Systèmes d'équations linéaire ; opérations élémentaires, aspects algorithmiques, conséquences théoriques

Gabriel LEPETIT

Mémoire de master 2 MEEF dirigé par Guy CASALE *sur une leçon présentée en collaboration avec Antoine* DIEZ.

Table des matières

***

La résolution de systèmes linéaires sur un corps est un problème de base, aussi bien en algèbre qu'en analyse. En effet, la résolution numérique d'un problème d'équations aux dérivées partielles passe souvent par une discrétisation nécessitant la résolution d'un système linéaire. En outre, le calcul de la transformée de Fourier rapide, central en traitement numérique du signal, fait appel à cette notion. L'étude de systèmes linéaires sur un anneau dans le cadre de la théorie des modules donne lieu à de riches développements en arithmétique.

La recherche de méthodes efficaces pour résoudre ce genre de systèmes est cruciale car les systèmes apparaissant dans la pratique sont souvent de très grande taille.

Dans ce mémoire, on présentera tout d'abord une théorie générale des systèmes linéaires, notamment le théorème de Rouché-Fontené qui donne des conditions d'existence des solutions, puis on développera une méthode algorithmique de résolution par opérations élémentaires, le pivot de Gauss, qu'on interprétera en termes d'actions de groupe sur l'espace des matrices. Enfin, la troisième partie sera l'occasion de s'intéresser à des méthodes directes ou itératives efficaces de résolution.

1 Théorie générale des systèmes linéaires

On se donne un corps K et $A \in \mathscr{M}_{m,n}(K)$, $b \in \mathscr{M}_{m,1}(K)$. On cherche à étudier le système linéaire $Ax = b$. On se demande d'abord s'il admet des solutions, quelle est la structure de l'espace des solutions, s'il est possible de calculer ces solutions explicitement. La recherche de l'efficacité dans cette démarche fera l'objet des parties suivantes.

Définition 1.1

Le système $Ax = b$ est dit compatible s'il admet au moins une solution $x \in \mathscr{M}_{m,1}(K)$. Le rang du système est le rang de A.

1.1 Systèmes de Cramer

Définition 1.2

On parle de système de Cramer quand A est dans $\mathrm{GL}_n(K)$.

Un système de Cramer admet donc une unique solution $x = A^{-1}b$.

Théorème 1.3 (*Formules de Cramer*)

Si $A = (a_{ij})_{1 \leqslant i,j \leqslant n}$, l'unique solution du système de Cramer $Ax = b$ est donnée par $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ *où*

$$\forall i \in [\![1,n]\!], x_i = \frac{\begin{vmatrix} a_{11} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{n,n} \end{vmatrix}}{\det A}$$

Démonstration. En notant $C_1, \dots, C_n$ les colonnes de A, on a $b = \sum_{i=1}^{n} x_i C_i$.

On note A_k la matrice obtenue en remplaçant la k-ème colonne par b, et $B_{i,k}$ la matrice obtenue en remplaçant la k-ème colonne par C_i. Par linéarité du déterminant par rapport à la k-ème colonne et en utilisant son caractère alterné, on a :

$$\det(A_k) = \sum_{i=1}^{n} x_i \det(B_{i,k}) = x_k \det B_{k,k} = x_k \det A$$

□

Le calcul de la solution demande de l'ordre de $(n+1)!$ opérations. En effet, il faut calculer $n+1$ déterminants, chacun se calculant en $n!$ opérations.

Exemple 1.4. Considérons $A = \begin{pmatrix} 2 & -5 & 2 \\ 1 & 2 & -4 \\ 3 & -4 & -6 \end{pmatrix}$ et $b = \begin{pmatrix} 7 \\ 3 \\ 5 \end{pmatrix}$. Alors on a $x_1 = \frac{-1}{46} \begin{vmatrix} 7 & -5 & 2 \\ 3 & 2 & -4 \\ 5 & -4 & -6 \end{vmatrix} = 5$, $x_2 = \frac{-1}{46} \begin{vmatrix} 2 & 7 & 2 \\ 1 & 3 & -4 \\ 3 & 5 & -6 \end{vmatrix} = 1$, $x_3 = \frac{-1}{46} \begin{vmatrix} 2 & -5 & 7 \\ 1 & 2 & 3 \\ 3 & -4 & 5 \end{vmatrix} = 1$.

1.2 Cas général : le théorème de Rouché-Fontené

Soit $A \in \mathscr{M}_{m,n}(K)$, on note $r \leqslant \min(n,m)$ le rang du système. Quitte à permuter et renuméroter les équations, on peut supposer que $\delta = \begin{vmatrix} a_{11} & \dots & a_{1,r} \\ \vdots & & \vdots \\ a_{r,1} & \dots & a_{r,r} \end{vmatrix} \neq 0$

Définition 1.5

On appelle déterminants caractéristiques du système $Ax = b$ les déterminants

$$\Delta_s = \begin{vmatrix} a_{11} & \dots & a_{1,r} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{r,1} & \dots & a_{r,r} & b_r \\ a_{s,1} & \dots & a_{s,r} & b_s \end{vmatrix}$$

pour $s \in [\![r+1, n]\!]$

Théorème 1.6 (*Rouché-Fontené*)

Le système $Ax = b$ est compatible si et seulement si $\Delta_s = 0$ pour tout $s \in [\![r+1, n]\!]$. Dans ce cas, les solutions forment un espace affine de dimension $n-r$. Une solution x est obtenue en donnant à $x_{r+1}, \dots, x_n$ des valeurs arbitraires puis en calculant l'unique solution du système

$$\begin{cases} a_{11}x_1 + \dots + a_{1,r}x_r &= b_1 - a_{1,r+1}x_{r+1} - \dots - a_{1,n}x_n \\ \vdots & \vdots \\ a_{r,1}x_1 + \dots + a_{r,r}x_r &= b_r - a_{r,r+1}x_{r+1} - \dots - a_{r,n}x_n \end{cases}$$

La preuve présentée est issue de [5], p. 144.

Démonstration. Notons $C_1, \dots, C_n$ les colonnes de A. Comme δ est un mineur non nul de taille r de $(C_1 \dots C_r)$, les r premières colonnes de A forment une famille libre, donc une base de $\mathrm{Im}A$. Ainsi, le système $Ax = b$ est compatible si et seulement si $(C_1, \dots C_r, b)$ est liée, c'est à dire de rang r, ce qui équivaut à la nullité de tous les mineurs de taille $r+1$ de $(C_1 \dots C_r b)$, qui ne sont autre que les déterminants caractéristiques Δ_s, pour $s \in [\![r+1, n]\!]$.

Sous ces conditions, $B = (C_1 \dots C_n b)$ est de rang r donc ses lignes constituent une famille de rang r. Mais $B = \begin{pmatrix} a_{11} & \dots & a_{1r} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & & \vdots & \vdots \\ a_{r1} & \dots & a_{rr} & \dots & a_{rn} & b_r \\ \vdots & & \vdots & & & \vdots \\ a_{p1} & \dots & a_{pr} & \dots & a_{pn} & b_p \end{pmatrix}$ donc puisque $\delta \neq 0$, les r premières lignes sont indépendantes. Les $p-r$ dernières lignes sont donc combinaisons linéaires des r premières, ce qui permet leur élimination du système d'équations linéaires, qui devient équivalent au système d'*équations principales* suivant :

$$\begin{cases} a_{11}x_1 + \dots + a_{1r}x_r + a_{1,r+1}x_{r+1} + \dots + a_{1,n}x_n &= b_1 \\ \vdots & \vdots \\ a_{r,1}x_1 + \dots + a_{r,r}x_r + a_{r,r+1}x_{r+1} + \dots + a_{r,n}x_n \dots &= b_r \end{cases}$$

A nouveau, le fait que $\delta \neq 0$ garantit qu'une fois $x_{r+1}, \dots, x_n$ arbitrairement choisis, il existe un unique $(x_1, \dots, x_r)$ tel que $(x_1, \dots, x_n)$ soit solution de $Ax = b$. En particulier, l'espace des solutions est un espace affine de dimension $n-r$. □

Exemple 1.7. Soit la matrice réelle de rang 2 $A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 1 & 0 & -1 & -1 \\ -1 & 1 & 1 & 2 \end{pmatrix}$ et $b = \begin{pmatrix} 1 \\ 1 \\ m \end{pmatrix}, m \in \mathbb{R}$. Le seul déterminant caractéristique est $\Delta = \begin{vmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ -1 & 1 & m \end{vmatrix} = -2(m+1)$. Le système admet donc des solutions si et seulement si $m = -1$ et dans ce cas, il est équivalent à $\begin{cases} x_1 = 1 + x_3 + x_4 \\ x_2 = -x_4 \end{cases}$.

1.3 Le cas des systèmes homogènes

Proposition 1.8

Soit $A \in \mathscr{M}_{m,n}(K)$. Si A est de rang r, alors les solutions de $Ax = 0$ constituent un espace vectoriel de dimension $n - r$.

Ce résultat de base peut être appliqué par exemple pour prouver le théorème d'Artin, résultat de théorie de Galois. On s'appuie sur la démonstration de [6].

Lemme 1.9 (*Dedekind*)

Soient K, L deux corps et $(\varphi_1, \ldots, \varphi_n)$ une famille de morphismes de corps distincts de K dans L. Alors elle est linéairement indépendante sur L : si $\forall x \in K, \sum_{i=1}^{n} \alpha_i \varphi_i(x) = 0$, on a $\alpha_1 = \cdots = \alpha_n = 0$.

Démonstration. Supposons qu'il existe $\alpha_1, \ldots, \alpha_n$ non tous nuls tels que $\forall x \in K, \sum_{i=1}^{n} \alpha_i \varphi_i(x) = 0$. Quitte à réordonner les termes de la somme, on peut trouver $r \in \mathbb{N}^*$ tels que $\alpha_1, \ldots, \alpha_r$ sont non nuls et $\alpha_{r+1} = \cdots = \alpha_n = 0$, et on peut choisir $(\alpha_i)_i$ de sorte que ce r soit minimal.

Soit $y \in K$ tel que $\varphi_1(y) \neq \varphi_r(y)$. On a pour tout $x \in K$, $0 = \sum_{i=1}^{r} \alpha_i \varphi_i(yx) = \sum_{i=1}^{r} \alpha_i \varphi_i(y) \varphi_i(x)$.

De plus, $\varphi_1(y) \sum_{i=1}^{r} \alpha_i \varphi_i(x) = 0$ donc en soustrayant ces deux équations,

$$\forall x \in K, \sum_{i=2}^{r} \alpha_i (\varphi_i(y) - \varphi_1(x)) \varphi_i(x) = 0$$

Comme par hypothèse $\alpha_r(\varphi_r(y) - \varphi_1(y)) \neq 0$, on a une combinaison linéaire à coefficients non tous nuls de $r-1$ termes de $(\varphi_i)_i$ ce qui contredit la minimalité de r. □

Théorème 1.10

Si L est un corps et H est un groupe fini du groupe des automorphismes de L, alors si $L^H = \{x \in L : \forall \sigma \in H, \sigma(x) = x\}$, L/L^H est une extension finie, $|H| = [L : L^H]$ et H est le groupe des L^H-automorphismes de L.

Démonstration. On note $m = [L : L^H]$ (éventuellement égal à ∞) et $n = |H|$. On va vérifier dans un premier temps que $m = n$.

1 Supposons que $m < n < +\infty$. Fixons $x_1, \ldots, x_m$ une base de L sur L^H et notons $H = \{\sigma_1, \ldots, \sigma_n\}$. Considérons le système de m équations à n inconnues dans L, $Y_1, \ldots, Y_n$ défini par :

$$\forall j \in [\![1, m]\!], \sigma_1(x_j) Y_1 + \ldots \sigma_n(x_j) Y_n = 0$$

C'est un système surdéterminé donc il admet une solution non nulle $(y_1, \dots, y_n)$. Par suite, pour tout $x = \sum_{j=1}^{m} \alpha_j x_j \in L$, où $\alpha_j \in L^H$, on a

$$\sum_{i=1}^{n} \sigma_i(x) y_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j \sigma_i(x_j) y_i = \sum_{j=1}^{m} \alpha_j \left(\sum_{i=1}^{m} \sigma_i(x_j) y_i \right) = 0$$

On a donc $\sum_{i=1}^{n} y_i \sigma_i = 0$ avec les y_i non tous nuls ce qui contredit le lemme d'indépendance de Dedekind ci-dessous. Donc $m \geqslant n$.

2 Supposons que $m > n$. Il existe donc une famille $(x_1, \dots, x_{n+1})$ d'éléments de L libre sur L^H. Selon le même argument que pour le premier point, on peut trouver une famille non nulle $(y_1, \dots, y_{n+1}) \in L^{n+1})$ vérifiant (S) :

$$\forall i \in [\![1, n]\!], \sigma_i(x_1) y_1 + \dots + \sigma_i(x_{n+1}) y_{n+1} = 0$$

Sans perte de généralité, on peut supposer que parmi toutes les solutions non nulles, $(y_1, \dots, y_{n+1})$ a un nombre minimal r de termes non nuls. Alors quitte à renuméroter, on peut supposer que $\forall i \leqslant r, y_i \neq 0$ et $\forall i > r, y_i = 0$. Ainsi, (S) se réécrit :

$$\forall i \in [\![1, n]\!], \sigma_i(x_1) y_1 + \dots + \sigma_i(x_r) y_r = 0$$

Soit $\sigma \in H$, appliquons σ au système (S) : $\forall i \in [\![1, n]\!], (\sigma \circ \sigma_i)(x_1)\sigma(y_1) + \dots + (\sigma \circ \sigma_i)\sigma(y_r) = 0$. Comme $\tau \mapsto \sigma \circ \tau$ est une permutation de l'ensemble fini H, on a donc (Δ) :

$$\forall i \in [\![1, n]\!], \sigma_i(x_1) y_1 + \dots + \sigma_i(x_r) y_r = 0$$

En multipliant (S) par $\sigma(y_1)$, (Δ) par y_1 et en additionnant les deux systèmes, on obtient

$$\forall i \in [\![1, n]\!], \sigma_i(x_2)(\sigma(y_1)y_2 - \sigma(y_2)y_1) + \dots + \sigma_i(x_r)(\sigma(y_1)y_r - \sigma(y_r)y_1) = 0$$

L'entier r étant le nombre minimal de termes non nuls d'une solution non triviale de (S), on a $\forall j \in [\![2, r]\!], \sigma(y_1)y_j - y_1\sigma(y_j) = 0$, soit $\sigma(y_1 y_j^{-1}) = y_1 y_j^{-1}$ donc $\forall j \in [\![2, r]\!], y_1 y_j^{-1} \in L^H$. Ainsi pour tout $2 \leqslant j \leqslant r$, il existe $z_j \in (L^H)^*$ tel que $y_j = z_j y_1$.

La ligne de (S) correspondant à $\sigma_i = \mathrm{id}_L$ devient alors : $x_1 y_1 + x_2 z_2 y_1 + \dots + x_r z_r y_1 = 0$ donc comme $y_1 \neq 0$, on a $x_1 + x_2 z_2 + \dots + x_r z_r = 0$, de sorte que $(x_1, \dots, x_r)$ est une famille liée, ce qui contredit l'hypothèse initiale. Donc $m \leqslant n < +\infty$ et finalement $m = n$.

3 Notons G le groupe des L^H-automorphismes de L. Il contient H de manière évidente. Montrons que G est fini. Soit $(a_1, \dots, a_n)$ une base de L sur L^H, $\Pi_1, \dots, \Pi_r$ les polynômes minimaux respectifs des a_i sur L^H et $f = \Pi_1 \dots \Pi_r \in L^H[X]$. Soit R l'ensemble (fini) des racines de f dans L. Comme $\Pi_j(a_j) = 0$ pour tout j, R contient $\{a_1, \dots a_n\}$.

De plus, si $x = \sum_{i=1}^{n} \alpha_i a_i \in L$, où $\alpha_i \in L^H$, alors, pour tout élément σ de G, on a $\sigma(x) = \sum_{i=1}^{n} \alpha_i \sigma(a_i)$. Cela nous assure que $\psi : \begin{array}{ccc} G & \longrightarrow & \mathfrak{S}(R) \\ \sigma & \longmapsto & \sigma_{|R} \end{array}$ est injective et donc que G est fini.

On a $L^H \subset L^G \subset L$ par définition de G, et $L^G \subset L^H \subset L$ car $H \subset G$ donc $L^H = L^G$. Selon la conclusion du deuxième point, on a $|G| = [L : L^H] = [L : L^G] = |H|$ donc $G = H$.

□

Donnons quelques précisions supplémentaires sur la théorie de Galois. Étant donné une extension de corps L/K, on s'intéresse à son *groupe de Galois* $\mathrm{Gal}(L/K)$ qui est le groupe des K-automorphismes de corps de L. Le résultat majeur de cette théorie est la correspondance de Galois entre les corps intermédiaires $K \subset M \subset L$ et les sous-groupes H de $\mathrm{Gal}(L/K)$:

Théorème 1.11

Si L/K est une extension galoisienne, les applications $\mathrm{Fix} : H \mapsto L^H$ *et* $\mathrm{Gal} : M \mapsto \mathrm{Gal}(L/M)$ *sont réciproques l'une de l'autre, où L^H, comme défini dans l'énoncé du théorème d'Artin est appelé sous-corps fixe de L associé à H*

Il est remarquable qu'en vertu du théorème d'Artin, toute extension finie vérifie $\mathrm{Gal} \circ \mathrm{Fix} = \mathrm{id}$.

Définition 1.12

Soit L/K une extension algébrique. On dit que c'est une extension galoisienne si $L^{\mathrm{Gal}(L/K)} = K$.

On suppose à présent que K est un corps parfait, c'est-à-dire que si L/K est une extension algébrique, alors tout polynôme de $L[X]$ n'admet que des racines simples dans son corps de décomposition – L est dit *séparable*. La plupart des corps usuels sont parfaits : $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, les corps finis. En revanche pour p premier, $\mathbb{F}_p(T)$ n'est pas parfait.

Définition 1.13

L'extension algébrique L/K est dite normale si tout polynôme ***irréductible*** *$f \in K[X]$ admettant une racine dans L se décompose en produit de facteurs de degré 1 dans L.*

Par exemple $\mathbb{C}/\mathbb{R}$ est une extension normale.

Proposition 1.14

Soit L/K une extension finie, alors on a l'équivalence entre :

1 *L/K est galoisienne ;*

2 *L/K est normale ;*

3 *L est le corps de décomposition d'un polynôme $f \in K[X]$;*

4 $\mathrm{Gal}(L/K)$ *est d'ordre $[L : K]$;*

Démonstration. Montrons la première implication. Comme l'extension L/K est finie, son groupe de Galois G est fini (vu dans le troisième point de la démonstration du théorème d'Artin). Soit $f \in K[X]$ irréductible admettant une racine a dans L. Introduisons $P(X) = \prod_{\sigma \in G} (X - \sigma(a))$. Si $\tau \in G$, en notant encore τ son prolongement naturel à $L[X]$, on a $\tau(P) = \prod_{\sigma \in G} (X - (\tau \circ \sigma)(a)) = P$ car $\sigma \mapsto \tau \circ \sigma$ est une permutation de G. Donc $P \in L^G[X] = K[X]$ car L/K est galoisienne ; et de plus $P(a) = 0$. Mais f étant irréductible, c'est à constante près le polynôme minimal de a sur K donc f divise P dans $K[X]$. Ainsi, f est scindé à racines simples sur L. □

En particulier si L/K est galoisienne finie et $K \subset M \subset L$ est un corps intermédiaire, alors L/M est galoisienne puisque L est le corps de décomposition de $f \in K[X] \subset M[X]$, ce qui

prouve la correspondance de Galois.

Voir également [11] pour un point de vue concis sur la théorie de Galois.

2 Sur les opérations élémentaires

2.1 Définitions

Une *matrice de dilatation* est une matrice de la forme $D_i(\alpha) = \mathrm{Diag}(1,\dots,1,\alpha,1,\dots,1)$ où $\alpha \neq 0$ est en i-ème position dans la diagonale. Son inverse est $D_i(\alpha^{-1})$.

Une *matrice de transvection* est une matrice de $\mathrm{SL}_n(K)$ de la forme $T_{ij}(\lambda) = \begin{pmatrix} 1 & & & \\ & 1 & \lambda & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$ où λ est en position (i,j) et $i \neq j$. L'inverse de $T_{ij}(\lambda)$ est $T_{ij}(-\lambda)$.

Une *matrice de permutation* est une matrice P_{ij} de la forme $\begin{pmatrix} I_{i-1} & & 0 \\ & U & \\ 0 & & I_{n-1-i-j} \end{pmatrix}$ avec $U = \begin{pmatrix} 0 & & & 1 \\ & 1 & & \\ & & \ddots & \\ 1 & & & 0 \end{pmatrix}$ bloc carré de taille $i+j+1$. La matrice P_{ij} est inversible d'inverse P_{ji}.

Proposition 2.1

Soit $A \in \mathscr{M}_{m,n}(K)$. On note L_i la i-ème ligne de A.

- *Calculer $D_i(\alpha)A$, c'est faire l'opération élémentaire $L_i \leftarrow \alpha L_i$.*
- *Calculer $T_{i,j}(\lambda)A$, c'est faire l'opération élémentaire $L_i \leftarrow L_i + \lambda L_j$.*
- *Calculer $P_{ij}A$, c'est faire l'opération d'échange $L_i \leftrightarrow L_j$.*

Remarque. Par multiplication à droite par ces matrices, on obtient les opérations élémentaires analogues sur les colonnes.

Le théorème suivant peut être prouvé de plusieurs manières. C'est soit une conséquence de la terminaison de l'algorithme du pivot de Gauss qu'on verra dans la section suivante, soit un résultat théorique comme montré dans [8], chapitre IV.

Théorème 2.2

- *Les matrices de transvection de taille n engendrent* $\mathrm{SL}_n(K)$.
- *Les matrices de transvection et de dilatation de taille n engendrent* $\mathrm{GL}_n(K)$.

Corollaire 2.3

Si K est de caractéristique différente de 2 ou 3, $\mathrm{SL}_n(K))$ *est son propre groupe dérivé.*

Exemple 2.4. Une application de ce résultat est le théorème de Frobenius-Zolotarev : si p est un nombre premier supérieur à 5, et $u \in \mathrm{GL}_n(\mathbb{F}_p)$ alors la signature de u est $\varepsilon(u) = \left(\frac{\det u}{p}\right)$.

2.2 L'algorithme du pivot de Gauss

Ce lemme donne l'idée de base de l'algorithme, qui est utilisé au moins depuis le I^er siècle en Chine, mais a été formalisé par Gauss en 1810.

Lemme 2.5

Toute matrice $A \in \mathscr{M}_{m,n}(K)$ à première colonne non nulle peut être transformée par multiplication à gauche par des transvections en une matrice de la forme $\begin{pmatrix} \beta & \star \\ 0 & \tilde{A} \end{pmatrix}$ où $\tilde{A} \in \mathscr{M}_{m-1,n-1}(K)$.

Démonstration. En effet, il suffit d'effectuer pour $i \geqslant 2$, l'opération élémentaire $L_i \leftarrow L_i - \frac{a_{i,1}}{a_{11}} L_1$, ce qui revient à multiplier par la matrice de transvection $T_{i,1}\left(-\frac{a_{i,1}}{a_{11}}\right)$. □

Corollaire 2.6

Soit $A \in \mathscr{M}_{m,n}(K)$. Il existe $M \in \mathrm{GL}_n(K)$ produit de matrices de transvections triangulaires supérieures et de matrices de permutation telle que $T = MA$ est triangulaire supérieure.

Remarque. Ainsi, $AX = b$ est équivalent à $TX = Mb$.

L'algorithme du pivot de Gauss détermine si un système du type $AX = B, A \in \mathscr{M}_{m,n}(K), B \in \mathscr{M}_{m,1}(K)$ est compatible, et s'il l'est le réduit à un système d'équations du type

$$\begin{pmatrix} a_{11} & & \star & \\ 0 & \ddots & & T \\ 0 & 0 & a_{r,r} & \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix} = B'$$

où $r \leqslant \inf(m,n)$ est le rang de A, les $a_{i,i}$ sont tous non nuls et $T \in \mathscr{M}_{r,n-r}(K)$, $B' \in \mathscr{M}_{r,1}(K)$. En vertu du théorème de Rouché-Fontené, l'espace des solutions de $AX = B$ est de dimension $n-r$ et une solution $x = (x_1, \ldots, x_n)$ s'obtient en donnant une valeur arbitraire à $x_{r+1}, \ldots, x_n$ et en résolvant par remontée le système triangulaire supérieur

$$\begin{cases} a_{11}x_1 + \cdots + a_{1,r}x_r &= b'_1 - a_{1,r+1}x_{r+1} - \cdots - a_{1,n}x_n \\ \vdots & \vdots \\ a_{r,r}x_r &= b'_r - a_{r,r+1}x_{r+1} - \cdots - a_{r,n}x_n \end{cases}$$

Description de l'algorithme

Entrées : une matrice $A = (a_{i,j})_{i,j}$ de taille $m \times n$, un vecteur colonne B de taille m. On note $L_1, \ldots, L_m$ les lignes de la matrice $\begin{pmatrix} A & B \end{pmatrix}$.

Pour $1 \leqslant k \leqslant n-1$:
Si il existe $r \geqslant k$ tel que $a_{r,k} \neq 0$:
Trouver $r \geqslant k$ tel que $|a_{r,k}| = \max_{k \leqslant i \leqslant n} |a_{i,k}|$.
Échanger L_k et L_r
Pour $i \geqslant k+1$:
Effectuer $L_i \leftarrow L_i - \frac{a_{i,k}}{a_{k,k}} L_k$.
Si L_i est de la forme $(0 \ldots 0, b'_i)$:
Si $b'_i \neq 0$: terminer et renvoyer ("non compatible",A,B).
Sinon, supprimer L_i de la matrice.

Renvoyer ("compatible",A', B'), où B' est la dernière colonne de A et A' la matrice constituée des n premières colonnes de A.

Remarque.
- Cet algorithme nécessite $O(n^3)$ opérations.
- Sa bonne terminaison est assurée par le lemme.
- On choisit le pivot de module maximal pour des raisons de stabilité numérique : un exemple spectaculaire est donné dans [3], p. 78.

Le résultat suivant est une application de la méthode du pivot :

Théorème 2.7 (*décomposition de Bruhat*)

Si $\mathcal{T}_s$ est le groupe des matrices triangulaires supérieures inversibles de $\mathcal{M}_n(K)$, on a $\mathrm{GL}_n(K) = \coprod_{\sigma\in\mathfrak{S}_n} \mathcal{T}_s P_\sigma \mathcal{T}_s$ où $P_\sigma = (\delta_{i,\sigma(j)})_{i,j}$ est la matrice de permutation associée à σ.

En utilisant le pivot de Gauss sur une matrice rationnelle, on peut obtenir des résultats sur les systèmes diophantiens, comme par exemple cette proposition :

Proposition 2.8

Soit $A \in \mathcal{M}_n(\mathbb{Z})$. Le système diophantien $Ax = 0$ admet une solution non nulle dans $\mathbb{N}^n$ si et seulement $0_{\mathbb{R}^n}$ est dans l'enveloppe convexe des colonnes de A.

Démonstration. On note A_i la i-ème colonne de A.

$\Leftarrow$: soit x solution non nulle dans $\mathbb{N}^n$, alors $0 = (A_1 \dots A_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i A_i$ donc en divisant par n, on obtient le résultat.

$\Rightarrow$: soit l minimal tel que 0 s'écrive comme combinaison convexe a l termes $\sum_{j=1}^{l} x_j A_{i_j}$ des colonnes de A. Selon le théorème de Carathéodory, en notant r le rang sur $\mathbb{Q}$ de la matrice $(A_{i_1}, \dots, A_{i_l})$, on a $l \leqslant r+1$. Mais puisqu'on a exhibé une relation de dépendance linéaire entre ces colonnes, $r < l$. Ainsi $r = l-1$ et par l'algorithme du pivot de Gauss sur $\mathbb{Q}$, on peut trouver $P \in \mathrm{GL}_m(\mathbb{Q})$ tel que $P(A_{i_1} \dots A_{i_r}) = \begin{pmatrix} M \\ 0 \end{pmatrix}$ où $M \in \mathcal{M}_{r,r+1}(\mathbb{Z})$ est de rang r. Donc $\ker_{\mathbb{Q}} M$ est de dimension 1 sur $\mathbb{Q}$ et de plus $M\begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix} = 0$ donc $x' = (x_1, \dots, x_r)$ est un vecteur directeur à coefficients positifs de $\ker_{\mathbb{Q}} M$.

Or $\ker_{\mathbb{Q}} M \subset \ker_{\mathbb{R}} M$ donc x' est également un vecteur directeur de $\ker_{\mathbb{R}} M$, de sorte que tous ses éléments ont leurs coefficients tous positifs ou tous négatifs. En multipliant x' par un coefficient bien choisi, on peut donc trouver $y' \in \mathbb{N}^r$ tel que $(A_{i_1} \dots A_{i_r})y' = 0$. On obtient $y \in \mathbb{N}^n$ tel que $Ay = 0$ en complétant y' avec des 0. □

2.3 Interprétation en termes d'actions de groupes

Dans cette sous-section, on cherche à étudier les orbites de l'action de $\mathrm{GL}_m(K)$ sur $\mathcal{M}_{m,n}(K)$ par multiplication à gauche, ce qui correspond à des opérations sur les lignes. Dans le contexte de la résolution de systèmes linéaires, cela est plus naturel que la multiplication à droite, qui correspond à un changement de variable affine des inconnues.

Des précisions supplémentaires ainsi qu'une application à l'étude des grassmanniennes peuvent être trouvés dans [2], chapitre IV.

Définition 2.9

On appelle pivot d'une ligne le coefficient non nul (s'il existe) situé dans la colonne la plus à gauche. Une matrice est dite échelonnée réduite lorsqu'elle vérifie les conditions suivantes :

- *Si une ligne est nulle, toutes les lignes suivantes sont nulles ;*
- *Le pivot d'une ligne est strictement plus à droite que ceux des lignes précédents ;*
- *Tous les pivots sont égaux à 1 et sont les seuls coefficients non nuls de leur colonne.*

Une matrice échelonné réduite est déterminée par la liste $(j_1, \dots, j_r)$ *des colonnes des pivots, qu'on appelle son type.*

Exemple 2.10. La matrice $\begin{pmatrix} 0 & 1 & 0 & \star & 0 & \star & \star \\ 0 & 0 & 1 & \star & 0 & \star & \star \\ 0 & 0 & 0 & 0 & 1 & \star & \star \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ est échelonnée réduite.

On note $\mathscr{E}_{m,n}$ l'ensemble des matrices échelonnées réduites de $\mathscr{M}_{m,n}(K)$. Le théorème suivant montre qu'elles constituent une classe de représentations pour l'action considérée.

Théorème 2.11

Soient $A, A' \in \mathscr{M}_{m,n}(K)$. *Alors* A *et* A' *sont dans la même orbite pour l'action par multiplication à gauche si et seulement si* $\ker A = \ker A'$. *On a* $\mathscr{M}_{m,n}(K) = \coprod_{E \in \mathscr{E}_{m,n}} \mathrm{Orb}(E)$.

Démonstration.

- Supposons $\ker A = \ker A' = F \subset K^m$. Soit une base $(e_1, \dots, e_p)$ de F, qu'on complète en une base $(e_1, \dots, e_p, f_{p+1}, \dots, f_m)$ de K^m. Alors $(v_{p+1}, \dots, v_m) = (A(f_{p+1}), \dots, A(f_m))$ (resp. $(v'_{p+1}, \dots, v'_m) = (A'(f_{p+1}), \dots, A'(f_m))$) est une base de $\mathrm{Im}A$ (resp. de $\mathrm{Im}A'$). En complétant chacune de ces deux familles libres en des bases $\mathscr{B}$, $\mathscr{B}'$ de K^m et en notant P la matrice de passage de $\mathscr{B}$ à $\mathscr{B}'$, on a $A' = PA$.

- Le fait que toute matrice est dans l'orbite d'une matrice échelonnée réduite résulte de la méthode du pivot de Gauss : en effet, partant d'une matrice réduite selon l'algorithme décrit ci-dessus, quitte à multiplier par des matrices de dilatation, on peut supposer que le pivot de chaque ligne est 1 ; de plus, si le pivot d'une ligne i est situé en $j^{\text{ième}}$ position, on peut annuler par des opérations $L_k \leftarrow L_k - \star L_i$ tous les coefficients $a_{k,j}$ pour $k \neq i$.

- Montrons enfin que deux matrices distinctes $E, E' \in \mathscr{E}_{m,n}$ ne peuvent être dans la même orbite.

  Procédons par récurrence sur m. Supposons que le résultat soit vrai pour toutes matrices échelonnées réduites à $m-1$ lignes. Supposons qu'il existe P inversible telle que $E = PE'$. Notons $(j_1, \dots, j_r)$ et $(j'_1, \dots, j'_s)$ leurs types respectifs. Alors $s = r$ est le rang de E, E'. De plus, les j_1-1 premières colonnes de PE sont nulles et celle d'indice j_1 est la première colonne de P. Donc $j_1 = j'_1$. Prenons $P = \begin{pmatrix} 1 & p_{12} & \cdots & p_{1m} \\ 0 & & & \\ \vdots & & Q & \\ 0 & & & \end{pmatrix}$, $E = \begin{pmatrix} L \\ F \end{pmatrix}$, $E' = \begin{pmatrix} L \\ F \end{pmatrix}$ $L, L' \in \mathscr{M}_{1,n}(K)$, des décompositions par blocs de P, E, E'. Alors F et F' sont

échelonnées réduites et $PE = \begin{pmatrix} L + (p_{12} \dots p_{1m})F \\ QF \end{pmatrix} = E'$ d'où $QF = F'$ et par hypothèse de récurrence $F = F'$, d'où $Q = I_{m-1}$. De plus, comme F est de type $(j_2, \dots, j_r)$, pour $k \geqslant 2$, le coefficient d'indice $(1, j_k)$ de PE est p_{1k} et d'autre part, le coefficient d'indice $(1, j_k)$ de E' est 0 car $j_1 = j'_1$. D'où $P = I_m$ et $E = E'$.

□

Remarque. On peut conduire la même étude pour l'action par multiplication à droite. Dans ce cas, deux matrices sont dans la même orbite si et seulement si elles ont la même image.

En utilisant la notion duale de *matrice échelonnée réduite en lignes*, on peut obtenir une description pratique des grassmanniennes de taille r de K^n, c'est à dire les sous-espaces vectoriels de dimension r de K^n :

Proposition 2.12

Si F est un sous-espace vectoriel de K^n de dimension r, il existe une unique suite strictement croissante $i = (i_1, \dots, i_r)$ et une matrice échelonnée A_F réduite en lignes de type i dans $\mathcal{M}_{n,r}(K)$ telle que A_F est la matrice d'une base de F dans la base canonique de K^n.

Démonstration. Voir [2] p. 137. □

L'étude du *plongement de Plücker* permet de prolonger cette étude et de réaliser l'espace des grassmanniennes comme un sous-ensemble d'un espace projectif.

Notation : On note $\Lambda(r, n)$ l'ensemble des parties à r éléments de $[\![, n]\!]$, et $\mathrm{Gr}_{r,n}$ l'espace des grassmanniennes de taille r de K^n.

En utilisant la multilinéarité par rapport aux colonnes du déterminant, on obtient la formule suivante :

Proposition 2.13 (*formule de Binet-Cauchy*)

Si $A \in \mathcal{M}_{m,n}, B \in \mathcal{M}_{n,l}$, alors pour tout $r \leqslant \min(m, n, l)$, $I \in \Lambda(r, m), J \in \Lambda(r, l)$, on a la relation sur les mineurs suivante :

$$\Delta_{I,J}(AB) = \sum_{K \in \Lambda(r,n)} \Delta_{I,K}(A)\Delta_{K,J}(B)$$

Si $A \in \mathcal{M}_{m,n}(K)$, $r \leqslant \min(m, n)$ on définit $\Lambda^r(A) = (\Delta_{I,J})_{\substack{I \in \Lambda(r,m) \\ J \in \Lambda_{(r,n)}}}$. On peut voir cet objet comme une matrice dans $\mathcal{M}_{\binom{r}{m},\binom{r}{n}}$. La formule ci-dessus se réécrit alors $\Lambda^r(AB) = \Lambda^r(A)\Lambda^r(B)$.

En particulier, si $A \in \mathcal{M}_{n,r}(K)$, $Q \in \mathrm{GL}_r(K)$, alors $\Lambda^r(AQ) = \det Q \Lambda^r(A)$.

Théorème 2.14

Soit $F \in \mathrm{Gr}_{r,n}$, $A_F \in \mathcal{M}_{n,r}(K)$ matrice d'une base $\mathcal{B}$ de F dans la base canonique de K^n. La droite engendrée par $\Lambda^r(A_F) \in \mathcal{M}_{\binom{r}{n},1}(K)$ ne dépend que de F, l'application

$$\begin{array}{rccc} \psi : & \mathrm{Gr}_{r,n} & \longrightarrow & \mathbb{P}^{\binom{r}{n}-1}(K) \\ & F & \longmapsto & [\Lambda^r(A_F)] \end{array}$$

est bien définie, et injective. De plus, ψ commute aux actions naturelles de $\mathrm{GL}_n(K)$ sur $\mathrm{Gr}_{r,n}$ et $\mathbb{P}^{\binom{r}{n}-1}(K)$, son image est une orbite de $\mathbb{P}^{\binom{r}{n}-1}(K)$ sous cette action.

Démonstration. Puisque A_F est de rang r, $\Lambda^r(A_F) \neq 0$. Soit $\mathcal{B}'$ autre base de F et A'_F la matrice associée, soit $Q \in \mathrm{GL}_r(K)$ matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$. Alors $A'_F = A_F Q$ donc

$\Lambda^r(A'_F) = \det Q \Lambda^r(A_F)$, de sorte que les droites engendrées par $\Lambda^r(A'_F)$ et $\Lambda^r(A_F)$ sont les mêmes. Donc ψ est bien définie.

Pour tout $v \in K^n$, $v \in F$ si et seulement si la matrice $(A_F \ v) \in \mathscr{M}_{n,r+1}(K)$ est de rang r, c'est à dire si tous ses mineurs de taille $r+1$ sont nuls.

En développant pour tout $I \in \Lambda(r+1,n)$ par rapport à la dernière colonne le mineur de taille $r+1$ associé à I de $(A_F \ v)$, on obtient

$$\forall v \in K^n, v \in F \iff \forall I \in \Lambda(r+1,n), \Delta_I = \sum_{i \in I} (-1)^i v_i \Delta_{I \setminus \{i\}}(A_F) = 0$$

Cette équation est homogène donc ne dépend pas de l'élément de la droite engendrée par $\Lambda^r(A_F)$ choisi. Ainsi, ψ est injectif.

Le groupe linéaire $\mathrm{GL}_n(K)$ agit sur $\mathrm{Gr}_{r,n}$ via $P \cdot F = P(F)$ et sur $\mathbb{P}^{\binom{r}{n}-1}(K)$ via $P \cdot [v] = [\Lambda^r(P)v]$. Alors $\forall P \in \mathrm{GL}_n(K)$, $\psi(P \cdot F) = [\Lambda^r(PA_F)] = [\Lambda^r(P)\Lambda^r(A_F)] = P \cdot [\Lambda^r(A_F)]$.

En outre, l'action de $\mathrm{GL}_n(K)$ sur $\mathrm{Gr}_{r,n}$ est évidemment transitive donc l'image de ψ est une orbite dans $\mathbb{P}^{\binom{r}{n}-1}(K)$.

□

2.4 Sur un anneau euclidien : l'algorithme des facteurs invariants

On s'appuie sur [1], pp.286-287.

Théorème 2.15

Soit A un anneau euclidien, $U \in \mathscr{M}_{m,n}(A)$. Alors il existe $s \in \mathbb{N}$ et $(d_1, \dots, d_s) \in A^s$ tels que $d_s | \dots | d_1$ et $(P,Q) \in \mathrm{GL}_m(A) \times \mathrm{GL}_n(A)$ tels que

$$U = P \begin{pmatrix} d_1 & & & (0) \\ & \ddots & & \\ & & d_s & \\ (0) & & & 0 \end{pmatrix} Q$$

De plus, si U vérifie une relation de cette forme avec $(d'_1, \dots, d'_r)$, alors $r = s$ et pour tout i, d'_i et d_i sont associés.

Remarque. Le résultat est encore vrai si A est un anneau principal, cependant, la preuve n'est alors plus algorithmique.

Décrivons l'algorithme permettant d'obtenir une telle réduction.

Algorithme des facteurs invariants

Entrées : A un anneau euclidien de stathme φ, $U \in \mathscr{M}_{m,n}(A)$. On note C_j sa $j^{\text{ème}}$ colonne et L_i sa $i^{\text{ème}}$ ligne.

Si $U \neq 0$:

1 Choisir (i_0, j_0) tel que $\varphi(u_{i_0,j_0})$ est minimal parmi les stathmes des coefficients non nuls de U.

Échanger C_1 et C_{j_0}, L_1 et L_{i_0}.

2 **Pour** $2 \leqslant i \leqslant m$:

(a) Effectuer la division euclidienne $u_{i,1} = u_{11}q + r_i$.
Effectuer $L_i \leftarrow L_i - qL_1$

(b) Si $r_i \neq 0$, échanger L_i et L_1 et retourner en **2** (a)

3 Pour $2 \leqslant j \leqslant n$:

(a) Effectuer la division euclidienne $u_{1,j} = u_{11}q + s_j$.
Effectuer $C_j \leftarrow C_j - qC_1$

(b) Si $s_j \neq 0$, échanger C_j et C_1, retourner en **2**

4 (a) Si il existe $i_1 \geqslant 2$, $j_1 \geqslant 2$ tels que u_{11} ne divise pas u_{i_1,j_1}, faire $C_1 \leftarrow C_1 + C_{j_1}$ et retourner en **2**.

(b) Sinon, appliquer l'algorithme à $(u_{i,j})_{2\leqslant i,j\leqslant n}$

Expliquons le principe : les étapes **1** à **3** visent à obtenir une matrice de la forme $\begin{pmatrix} \beta & 0 \\ 0 & \tilde{A} \end{pmatrix}$ où $\tilde{U} \in \mathscr{M}_{m-1,n-1}(A)$. Pour ce faire, on procède de manière analogue au pivot de Gauss en utilisant cette fois la terminaison de l'algorithme d'Euclide en un nombre fini d'étapes. En effet, pour annuler $u_{i,1}$, on retranche à L_i le quotient de la division euclidienne par u_{11} et on échange L_i et L_1 puis on recommence, ce qui revient à placer à l'étape k le $k^{\text{ième}}$ reste de l'algorithme d'Euclide en place $(1,1)$.

Exemple 2.16 (résolution de systèmes diophantiens)**.** Soit $A \in \mathscr{M}_{m,n}(\mathbb{Z})$, $b \in \mathscr{M}_{m,1}(\mathbb{Z})$. Soit $(P,Q) \in \mathrm{GL}_m(\mathbb{Z}) \times \mathrm{GL}_n(\mathbb{Z})$ tel que $PAQ = D = \begin{pmatrix} d_1 & & (0) \\ & \ddots & \\ & & d_r \\ (0) & & 0 \end{pmatrix}$ où $d_1|\ldots|d_r$. Alors $\forall x \in \mathscr{M}_{n,1}(\mathbb{Z})$, $Ax = b \iff Dy = b'$ où $y = U^{-1}x$, $b' = Ub$. Le système admet donc une solution si et seulement si $\forall 1 \leqslant i \leqslant r, d_i|b'_i$.

Proposition 2.17 (*théorème de la base adaptée*)

Soient A un anneau euclidien, L un A-module de rang n et $R \subset L$ un sous A-module. Il existe une base $(e_1,\ldots,e_n)$ de L et $(d_1,\ldots,d_k) \in A^k$ tels que $d_1|\ldots|d_k$ et $(d_1e_1,\ldots,d_ke_k)$ est une base de R.

Dans le cas des groupes abéliens finis, qui sont des $\mathbb{Z}$-modules, on obtient le résultat suivant :

Proposition 2.18

Si G est un groupe abélien fini, il existe un unique $(d_1,\ldots,d_r) \in \mathbb{Z}^r$ tel que $d_r|\ldots|d_1$ et $G \simeq \mathbb{Z}/d_1\mathbb{Z} \times \ldots \mathbb{Z}/d_r\mathbb{Z}$.

Mentionnons également que, si K est un corps, l'utilisation de l'algorithme des facteurs invariants sur un certain $K[X]$-module de type fini permet d'obtenir les invariants de similitude d'un endomorphisme de u de K^n, c'est-à-dire l'unique $(P_1,\ldots,P_r) \in K[X]^r$ tel que $P_r|\ldots P_1$ et dans une base $\mathscr{B}$ bien choisie, $M_{\mathscr{B}}(u) = \begin{pmatrix} C_{P_1} & & 0 \\ & \ldots & \\ 0 & & C_{P_r} \end{pmatrix}$, avec C_{P_i} la matrice compagnon associée à P_i.

3 Résolution effective des systèmes linéaires

3.1 Méthodes directes

Commençons par remarquer le fait suivant : la résolution d'un système triangulaire supérieur de taille n demande $O(n^2)$ opérations. En effet, si $\begin{cases} t_{11}x_1 + \cdots + t_{1,n}x_n &= b_1 \\ \vdots & \vdots \\ t_{n-1,n-1}x_{n-1} + t_{n-1,n}x_n &= b_{n-1} \\ t_{n,n}x_n &= b_n \end{cases}$, on calcule x_n en une opération, puis x_{n-1} en trois opérations (une multiplication, une soustraction, une division), puis x_{n-2} en cinq, ..., x_1 en $1 + 2(n-1)$ opérations. Comme $\sum_{i=1}^{n} 1 + 2(n-i) = n + n(n-1) = n^2$, c'est bien le nombre d'opérations nécessaires.

Définition 3.1

Le conditionnement de $A \in \mathcal{M}_n(\mathbb{C})$ relativement à la norme subordonnée $\|\cdot\|$ est $\mathrm{cond}(A) = \|A\|\|A^{-1}\| \geqslant 1$.

Le conditionnement quantifie la dépendance de la solution de $Ax = b$ par rapport à une petite variation de b ou de A, comme le montre la proposition suivante :

Proposition 3.2

- *Si $Ax = b$ et $A(x + \delta x) = b + \delta b$, alors* $\dfrac{\|\delta x\|}{\|x\|} \leqslant \mathrm{cond}(A)\dfrac{\|\delta b\|}{\|b\|}$
- *Si $(A + \delta A)(x + \delta x) = b$, alors* $\dfrac{\|\delta x\|}{\|x + \delta x\|} \leqslant \mathrm{cond}(A)\dfrac{\|\delta A\|}{\|A\|}$

Ces inégalités sont optimales.

Une méthode directe vise à obtenir une solution exacte en un nombre fini d'étapes. On s'appuie pour cela sur des décompositions de A permettant de se ramener à la résolution d'un système triangulaire supérieure.

Théorème 3.3 (*décomposition LU*)

Soit $A \in \mathrm{GL}_n(\mathbb{C})$ telle que tous ses mineurs principaux sont non nuls. Il existe un unique couple (L, U) avec U triangulaire supérieure et L triangulaire inférieure ayant uniquement des 1 sur la diagonale tel que $A = LU$.

Par conséquent, pour tout $x \in \mathbb{R}^n$, $Ax = b \Leftrightarrow \begin{cases} Ux = y \\ Ly = b \end{cases}$. La décomposition LU et la résolution du système se font en $O(n^3)$ opérations.

Définition 3.4

Une matrice bande p est une matrice $A = (a_{ij})_{i,j}$ telle que $a_{ij} = 0$ pour $|i - j| > p$.

Proposition 3.5

La décomposition LU préserve la structure bande des matrices.

Théorème 3.6

Soit $A \in \mathcal{M}_n(\mathbb{C})$. Il existe $Q \in \mathcal{M}_n(\mathbb{C})$ unitaire et $R \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure telle que $A = QR$.

Démonstration. Essentiellement, pour une matrice A inversible, il s'agit du procédé d'orthonormalisation de Gram-Schmidt. En effet, soit $\langle \cdot, \cdot \rangle$ le produit scalaire canonique de $\mathbb{C}^n$.

Étant donné une base $(e_1,\dots,e_n)$ de $\mathbb{C}^n$, on construit par récurrence une base orthonormée $(f_1,\dots,f_n)$ e de $\mathbb{C}^n$ en posant $f_1=\dfrac{e_1}{\|e_1\|}$, et pour $i\geqslant 2$, $f_i=\dfrac{f_i'}{\|f_i'\|}$ où $f_i'=e_i-\sum_{j=1}^{i-1}\langle e_j,f_j\rangle f_j$. La matrice de passage de $(e_1,\dots,e_n)$ à $(f_1,\dots,f_n)$ est clairement triangulaire supérieure, ce qui donne le résultat. □

L'intérêt de la méthode est que l'inverse d'une matrice unitaire est sa transconjuguée, ce qui se calcule avec un coût faible et de plus si $x\in\mathbb{R}^n$, $Ax=b\Leftrightarrow Rx=Q^*b$.

3.2 Méthodes itératives de résolution d'un système linéaire

Soit $A\in \mathrm{GL}_n(\mathbb{R})$, $b\in\mathbb{R}^n$. On étudie le système $Ax=b$.

Définition 3.7

Si $(M,N)\in \mathrm{GL}_n(\mathbb{R})\times\mathscr{M}_n(\mathbb{R})$ est tel que $A=M-N$, on dit que la méthode itérative associée à (M,N) converge si pour tout $u_0\in R^n$, la suite de premier terme u_0 et définie par $\forall k\in\mathbb{N}, u_{k+1}=M^{-1}(Nu_k+b)$ converge.

Théorème 3.8

La méthode itérative associée à (M,N) converge si et seulement si $\rho(M^{-1}N)<1$.

Commençons par montrer un lemme :

Lemme 3.9

Soit $A\in\mathscr{M}_n(\mathbb{C})$, $\epsilon>0$. Alors il existe une norme subordonnée $|||\cdot|||$ telle que $|||A|||\leqslant\rho(A)+\varepsilon$.

Démonstration. Comme A est à coefficients dans $\mathbb{C}$, elle est trigonalisable : on se donne donc P inversible et $T=(t_{ij})_{1\leqslant i,j\leqslant n}$ triangulaire supérieure tels que $A=PTP^{-1}$.

Notons $(e_1,\dots,e_n)$ la base canonique de $\mathbb{C}^n$. Pour $\delta>0$, on pose $e_1'=\delta^{i-1}e_i$ et $D_\delta=\mathrm{Diag}(1,\delta,\cdots,\delta^{n-1})$.

On a donc $\forall j\in[\![1,n]\!], Te_j'=\delta^{j-1}Te_j=\delta^{j-1}\sum_{i=1}^{j}t_{ij}e_i=\sum_{i=1}^{j}\delta^{j-i}t_{ij}e_i'$, de sorte que $T_\delta=D_\delta^{-1}TD_\delta$ est la matrice $\begin{pmatrix} t_{11} & \delta t_{12} & \dots & \delta^{n-1}t_{1n}\\ & \ddots & \ddots & \dots\\ (0) & & \ddots & \delta t_{n-1n}\\ & & & t_{nn}\end{pmatrix}$.

On définit pour $x\in\mathbb{R}^n$, $\|x\|=\|(PD_\delta)^{-1}x\|_\infty$, et on note $|||\cdot|||$ la norme subordonnée associée. On vérifie aisément que $\forall B\in\mathscr{M}_n(\mathbb{R}), |||B|||=|||(PD_\delta)^{-1}BPD_\delta|||_\infty$.

Or (admis ici), pour tout $B=(b_{ij})_{i,j}\in\mathscr{M}_n(\mathbb{R})$, on a $|||B|||_\infty=\sup_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|b_{ij}|$. En choisissant $\delta>0$ tel que pour tout $1\leqslant i\leqslant n-1$, $\sum_{j=i+1}^{n}\delta^{j-i}|t_{ij}|\leqslant\varepsilon$, on obtient donc, puisque $\rho(A)=\sup_{1\leqslant i\leqslant n}|t_{ii}|$, $|||A|||=|||T_\delta|||_\infty\leqslant\rho(A)+\varepsilon$. □

Démonstration (du théorème)**.** Soit $u\in\mathbb{R}^n$ tel que $Au=b$, c'est à dire $Mu=Nu+b$. Posons $e_k=u_k-u$ en reprenant les notations du théorème. Alors :

$$e_{k+1}=M^{-1}(Nu_k+b)-M^{-1}Nu-M^{-1}b=M^{-1}N(u_k-u)=M^{-1}Ne_k$$

Ainsi, par une récurrence immédiate, $\forall k \in \mathbb{N}, e_k = (M^{-1}N)^k e_0$. Dès lors, deux cas se présentent :

- Si $\rho(M^{-1}N) < 1$, on fixe $\varepsilon = \dfrac{1-\rho(M^{-1}N)}{2}$ et le lemme nous fournit une norme subordonnée $|||\cdot|||$ telle que $|||M^{-1}N||| \leqslant \rho(M^{-1}N)+\varepsilon < 1$. Donc pour la norme $\|\cdot\|$ associée, on a pour tout k, $\| e_k \| \leqslant |||M^{-1}N|||^k \| e_0 \|$ donc $\lim\limits_{k\to+\infty} e_k = 0$ si bien que $(u_k)_k$ converge vers u.
- Si $\rho(M^{-1}N) \geqslant 1$, soit λ valeur propre complexe de module supérieur ou égal à 1, et $\tilde{u} = \tilde{u}_1 + i\tilde{u}_2$ un vecteur propre associé. Comme pour tout k, $(M^{-1}N)^k\tilde{u} = \lambda^k\tilde{u}$, la méthode itérative ne converge pas pour $u_0 = u + \tilde{u}_1$.

□

Décrivons maintenant quelques cas particuliers de méthodes itératives :

- Méthode de Jacobi : $M = \mathrm{Diag}(a_{11},\dots,a_{nn}) = D$ et $N = D-A$. On note $J = D^{-1}(D-A)$.
- Méthode de Gauss-Seidel : $M = D-E$ où $D = \mathrm{Diag}(a_{11},\dots,a_{nn})$ et $E = -A_{\mathrm{inf}}$, partie triangulaire inférieure stricte de A. $N = -A_{\mathrm{sup}} = F$. On note $\mathscr{L}_1 = (D-E)^{-1}F$.
- Méthode de relaxation : $M = \dfrac{D}{\omega} - E$ et $N = \dfrac{1-\omega}{\omega}D + F$, $\mathscr{L}_\omega = \left(\dfrac{D}{\omega} - E\right)^{-1}\left(\dfrac{1-\omega}{\omega}D + F\right)$.

Proposition 3.10

Si A est une matrice tridiagonale, $\rho(\mathscr{L}_1) = (\rho(J))^2$. La méthode de Gauss-Seidel a donc une vitesse de convergence double de celle de la méthode de Jacobi.

Démonstration. Remarque préliminaire : introduisons pour $\mu \neq 0$,

$$A(\mu) = \begin{pmatrix} b_1 & \mu^{-1}c_2 & & (0) \\ \mu a_2 & b_2 & \ddots & \\ & \ddots & \ddots & \mu^{-1}c_n \\ (0) & & \mu a_n & b_n \end{pmatrix}$$

où $A = A(1)$. Alors $A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$ où $Q(\mu) = \mathrm{Diag}(\mu,\mu^2,\dots,\mu^n)$, donc $\det A(\mu) = \det A(1)$.

Les valeurs propres de J sont les racines du polynôme caractéristique $p_J(\lambda) = \det(D^{-1}(E+F)-\lambda I)$, ce sont aussi celles de $q_J(\lambda) = \det(\lambda D - E - F)$. De même, les valeurs propres de $\mathscr{L}_1$ sont les racines de $p_{\mathscr{L}_1}(\lambda) = \det((D-E)^{-1}F - \lambda I)$, et celles de $q_{\mathscr{L}_1}(\lambda) = \det(\lambda D - \lambda E - F)$.

Mais selon la remarque préliminaire, pour tout $\lambda \in \mathbb{C}^*$, $q_{\mathscr{L}_1}(\lambda^2) = \det(\lambda^2 D - \lambda^2 E - F) = \lambda^n \det(\lambda D - \lambda E - \lambda^{-1}F) = \lambda^n \det(\lambda D - E - F) = \lambda^n q_J(\lambda)$.

Donc les valeurs propres non nulles de $\mathscr{L}_1$ sont les carrés de valeurs propres non nulles de J, ce qui permet de conclure. □

Proposition 3.11

Le rayon spectral de $\mathscr{L}_\omega$ est strictement supérieur à $|\omega - 1|$. La méthode de relaxation ne peut donc converger que si $\omega \in]0,2[$.

Démonstration. La matrice $\mathscr{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$ est trigonalisable comme produit de matrices trigonalisables et en notant $\lambda_1, \ldots, \lambda_n$ ses valeurs propres avec multiplicité, on a

$$\prod_{i=1}^{n} \lambda_i = \det(\mathscr{L}_\omega) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = \frac{\prod_{i=1}^{n} \frac{1-\omega}{\omega} a_{ii}}{\prod_{i=1}^{n} \frac{a_i i}{\omega}} = (1-\omega)^n$$

Donc $\rho(\mathscr{L}_\omega)^n \geqslant |\det(\mathscr{L}_\omega)| = |1-\omega|^n$ de sorte que $\rho(\mathscr{L}_\omega) \geqslant |\omega - 1|$. □

Remarque. Par des techniques similaires, on montre que si A est tridiagonale et J a un spectre réel, la méthode de Jacobi et la méthode de relaxation pour $0 < \omega < 2$ convergent ou divergent simultanément. De plus, $\omega_0 = \dfrac{1}{1 + \sqrt{1 - \rho(J)^2}}$ est un paramètre de relaxation tel que $\rho(\mathscr{L}_{\omega_0})$ est minimal. La preuve de ce fait et d'autres détails se trouvent dans [3], pp. 95-107.

3.3 Méthodes de gradient

Ici, on prend $A \in \mathscr{S}_n^{++}(\mathbb{R})$.

Proposition 3.12

La résolution de $Ax = b$ équivaut à trouver le point qui minimise la fonctionnelle :

$$\Phi(y) = \frac{1}{2} y^T A y - y^T b.$$

Démonstration. En effet, il est facile de voir que

$$\nabla\Phi(y) = \frac{1}{2}(A^T + A)y - b = Ay - b. \tag{1}$$

Et si x est solution du système linéaire, alors $\Phi(y) = \Phi(x + (y-x)) = \Phi(x) + \frac{1}{2}(y - x)^T A(y-x)$, *i.e.* $\frac{1}{2}\|y-x\|_A^2 = \Phi(y) - \Phi(x)$, où $\|z\|_A^2 = z^T A z$ est la norme associée à A que l'on utilisera toujours par la suite.

□

Définition 3.13

Une méthode de gradient consiste à partir d'un point $x_0 \in \mathbb{R}^n$ et à construire la suite

$$x_{k+1} = x_k + \alpha_k d_k \tag{2}$$

où $d_k \in \mathbb{R}^n$ est une direction à choisir et $\alpha_k \in \mathbb{R}$.

Une idée naturelle est de choisir α_k de sorte à optimiser $\Phi(x_{k+1})$ dans la direction d_k, c'est à dire tel que $\dfrac{d}{d\alpha_k}\Phi(x_k + \alpha_k d_k) = -d_k^T r_k + \alpha_k d_k^T A d_k = 0$, où $-r_k := \nabla\Phi(x_k) = Ax_k - b$. On trouve :

$$\alpha_k = \frac{\langle d_k, r_k\rangle}{\|d_k\|_A^2} \tag{3}$$

(c'est bien défini lorsque $d_k \neq 0$ car $A \in S_n^{++}(\mathbb{R})$).

Théorème 3.14

Soit x la solution du système $Ax = b$ ou de façon équivalente, la solution du problème de minimisation (1)*. Si α_k est choisi comme dans* (3)*, alors la suite* (2) *vérifie :*

$$\|x_{k+1} - x\|_A^2 = (1 - \sigma_k)\|x_k - x\|_A^2$$

où $\sigma_k = \dfrac{\langle d_k, r_k\rangle^2}{\|d_k\|_A^2 \|r_k\|_{A^{-1}}^2} \in (0, 1]$.

Démonstration. Il suffit de calculer :

$$\begin{aligned} \|x_{k+1} - x\|_A^2 &= \|x_k - x + \alpha_k d_k\|_A^2 = \|x_k - x\|_A^2 + \alpha_k^2 \|d_k\|_A^2 + 2\alpha_k \langle d_k, A(x_k - x)\rangle \\ &= \|x_k - x\|_A^2 + \alpha_k^2 \|d_k\|_A^2 - 2\alpha_k \langle d_k, r_k\rangle \end{aligned}$$

car $A(x_k - x) = Ax_k - b = -r_k$ et $\|x_k - x\|_A^2 = \|r_k\|_{A^{-1}}^2$. Et en remplaçant α_k par son expression :

$$\|x_{k+1} - x\|_A^2 = \left(1 - \frac{\langle d_k, r_k\rangle^2}{\|d_k\|_A^2 \|r_k\|_{A^{-1}}^2}\right) \|x_k - x\|_A^2.$$

□

3.3.1 Méthode de gradient conjugué

Remarquons que pour tout $k \in \mathbb{N}$:

$$r_{k+1} = r_k - \alpha_k A d_k \tag{4}$$

et α_k est choisi de sorte à ce que

$$\langle r_{k+1}, d_k\rangle = 0. \tag{5}$$

Idée. Construire des directions (d_k) deux à deux A-orthogonales ; ainsi, r_{k+1} sera orthogonal à $\mathrm{Vect}(d_0, \ldots, d_k)$.

Notations. Pour $x, y \in \mathbb{R}^n$, on note $x \perp y$ lorsque x et y sont orthogonaux pour le produit scalaire euclidien et $x \perp_A y$ lorsque x et y sont orthogonaux pour le produit scalaire donné par A. On étend naturellement cette notation à des sous-espaces de $\mathbb{R}^n$.

On pose $d_0 = r_0$ et pour $k \in \mathbb{N}$, on construit d_{k+1} comme l'orthogonalisé de Gram-Schmidt pour le produit scalaire donné par A de r_{k+1} relativement à $\mathrm{Vect}(d_k)$:

$$d_{k+1} = r_{k+1} - \beta_k d_k \tag{6}$$

où

$$\beta_k = \frac{\langle r_{k+1}, A d_k\rangle}{\|d_k\|_A^2} \quad \text{si } d_k \neq 0, \quad \beta_k = 0 \text{ sinon.} \tag{7}$$

Remarquons que si $d_k = 0$ alors r_k et d_{k-1} sont colinéaires et comme ils sont aussi orthogonaux par (5), $r_k = 0$.

Lemme 3.15

Avec le choix (7)*, les directions* (6) *vérifient pour tout* $k \in \mathbb{N}$ *la propriété suivante : si* $r_0, \dots, r_k$ *ne sont pas nuls alors,*

1 $\text{Vect}(r_0, \dots, r_k) = \text{Vect}(d_0, \dots, d_k)$

2 $r_{k+1} \perp \text{Vect}(d_0, \dots, d_k)$

3 $d_{k+1} \perp_A \text{Vect}(d_0, \dots, d_k)$

Démonstration. On procède par récurrence sur $k \in \mathbb{N}$. Lorsque $k = 0$, 1, 2 et 3 sont vrais grâce aux relations $r_0 = d_0$, (5) et (6) et bien sûr $r_0 \neq 0$ sinon il n'y a rien à faire. Supposons donc le résultat vrai au rang $k-1$, $k \in \mathbb{N}^*$.

1 Par (6), on a : $d_k = r_k - \beta_{k-1} d_{k-1}$.

2 Par (5), on a déjà $r_{k+1} \perp d_k$ et si $j \in \{0, \dots, k-1\}$, la relation (4) couplée à l'hypothèse de récurrence 2 et 3 donne $r_{k+1} \perp d_j$.

3 Par (6), on a déjà $d_{k+1} \perp_A d_k$ (c'est la définition) et si $j \in \{0, \dots, k-1\}$, la relation (6) couplée à l'hypothèse de récurrence 3 donne :

$$\langle d_{k+1}, Ad_j \rangle = \langle r_{k+1}, Ad_j \rangle.$$

Montrons que $Ad_j \in \text{Vect}(r_0, \dots, r_k)$, ce qui conclura grâce aux relations 1 et 2 que l'on vient de prouver. Grâce à la relation (4) avec $k = j$, il suffit de montrer que $\alpha_j \neq 0$, ce qui est la cas car :

$$\alpha_j = 0 \overset{(3)}{\iff} \langle r_j, d_j \rangle = 0 \overset{(6)}{\iff} r_j = 0$$

et on a justement supposé le contraire.

□

Théorème 3.16

La méthode de gradient associée aux directions (6) *avec le choix* (7) *converge vers la solution* x *du problème* $Ax = b$ *en au plus* n *itérations.*

Démonstration. Les conditions 1 et 2 du lemme précédent assurent que tant que $r_l \neq 0$, la famille $(r_0, \dots, r_l)$ est une famille orthogonale donc libre. On est en dimension n donc nécessairement $l + 1 \leqslant n$ et si $r_l = 0$, x_l est solution du système. □

3.3.2 Méthode de gradient à pas optimal

On choisit pour direction la "plus grande pente", c'est à dire :

$$d_k = -\nabla\Phi(x_k) = -Ax_k + b = r_k.$$

Dans ce cas, $d_k \neq 0$ tant que la solution n'est pas atteinte. La convergence découle essentiellement de l'inégalité de Kantorovich :

Lemme 3.17 (*Inégalité de Kantorovich*)

En notant $0 < \lambda_1 \leq \dots \leq \lambda_n$ *les valeurs propres de* A*, on a pour tout* $y \in \mathbb{R}^n$*,*

$$\frac{\|y\|^4}{\|y\|_A^2 \|y\|_{A^{-1}}^2} \geq \frac{4\lambda_n \lambda_1}{(\lambda_n + \lambda_1)^2}$$

Démonstration. On va montrer l'inégalité équivalente :

$$\forall y \in \mathbb{R}^n,\ \|y\|^4 \leq \frac{1}{4}\left(\sqrt{\frac{\lambda_n}{\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2$$

On peut même supposer que $\|y\| = 1$ et commencer par remarquer :

$$1 = \|y\|^2 = \langle y, AA^{-1}y\rangle \leq \|y\|_A \|A^{-1}y\|_A = \|y\|_A \|y\|_{A^{-1}}$$

Et dans une base orthonormale de vecteurs propres :

$$\begin{aligned}
\|y\|_A\|y\|_{A^{-1}} &= \sqrt{\left(\sum_{i=1}^n \lambda_i y_i^2\right)\left(\sum_{i=1}^n \frac{1}{\lambda_i} y_i^2\right)} = \sqrt{\frac{\lambda_1}{\lambda_n}\left(\sum_{i=1}^n \frac{\lambda_i}{\lambda_1} y_i^2\right)\left(\sum_{i=1}^n \frac{\lambda_n}{\lambda_i} y_i^2\right)} \\
&\leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}}\left(\left(\sum_{i=1}^n \frac{\lambda_i}{\lambda_1} y_i^2\right) + \left(\sum_{i=1}^n \frac{\lambda_n}{\lambda_i} y_i^2\right)\right) \\
&\leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}}\left(\sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_1} + \frac{\lambda_n}{\lambda_i}\right) y_i^2\right)
\end{aligned}$$

La fonction $x \mapsto \dfrac{x}{\lambda_1} + \dfrac{\lambda_n}{x}$ admet un maximum en λ_1 ou en λ_n et il vaut dans les deux cas $1 + \dfrac{\lambda_n}{\lambda_1}$. Ainsi,

$$\|y\|_A\|y\|_{A^{-1}} \leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}}\left(\sum_{i=1}^n \left(1 + \frac{\lambda_n}{\lambda_1}\right) y_i^2\right) \leq \frac{1}{2}\left(\sqrt{\frac{\lambda_n}{\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n}}\right)$$

et le résultat suit en élevant au carré. □

Et sachant que $\text{cond}(A) = \lambda_n/\lambda_1$, on obtient le résultat suivant

Théorème 3.18

Avec les choix précédents et $d_k = r_k$, la suite (2) *converge vers x avec :*

$$\|x_k - x\|_A \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\|x_k - x\|_A$$

Plus précisément,

$$\|x_k - x\| \leq \sqrt{\text{cond}(A)}\left(\frac{\text{cond}(A) - 1}{\text{cond}(A) + 1}\right)^k \|x_0 - x\|$$

Démonstration. La première inégalité découle directement de l'inégalité de Kantorovich. Pour la seconde, on remarque que pour tout $y \in \mathbb{R}^n$, $\lambda_1\|y\|^2 \leq \|y\|_A^2 \leq \lambda_n\|y\|^2$ □

Avec la dernière inégalité, on voit que la convergence peut être lente lorsque la matrice est mal conditionnée.

Références

[1] Vincent Beck, Jérome Malick, and Gabriel Peyré. *Objectif agrégation*. H et K, 2ème edition, 2005.

[2] Philippe Caldero and Jérome Germoni. *Histoires hédonistes de groupes et de géométrie*, volume 1. Calvage et Mounet, 2013.

[3] Philippe Ciarlet. *Introduction à l'analyse numérique et à l'optimisation*. Masson, 1988.

[4] Xavier Gourdon. *Les maths en tête : algèbre*. Ellipses, 2ème edition, 2009.

[5] Joseph Grifone. *Algèbre linéaire*. Editions Cépaduès, 4ème edition, 2011.

[6] Alain Jeanneret and Daniel Lines. *Invitation à l'algèbre*. Editions Cépaduès, 2008.

[7] Peter Lax. *Linear algebra and its applications*. Wiley Intersciences, 2ème edition, 2007.

[8] Daniel Perrin. *Cours d'algèbre*. Ellipses, 1996.

[9] Alfio Quarteroni, Ricardo Sacco, and Fausto Saleri. *Numerical Mathematics*. Springer.

[10] Edouard Ramis, Claude Deschamps, and Jacques Odoux. *Cours de mathématiques spéciales*, volume 1. Dunod, 2001.

[11] Pierre Samuel. *Théorie algébrique des nombres*. Hermann, 1967.