m,n}(K)$ et l'action par multiplication à droite de $\mathrm{GL}_n(K)$ sur $\mathcal{M}_{m,n}(K)$, ce qui donne une action du groupe $\mathrm{GL}_m(K) \times \mathrm{GL}_n(K)$ sur $\mathcal{M}_{m,n}(K)$ donnée par $(P, Q) \cdot M = PMQ^{-1}$. On va voir comment l'algorithme du pivot de Gauss permet d'exhiber des invariants totaux pour ces actions.

1.1 Déterminant

Avant toute chose, rappelons quelques résultats bien connus sur le déterminant d'une matrice.

Soit V un espace vectoriel, et $r \geqslant 1$.

Une forme r-linéaire f est dite alternée si pour $(x_1, \ldots, x_r) \in V^r$, on a

$$\exists i \neq j,\ x_i = x_j \Rightarrow f(x_1, \ldots, x_n) = 0.$$

Théorème 1.1. *L'ensemble des formes r-linéaires alternées est un sous-espace vectoriel des formes r-linéaires. Si $r = \dim(V)$, alors il est de dimension 1.*

Ceci permet de définir, à un scalaire près, le déterminant $\det_{\mathcal{B}}$ dans la base $\mathcal{B}$ comme l'unique forme n-linéaire alternée qui vaut 1 en $\mathcal{B}$ et, a fortiori, le déterminant d'une matrice dans la base canonique, de sorte que $\det(I_n) = 1$.

Théorème 1.2.

$$\forall P, Q \in \mathcal{M}_n(A),\ \det(PQ) = \det(P)\det(Q)$$

Ceci permet en particulier de montrer que l'ensemble des matrices $M \in \mathcal{M}_n(A)$ telles que $\det(M) \in A^\times$ est un groupe, noté $\mathrm{GL}_n(A)$. Le déterminant est un morphisme de groupes $\mathrm{GL}_n(A) \to A^\times$. On a également la formule :

Théorème 1.3.

$${}^t\mathrm{Com}(M)M = M{}^t\mathrm{Com}(M) = \det(M) \cdot \mathrm{I}_n.$$

Ce qui montre que $\mathrm{GL}_n(A)$ est aussi l'ensemble des matrices inversibles de l'algèbre $\mathcal{M}_n(A)$.

1.2 Matrices élémentaires

On désigne par $E_{i,j}$ la matrice ayant des coefficients nuls partout sauf celui à la ligne i et à la colonne j qui vaut alors 1. On note $\delta_{i,j} = \begin{cases} 1 & \text{si } i=j \\ 0 & \text{sinon} \end{cases}$ le symbole de KRONECKER, de sorte que

$$E_{i,j} = (\delta_{i,k}\delta_{j,\ell})_{\substack{1\leqslant k\leqslant m \\ l\leqslant \ell\leqslant n}} \qquad \text{et} \qquad E_{i,j}E_{k,\ell} = \delta_{j,k}E_{i,\ell}.$$

Fait 1.4. *La famille $(E_{i,j})$ est une base du A-module libre $\mathcal{M}_{m,n}(A)$.*

Définition 1.5. Pour $n \in \mathbb{N}$, on définit dans $\mathcal{M}_n(A)$, la

- *matrice de transvection* d'indice (i,j) avec $1 \leq i,j \leq n$ et $i \neq j$ et de rapport $\lambda \in A$ la matrice $T_{i,j}(\lambda) = I_n + \lambda E_{i,j}$;
- *matrice de dilatation* d'indice $i \in [\![1,n]\!]$ et de rapport $\mu \in A^\times$ la matrice $D_i(\mu) = I_n + (\mu-1)E_{i,i}$;
- *matrice de permutation* associée à $\sigma \in \mathfrak{S}_n$ la matrice $P_\sigma = (\delta_{i,\sigma(j)})_{1\leqslant i,j\leqslant n} = \sum_{k=1}^n E_{\sigma(k),k}$.

Ces définitions n'ont d'intérêt que pour $\lambda \in A \setminus \{0\}$ et $\mu \in A^\times \setminus \{1\} \subset A \setminus \{0,1\}$. En particulier, le cas $A = \mathbb{F}_2$ est à traiter séparément chaque fois qu'il est nécessaire de le considérer !

Fait 1.6. *On a, pour tous i,j,λ,μ,σ :*

- $\det D_i(\mu) = \mu$ *et donc* $D_i(\mu) \in \mathrm{GL}_n(A)$, *de plus* $D_i(\mu)^{-1} = D_i\left(\frac{1}{\mu}\right)$;
- $\det T_{i,j}(\lambda) = 1$ *et donc* $T_{i,j}(\lambda) \in \mathrm{SL}_n(A)$, *de plus* $T_{i,j}(\lambda)^{-1} = T_{i,j}(-\lambda)$;
- $\det P_\sigma = \varepsilon(\sigma)$ *et donc* $P_\sigma \in \mathrm{GL}_n(A)$, *de plus* $P_\sigma^{-1} = P_{\sigma^{-1}}$.

On n'a jamais $D_i(\mu) \in \mathrm{SL}_n(A)$ sauf si $\mu = 1$ (attention à la caractéristique 2 où $-1 = 1$) ce qui n'est pas un paramètre intéressant donc si on veut ne travailler qu'avec des matrices de $\mathrm{SL}_n(A)$, il faut se passer des matrices de dilatation. En revanche, même dans $\mathrm{SL}_n(A)$ on peut travailler avec des matrices de permutations « signées » en posant plutôt

$$\widetilde{P_{(i,i+1)}} = P_\sigma D_i(-1) = \sum_{k=1}^n \varepsilon_k E_{\sigma(k),k} \qquad \text{où} \qquad \varepsilon_k = \begin{cases} -1 & \text{si } k=i \\ 1 & \text{si } k\neq i \end{cases} \qquad \text{pour} \qquad 1 \leqslant i < n$$

et en considérant ensuite les matrices $\widetilde{P_\sigma}$ dans le sous-groupe de $\mathrm{SL}_n(A)$ engendré par les $\widetilde{P_{(i,i+1)}}$ qui est isomorphe à $\mathfrak{S}_n$.

Exercice 1. Pourquoi est-ce difficile de donner une formule des $\widetilde{P_\sigma}$?

Quelques calculs : Soit $M \in \mathcal{M}_{m,n}(A)$ une matrice quelconque. Notons $L_1,\dots,L_m$ ses lignes et $C_1,\dots,C_n$ ses colonnes. On a :

- $T_{i,j}(\lambda)M$ est la matrice dont les lignes sont $L_1,\dots,L_{i-1},L_i+\lambda L_j,L_{i+1},\dots,L_m$, opération associée : $L_i \leftarrow L_i + \lambda L_j$;
- $D_i(\mu)M$ est la matrice dont les lignes sont $L_1,\dots,L_{i-1},\mu L_i,L_{i+1},\dots,L_m$, opération associée : $L_i \leftarrow \mu L_i$;
- $P_{(i\ j)}M$ est la matrice dont les lignes sont $L_1,\dots,L_{i-1},L_j,L_{i+1},\dots,L_{j-1},L_i,L_{j+1},\dots,L_m$, opération associée : $L_i \leftrightarrow L_j$;
- $MT_{i,j}(\lambda)$ est la matrice dont les colonnes sont $C_1,\dots,C_{j-1},C_j+\lambda C_i,C_{j+1},\dots C_n$, opération associée : $C_j \leftarrow C_j + \lambda C_i$;
- $MD_j(\mu)$ est la matrice dont les colonnes sont $C_1,\dots,C_{j-1},\mu C_j,C_{j+1},\dots,C_n$, opération associée : $C_j \leftarrow \mu C_j$;
- $MP_{(i\ j)}$ est la matrice dont les colonnes sont $C_{(i\ j)(1)},\dots,C_{(i\ j)(n)}$, opération associée : $C_j \leftrightarrow C_i$.

Ce qu'il faut retenir :

(1) agir à par multiplication à gauche (L = Left = Lignes) c'est modifier les **lignes**, dans les formules, c'est la ligne i qui est changée ;

(2) agir à par multiplication à droite c'est modifier les **colonnes**, dans les formules, c'est la colonne j qui est changée.

(3) Pour retrouver les formules, il suffit de faire le calcul avec une matrice 2×2 sur son brouillon.

Lemme 1.7. *(1) Les opérations élémentaires par multiplication à gauche ne changent pas le noyau d'une matrice.*

(2) Les opérations élémentaires par multiplication à droite ne changent pas l'image d'une matrice.

(3) Les opérations élémentaires par multiplication à droite ou à gauche ne changent pas le rang d'une matrice.

Démonstration. (1) Il suffit d'observer que $\ker A = \bigcap_{i=1}^{n} \ker L_i = \mathrm{Vect}(L_1^*, \dots, L_n^*)^\top$.

(2) Il suffit d'observer que $\mathrm{im}(A) = \mathrm{Vect}(C_1, \dots, C_n)$.

(3) C'est une conséquence du théorème du rang et des points (1) et (2). □

1.3 Algorithmes d'échelonnement

Théorème 1.8 (Pivot de Gauss)**.** *On suppose que $A = K$ est un corps ! Soit $M \in \mathcal{M}_{m,n}(K)$ une matrice de rang r. Alors il existe des familles de matrices de transvection, permutation et dilatations $(A_1, \dots, A_s)$ et $(B_1, \dots, B_t)$ telles que $A_s \cdots A_1 M B_1 \dots B_t = J_r$.*

Démonstration. On procède par récurrence sur n. Si $n \leq 1$, il suffit de faire une dilatation.

Hérédité : On considère la première ligne et la première colonne de la matrice M.

Si toutes deux sont nulles, alors on prend $A_1 = P_{(1\ n)} = B_1$ de sorte que la dernière ligne et la dernière colonne de $A_1 M B_1$ sont nulles et on considère la matrice extraire $M' \in \mathcal{M}_{n-1}(K)$ des $n-1$ premières lignes et colonnes de M. L'hypothèse de récurrence permet de trouver des matrices de transvection, dilatation et permutation qui transforment M' en une matrice $J_{r'} \in \mathcal{M}_{n-1}(K)$. Quitte à compléter ces matrices par une ligne et une colonne dont seul le dernier coefficient est 1, ce sont toujours des matrices de transvection, dilatation et permutation qui transforment $A_1 M B_1$ en $J_{r'} \in \mathcal{M}_n(K)$.

Si l'une des deux est non nulle, disons la première ligne, alors on peut opérer sur les colonnes par une matrice de permutation A_1 pour que le premier coefficient de $A_1 M$ soit non nul. Par une matrice de dilatation B_1, on peut alors faire en sorte que le premier coefficient de $A_1 M B_1$ soit égal à 1. Par des opérations de soustraction via les matrices de transvection des lignes et colonnes, on peut alors se ramener au cas où la première ligne et la première colonne de $A_s \cdots A_1 M B_1 \cdots B_t$ n'ont que le premier coefficient non nul. Soit M' la matrice extraite des $n-1$ dernières lignes et colonnes de cette dernière matrice. Par hypothèse de récurrence, on trouve des opérations élémentaires qui transforment M' en $A_{s'} \cdots A_{s+1} M' B_{t+1} \cdots B_{t'} = J_{r'} \in \mathcal{M}_{n-1}(K)$. En complétant comme précédemment les matrices, on obtient que $A_{s'} \cdots A_1 M B_1 \cdots B_{t'} = J_{r'+1} \in \mathcal{M}_n(K)$.

Pour conclure, il suffit d'appliquer le point (3) du lemme précédent. □

Cet algorithme permet notamment le calcul de l'inverse d'une matrice comme suit :

Corollaire 1.9. *Si $A = K$ est un corps et $M \in \mathcal{M}_n(K)$ est une matrice de rang n, soit $(A_1, \dots, A_s)$ et $(B_1, \dots, B_t)$ les matrices données par un pivot de Gauss de M. Alors l'inverse de M est $B_1 \cdots B_t A_s \cdots A_1$.*

Remarque 1.10. En pratique, on calcule directement l'inverse de M en appliquant les mêmes opérations sur les lignes et les colonnes aux matrices M et I_n.

Estimons grossièrement le coût d'un tel algorithme. À l'étape n, il faut éventuellement faire une permutation : ce qui ne coûte aucun calcul dans le corps K puis appliquer jusqu'à $2(n-1)$ matrices de transvection à gauche et à droite. Chaque multiplication par une matrice de transvection coûte exactement n multiplications et n additions, soit au total $2n(n-1)$ additions et autant de multiplications. En sommant sur le nombre d'étape, cela donne un coût en $\sum_{k=1}^{n} 2k(k-1) = \frac{2}{3}\Big(n(n-1)(n-2)\Big)$ opérations de chaque type pour trouver les matrices $(A_i)_i, (B_j)_j$. Il faut multiplier par 2 cette quantité pour avoir une estimation du coût du calcul de l'inverse. C'est un ordre $O(n^3)$, ce qui est beaucoup même quand n n'est pas très grand. Mais pensez à la formule que vous connaissez également :

$$\det(M) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} m_{i,\sigma(i)}$$

qui coûte $n!$ additions dont chaque terme coûte n multiplications, ou encore à l'autre formule :

$$M^t \operatorname{Com} M = (\det M)\, I_n$$

où le calcul de la comatrice coûte à lui seul n^2 calculs de déterminants qui coûtent eux-mêmes $(n-1)!$ si on n'utilise pas l'algorithme de Gauss. Ainsi, le pivot de Gauss est un moyen nettement plus efficace d'inverse les matrices, de calculer les déterminants, etc...

Néanmoins cet algorithme est relativement long si l'on veut seulement calculer un déterminant ou résoudre un système linéaire puisqu'on change à la fois le noyau et l'image de la matrice. On peut améliorer légèrement les choses ainsi :

Définition 1.11. On considère une matrice $M = (m_{i,j})_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} \in \mathcal{M}_{m,n}(A)$. Pour tout $1 \leqslant i \leqslant m$, on note

$$d_i(M) = \inf\{j \in [\![1,n]\!],\ m_{i,j} \neq 0\} \in [\![1,n]\!] \cup \{+\infty\}.$$

C'est le « premier indice » où se trouve un coefficient non nul à la ligne i. On dit que M est *échelonnée supérieurement* si pour tout $i \in [\![2,n]\!]$, on a $d_i(M) > d_{i-1}(M)$ ou $d_i(M) = +\infty$.

On dit qu'une matrice M échelonnée supérieurement est *réduite* si pour tout $i \in [\![1,m]\!]$ tel que $j = d_i(M) < +\infty$, on a $m_{i,j} = 1$, autrement dit, les premiers coefficients non nuls de chaque ligne sont des 1.

Théorème 1.12 (Élimination de Gauss-Jordan)**.** *On suppose que $A = K$ est un corps ! Soit $M \in \mathcal{M}_{m,n}(K)$ une matrice. Alors il existe des familles de matrices de transvection, permutation (resp. et dilatation) $(A_1, \dots, A_s)$ telles que $A_s \cdots A_1 M$ est échelonnée supérieurement (resp. réduite).*

Idée de preuve. Dans l'algorithme de Gauss traité précédemment, il suffit de ne pas appliquer les considérations sur les colonnes. □

La conséquence est qu'on peut facilement résoudre un système linéaire lorsqu'on dispose d'une matrice échelonnée réduite, qu'on peut facilement calculer un déterminant lorsqu'on dispose d'une matrice échelonnée.

1.4 Généralisations (hors-programme)

Lorsque A est un anneau euclidien (ou même principal), on dispose encore de versions faibles de ces résultats.

Attention, ces deux derniers énoncés sont purement culturels à moins que vous ne sachiez les démontrer.

Théorème 1.13 (Forme normale de Hermite)**.** *On suppose que A est un anneau euclidien. Soit $M \in \mathcal{M}_{m,n}(A)$ une matrice. Alors il existe des familles de matrices de transvection et permutation $(A_1, \dots, A_s)$ telles que :*

- *$M' = A_s \cdots A_1 M$ est échelonnée supérieurement ;*
- *les coefficients $d_i = m'_{i,j}$ pour $j = d_i(M)$ vérifient l'inégalité $N(d_i) > N(m_{i',j})$ pour tout que pour tout $i' < i$.*

De plus, si $A = \mathbb{Z}$, alors on peut demander en outre que les $d_i(M)$ sont tous positifs, de même que les coefficients $m'_{i',j}$ pour $i' < i$ et $j = d_i(M)$. La matrice M' est alors uniquement déterminée dans ce cas.

On a également vu, dans le cadre de la réduction des endomorphismes, la

Théorème 1.14 (Forme normale de Smith)**.** *On suppose que A est un anneau principal. Soit $M \in \mathcal{M}_{m,n}(A)$ une matrice. Alors il existe des familles de matrices de transvection et permutation $(A_1, \dots, A_s)$ et $(B_1, \dots, B_t)$ telles que :*

- *$M' = A_s \cdots A_1 M B_1 \cdots B_t$ est diagonale ;*
- *les coefficients $d_i(M) = m'_{i,i}$ pour $i < \min(m,n)$ vérifient la relation de divisivilité $d_i(M) | d_{i+1}(M)$.*

De plus, les $d_i(M)$ sont uniquement déterminés.

2 Groupes linéaires

On se donne A un anneau commutatif (unitaire*) et E un A-module.

2.1 Généralités

Définition 2.1. On appelle *groupe général linéaire* sur E, noté $\mathrm{GL}(E)$, le groupe des automorphismes de A-modules de E. C'est aussi le groupe des éléments inversibles de la A-algèbre $\mathrm{End}_A(E)$.

Si E est un A-module libre de type fini, de rang n et de base $\mathcal{B} = (e_1, \dots, e_n)$, on a un isomorphisme de groupes

$$\begin{array}{ccc} \mathrm{GL}(E) & \simeq & \mathrm{GL}_n(A) \\ g & \mapsto & \mathrm{Mat}_{\mathcal{B}}(g) \end{array} \qquad \text{où } \mathrm{Mat}_{\mathcal{B}}(g) = (m_{i,j})_{1 \leqslant i,j \leqslant n} \qquad \text{avec} \qquad g(e_j) = \sum_{k=1}^{n} m_{k,j} e_k$$

Plus simplement, si A est un corps (on le notera plutôt K), alors E est un espace vectoriel et c'est donc en particulier un module libre*. Dans toute la suite on suppose désormais que E est un A-module **libre de type fini**, de rang $n \in \mathbb{N}^*$.

En pratique, on s'intéresse le plus souvent au cas où $A = K$ est un corps et E est un espace vectoriel, mais il est parfois intéressant de considérer des cas légèrement plus généraux : lorsque $A = \mathbb{Z}$ pour étudier les groupes abéliens comme $\mathbb{Z}$-modules, résoudre des systèmes d'équations diophantiens, travailler avec le groupe modulaire $\mathrm{PGL}_2(\mathbb{Z})$, ou encore $A = k[X]$ pour étudier les endomorphismes d'un k-espace vectoriel de dimension finie. Remarquez que dans ces cas, l'anneau A est principal.

Définition 2.2. Un groupe G est dit *linéaire* s'il existe un anneau commutatif A et un A-module libre de type fini E tel que G est un sous-groupe de $\mathrm{GL}(E)$ pour un certain A.

Autrement dit, un groupe linéaire est un certain sous-groupe d'un groupe général linéaire.

On remarquera que si A est intègre, on peut se ramener au cas des corps en considérant $K = \mathrm{Frac}(A)$ et $\mathrm{GL}_n(A)$ comme sous-groupe de $\mathrm{GL}_n(K)$. Ceci est typiquement le cas si on travaille sur l'anneau $A = k[X]$ des polynômes en une variable à coefficients dans un corps k ou sur l'anneau $A = \mathbb{Z}$.

Remarque 2.3. On a déjà vu que tout groupe fini est un groupe linéaire.

On dispose par ailleurs d'un morphisme de groupes $\det : \mathrm{GL}_n(A) \to A^\times$ qui nous permet de définir, via l'isomorphisme $\mathrm{GL}(E) \simeq \mathrm{GL}_n(A)$, un morphisme de groupes $\mathrm{GL}(E) \to A^\times$. Ce dernier ne dépend pas de la base choisie et on le note encore $\det : \mathrm{GL}(E) \to A^\times$.

Définition 2.4. On appelle *groupe spécial linéaire*, le noyau du morphisme $\det : \mathrm{GL}(E) \to A^\times$. On le note $\mathrm{SL}(E)$, ou encore $\mathrm{SL}_n(A)$.

On sait d'emblée que c'est un sous-groupe distingué de $\mathrm{GL}(E)$ et que $\mathrm{GL}(E)/\mathrm{SL}(E) \simeq \mathrm{im}(\det) = A^\times$ (la surjectivité se vérifie grâce aux matrices de dilatation).

Remarque 2.5. On observe également que le groupe $\mathrm{GL}_n(A)$ est un sous-groupe de $\mathrm{SL}_{n+1}(A)$ via le morphisme de groupes injectif $M \mapsto \left(\begin{array}{c|c} M & 0 \\ \hline 0 & (\det(M))^{-1} \end{array}\right)$, ce qui est parfois commode.

Dans la suite, on supposera systématiquement que $A = K$ est un corps. Pour effectuer convenablement un pivot de Gauss, il suffit en fait de supposer que A est principal (par exemple Euclidien). Attention : l'algorithme est mis en défaut si l'anneau A ne dispose pas des identités de Bézout ! Sur les anneaux factoriels, on ne peut donc pas dire grand chose.

2.2 Générateurs

L'algorithme du pivot de Gauss a des conséquences immédiates sur l'étude des groupes linéaires.

Corollaire 2.6. *(1) Le groupe* $\mathrm{GL}_n(K)$ *est engendré par les matrices de transvection, dilatation et permutation.*

(2) Le groupe $\mathrm{SL}_n(K)$ *est engendré par les matrices de transvection.*

∗. On s'en tiendra à la définition Bourbakiste suivant laquelle tout anneau est par définition unitaire.

∗. Si E est de dimension infinie, on peut alors réaliser $\mathrm{GL}(E)$ comme limite inductive de groupes linéaires finis sur l'ensemble inductif partiellement ordonné des sous-espaces vectoriels de dimension finie de E.

Démonstration. (1) découle du pivot de Gauss et du fait que les matrices de $\mathrm{GL}_n(K)$ sont surjectives donc de rang n.

(2) On observe que $P_{(i\ j)} = T_{i,j}(1)T_{j,i}(-1)D_i(-1)$ et que $D_i(\mu)T_{i,j}(\lambda) = T_{i,j}(\lambda\mu)D_i(\mu)$. Donc dans une écriture d'une matrice $A = A_1 \dots A_r \in \mathrm{SL}_n(K)$ où les matrices A_k sont des matrices élémentaires, on peut se ramener au cas où les A_k ne sont pas des matrices de permutations, puis au cas où $A = A_1 \cdots A_s B_1 \cdots B_t$ où les A_i sont des matrices de transvection et les B_j sont des matrices de dilatation. Donnons-nous une telle écriture avec t minimal. On doit ensuite observer le calcul suivant : $m_\lambda = T_{i,j}(\lambda)T_{j,i}(-1/\lambda)T_{i,j}(\lambda)T_{j,i}(1/\lambda) = \begin{pmatrix} 0 & \lambda \\ -1/\lambda & 0 \end{pmatrix}$. On a alors $m_\lambda m_{-1} = D_i(\lambda)D_j(1/\lambda)$, ce qui permet d'écrire $B_1B_2 = \Pi B'$ où Π est un produit de matrices de transvection et B' une matrice de dilatation. Ainsi $t = 1$. Mais alors $\det A = \det B_1 = 1$, ce qui nous dit qu'en fait $B_1 = I_n$, ce qui est exclu, et donc $t = 0$. □

Corollaire 2.7. *(1) Le noyau est un invariant total pour l'action de* $\mathrm{GL}_n(K)$ *sur* $\mathcal{M}_n(K)$ *par multiplication à gauche.*

(2) L'image est un invariant total pour l'action de $\mathrm{GL}_n(K)$ *sur* $\mathcal{M}_n(K)$ *par multiplication par l'inverse à droite.*

(3) Le rang est un invariant total pour l'action par équivalence de $\mathrm{GL}_n(K) \times \mathrm{GL}_n(K)$ *sur* $\mathcal{M}_n(K)$.

Démonstration. (1), (2) et (3) découlent du fait que $\mathrm{GL}_n(K)$ est engendré par les matrices élémentaires et respectivement des points (1), (2) et (3) d'un lemme précédent. □

2.3 Interprétation géométrique des générateurs

Soit K un corps et E un K-espace vectoriel de dimension $n \in \mathbb{N}^*$.

Dilatations

Définition 2.8. On suppose que le corps K contient au moins 3 éléments. On appelle *dilatation* de E d'hyperplan H, de droite D et de rapport λ, un endomorphisme $u \in \mathrm{GL}(E)$ tel que $H = \ker(u - \mathrm{id}_E)$ est un hyperplan, $D = \ker(u - \lambda\,\mathrm{id}_E)$ et $\det(u) = \lambda \neq 1$.

Fait 2.9. *(1) Si* $E = H \oplus D$ *avec* H *hyperplan,* D *droite et si* $\lambda \in K \setminus \{0, 1\}$*, alors il existe une unique dilatation de* E *d'hyperplan* H*, de droite* D *et de rapport* λ.

(2) La droite D *est entièrement déterminée par* H *et* λ.

(3) Si $\lambda = -1 \neq 1$ [*]*, alors* u *est la symétrie de* E *par rapport à* H *parallèlement à* D.

(4) u *est une dilatation de rapport* λ *si, et seulement si, il existe une base* $\mathcal{B}$ *de* E *telle que*

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & \lambda \end{pmatrix}.$$

Démonstration. (1) L'existence est claire sur une base adaptée à la décomposition $E = H \oplus D$ en posant $u(h) = h$ pour $h \in H$ et $u(d) = \lambda d$ pour $d \in D$. L'unicité se vérifie par égalité sur les sous-espaces H et D.

(2) Soit $u \in \mathrm{GL}(E)$ tel que $\ker(u - \mathrm{id}_E) = H$ et $\det u = \lambda \neq 1$. Soit $D' = \mathrm{im}(u - \mathrm{id}_E)$. Alors par le théorème du rang, on sait que D' est une droite de E ; de plus, on a $u(D') \subseteq D'$. Si on avait $D' \subset H$, alors on aurait $(u - \mathrm{id})^2 = 0$, donc $\mu_u | (X-1)^2$ et, en particulier, u serait trigonalisable d'unique valeur propre égale à 1. Ceci contredit $\det u \neq 1$. Donc $D' \not\subset H$. Soit $\mu \in K^*$ tel que $u_{D'} = \mu\,\mathrm{id}_{D'}$. Alors $\det u = \mu = \lambda$ et nécessairement $D' = \ker(u - \lambda\,\mathrm{id}_E) = D$.

(3) On a $(u_H)^2 = \mathrm{id}_H$ et $(u_D)^2 = \mathrm{id}_D$, d'où $u^2 = \mathrm{id}_E$. Il suffit de conclure en observant que $\ker(u - \mathrm{id}_E) = H$ et $\ker(u + \mathrm{id}_E) = D$.

(4) H et D sont des sous-espaces propres de u et sont en somme directe. Donc u est diagonalisable et les multiplicités des valeurs propres sont données par les dimensions respectives de H et D. La réciproque se lit sur la matrice. □

∗. ce qui impose $\mathrm{car}(K) \neq 2$

Transvections

Définition 2.10. Soit H un hyperplan de E et $u \in \mathrm{GL}(E)$. On dit que u est **une** *transvection* d'hyperplan H de E si $\ker(u - \mathrm{id}_E) = H$ et si $\det u = 1$.

On appelle *droite de la transvection* u la droite $D = \mathrm{im}(u - \mathrm{id}_E)$.

Proposition 2.11. *Soit u une transvection d'hyperplan H et de droite D. Alors :*

(1) $D \subset H$;

(2) il existe un vecteur $v \in E$ et une forme linéaire non nulle $f \in E^*$ tels que $u = \mathrm{id}_E + f(\cdot)v$;

(3) il existe une base $\mathcal{B}$ de E telle que

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & 1 \\ 0 & & & 1 \end{pmatrix}.$$

(4) Si H est un hyperplan de E et $x \notin H$, alors tout transvection de H s'étend en une transvection de E qui est l'identité sur Kx.

Démonstration. (1) On observe que D est une droite u-stable, par définition de D, et on pose $\mu \in K$ tel que $u_D = \mu \,\mathrm{id}_D$. Si on avait $\mu \neq 1$, alors D serait le sous-espace propre supplémentaire de H dans E et donc on aurait $\det(u) = \mu \neq 1$, ce qui est exclu. Donc $\mu = 1$ et $D \subset H$.

(2) En particulier, le polynôme $(X-1)^2$ annule u et c'en est en fait le polynôme minimal car u n'est pas l'identité. Soit $v \in D \setminus \{0\}$. Pour tout $x \in E$, il existe un scalaire $\mu_x \in K$ tel que $u(x) - x = \mu_x v$. Soit $f : x \mapsto \mu_x$. C'est une forme linéaire non nulle, de noyau H vérifiant l'identité souhaitée.

(3) On pose $v = e_{n-1}$ et e_n le vecteur antédual de f. On complète e_{n-1} en une base $(e_1, \dots, e_{n-1})$ de H. Alors la famille $\mathcal{B} = (e_1, \dots, e_n)$ est une base car $e_n \notin H = \ker f$ et elle convient.

(4) Par définition $E = H \oplus Kx$. Soit u une transvection de H que l'on étend en un endomorphisme v de E par linéarité en posant $v(h) = h \ \forall h \in H$ et $v(x) = x$. Alors $ker(v - \mathrm{id}_E) = \ker u \oplus Kx$ est un hyperplan de E et $\det v) = \det u \det \mathrm{id}_{Kx} = 1$. Donc v est une transvection de E. □

Théorème

Théorème 2.12. *Soit E un k-espace vectoriel de dimension n. Tout élément de $\mathrm{SL}(E)$ est le produit d'au plus $2n$ transvections.*

Lemme 2.13. *On suppose $n \geq 2$ et $x, y \in E \setminus \{0\}$. Il existe une ou deux transvections $u, v \in E$ telles que $u(x) = y$ ou $vu(x) = y$.*

Démonstration. Supposons que x et y ne sont pas colinéaires. Soit $z = y - x$. La famille (x, z) est libre donc il existe $f \in E^*$ telle que $f(x) = 1$ et $f(z) = 0$. Soit $u = \mathrm{id}_E + f(\cdot)z$. Alors $u(x) = x + z = y$.

Si x et y sont colinéaires, soit $z \in E \setminus Kx$, ce qui existe car $n \geq 2$. Alors il existe x, z ne sont pas colinéaires et y, z ne sont pas colinéaires, donc il existe u, v des transvections telles que $u(x) = z$ et $v(z) = y$. □

Lemme 2.14. *Soit H_1, H_2 deux hyperplans distincts de E et $x \notin H_1 \cup H_2$. Alors il existe une transvection u de E telle que $u(x) = x$ et $u(H_1) = H_2$.*

Démonstration. Soit $H = H_1 \cap H_2 + Kx$. C'est un hyperplan de E car $\dim H_1 \cap H_2 = n - 2$ comme intersection de deux hyperplans distincts et la somme $H_1 \cap H_2 + Kx$ est directe car $x \notin H_1 \cap H_2$.

Soit $z_1 \in H_1 \setminus H_1 \cap H_2$. Alors $z_1 \notin H$, donc il existe une forme linéaire $f \in E^*$ telle que $f(H) = 0$ et $f(z_1) = 1$. D'autre part, comme $x \notin H_2$, on a $E = H_2 \oplus Kx$. Donc il existe $z_2 \in H_2$ tel que $z = z_2 - z_1 \in Kx$.

La transvection $u = \mathrm{id}_E + f(\cdot)z$ convient. En effet, $u(z) = z + f(z)z = z$ car $z \in Kx \subset H$ et $z \neq 0$ car $z_1 \notin H_2$. Donc $u_{Kx} = \mathrm{id}_{Kx}$. De plus, si $y_1 \in H_1$, on écrit $y_1 = \lambda z_1 + y_2$ avec $y_2 \in H_1 \cap H_2$ et $\lambda \in K$. Alors $u(y_1) = y_1 + f(y_1)z = y_1 + \lambda z = \lambda z_1 + y_2 + \lambda z_2 - \lambda z_1 = y_2 + \lambda z_2 \in H_2$. □

Démonstration du théorème. On démontre par récurrence sur $n \in \mathbb{N}^*$ que pour tout corps K, le groupe $\mathrm{SL}_n(K)$ est engendré par les transvections. Le résultat est évident si $n = 1$ car le groupe est trivial.

Soit $n \geq 2$ et supposons le résultat connu en dimension $n - 1$. Soit $g \in \mathrm{SL}(E)$ et $x \in E \setminus \{0\}$. Posons $y = g(x)$. Le lemme 2.13 donne l'existence d'un produit d'une ou deux transvections $w \in \mathrm{SL}(E)$ tel que $w(y) = x$. Ainsi, $wg(x) = x$ et il suffit d'écrire $h = wg$ comme produit de transvections.

Soit H_1 un supplémentaire de Kx dans E et $H_2 = h(H_1)$. Alors $x \notin H_1$ par construction et $x \notin h(H_1)$ car sinon, on aurait $h^{-1}(x) = x \in H_1$. On peut donc appliquer le lemme 2.14 pour trouver une transvection $v \in \mathrm{SL}(E)$ telle que $v(x) = x$ et $v(H_2) = H_1$.

Ainsi, l'élément $k = vh = vwg$ stabilise la décomposition $E = H_1 \oplus Kx$ et $\det(k) = \det(\mathrm{id}_{Kx}) \cdot \det(k_{H_1})$. Par hypothèse de récurrence, k s'écrit comme produit de transvections de H_1, et donc $g = w^{-1}v^{-1}k$ s'écrit comme produit de transvections de E. □

On observe en particulier que tout élément de $\mathrm{GL}(E)$ est le produit d'une dilatation et d'au plus $2n$ transvections. En effet, il suffit de voir que pour $g \in \mathrm{GL}(E)$, en posant $\lambda = \det g$ et $d \in \mathrm{GL}(E)$ une dilatation de rapport $\frac{1}{\lambda}$, on a $gd \in \mathrm{SL}(E)$ et donc g s'écrit comme le produit d'une dilatation et de transvections.

2.4 Étude du groupe $\mathrm{GL}(E)$

Résultats de conjugaison

Fait 2.15. *Soit H un hyperplan de E, soit D une droite de E, soit $\lambda \in K^*$ et $h \in \mathrm{GL}(E)$ un endomorphisme inversible.*

Si $h \in \mathrm{GL}(E)$ est une transvection (resp. dilatation) d'hyperplan H, de droite D (resp. et de rapport λ), alors ghg^{-1} est une transvection (resp. dilatation) d'hyperplan $g(H)$, de droite $g(D)$ (resp. et de rapport λ).

Démonstration. Ce résultat est à traiter en exercice. □

Proposition 2.16.
1. *Les transvections dans $\mathrm{GL}(E)$ sont deux à deux conjuguées.*
2. *Deux dilatations de $\mathrm{GL}(E)$ sont conjuguées si, et seulement si, elles ont même rapport.*
3. **Si** $\dim E \geq 3$, *alors les transvections dans $\mathrm{SL}(E)$ sont deux à deux conjuguées.*
4. *Dans $\mathrm{SL}_2(K)$, toute transvection est conjuguée à $\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ pour un certain $\lambda \in K^*$.*
5. *Dans $\mathrm{SL}_2(K)$, les matrices $\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ et $\begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix}$ sont conjuguées si, et seulement si, le rapport $\frac{\lambda}{\mu}$ est un carré dans $K^\times$.*

Démonstration. Ce résultat est à traiter en exercice. □

Centre, quotient et espace projectif

Définition 2.17. On appelle *homothétie* de E un endomorphisme $u \in \mathrm{GL}(E)$ tel qu'il existe $\lambda \in K^*$ tel que $u = \lambda \,\mathrm{id}_E$.

Fait 2.18. *L'ensemble des homothéties est un sous-groupe, noté $K^\times \,\mathrm{id}_E$ de $\mathrm{GL}(E)$ isomorphe à K^*.*

Définition 2.19. Soit E un K-espace vectoriel. On appelle *espace projectif* sur E, noté $\mathbb{P}(E)$ l'ensemble des droites vectorielles de E.

Le groupe $\mathrm{GL}(E)$ agit naturellement sur $\mathbb{P}(E)$, de même que ses sous-groupes et, en particulier le groupe $\mathrm{SL}(E)$.

Lemme 2.20. *Le centre de l'action de $\mathrm{GL}(E)$ sur $\mathbb{P}(E)$ est le groupe des homothéties.*

Démonstration. On considère un élément $u \in \mathrm{GL}(E)$ dans le centre de cette action, de sorte que $u(Kx) = Kx$. Ceci définit donc des scalaires λ_x tels que $u(x) = \lambda_x x$.

Soient $x, y \in E \setminus \{0\}$. Si x et y sont colinéaires avec $y = \mu x$, alors $u(y) = u(\mu x) = \mu u(x)$ donc $\lambda_y = \lambda_x$ car $x \neq 0$. Si x et y ne sont pas colinéaires, alors (x, y) libre donc $u(x + y) = u(x) + u(y)$ donne $\lambda_x = \lambda_{x+y} = \lambda_y$. Ainsi, l'application $\lambda : E \setminus \{0\} \to K^*$ vérifiant $\lambda(x) = \lambda_x$ est constante. □

Théorème 2.21. *On a $\mathcal{Z}(\mathrm{GL}(E)) = K^\times \,\mathrm{id}_E$ et $\mathcal{Z}(\mathrm{SL}(E)) = \mu_n(K) \,\mathrm{id}_E$.*

Démonstration. Les homothéties sont dans le centre de $\mathrm{GL}(E)$. Réciproquement, si un élément $u \in \mathrm{GL}(E)$ (resp. $u \in \mathrm{SL}(E)$) est dans le centre du groupe, alors son action par conjugaison sur les transvections préserve les droites associées au transvection. Donc u est dans le centre de l'action de $\mathrm{GL}(E)$ sur $\mathbb{P}(E)$. □

Définition 2.22. On appelle *groupe projectif linéaire* le groupe quotient $\mathrm{PGL}(E) = \mathrm{GL}(E)/\mathcal{Z}(\mathrm{GL}(E))$. On appelle *groupe projectif spécial linéaire* le groupe quotient $\mathrm{PSL}(E) = \mathrm{SL}(E)/\mathcal{Z}(\mathrm{SL}(E))$.

Fait 2.23. *Les groupes projectifs et projectifs spéciaux agissent fidèlement et transitivement sur l'espace projectif $\mathbb{P}(E)$.*

De plus, si $n = 2$, alors l'action de ces groupes sur l'espace projectif est exactement 3-transitive.

Remarque 2.24. En général, il faudra travailler davantage pour définir la notion de repère projectif : ce sont les familles de points $p_0, \dots, p_n \in \mathbb{P}(E)$ tels qu'il existe une base $(e_1, \dots, e_n)$ de E telle que $p_i = Ke_i$ et $p_0 = K(e_1 + \cdots + e_n)$.

L'action de $\mathrm{PGL}(E)$ sera alors transitive sur les repères projectifs de $\mathbb{P}(E)$.

On peut alors définir le birapport de quatre points (a, b, c, d) de $\mathbb{P}_1(K) = \mathbb{P}(K^2)$ comme l'unique $\delta \in K \cup \{\infty\} = \mathbb{P}_1(K)$ tel que (a, b, c, d) est dans l'orbite de $(0 = [1 : 0], 1 = [1 : 1], \infty = [0 : 1], \delta)$ sous l'action de $\mathbb{PGL}_2(K)$ sur $\mathbb{P}_1(K)^4$.

Il est plus délicat de généraliser cette notion en dimension supérieur car quatre droites distinctes en position quelconque ne forment plus nécessairement un repère projectif de $\mathbb{P}_2(K) = \mathbb{P}(K^3)$.

Groupe dérivé et simplicité

Théorème 2.25. *On a les égalités suivantes :*

1. *Si $n \geq 3$ ou $\mathrm{Card} K \geq 3$, alors $\mathcal{D}(\mathrm{GL}_n(K)) = \mathrm{SL}_n(K)$.*
2. *Si $n \geq 3$ ou $\mathrm{Card} K \geq 4$, alors $\mathcal{D}(\mathrm{SL}_n(K)) = \mathrm{SL}_n(K)$ (on dit que le groupe est parfait).*
3. *Si $n \geq 3$ ou $\mathrm{Card} K \geq 4$, alors $\mathrm{PSL}_n(K)$ est simple.*

Démonstration. Ces différents résultats qui peuvent faire l'objet de développements sont proposés en exercices. □

2.5 Premiers résultats topologiques de connexité par arcs

Dans cette dernière partie, on se restreint au cas plus spécifique du corps $K = \mathbb{R}$. La topologie de $\mathbb{R}$ fait de $\mathrm{GL}_n(\mathbb{R})$, de ses sous-groupes et de ses quotients des espaces topologiques car $\mathrm{GL}_n(\mathbb{R})$ est une partie (ouverte) du $\mathbb{R}$-espace vectoriel $\mathcal{M}_n(\mathbb{R})$.

Proposition 2.26. *L'ensemble des matrices de transvection est étoilé par rapport à l'identité. Le groupe $\mathrm{SL}_n(\mathbb{R})$ est connexe par arcs.*

Démonstration. Toute transvection est naturellement connectée par un arc formé de matrices de transvections à l'identité via la formule $t \mapsto u_t = \mathrm{id}_E + f(\cdot)tz$.

Le groupe $\mathrm{SL}_n(\mathbb{R})$ est engendré par les transvections, et un produit de transvection reste dans $\mathrm{SL}_n(\mathbb{R})$. □

Proposition 2.27. *Le groupe $\mathrm{GL}_n(\mathbb{C})$ est connexe par arcs.*

Le groupe $\mathrm{GL}_n(\mathbb{R})$ admet deux composantes connexes par arcs qui sont $\mathrm{GL}_n^+(\mathbb{R})$ et $\mathrm{GL}_n^-(\mathbb{R})$. De plus, $\mathrm{GL}_n^+(\mathbb{R})$ est la composante neutre de $\mathrm{GL}_n(\mathbb{R})$ et, en particulier, c'est un sous-groupe distingué.

Démonstration. La décomposition polaire des nombres complexes permet de ramener toute matrice de dilatation complexe à la matrice identité.

Le groupe $\mathrm{GL}_n(\mathbb{R})$ n'est pas connexe comme l'indique l'image du morphisme continu induit par le déterminant. Le groupe $\mathrm{GL}_n^+(\mathbb{R})$ est connexe par arcs car il est engendré par des matrices de dilatation et de transvection et toute matrice de dilatation réelle de paramètre positif peut être reliée par un arc à l'identité.

L'ensemble $\mathrm{GL}_n^-(\mathbb{R})$ est l'image de $\mathrm{GL}_n^+(\mathbb{R})$ par l'application continue de multiplication par une matrice de dilatation de paramètre -1. On a donc écrit $\mathrm{GL}_n(\mathbb{R})$ comme réunion disjointe de deux parties connexes. Comme $\mathrm{GL}_n(\mathbb{R})$ n'est pas connexe, c'en sont ses composantes connexes. □

Remarque 2.28. Pour les plus farouches d'entre vous, vous pouvez envisager également des résultats de compacité ou une approche élémentaire des groupes de Lie réels via l'exponentielle matricielle.

3 Matrices dans les cas réels et complexes

On a, dans un cours précédent, défini la notion d'espace euclidien et hermitien. Ce sont des cadres naturels de géométrie, qu'on généralise ensuite en dimension infinie pour faire de l'analyse. C'est le cadre utilisé, par exemple, pour les séries de FOURIER.

Tous les résultats qui vont suivre sont aussi bien valable dans les deux contextes réels et complexes. Comme il est attendu du lecteur qu'il soit déjà familiarisé avec le cadre réel, on énoncera les résultats et démonstrations dans le cadre complexe uniquement. Il est attendu du lecteur qu'il sache écrire les énoncés et démonstrations analogues dans le cadre réel.

On rappelle que $\mathcal{H}_n(\mathbb{C})$ (resp. $\mathcal{S}_n(\mathbb{R})$) l'espace vectoriel des matrices hermitiennes (i.e. $M^* = {}^t\overline{M} = M$, que $\mathcal{H}_n^+(\mathbb{C})$ (resp. $\mathcal{S}_n^+(\mathbb{R})$) le cône des matrices hermitiennes positives (i.e. $X^*MX \geqslant 0$ pour tout $X \in \mathcal{M}_{n,1}(\mathbb{C})$) et que $\mathcal{H}_n^{++}(\mathbb{C})$ (resp. $\mathcal{S}_n^{++}(\mathbb{R})$) l'ensemble des matrices hermitiennes définies positives. On notera $\mathcal{U}_n(\mathbb{C})$ (resp. $\mathcal{O}_n(\mathbb{R})$) le groupe des matrices unitaires (resp. orthogonales) de taille n, c'est-à-dire des matrices vérifiant $M^*M = I_n$ et $\mathcal{SU}_n(\mathbb{C})$ (resp. $\mathcal{SO}_n(\mathbb{R})$) le sous-groupe distingué des matrices unitaires de déterminant 1.

La topologie métrique du corps de base nous permet alors d'introduire des métriques sur les espaces matriciels considérés.

3.1 Quelques rappels sur les normes matricielles

Puisqu'il s'agit, à terme, de faire de l'analyse sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, on a besoin de munir l'espace des matrices $\mathcal{M}_n(\mathbb{K})$ d'une bonne topologie. Les résultats de convergence que vous verrez alors en option B s'expriment en fonction des normes qu'on va introduire et les propriétés algébriques de ces normes vont jouer un rôle essentiel dans les démonstrations.

Par ailleurs, certains résultats topologique (de compacité par exemple) offrent, en retour, des informations supplémentaires de structure de certains groupes linéaires considérés.

Rappelons que dans un $\mathbb{K}$-espace vectoriel de dimension finie, toutes les normes sont équivalentes, toutes les boules fermées sont compactes (locale compacité de $\mathbb{K}$) et les parties compactes sont exactement les parties fermées et bornées pour n'importe quelle norme induisant la topologie.

Une première information possible sur les matrices est la suivante :

Définition 3.1. Soit M une matrice de $\mathcal{M}_n(\mathbb{K})$. On appelle rayon spectral de M, et on note $\rho(M)$, la quantité :

$$\rho(M) = \inf\{r > 0,\ \forall \lambda \in \mathrm{sp}(M),\ |\lambda| < r\}.$$

Fait 3.2. *Si* $\mathrm{sp}(M) \neq \emptyset$*, et en particulier, si* $\mathbb{K} = \mathbb{C}$*, alors* $\rho(M) = \max_{\lambda \in \mathrm{sp}(M)} |\lambda|$.

Ce n'est pas une norme car, par exemple, toutes les matrices nilpotentes ont pour rayon spectral 0.

Notation 3.3. Si $|\cdot|$ est une norme sur $\mathbb{K}^n$, on notera $\mathbb{S}^{n-1}_{|\cdot|} = \{X \in \mathbb{K}^n,\ |X| = 1\}$ la sphère centrée en 0 de rayon 1 de $\mathbb{K}^n$ pour la norme $|\cdot|$.

Proposition-définition 3.4. Soit $m, n \in \mathbb{N}$. Soient $|\cdot|_m$ et $|\cdot|_n$ deux normes respectivement sur $\mathbb{K}^m$ et $\mathbb{K}^n$. Soit $M \in \mathcal{M}_{m,n}(\mathbb{K})$.

(1) La borne supérieure suivante est finie et atteinte (donc c'est un maximum) :

$$\|M\| = \sup_{X \in \mathbb{K}^n \setminus \{0\}} \frac{|MX|_m}{|X|_n} = \sup_{X \in \mathbb{S}^{n-1}_{|\cdot|_n}} |MX|_m$$

(2) L'application $\|\cdot\| : \mathcal{M}_{m,n}(\mathbb{K}) \to \mathbb{R}_+$ est une norme.

On dit que $\|\cdot\|$ est la *norme subordonnée* aux normes $|\cdot|_m$ et $|\cdot|_n$. Si $m = n$ et $|\cdot|_n = |\cdot|_m$, on dit juste que $\|\cdot\|$ est subordonnée à $|\cdot|_n$.

Démonstration. (1) Par linéarité, on a l'égalité des deux sup. Comme la sphère unité $\mathbb{S}^{n-1}$ est un compact, l'application continue (par composition) $X \mapsto |MX|_m$ est bornée et atteint ses bornes.

(2) Pour tout $M, N \in \mathcal{M}_{m,n}(\mathbb{K}$ et tout $\lambda \in \mathbb{K}$, on a :

— séparation : $\|M\| = 0 \iff \forall X \in \mathbb{R}^n,\ |MX|_m = 0 \iff \forall X \in \mathbb{R}^n,\ MX = 0 \iff M = 0$;

— absolue homogénéité : $\|\lambda M\| = |\lambda| \|M\|$;

— inégalité triangulaire : pour tout $X \in \mathbb{S}^{n-1}$, on a $|(M+N)X|_m \leqslant |MX|_m + |NX|_m \leqslant \|M\| + \|N\|$. En passant au sup, cela donne $\|M + N\| \leqslant \|M\| + \|N\|$.

□

Exercice 2. Soit $n \in \mathbb{N}^*$ et $M = (m_{i,j}) \in \mathcal{M}_n(\mathbb{K})$. Alors

1. $\|M\|_1 = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{n} |m_{i,j}|$ est la norme subordonnée à $|X|_1 = \sum\limits_{i=1}^{n} |x_i|$;
2. $\|M\|_2 = \max\limits_{\lambda\in \mathrm{sp}(M^*M)} |\lambda|$ (rayon spectral de M^*M) est la norme subordonnée à $|X|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$;
3. $\|M\|_\infty = \max\limits_{1\leqslant i\leqslant n} \sum\limits_{j=1}^{n} |m_{i,j}|$ est la norme subordonnée à $|X|_\infty = \max\limits_{1\leqslant i\leqslant n} |x_i|$;

Définition 3.5. Une norme $\|\cdot\|$ sur $\mathcal{M}_n(\mathbb{K})$ est dite *matricielle* si pour toutes matrices $M, N \in \mathcal{M}_n(\mathbb{K})$, on a $\|MN\| \leqslant \|M\|\|N\|$.

Fait 3.6. *Si $\|\cdot\|$ est une norme sur $\mathcal{M}_n(\mathbb{K})$ subordonnée à une norme $|\cdot|$ sur $\mathbb{K}^n$, alors c'est une norme matricielle. En particulier, pour tout $M \in \mathcal{M}_n(\mathbb{K})$ et tout $k \in \mathbb{N}$, on a $\|M^k\| \leqslant \|M\|^k$.*

Démonstration. Soient $M, N \in \mathcal{M}_n(\mathbb{K})$. Pour tout $X, Y \in \mathbb{M}_n(\mathbb{K})$, on a par définition $|MX| \leqslant \|M\||X|$ et $|NY| \leqslant \|N\||Y|$. En particulier, en prenant $X = NY$, cela donne $|MNY| \leqslant \|M\||NY| \leqslant \|M\|\|N\||Y|$. En passant au sup pour $Y \in \mathbb{S}^{n-1}$, on a $\|MN\| \leqslant \|M\|\|N\|$. □

Remarque 3.7. Attention : il existe des normes matricielles qui ne sont pas des normes subordonnées. Par exemple, la norme de Frobenius définie par $\|M\| = \sqrt{\mathrm{tr}(M^*M)}$ est une norme matricielle (appliquer l'inégalité de Cauchy-Schwarz) mais n'est pas une norme subordonnée (vérifier ce qu'il se passe pour $M = I_n$).

Proposition 3.8. *Soit $|\cdot|$ une norme quelconque sur $\mathbb{K}^n$ et $\|\cdot\|$ la norme subordonnée à $|\cdot|$. Alors, pour toute matrice $M \in \mathcal{M}_n(\mathbb{K})$, on a $\rho(M) \leqslant \|M\|$.*

Démonstration. Si $\mathrm{sp}(M) = \emptyset$, alors $\rho(M) = 0 \leqslant \|M\|$. Sinon, soit $\lambda \in \mathrm{sp}(M)$ une valeur propre et X un vecteur propre associé. Alors $\frac{|MX|}{|X|} = |\lambda|$ donc $|\lambda| \leqslant \|M\|$. Ainsi $\rho(M) = \max\limits_{\lambda\in\mathrm{sp}(M)} |\lambda| \leqslant \|M\|$. □

Concluons avec un lemme qui pourra être utile.

Lemme 3.9. *Soit $C \in \mathrm{GL}_n(\mathbb{K})$. Alors*
(1) $X \mapsto |CX|_\infty$ est une norme sur $\mathbb{K}^n$ notée $|\cdot|_C$;
(2) la norme $\|\cdot\|_C$ subordonnée à $|\cdot|_C$ est donnée par $\|M\|_C = \|CMC^{-1}\|_\infty$.

Démonstration. (1) est évident.
(2) Soit $M \in \mathcal{M}_n(\mathbb{K})$. Pour tout $X \in \mathbb{S}^{n-1}(|\cdot|_C)$, on a

$$|MX|_C = |CMX|_\infty = |CMC^{-1}CX|_\infty \leqslant \|CMC^{-1}\|_\infty |CX|_\infty = \|CMC^{-1}\|_\infty$$

Donc, en passant au sup, on a $\|M\|_C \leqslant \|CMC^{-1}\|_\infty$.

Réciproquement, soit $Y \in \mathbb{S}^{n-1}(|\cdot|_\infty)$ réalisant le supremum de la norme subordonnée $\|\cdot\|_\infty$ pour la matrice CMC^{-1}, c'est-à-dire $|CMC^{-1}Y|_\infty = \|CMC^{-1}\|_\infty$. Posons $X = C^{-1}Y \in \mathbb{K}^n$. Alors

$$\|CMC^{-1}\|_\infty = \frac{|MC^{-1}Y|_C}{|Y|_\infty} = \frac{|MX|_C}{|CX|_\infty} = \frac{|MX|_C}{|X|_C} \leqslant \|M\|_C.$$

D'où l'égalité $\|CMC^{-1}\|_\infty = \|M\|_C$ pour toute matrice M. □

Remarquons que la norme $\|M\|_C = \|CMC^{-1}\|_\infty$ est, en particulier, une norme matricielle, ce qui pourra s'avérer utile.

3.2 Interprétation matricielle du théorème spectral

On dispose sur $\mathbb{K}^n$ d'un produit scalaire canonique, c'est-à-dire la forme hermitienne définie positive $\varphi = \langle \cdot, \cdot \rangle$ donnée par :

$$\langle (x_i)_{1 \leqslant i \leqslant n}, (y_i)_{1 \leqslant i \leqslant n} \rangle = \sum_{i=1}^{n} \overline{x_i} y_i$$

qui induit une norme $|\cdot|_2$ sur $\mathbb{K}^n$ donnée par

$$|(x_i)_{1 \leqslant i \leqslant n}|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2};$$

autrement dit, $|x|_2 = \sqrt{\langle x, x \rangle}$. Attention : la norme subordonnée $\|\cdot\|_2$ sur $\mathcal{M}_n(\mathbb{K})$ n'est pas la norme canonique $\|\cdot\|$ donnée par $\|M\| = \mathrm{tr}(M^*M)$ mais une autre norme donnée par $\|M\|_2 = \rho(M^*M)$.

Pour cette forme hermitienne définie positive, on réinterprète naturellement $\mathrm{End}(\mathbb{K}^n) = \mathcal{M}_n(\mathbb{K})$, $\mathcal{H}(\varphi) = \mathcal{H}_n(\mathbb{K})$, $\mathcal{U}(\varphi) = \mathcal{U}_n(\mathbb{K})$, etc...

Lemme 3.10. *Le groupe $\mathcal{U}_n(\mathbb{K})$ est compact.*

Démonstration. C'est un fermé comme image inverse de $\{I_n\}$ par l'application continue $M \mapsto M^*M$. C'est un borné car pour $U \in \mathcal{U}_n(\mathbb{K})$, on a $\|U\|_2 = \sqrt{\rho(U^*U)} = 1$, ou encore $\|U\|_2 = \sqrt{\mathrm{tr}(U^*U)} = \sqrt{n}$ qui ne dépend pas de U. □

Le théorème spectral, vu dans le cadre des formes sesquilinéaires et hermitiennes, se réécrit alors matriciellement comme suit :

Théorème 3.11 (Théorème spectral matriciel et réduction simultanée)**.**
(1) Toute matrice $M \in \mathcal{H}_n(\mathbb{C})$ (resp. $\mathcal{S}_n(\mathbb{R})$) s'écrit $M = P^{-1}DP$ avec $P \in \mathcal{U}_n(\mathbb{C})$ (resp. $\mathcal{O}_n(\mathbb{R})$) et D diagonale **à coefficients réels**.
(2) Si $M \in \mathcal{H}_n(\mathbb{C})$ et $Q \in \mathcal{H}_n^{++}(\mathbb{C})$, alors il existe des matrices $P \in \mathrm{GL}_n(\mathbb{C})$ et D diagonale réelle telles que $Q = P^*P$ et $M = P^*DP$.

Démonstration. (1) On montre d'abord que toutes les valeurs propres complexes de M sont réelles. En effet, si $\lambda \in \mathbb{C}$ est valeur propre, de vecteur propre complexe $X \in \mathcal{M}_{n,1}(\mathbb{C})$, alors $MX = \lambda X$ donc $\overline{\lambda} X^*X = (\lambda X)^* X = (MX)^* X = X^*MX = \lambda X^*X$. Comme $X^*X = \|X\| > 0$, on en déduit que $\overline{\lambda} = \lambda$ est réelle.

Ainsi, M admet un vecteur propre X (réel si $\mathbb{K} = \mathbb{R}$) et comme M est normal, on a également une décomposition en somme directe orthogonale de sous-espaces M-stables $\mathbb{K}X \overset{\perp}{\oplus} X^\perp = \mathbb{K}^n$. On conclut par récurrence sur la dimension.

(2) Soit $M \in \mathcal{H}_n(\mathbb{K})$ et $Q \in \mathcal{H}_n^{++}(\mathbb{K})$. Soit $\mathcal{B}$ la base canonique de $\mathbb{K}^n$ et φ, ψ les formes bilinéaires symétriques de matrices respectives $\mathrm{Mat}_{\mathcal{B}}(\varphi) = Q$ et $\mathrm{Mat}_{\mathcal{B}}(\psi) = M$. Soit $u \in \mathrm{End}(\mathbb{K}^n)$ l'endomorphisme tel que $\mathrm{Mat}_{\mathcal{B}}(u) = Q^{-1}M = R$. Alors

$$\begin{aligned} R =& Q^{-1}MQ^{-1}Q \\ =& Q^{-1}(Q^{-1}M)^*Q && \text{car } Q, M \text{ sont hermitiennes} \\ =& Q^{-1}R^*Q \end{aligned}$$

Ainsi, l'endomorphisme u est φ-autoadjoint (ou encore φ-hermitien). Donc les sous-espaces propres de u sont φ-orthogonaux. Sur chaque sous-espace propre $E_u(\lambda)$ pour $\lambda \in \mathrm{sp}(u)$, on fixe une base φ-orthonormée $\mathcal{B}'_\lambda$ de vecteurs propres de u, ce qui est possible par le procédé d'orthonormalisation de Gram-Schmidt. On note $\mathcal{B}' = \bigsqcup_{\lambda \in \mathrm{sp}(u)} \mathcal{B}'_\lambda$ la base φ-orthonormée ainsi obtenue. Matriciellement, cela donne une matrice de passage P de $\mathcal{B}$ à $\mathcal{B}'$ telle que $P^{-1}RP = D$ et $P^*QP = I_n$ avec D diagonale. Ainsi

$$\begin{aligned} D =& D^* \\ =& P^*M^*(Q^{-1})^*(P^{-1})^* \\ =& P^*M(P^{-1}Q^{-1})^* && \text{car } M \text{ hermitienne} \\ =& P^*M(P^*)^* \\ =& P^*MP \end{aligned}$$

Ce qui conclut. □

3.3 Une première décomposition matricielle remarquable : la décomposition polaire

Une première décomposition matricielle très utile pour résoudre certains problèmes théoriques est la décomposition polaire. Elle généralise naturellement la forme polaire des nombres complexes inversibles $z = re^{i\theta}$ où $r \in \mathbb{R}_+^* \simeq \mathcal{H}_1^{++}(\mathbb{C})$ et $e^{i\theta} \in \mathbb{U} \simeq \mathcal{U}_1(\mathbb{C}) \simeq \mathcal{SO}_2(\mathbb{R})$ s'identifie à une rotation de $\mathbb{C}$ vu comme $\mathbb{R}$-espace vectoriel de dimension 2.

Corollaire 3.12 (Unicité de la racine carrée)**.** *Toute matrice $M \in \mathcal{H}_n^+(\mathbb{K})$ admet une racine carrée $N \in \mathcal{H}_n^+(\mathbb{K})$, c'est-à-dire $N^2 = M$.*

Si, de plus, $M \in \mathcal{H}_n^{++}(\mathbb{K})$, alors $N \in \mathcal{H}_n^{++}(\mathbb{K})$ est unique.

Démonstration. On écrit $M = P^{-1}DP$ avec $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ diagonale constituée des valeurs propres réelles de M, donc (resp. strictement) positives.

Existence : $N = P^{-1}\operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})P$ convient.

Unicité : Soit $H \in \mathcal{H}_n^{++}(\mathbb{K})$ telle que $H^2 = M$. Alors H est diagonalisable et commute à M. Soit Q le polynôme interpolateur de Lagrange donné par $Q(\lambda_i) = \sqrt{\lambda_i}$. Alors $Q(M) = Q(P^{-1}DP) = P^{-1}Q(D)P = P^{-1}\sqrt{D}P = N$. Comme H et $Q(H^2) = Q(M) = N$ commutent, elles sont codiagonalisables. Soit $U = H^{-1}N$. Alors $U \in \mathcal{H}_n^{++}$ car $(H^{-1}N)^* = N^*(H^{-1})^* = NH^{-1} = H^{-1}N$ et les valeurs propres de U sont des produits d'une valeur propre de N et d'une valeur propre de H^{-1} donc strictement positives. De plus $U^2 = H^{-2}N^2 = M^{-1}M = I_n$, donc les valeurs propres de U sont dans $\{\pm 1\}$. Ainsi $U = I_n$ donc $H = N$. □

Théorème 3.13 (Décomposition polaire (réelle ou complexe))**.** *Pour toute matrice $M \in \mathrm{GL}_n(\mathbb{K})$, il existe un unique couple $U \in \mathcal{U}_n(\mathbb{K})$ et $H \in \mathcal{H}_n^{++}(\mathbb{K})$ tel que $M = UH$.*

De plus, l'application $\begin{array}{ccc} \mathcal{U}_n(\mathbb{K}) \times \mathcal{H}_n^{++}(\mathbb{K}) & \to & \mathrm{GL}_n(\mathbb{K}) \\ (U, H) & \mapsto & UH \end{array}$ *est un homéomorphisme.*

Pour toute matrice $M \in \mathcal{M}_n(\mathbb{K})$, il existe un couple $U \in \mathcal{U}_n(\mathbb{K})$ et $H \in \mathcal{H}_n^+(\mathbb{K})$ tel que $M = UH$, ce couple n'est pas nécessairement unique.

Démonstration. **Unicité :** Si $M = UH$, alors $M^*M = H^*U^*UH = H^*H = H^2$. Ainsi nécessairement $H = \sqrt{M^*M}$ et donc $U = MH^{-1}$.

Existence : Il reste à montrer que $U = MH^{-1}$ est unitaire.

$$UU^* = (MH^{-1})(MH^{-1})^* = MH^{-1}(H^{-1})^*M^* = MH^{-2}M^* = M\left(M^*M\right)^{-1}M^* = I_n$$

Ainsi l'application $(U, H) \mapsto UH$ est bijective, continue. Il reste à montrer que son inverse est continue. On utilise un critère métrique de caractérisation. Soit $M_k \underset{k\to\infty}{\longrightarrow} M$. On écrit $M_k = U_k H_k$ avec $U_k \in \mathcal{U}_n(\mathbb{K})$ et $H_k \in \mathcal{H}_n^{++}(\mathbb{K})$. Comme $\mathcal{U}_n(\mathbb{K})$ est compact, on peut extraire une sous-suite convergente $U_{\varphi(k)} \underset{k\to\infty}{\longrightarrow} U'$. Alors $\lim_{k\to\infty} U_{\varphi(k)}^{-1} M_{\varphi(k)} = U'^{-1}M$ est hermitienne comme limite de matrices dans l'espace vectoriel des matrices hermitiennes (fermé de $\mathrm{GL}_n(\mathbb{K})$ donc, par unicité $U' = U$ est la seule valeur d'adhérence de la suite (U_k). Ainsi $U_k \underset{k\to\infty}{\longrightarrow} U$ et donc $H_k = U_k^{-1}M_k \underset{k\to\infty}{\longrightarrow} H$. □

Corollaire 3.14. *Le groupe $\mathcal{U}_n(\mathbb{K})$ est un sous-groupe compact maximal de $\mathrm{GL}_n(\mathbb{K})$.*

Remarque 3.15. On peut également définir un homéomorphisme

$$\begin{array}{ccc} \mathcal{H}_n^{++}(\mathbb{K}) \times \mathcal{U}_n(\mathbb{K}) & \to & \mathrm{GL}_n(\mathbb{K}) \\ (H, U) & \mapsto & HU. \end{array}$$

On pourra montrer que ces homéomorphismes sont en fait des $\mathcal{C}^\infty$-difféomorphismes.

Remarque 3.16. On dispose en outre d'un algorithme, peu efficace, qui permet de déterminer la décomposition polaire d'une matrice. En effet, si $M \in \mathrm{GL}_n(\mathbb{K})$, alors $M^*M \in \mathcal{H}_n^{++}(\mathbb{K})$ se calcul directement en $O(n^3)$ opérations (multiplication matricielle), puis on en trouve une base orthonormée par un procédé de Gram-Schmidt, ce qui donnera en particulier les valeurs propres de M^*M, à nouveau en $O(n^3)$ opérations. On en déduit alors le polynôme interpolateur des racines des valeurs propres Q, de degré n qui permet de calculer $\sqrt{M^*M} = H = Q(M)$ en $O(n^4)$ opérations. Puis $U = MH^{-1}$ se calcule en $O(n^3)$ opération à nouveau. Ainsi, c'est le calcul de H qui est coûteux et donne un algorithme en $O(n^4)$.

4 Résolution de systèmes linéaires : méthodes directes

Souvent, en analyse, on ramène un problème difficile à un système d'équations linéaire qui « approxime bien » le problème de départ. On n'a alors pas besoin de résoudre le système linéaire de manière exacte mais seulement de manière approchée. En revanche, on s'intéresse à la « robustesse » de la résolution, c'est-à-dire qu'on veut que la méthode de résolution du système linéaire donne des solutions proches lorsqu'on perturbe un petit peu les coefficients des équations.

Par exemple, lorsqu'on résout des équations différentielles linéaires, on est amené à calculer des exponentielles de matrices et il est alors commode de pouvoir diagonaliser une matrice : ce qui est difficile en général quand on ne connait pas les valeurs propres.

Attention, ceci est très important ! Même si on ne sait pas faire correctement de théorie de Galois avec les outils de l'agrégation, on sait néanmoins que le groupe $\mathcal{A}_n$ est simple pour $n \geqslant 5$ ce qui a pour conséquence qu'il est impossible de résoudre par radicaux les équations polynomiales de degré supérieur ou égal à 5. Ceci étant, même en degré 3 ou 4, déterminer les racines d'un polynôme ça n'est pas si facile que ça. En pratique, quand on veut approximer les racines d'un polynôme réel ou complexe, on utilise en fait les techniques qui vont suivre appliquées à une matrice compagnon.

On veut donc résoudre un système d'équations linéaires, c'est-à-dire trouver l'ensemble des $(x_j)_{1\leqslant j\leqslant n} \in \mathbb{K}^n$ vérifiant

$$\begin{cases} a_{1,1}x_1 + \cdots + a_{1,n}x_n &= b_1 \\ &\vdots \\ a_{m,1}x_1 + \cdots + a_{m,n}x_n &= b_m \end{cases}$$

pour des $(a_{i,j})_{\substack{1\leqslant i\leqslant m \\ 1\leqslant j\leqslant n}}$ et $(b_i)_{1\leqslant i\leqslant m}$, autrement dit, résoudre $AX = B$ avec $A \in \mathcal{M}_{m,n}(\mathbb{K})$, $X \in \mathcal{M}_{n,1}(\mathbb{K})$ et $B \in \mathcal{M}_{m,1}(\mathbb{K})$.

Ce que prédit l'algèbre linéaire, c'est que ce système n'admet des solutions que si $b \in \mathrm{im}(A)$ et dans ce cas, l'ensemble des solutions est un espace affine de dimension $\dim(\ker A)$. En effet, si $b = Ax_0$ et x est solution, alors $A(x - x_0) = b - b = 0$ donc $x \in x_0 + \ker A$. Il s'agit donc de mettre en place des algorithmes qui permettent d'exhiber un tel x_0 et, éventuellement, de décrire $\ker A$.

Comme il s'agit de présenter des méthode de résolution approchées, même si le cours d'algèbre ne traitera que de situations de résolution exactes, on se placera désormais uniquement dans le cas du corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Les méthodes itératives de résolution seront traitées en option B, pour tous, et pourront vous être utile également dans certaines leçons d'analyse.

4.1 Le cas échelonné

Supposons que $A \in \mathcal{M}_{m,n}(\mathbb{K})$ est échelonnée supérieurement.

On a des lieux de pivot $(i, d_i(A))$. Posons $I = \{i \in [\![1, m]\!],\ d_i(A) \neq +\infty\}$. Alors I est un intervalle d'entiers $I = [\![1, m']\!]$ et le système $AX = B$ n'a de solutions que si $b_i = 0$ pour $i > m'$, condition qu'on suppose réalisée. Pour simplifier, on suppose $m = m'$.

Quitte à multiplier A par $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$ où $\alpha_i = \frac{1}{a_{m,d_m(A)}}$, on obtient une matrice $A' = DA$ échelonnée supérieurement réduite, et cela change B en $B' = (\alpha_1 b_1, \ldots \alpha_m b_m)$. On suppose donc que la matrice échelonnée A est réduite.

Posons $J = \{d_i(A), i \in I\}$. Alors la matrice extraire $A' = A_{I,J}$ est triangulaire supérieure inversible par construction et son inverse se calcule par un pivot de Gauss, ligne à ligne en opérant sur les colonnes :

- Pour la ligne 1, on fait :
  - $C_2 \leftarrow C_2 - a'_{1,2}C_1$, i.e. $a'_{1,2}$ devient 0, et on change b_1 en $b_1 - a'_{1,2}b_2$;
  - ...
  - $C_m \leftarrow C_m - a'_{1,m}C_1$, i.e. $a'_{1,m}$ devient 0, et on change b_1 en $b_1 - a'_{1,m}b_m$;
- pour la ligne 2, on fait :
  - $C_3 \leftarrow C_3 - \frac{a'_{2,3}}{a'_{2,1}}C_2$, i.e. $a'_{2,3}$ devient 0, et on change b_2 en $b_2 - a'_{2,3}b_3$;
  - ...
  - $C_m \leftarrow C_m - \frac{a'_{2,m}}{a'_{2,m}}C_2$, i.e. $a'_{2,m}$ devient 0, et on change b_2 en $b_2 - a'_{2,m}b_m$;
- ...

On n'a donc pas besoin de vraiment calculer l'inverse, ni de stocker les opérations effectuée. On n'a pas non plus besoin de déterminer A' mais seulement $A'_{I,J}$, ce qui coûte $\frac{m(m-1)}{2}$ multiplications pour déterminer A' et encore m multiplications pour déterminer B'.

Ainsi, on peut calculer un X'_0 en $\frac{m(m-1)}{2}$ soustractions et autant multiplications. On pose $(X_0)_j = (X'_0)_i$ pour $i \in [\![1,m]\!]$ et $j = d_i(A) \in J$; et $(X_0)_j = 0$ si $j \notin J$. Soit au total $\frac{m(m-1)}{2} \times 2 + m = m^2$ multiplications et $\frac{m(m-1)}{2}$ soustractions.

Remarque 4.1. Si A est échelonnée inférieurement, alors on peut se ramener au cas précédent par conjugaison avec la matrice antidiagonale $J = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix}$, qui est aussi la matrice d'une permutation (laquelle ?) d'ordre 2, de sorte que $J^2 = I_n$. En effet $JBJ = JAXJ = (JAJ)(JXJ)$ change $B = (b_1, \dots, b_m)$ en $(b_m, \dots, b_1)$, $X = (x_1, \dots, x_n)$ en $(x_n, \dots, x_1)$ et JAJ est échelonnée supérieurement.

Attention : il serait FAUX de dire qu'on se ramène au cas précédement par transposition !

4.2 Le principe d'une méthode de résolution directe

Comme on peut le voir, résoudre un système linéaire est moins couteux lorsque la matrice est triangulaire. L'algorithme d'échelonnement de GAUSS-JORDAN donne l'existence de matrices de permutations et transvections de sorte que le produit de ces matrices P est tel que PA est échelonnée supérieurement. On résout alors facilement $AX = B$ en résolvant $PAX = PB$.

Plutôt que de résoudre le problème linéaire directement d'une matrice quelconque, une idée est de d'abord la décomposer en quelques (deux ou trois) matrices pour lesquelles on sait « rapidement » résoudre le système linéaire : Si $A = PQ$, alors résoudre $AX = B$ équivaut à résoudre :

$$\begin{cases} PY = B \\ QX = Y \end{cases}$$

et on espère que les problèmes $PY = B$ et $QX = Y$ sont eux-mêmes rapides à résoudre. C'est une sorte de principe « diviser pour régner » comme ceux qu'on applique dans les algorithmes de tri par exemple.

4.3 Décomposition LU

Si par exemple P et Q sont échelonnées, alors on résout $AX = B$ en $2m^2$ multiplications. C'est pour cette raison qu'on s'intéresse à la décomposition LU. Ce coût est similaire à celui du calcul $A^{-1}B$ si, par exemple, A est inversible et qu'on a pu calculer son inverse. Il s'agit donc de comparer le coût du calcul de l'inverse par pivot de GAUSS ($\simeq \frac{4}{3}n^3$), et de l'échelonnement de GAUSS-JORDAN ($\simeq \frac{2}{3}n^3$) avec celui du calcul de la décomposition LU.

On utilise la notation L pour « lower » et U pour « upper ».

Un cadre particulier pour les méthodes directes — et à connaître dans le cadre de l'agrégation — est le suivant :

Théorème 4.2 (Décomposition LU : condition suffisante d'existence et d'unicité)**.** *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On suppose que tous les mineurs principaux de A sont inversibles, c'est-à-dire que pour $m \in [\![1,n]\!]$ et $I = [\![1,m]\!]$, on a $\det(A_{I,I}) \neq 0$. Alors la factorisation* LU *de $A = LU$ avec L unitriangulaire inférieure et U triangulaire supérieure existe et est unique.*

Remarque 4.3. En particulier, on notera que A est inversible par hypothèse !

Démonstration. On démontre par récurrence forte, simultanément, l'existence et l'unicité. Cette démonstration est en même temps constructive car elle décrit un algorithme permettant de calculer la factoriation LU. On note A_m au lieu de $A_{I,I}$ pour $1 \leqslant m \leqslant n$ et $I = [\![1,m]\!]$.

Pour $n = 1$, on a $A_1 = (a_{1,1})$ avec $a_{1,1} \neq 0$, donc $L = (1)$ et $U = (a_{1,1})$ conviennent.

Hérédité : On suppose donc que la sous-matrice A_m se décompose de manière unique en $A_m = L_m U_m$ pour tout $1 \leqslant m < n$, avec L_m unitriangulaire inférieure et U_m triangulaire supérieure. Décomposons $A = A_n$ en blocs

$$A = \left(\begin{array}{c|c} A_{m-1} & B \\ \hline {}^tC & a_{n,n} \end{array}\right)$$

avec $B, C \in \mathcal{M}_{n-1,1}(\mathbb{K})$. On cherche alors une décomposition de $A = LU$ de la forme :

$$L = \left(\begin{array}{c|c} L_{m-1} & 0 \\ \hline {}^tC' & 1 \end{array}\right) \qquad \text{et} \qquad U = \left(\begin{array}{c|c} U_{m-1} & B' \\ \hline 0 & d \end{array}\right)$$

avec $B', C' \in \mathcal{M}_{n-1,1}(\mathbb{K})$. Le produit matriciel par blocs donne

$$\left(\begin{array}{c|c} L_{m-1} & 0 \\ \hline {}^tC' & 1 \end{array}\right)\left(\begin{array}{c|c} U_{m-1} & B' \\ \hline 0 & d \end{array}\right) = \left(\begin{array}{c|c} L_{m-1}U_{m-1} & L_{m-1}B' \\ \hline {}^tC'U_{m-1} & {}^tC'B' + d \end{array}\right) = \left(\begin{array}{c|c} A_{m-1} & B \\ \hline {}^tC & a_{n,n} \end{array}\right)$$

Par hypothèse, on a $\det(A_{m-1}) = \det(L_{m-1})\det(U_{m-1})$ donc les matrices L_{m-1} et U_{m-1} sont inversibles, ce qui permet de conclure qu'il suffit de poser :

$$B' = L_{m-1}^{-1}B, \qquad C' = {}^tU_{m-1}^{-1}C, \qquad d = a_{n,n} - {}^tC'B'.$$

De plus, ces matrices sont ainsi uniquement déterminées. □

Remarque 4.4. On calcule facilement le déterminant de A à partir de sa décomposition LU car $\det(A) = \det(U) = \prod_{i=1}^{n} U_{i,i}$.

Exercice 3. Montrer que la décomposition de $A = LU$ existe encore si $\det(A) = 0$ et les autres mineurs principaux sont inversibles.

Remarque 4.5. Génériquement (au sens de la densité), une matrice $A \in \mathcal{M}_n(\mathbb{K})$ vérifie les hypothèses du théorème puisque les équations $\det(A_{I,I}) = 0$ définissent des fermés d'intérieur vide de $\mathcal{M}_n(\mathbb{K})$.

En ce sens, on peut dire qu'on a, jusque-là, pas fait d'hypothèse forte sur la matrice A.

On dispose du théorème général (admis) :

Théorème 4.6 (Culturel)**.** *Si $A \in \mathrm{GL}_n(\mathbb{K})$, alors il existe une matrice de permutation P, une matrice unitriangulaire inférieure L et une matrice triangulaire supérieure U telles que $PA = LU$.*

Remarque 4.7. Attention : ce théorème ne statue aucune forme d'unicité en général.

4.4 Factorisation de Cholesky

Remarque 4.8. Soit A une matrice dont les mineurs principaux sont inversibles. Quitte à multiplier U par une matrice diagonale, on peut écrire $U = DM$ avec M unitriangulaire supérieure.

Supposons de plus que la matrice A est hermitienne (ou symétrique si $\mathbb{K} = \mathbb{R}$), ce qui est une hypothèse assez forte au sens où l'ensemble des matrices hermitiennes est un $\mathbb{K}$-sous-espace vectoriel de dimension $\frac{n(n+1)}{2}$ de l'espace des matrices carrées. Alors $A^* = M^*D^*L^*$ mais, par unicité, on aura également $M^* = L$ et $DM = D^*L^* = D^*M$, donc finalement $D^* = D \in \mathcal{M}_n(\mathbb{R})$ et $A = L^*DL$, ce qui est moins coûteux en stockage.

En fait, on évitera cette décomposition pour des raisons de stabilité qu'on évoquera plus tard et qui pourront – peut être – être précisées dans un texte d'option B.

Lemme 4.9 (Critère de Sylvester)**.** *Soit $A \in \mathcal{H}_n(\mathbb{K})$ et $\Delta_m = \det(A_{I,I})$ pour $1 \leqslant m \leqslant n$ et $I = [\![1, m]\!]$ les mineurs principaux de A. Alors A est définie positive si, et seulement si, $\Delta_m > 0$ pour tout $1 \leqslant m \leqslant n$.*

Démonstration. On procède par récurrence sur n. Si $n = 1$, c'est clair. Notons $\varphi = \langle \cdot, \cdot \rangle$ le produit scalaire canonique sur $\mathbb{K}^n$ et fixons $(e_1, \ldots, e_n)$ la base canonique de $V = \mathbb{K}^n$. Pour $1 \leqslant m \leqslant n$, on pose $V_m = \mathrm{Vect}(e_1, \ldots, e_m)$.

Hérédité : Si A est définie positive, alors pour tout $1 \leqslant m \leqslant n$, la restriction $A_m = A_{I,I}$ pour $I = [\![1, m]\!]$ de A à V_m est encore la matrice d'une forme hermitienne définie positive, donc $\Delta_m = \det(A_m) > 0$.

Réciproquement, supposons que pour $1 \leqslant m \leqslant n$, on a $\Delta_m > 0$. Si par l'absurde, il existe deux valeurs propres λ, μ de A qui sont strictement négatives. Or, on sait que A est hermitienne donc diagonalisable en base orthonormée pour φ. Soient v, w deux éléments de cette base, vecteurs propres, pour les valeurs propres respectives λ, μ. Il existe une combinaison linéaire non nulle $\alpha v + \beta w \neq 0$ telle que $e_n^*(\alpha v + \beta w) = 0$. Ainsi $\alpha v + \beta w \in V_{n-1}$. Par hypothèse de récurrence, on a $A_{n-1} \in \mathcal{H}_n^{++}(\mathbb{K})$. Ainsi

$$\begin{aligned} 0 &< \langle A\alpha v + \beta w, \alpha v + \beta w \rangle \\ &= \langle \lambda\alpha v + \mu\beta w, \alpha v + \beta w \rangle && \text{car ce sont des vecteurs propres} \\ &= \lambda\alpha^2 + \mu\beta^2 && \text{car on a choisit une base orthonormée} \\ &< 0 && \text{car } \lambda, \mu < 0 \end{aligned}$$

D'où une contradiction. Ainsi $\mathrm{sp}(A) \subset \mathbb{R}_+^*$ puisque $\det(A) = \Delta_n > 0$. Donc A est définie positive. □

Théorème 4.10 (Factorisation de Cholesky)**.** *Soit A une matrice hermitienne (ou symétrique) définie positive. Alors il existe une unique matrice triangulaire supérieure B, dont les coefficients diagonaux sont strictement positifs, telle que $A = B^*B$.*

Démonstration. **Existence :** Notons Δ_m les mineurs principaux de A, vérifiant $\Delta_m > 0$ d'après le Lemme. La matrice A admet donc une unique factorisation $A = LU$. De plus, les égalités successives données dans la preuve du théorème précédent $\Delta_m = \det(A_m) = \det(L_m)\det(U_m) = \det(U_m)$ assurent que les coefficients diagonaux u_i de U sont tous strictement positifs. Posons $D = \mathrm{diag}(\sqrt{u_1}, \dots, \sqrt{u_n})$. On a vu que $A = LD^2L^* = B^*B$ en posant $B = DL^*$.

Unicité : Si $A = B_1^*B_1 = B_2^*B_2$ sont deux décompositions de Cholesky de A, alors la matrice $D = B_2B_1^{-1} = B_1^*(B_2^{-1})^*$ est à la fois triangulaire supérieure et inférieure à diagonale positive, donc diagonale et à coefficients positifs. Mais alors $B_2 = DB_1$. Ce qui donne $A = B_1^*B_1 = B_2^*B_2 = B_1^*D^2B_1$. Ainsi $D^2 = I_n$ donc $D = I_n$ puisque ses coefficients sont positifs. □

Mise en œuvre de la factorisation de Cholesky (cas réel) : Si $A = (a_{i,j})$ et $B = (b_{i,j})$, l'égalité $A = {}^tBB$ donne :

$$a_{i,j} = \sum_{k=1}^{n} b_{k,i}b_{k,j} = \sum_{k=1}^{\min(i,j)} b_{k,i}b_{k,j} \qquad \text{car } b_{k,\ell} = 0 \quad \text{si } k > \ell.$$

La matrice A étant symétrique, on peut se contenter de vérifier les égalités pour $i \leqslant j$. Ce qui donne pour la première ligne de A :

$$\begin{array}{lll}
a_{1,1} = b_{1,1}^2 & \text{d'où} & b_{1,1} = \sqrt{a_{1,1}} \\
a_{1,2} = b_{1,1}b_{1,2} & \text{d'où} & b_{1,2} = \dfrac{a_{1,2}}{b_{1,1}} \\
\vdots & & \vdots \\
a_{1,n} = b_{1,1}b_{1,n} & \text{d'où} & b_{1,n} = \dfrac{a_{1,n}}{b_{1,1}}
\end{array}$$

Plus généralement, à la i-ème ligne de A, on a :

$$\begin{array}{lll}
a_{i,i} = \displaystyle\sum_{k=1}^{i} b_{k,i}^2 & \text{d'où} & b_{i,i} = \sqrt{a_{i,i} - \displaystyle\sum_{k=1}^{i-1} b_{k,i}^2} \\
\vdots & & \vdots \\
(j > i) \qquad a_{i,j} = \displaystyle\sum_{k=1}^{i} b_{k,i}b_{k,j} & \text{d'où} & b_{i,j} = \dfrac{a_{i,j} - \sum_{k=1}^{i-1} b_{k,i}b_{k,j}}{b_{i,i}} \\
\vdots & & \vdots
\end{array}$$

Ainsi, le calcul de la matrice B coûte n extractions de racines carrées, $\sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}$ divisions, $\sum_{i=1}^{n-1} i(i-1) = \frac{n(n-7/2)(n-1)}{6}$ additions et $\sum_{i=1}^{n-1} i(i-1) = \frac{n(n-7/2)(n-1)}{6}$ multiplications.

Ainsi, en nombres d'opérations, on peut estimer que cet méthode est d'un facteur 4 fois plus efficace que la décomposition LU en temps et 2 fois moins coûteuse en mémoire, bien qu'elle ne s'applique qu'aux matrices symétriques définies positives.

Remarque 4.11. En pratique, cet algorithme permet également de détecter si une matrice symétrique est, ou non, définie positive selon qu'on trouve une valeur $b_{k,k}^2 \leqslant 0$ ou non et on n'a donc pas besoin de le vérifier au préalable.

4.5 Factorisation QR

Le principe de cette dernière méthode est, non plus, d'écrire une matrice A comme produit de matrices triangulaires mais comme produit d'une matrice unitaire $Q \in \mathcal{U}_n(\mathbb{K})$ et d'une matrice triangulaire supérieure R.

Remarque 4.12. Résoudre le système $AX = B$ revient alors à résoudre le système $\begin{cases} Y = Q^*B \\ RX = Y \end{cases}$ puisque $Q^* = Q^{-1}$.

Cette méthode est, à nouveau, constructive en s'appuyant sur la démonstration du théorème suivant :

Théorème 4.13 (Factorisation QR)**.** *Soit $A \in \mathrm{GL}_n(\mathbb{K})$. Alors il existe $Q \in \mathcal{U}_n(\mathbb{K})$ et R triangulaire supérieure dont les éléments diagonaux sont strictement positifs, telles que $A = QR$. De plus, cette factorisation est unique.*

Démonstration. **Existence :** Comme $A \in \mathrm{GL}_n(\mathbb{K})$, ses colonnes $(A_i)_{1\leqslant i\leqslant n}$ forment une base de $\mathcal{M}_{n,1}(\mathbb{K}) \simeq \mathbb{K}^n$. Le procédé d'orthonormalisation de Gram-Schmidt donne alors une base orthonormée Q_i de $\mathbb{K}^n$ telle que $\mathrm{Vect}(A_1, \dots, A_i) = \mathrm{Vect}(Q_1, \dots, Q_i)$ pour tout $1 \leqslant i \leqslant n$. Ainsi $A_j = \sum_{i=1}^{j} \langle A_j, Q_i\rangle Q_i$. Si $A = QR$, on a alors $AE_j = A_j = QRE_j = Q\left(\sum_{i=1}^{j} r_{i,j}E_i\right) = \sum_{i=1}^{n} r_{i,j}Q_i$. On pose alors

$$r_{i,j} = \langle A_j, Q_i\rangle$$

Ce qui donne $r_{i,j} = 0$ si $i > j$ et $r_{i,i} = \langle A_i, Q_i\rangle = \dfrac{\|A_i\|_2^2 - \sum_{j=1}^{i-1} |\langle A_i, Q_j\rangle|^2}{\|Q_i\|_2^2} \in \mathbb{R} \setminus \{0\}$. Ainsi, quitte à changer Q_i en $-Q_i$, ce qui ne change pas l'orthonormalité des Q_i, on peut supposer $r_{i,i} > 0$.

Unicité : Si $A = Q_1R_1 = Q_2R_2$ sont deux décompositions Q, R de A, alors $D = Q_2^{-1}Q_1 = R_2R_1^{-1} \in \mathcal{U}_n(\mathbb{K})$ vérifie donc $D^{-1} = D^* = (R_1^{-1})^*R_2^*$ est à la fois triangulaire supérieure et inférieure donc diagonale à coefficients réels positifs. Mais alors $D^*D = I_n$ est une décomposition de Cholesky de la matrice identité, donc $D = I_n$. □

Remarque 4.14. En pratique, on n'utilise pas tel quel le procédé d'orthonormalisation de Gram-Schmidt parce que de petites erreurs liés aux approximations dans les calculs entrainent rapidement de grands écart avec, par exemple, le fait que la matrice Q doit être unitaire.

Mais qu'est-ce qu'une « petite erreur d'approximation » et comment la contrôler ?

4.6 Conditionnement d'une matrice

Supposons qu'on veuille résoudre un système linéaire $Ax = b$ avec $A \in \mathcal{M}_{m,n}(\mathbb{K})$ et $b \in \mathcal{M}_m(\mathbb{K})$. Si A n'est pas de rang m et qu'on perturbe légèrement b, alors il y a de fortes chances pour que le système se retrouve sans solutions car dans ce cas, l'image de A est un sous-espace vectoriel de $\mathbb{K}^m$ donc de mesure nulle et l'ensemble des b pour lesquels l'équation est sans solution est un ouvert dense. Mais si on perturbe légèrement A, alors il y a cette fois de fortes chances que le système ait des solutions.

Dans une modélisation d'un problème d'analyse numérique, on veut savoir à quel point la solution x au problème $Ax = b$ se comporte bien si on a fait des petites erreurs sur A et b, disons qu'on ne les connait qu'à ε près (pour l'instant, on ne met pas en cause les petites erreurs qui se rajoutent dans les calculs des algorithmes mis en place).

Exemple 4.15. Prenons

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

Alors $A \in \mathrm{GL}_4(\mathbb{R})$ et la seule solution est

$$x = A^{-1}b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Maintenant, si on garde A et qu'on perturbe légèrement b en

$$b + \delta b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} + \begin{pmatrix} 0,1 \\ -0,1 \\ 0,1 \\ -0,1 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$

alors la solution devient

$$x + \delta x = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}.$$

On observe que cette petite perturbation de b entraine une grande perturbation de x. On se doute que cela vient probablement de la taille des coefficients de A^{-1}. On va donc proposer un outil métrique qui prend en compte A^{-1} : le conditionnement.

Définition 4.16. Soit $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{K})$. On définit le *conditionnement* d'une matrice $A \in \mathrm{GL}_n(\mathbb{K})$ par :

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

Fait 4.17. *Pour $A, B \in \mathrm{GL}_n(\mathbb{K})$, tout $\lambda \in \mathbb{K})$, on a :*

(1) $\mathrm{cond}(A) = \mathrm{cond}(A^-1) = \mathrm{cond}(\lambda A)$ *;*

(2) $\mathrm{cond}(AB) \leqslant \mathrm{cond}(A)\,\mathrm{cond}(B)$ *;*

(3) si $\|\cdot\|$ est une norme subordonnée, alors $\mathrm{cond}(A) \geqslant 1$.

Démonstration. Ces faits élémentaires sont laissés en exercice au lecteur pour qu'il s'assure d'avoir bien retenu les définitions. □

Remarque 4.18. Si on travaille avec la norme p, alors la norme subordonnée est notée $\|\cdot\|_p$ et le conditionnement cond_p.

Dans la suite, on suppose qu'on s'intéresse à un système linéaire $Ax = b$ avec $A \in \mathrm{GL}_n(\mathbb{K})$ et $b \in \mathcal{M}_{n,1}(\mathbb{K}) \setminus \{0\}$ et on se fixe, une fois pour toutes, une norme $|\cdot|$ sur $\mathbb{K}^n$, et on considérera sa norme subordonnée $\|\cdot\|$. On supposera qu'on fait une erreur δA sur A et une erreur δb sur b qui induisent une erreur δx sur x.

Théorème 4.19. *Si $\delta A = 0$ (i.e. on fait une erreur sur b uniquement), alors*

$$\frac{|\delta x|}{|x|} \leqslant \mathrm{cond}(A)\frac{|\delta b|}{|b|}$$

avec un possible cas d'égalité (l'inégalité est dite optimale).

Démonstration. On a $Ax = b$ et $A(x+\delta x) = b+\delta b$ donc $A\delta x = \delta b$. Ainsi $|b| \leqslant \|A\||x|$ et $|\delta x| \leqslant \|A^{-1}\||\delta b|$, ce qui donne l'inégalité.

D'autre part, il existe y tel que $\|A\||y| = |Ay|$ car l'infimum de la norme d'opérateur est un minimum. De même, il existe d tel que $\|A^{-1}\||d| = |A^{-1}d|$. Si on choisit $b = Ay$ et $\delta b = d$, alors on a $x = y$ et $\delta x = A^{-1}d$ qui vérifient $|b| = \|A\||x|$ et $|\delta x| = \|A^{-1}\||\delta b|$, ce qui donne le cas d'égalité. □

Théorème 4.20. *Si $\delta b = 0$ (i.e. on fait une erreur sur A uniquement) et $x + \delta x \neq 0$, alors*

$$\frac{|\delta x|}{|x + \delta x|} \leqslant \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}$$

avec un possible cas d'égalité (l'inégalité est optimale).

Démonstration. On a $(A + \delta A)(x + \delta x) = b = Ax$, ce qui se réécrit $A\delta x = -\delta A(x + \delta x)$. On a ainsi

$$\begin{aligned} |\delta x| &= |A^{-1}\delta A(x + \delta x)| \\ &\leqslant \|A^{-1}\|\|\delta A\||x + \delta x| \\ &= \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}|x + \delta x| \end{aligned}$$

D'où l'inégalité annoncée.

D'un autre côté, si $\delta A = \varepsilon I_n$ pour $\varepsilon > 0$ est telle que $A+\delta A \in \mathrm{GL}_n(\mathbb{K})$, ce qui est le cas sur un ouvert proche de 0, et si $y \neq 0$ réalise l'égalité $|A^{-1}y| = \|A^{-1}\||y|$ alors, en choisissant $b = (A+\delta A)y = Ay+\varepsilon y$, on a d'une part $A^{-1}b = x = y+\varepsilon A^{-1}y$ et, d'autre part, $b = (A+\delta A)(x+\delta x) = (A+\delta A)y$, ce qui donne $y = x+\delta x$. Ainsi $\delta x = y - x = y - (y+\varepsilon A^{-1}y) = -\varepsilon A^{-1}y$. D'où le cas d'égalité

$$|\delta x| = \varepsilon\|A^{-1}\||y| = \frac{\|A\|}{\|A\|}\underbrace{\|\delta A\|}_{=\varepsilon}\|A^{-1}\|\underbrace{|x+\delta x|}_{=|y|} = \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}|x+\delta x|.$$

□

Remarque 4.21. Ce cas d'égalité conforte l'idée que, comme on sait que le conditionnement est toujours plus grand que 1, « ce sont les matrices d'homothétie qui sont les mieux conditionnées ». On pourra penser qu'une matrice est « bien conditionnée » si son conditionnement est « proche » de 1.

En général, on ne s'attend pas à connaître d'emblée l'erreur faite sur x, et donc de savoir que $x+\delta x \neq 0$. Voici donc un autre résultat utile :

Théorème 4.22. *Si $\delta b = 0$ (i.e. on fait une erreur sur A uniquement) et si $\|A^{-1}\delta A\| < 1$, alors*

$$\frac{|\delta x|}{|x|} \leqslant \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}\Big(1+O\big(\|\delta A\|\big)\Big).$$

Démonstration. On repars de $b = (A+\delta A)(x+\delta x) = Ax$ pour écrire cette fois

$$-\delta A x = (A+\delta A)\delta x = A(I_n + A^{-1}\delta A)\delta x.$$

La matrice $I_n + A^{-1}\delta A$ est inversible car $\|A^{-1}\delta A\| < 1$ par hypothèse. De plus, on a

$$\|(I_n + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1-\|A^{-1}\delta A\|} \leqslant \frac{1}{1-\|A^{-1}\|\|\delta A\|} = 1+O\big(\|\delta A\|\big).$$

Ce qui donne

$$\begin{aligned}|\delta x| \leqslant& |-(I_n + A^{-1}\delta A)^{-1}A^{-1}\delta A x| \\ &\leqslant \Big(1+O\big(\|\delta A\|\big)\Big)\|A^{-1}\|\|\delta A\||x| \\ &= \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}|x|\end{aligned}$$

□

Voici un dernier résultat qui mélange erreurs sur A et sur b, laissé en exercice :

Théorème 4.23.

$$\frac{|\delta x|}{|x|} \leqslant \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{|\delta b|}{|b|}\right)$$

4.7 Le cas particulier de la norme 2

Pour conclure ce cours, on se place dans le cas $\mathbb{K} = \mathbb{C}$ et on fixe $\langle\cdot,\cdot\rangle$ produit scalaire canonique sur $\mathbb{C}^n$, qui donne lieu à la norme $|\cdot|_2 = \sqrt{\langle\cdot,\cdot\rangle}$.

Lemme 4.24. *(1) Pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$, on a*

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

(2) Pour toute matrice **normale** *$A \in \mathcal{M}_n(\mathbb{C})$ (i.e. A et A^* commutent), on a :*

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

Démonstration. (1) La matrice A^*A est hermitienne et, pour $x \in \mathbb{C}^n$, on a $x^*A^*Ax = |Ax|_2^2 \geqslant 0$, donc A^*A est positive. Ainsi A^*A est diagonalisable en base orthonormée $(e_1 \leqslant \cdots \leqslant e_n)$ à valeurs propres réelles positives ou nulles. Soit $x = \sum_{i=1}^n x_i e_i$. Alors

$$\begin{aligned}
|Ax|_2^2 =& \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle \\
=& \sum_{i=1}^n \lambda_i |x_i|_2^2 && \text{car } e_i \text{ base orthonormée de vecteurs propres,} \\
\leqslant& \sum_{i=1}^n \rho(A^*A)|x_i|_2^2 && \text{car } |\lambda_i| \leqslant \rho(A^*A) \text{ pour tout } 1 \leqslant i \leqslant n, \\
=& \rho(A^*A)|x|_2^2
\end{aligned}$$

Ainsi

$$\|A\|_2 = \sup_{x\neq 0} \frac{|Ax|_2}{|x|_2} \leqslant \sqrt{\rho(A^*A)}$$

Mais, d'autre part, on a pour $x = e_n$

$$|Ae_n|_2^2 = \lambda_n^2 |e_n|_2^2 = \rho(A^*A)^2 |e_n|_2^2 \leqslant \|A\|_2^2 |e_n|_2^2$$

D'où l'égalité $\|A\|_2^2 = \rho(A^*A)^2$. En appliquant cette égalité à A^*, on trouve $\|A^*\|_2^2 = \rho(AA^*)^2$. Enfin, pour toutes matrices A, B on a $\chi_{AB} = \chi_{BA}$ donc, en particulier, $\mathrm{sp}(AB) = \mathrm{sp}(BA)$ et donc $\rho(AB) = \rho(BA)$, ce qui donne $\rho(AA^*)^2 = \rho(A^*A)^2$.

(2) Si A est normale, alors elle est diagonalisable en base orthonormée. Autrement dit, il existe D diagonale réelle et $P \in \mathcal{U}_n(\mathbb{C})$ telles que $A = P^*DP$. Ainsi, $\rho(A) = \rho(D)$ et $\rho(A^*A) = \rho(D^2)$, donc $\rho(A) = \sqrt{\rho(A^*A)} = \|A\|_2$. □

Voici enfin des résultats de conditionnement pour la norme 2 d'une matrice :

Théorème 4.25. *Soit $A \in \mathrm{GL}_n(\mathbb{C})$.*

(1) Si on écrit $\mathrm{sp}(A^*A) = \{\mu_1 \leqslant \cdots \leqslant \mu_n\}$, alors

$$\mathrm{cond}_2(A) = \sqrt{\frac{\mu_n}{\mu_1}} = \sqrt{\frac{\max_{1\leqslant i\leqslant n} \mu_i}{\min_{1\leqslant i\leqslant n} \mu_i}}.$$

(2) Si A est normale et qu'on écrit $\mathrm{sp}(A) = \{\lambda_1, \ldots, \lambda_n\}$ avec $|\lambda_1| \leqslant \cdots \leqslant |\lambda_n|$, alors

$$\mathrm{cond}_2(A) = \frac{|\lambda_n|}{|\lambda_1|} = \frac{\max_{1\leqslant i\leqslant n} |\lambda_i|}{\min_{1\leqslant i\leqslant n} |\lambda_i|}.$$

(3) $\mathrm{cond}_2(A) = 1$ si, et seulement si, A s'écrit $A = \lambda U$ avec $\lambda \in \mathbb{C}^*$ et $U \in \mathcal{U}_n(\mathbb{C})$.

Démonstration. (1) On a

$$\|A\|_2^2 = \rho(A^*A) = \max_{1\leqslant i\leqslant n} \mu_i = \mu_n$$

et

$$\|A^-1\|_2^2 = \rho((A^{-1})^*A^{-1}) = \rho((A^*A)^{-1}) = \max_{1\leqslant i\leqslant n} \frac{1}{\mu_i} = \frac{1}{\mu_1}.$$

D'où le résultat.

(2) On a

$$\|A\|_2 = \rho(A) = \max_{1\leqslant i\leqslant n} |\lambda_i| = |\lambda_n|$$

et

$$\|A^-1\|_2 = \rho(A^{-1}) = \max_{1\leqslant i\leqslant n} \frac{1}{|\lambda_i|} = \frac{1}{|\lambda_1|}.$$

D'où le résultat.

(3) $\Rightarrow$: si $\mathrm{cond}_2(A) = 1$, alors $\min \mathrm{sp}(A^*A) = \max \mathrm{sp}(A^*A) = \mu > 0$. Donc $A^*A - \mu I_n$ est à la fois nilpotente et hermitienne, donc nulle. En posant $\lambda = \sqrt{\mu}$ et $U = \frac{1}{\lambda}A$, on a alors $U^*U = \frac{A^*}{\lambda}\frac{A}{\lambda} = \frac{\mu I_n}{\mu} = I_n$. Donc $U \in \mathcal{U}_n(\mathbb{C})$.

$\Leftarrow$: découle des égalités pour $\lambda \in \mathbb{C}^*$ et $U \in \mathcal{U}_n(\mathbb{C})$:

$$\mathrm{cond}_2(\lambda U) = \mathrm{cond}_2(U) = \rho(U^*U) = \mathrm{cond}_2(I_n) = 1$$

□

On verra en exercices différents calculs de conditionnement.

4.8 Décomposition en valeurs singulières

Pour conclure ce chapitre, on va présenter la décomposition en valeurs singulières (en anglais SVD pour *Singular Value Decomposition*. C'est une décomposition qui est, d'une certaine façon, une généralisation aux matrices rectangulaires de la diagonalisation orthogonales des matrices dans le cas réel, en relaxant la condition de symétrie.

Soit $A \in \mathcal{M}_{m,n}(\mathbb{C})$. Alors la matrice $A^*A \in \mathcal{M}_n(\mathbb{C})$ est hermitienne et positive. Soit λ une valeur propre de A^*A et $s = \sqrt{\lambda}$. Si v est un vecteur propre de A pour la valeur propre λ, alors on peut considérer $u = \frac{1}{s}Av$ et on a

$$Av = su \qquad\qquad A^*u = sv \tag{1}$$

donc u est en particulier non nul car v l'est.

Réciproquement, si on a un tel couple de vecteurs non nuls (u, v) vérifiant 1, alors $A^*Av = sA^*u = s^2v$ donc s^2 est valeur propre de A^*A.

De plus, u et v ont même norme 2 car $u^*u = \frac{1}{s}v^*A^*u = \frac{1}{s}v^*A^*\frac{1}{s}Av = \frac{1}{s^2}(\lambda v^*v) = v^*v$.

Définition 4.26. Une *valeur singulière* de A est la racine carrée positive d'une valeur propre non nulle de A^*A.

Si (u, v) est un couple de vecteurs comme dans 1 de norme 1, alors on dit que u est un *vecteur singulier à gauche* et v est un *vecteur singulier à droite*.

Remarque 4.27. On montrera en exercice que A^*A et AA^* ont en fait les mêmes valeurs propres complexes, sauf éventuellement 0, donc que A^* et A ont mêmes valeurs singulières.

On dispose alors d'un résultat d'existence de diagonalisation en valeurs singulières :

Théorème 4.28 (Décomposition en valeurs singulières)**.** *Soient $m, n \in \mathbb{N}^*$. Soit $A \in \mathcal{M}_{m,n}(\mathbb{K})$ de valeurs singulières $s_1, \dots, s_r$ comptées avec multiplicité et $D = \mathrm{diag}(s_1, \cdots, s_r, 0, \cdots, 0) \in \mathcal{M}_n(\mathbb{R})$. Alors il existe des matrices $U, V \in \mathcal{U}_n(\mathbb{K})$ telles que $U^*AV = D$.*

En particulier, les r premières colonnes de U et V sont des vecteurs singulier, respectivement à gauche et à droite, de A.

Démonstration. Soit $\lambda_1, \dots, \lambda_r$ les valeurs propres non nulles de $H = A^*A$ qui sont donc réelles et strictement positives. Soit $s_i = \sqrt{\lambda_i}$ pour tout i les valeurs singulières de A correspondantes. Soit $\Delta = \mathrm{diag}(s_1, \dots, s_r)$.

La matrice $H \in \mathcal{M}_n(\mathbb{K})$ étant hermitienne, par diagonalisation en base orthonormée, il existe $V \in \mathcal{U}_n(\mathbb{K})$ telle que $V^*HV = D^2$. Soit $V_1 \in \mathcal{M}_{n,r}(\mathbb{K})$ la matrice extraite de V des r premières colonnes et $V_2 \in \mathcal{M}_{n,n-r}(\mathbb{K})$ la matrice extraite de V des $n-r$ dernières colonnes, de sorte que $V = \left(\begin{array}{c|c} V_1 & V_2 \end{array}\right)$. On a, d'une part

$$V_1^*HV_1 = \Delta^2$$

et d'autre part

$$V_2^*HV_2 = (AV_2)^*(AV_2) = 0$$

ce qui donne $AV_2 = 0$.

On pose $W_1 = \Delta^{-1}V_1^*A^* \in \mathcal{M}_{r,m}(\mathbb{K})$. On a alors

$$W_1AV_1 = \Delta^{-1}V_1^*A^*AV_1 = \Delta^{-1}\Delta^2 = \Delta.$$

De plus

$$W_1W_1^* = \Delta^{-1}V_1^*A^*AV_1\Delta^{-1} = \Delta^{-1}\Delta^2\Delta^{-1} = I_r.$$

Donc les r lignes de W_1 forment une famille libre orthonormée de r vecteurs de $\mathbb{K}^m$, qu'on peut compléter en une base orthonormée. Cela nous donne alors une matrice $W = W_2 \in \mathcal{M}_{m-r,r}(\mathbb{K})$ telle que $\left(\begin{array}{c} W_1 \\ \hline W_2 \end{array}\right) \in \mathcal{U}_m(\mathbb{K})$. On pose enfin $U = W^* = \left(\begin{array}{c|c} W_1^* & W_2^* \end{array}\right)$.

Calculons U^*AV :

$$U^*AV = WAV = \left(\begin{array}{c} W_1 \\ \hline W_2 \end{array}\right)\left(\begin{array}{c|c} AV_1 & AV_2 \end{array}\right) = \left(\begin{array}{c} W_1 \\ \hline W_2 \end{array}\right)\left(\begin{array}{c|c} AV_1 & 0 \end{array}\right) = \left(\begin{array}{c|c} \Delta & 0 \\ \hline W_2AV_1 & 0 \end{array}\right)$$

Or $W_2AV_1 = W_2\Delta^*W_1^* = 0$ car les colonnes de ΔW_1 constituent des vecteurs orthogonaux aux vecteurs des colonnes de W_2. D'où $U * AV = D$.

□