# ALGÈBRE - LEÇON 162 : SYSTÈMES D'ÉQUATIONS LINÉAIRES ; OPÉRATIONS ÉLÉMENTAIRES, ASPECTS ALGORITHMIQUES ET CONSÉQUENCES THÉORIQUES

SIMON RICHE

## 1. Commentaires du jury (rapport 2022)

Dans cette leçon, les techniques liées au simple pivot de Gauss constituent l'essentiel des attendus. Il est impératif de faire le lien avec la notion de système échelonné (dont on donnera une définition précise et correcte) et de situer l'ensemble dans le contexte de l'algèbre linéaire, sans oublier la dualité. Un point de vue opératoire doit accompagner l'étude théorique et l'intérêt algorithmique des méthodes présentées doit être expliqué, éventuellement en l'illustrant par des exemples simples (où l'on attend parfois une résolution explicite).

Parmi les conséquences théoriques, les candidats peuvent notamment donner des systèmes de générateurs de $\mathrm{GL}_n(\mathbf{K})$ et $\mathrm{SL}_n(\mathbf{K})$. Ils est aussi pertinent de présenter les relations de dépendance linéaire sur les colonnes d'une matrice échelonnée qui permettent de décrire simplement les orbites de l'action à gauche de $\mathrm{GL}_n(\mathbf{K})$ sur $\mathcal{M}_n(\mathbf{K})$ donnée par $(P, A) \mapsto PA$.

S'ils le désirent, les candidats peuvent exploiter les propriétés des systèmes d'équations linéaires pour définir la dimension des espaces vectoriels et obtenir une description de l'intersection de deux sous-espaces vectoriels donnés par des systèmes générateurs, ou d'une somme de deux sous-espaces vectoriels donnés par des équations.

De même, des discussions sur la résolution de systèmes sur $\mathbf{Z}$ et la forme normale de Hermite peuvent trouver leur place dans cette leçon. Enfin, il est possible de présenter les décompostions $LU$ et de Choleski, en évaluant le coût de ces méthodes ou encore d'étudier la résolution de l'équation normale associée aux problèmes des moindres carrés et la détermination de la solution de norme minimale par la méthode de décomposition en valeurs singulières.

## 2. Plan

### 2.1. Ce qui doit apparaître.

Étude théorique :

- Explication de la notation matricielle $Ax = b$.
- Description "théorique" des solutions en termes de $\ker(A)$.
- Systèmes de Cramer et formules associées.
- Théorème de Rouché–Fontené.

Méthodes "directes" de résolution :

- pivot de Gauss et interprétation en termes d'opérations élémentaires.
- résolution des systèmes triangulaires par remontée.
- décomposition $LU$.

*Date*: Année 2022–2023.

— méthode de Cholesky pour les matrices symétriques définies positives.

Méthodes itératives :
— étude théorique de la convergence.
— méthode de Jacobi.
— méthode de Gauss–Seidel.
— critères (ou exemples) de convergence pour ces méthodes.

Conséquences théoriques :
— générateurs de $\mathrm{GL}_n$ et $\mathrm{SL}_n$ ; conséquences.
— description des orbites de $\mathrm{GL}_n$ pour l'action $(P, M) \mapsto PM$.

2.2. **Ce qui peut apparaître.** Factorisation $QR$ (méthode directe).

Méthode de relaxation (méthode itérative).

Méthodes de descente/gradient.

Étude détaillée du cas des matrices tridiagonales.

Décomposition de Bruhat.

Résolution d'équations linéaires dans $\mathbb{Z}^n$. Lien avec la description des groupes abéliens de type fini.

Théorème d'Artin en théorie de Galois.

## 3. Quelques questions bêtes auxquelles il faut absolument savoir répondre rapidement

(1) Comparer la décomposition $LU$ de la matrice ${}^tA$ en fonction de celle de $A$ (sans oublier la question de l'existence).
(2) Comment passe-t-on de la description d'un sous-espace vectoriel en termes d'un système de générateurs à une description en termes d'équations ? Et inversement ? Conséquences pour la description d'une somme / d'une intersection.

## 4. Exercices

4.1. **Systèmes linéaires.**

**Exercice 1.** Soit $\mathbb{K}$ un corps, et soit $\mathbb{L}$ une extension de $\mathbb{K}$.

(1) Montrer que si $A \in \mathrm{M}_{n,m}(\mathbb{K})$ et $b \in \mathbb{K}^n$, alors le système $Ax = b$ admet une solution dans $\mathbb{K}^m$ si et seulement si il admet une solution dans $\mathbb{L}^m$. Quand cette condition est vérifiée, comparer les dimensions (comme espace affine) des espaces de solutions sur $\mathbb{K}$ et sur $\mathbb{L}$.
(2) Cette propriété est-elle encore vraie pour des équations polynomiales plus générales ?
(3) Application : montrer que si $A \in \mathrm{M}_n(\mathbb{K})$ les conditions suivantes sont équivalentes :
   (a) $A$ est diagonalisable sur $\mathbb{K}$ ;
   (b) $A$ est diagonalisable sur $\mathbb{L}$ et toutes ses valeurs propres (sur $\mathbb{L}$) appartiennent à $\mathbb{K}$.

*Indication* : On pourra interpréter l'existence d'une solution d'un système comme une condition sur le rang de certaines matrices.

4.2. **Décomposition $LU$.**

**Exercice 2.** (1) Montrer que si $k = \mathbb{R}$ ou $\mathbb{C}$, l'ensemble des matrices de $\mathrm{GL}_n(k)$ qui admettent une décomposition $LU$ est un ouvert dense de $\mathrm{GL}_n(k)$.

(2) En déduire que toute matrice $M$ de $\mathrm{GL}_n(k)$ peut s'écrire sous la forme $M_1 \cdot M_2 \cdot M_3$ avec $M_1$ et $M_3$ triangulaires inférieures et $M_2$ triangulaire supérieure.

*Indication* : pour (1), on pourra utiliser le fait que les zéros d'une application polynomiale sont isolés. Pour (2), on pourra utiliser qu'une intersection d'ouverts dense est non vide.

**Exercice 3.** Soit $M$ dans $\mathrm{M}_n(k)$, et notons $\Delta_l$ le déterminant de la matrice $(m_{i,j})_{1\leq i,j\leq l}$ pour tout $l \leq n$. Montrer que si $M$ admet une décomposition $LU$, alors la diagonale de la composante "$U$" est $\left(\Delta_1, \frac{\Delta_2}{\Delta_1}, \cdots, \frac{\Delta_n}{\Delta_{n-1}}\right)$.

Référence : [CG, Chap. IV, Ex. B.8].

**Exercice 4.** Soit $k$ un corps fini.

(1) Quelles valeurs peut prendre le nombre de solutions d'un système $Ax = b$ (avec $A \in \mathrm{M}_{n,m}(k)$ et $b \in k^n$) dans $k^m$ ?

(2) Quel est le nombre de matrices de $\mathrm{GL}_n(k)$ admettant une décomposition $LU$ ? En déduire la probabilité pour qu'une matrice arbitraire de $\mathrm{GL}_n(k)$ admette une décomposition $LU$.

(3) Déterminer un équivalent de cette probabilité quand $|k| \to +\infty$.

Référence : [CG, Chap. IV, Ex. B.9].

**Exercice 5.** Montrer que pour toute matrice $M$ de $\mathrm{M}_n(k)$, il existe une matrice de permutation $P$ telle que la matrice $PM$ admet une décomposition $LU$.
(*Indication* : on pourra raisonner par récurrence sur $n$.)

Référence : [CG, Chap. IV, Ex. B.11].

**Exercice 6.** (1) Montrer que si $A$ est une matrice symétrique dont tous les déterminants principaux sont non nuls, il existe un unique couple $(L, D)$ où $L$ est une matrice triangulaire inférieure avec des 1 sur la diagonale et $D$ une matrice diagonale tel que

$$A = L \cdot D \cdot {}^{\mathrm{t}}L.$$

(2) Donner des formules explicites permettant de calculer $D$ et $L$ par récurrence.

(3) En déduire un algorithme de calcul de la signature de $A$.

Référence : [Ro, Chap. 5, §8].

**Exercice 7** (Inégalité de Hadamard)**.** (1) Montrer que si $A \in \mathrm{M}_n(\mathbb{R})$ et si on note ses vecteurs colonne $X_1, \cdots, X_n$, alors

$$|\det(A)| \leq \prod_{i=1}^{n} \|X_i\|_2.$$

(*Indication* : on pourra se ramener au cas où $A$ est inversible, puis appliquer la décomposition de Cholesky à ${}^{\mathrm{t}}A \cdot A$ et considérer les coefficients diagonaux.)

(2) Montrer que si les $X_i$ sont tous non nuls, l'inégalité est une égalité si et seulement si $X_1, \cdots, X_n$ sont deux à deux orthogonaux.

(3) Application : Montrer que si $A \in \mathrm{M}_n(\mathbb{R})$ et si $|a_{i,j}| \leq c$ pour tous $i, j$, alors $|\det(A)| \leq c^n n^{n/2}$.

Référence : Pour une autre méthode de démonstration de (1) et (2) (basée sur le procédé d'orthonormalisation de Schmidt) voir [Go, Chap. 5, §3.3, Théorème 7]. Pour (3), voir [Go, Chap. 5, §3.5, exercice 1].

**Exercice 8.** Une matrice $M$ est dite *tridiagonale* si ses coefficients vérifient $m_{i,j} = 0$ si $|i-j| > 1$. Considérons une matrice $M$ tridiagonale, et notons ses coefficients sous-diagonaux, resp. diagonaux, resp. sur-diagonaux, $a_2, \cdots, a_n$, resp. $b_1, \cdots, b_n$, resp. $c_1, \cdots, c_{n-1}$.

(1) Pour $k \in \{0, \cdots, n\}$ on note $\delta_k$ le déterminant de la $k$-ème sous-matrice principale de $M$. Montrer que cette suite est déterminée par les conditions suivantes :

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \delta_k = b_k \delta_{k-1} - a_k c_{k-1} \delta_{k-2} \quad \text{si } k \geq 2.$$

(2) Calculer la décomposition $LU$ de $M$ en fonction des $\delta_k$ (en supposant que ces coefficients sont non nuls).

(3) En déduire un algorithme de résolution du système $Mx = b$.

Référence : [Ci, Théorème 4.3.2] ou [FGN2, Ex. 1.39].

**Exercice 9.** (1) Montrer que l'application envoyant $M \in \mathrm{GL}_n(\mathbb{C})$ sur $(Q, R)$, où $Q$ et $R$ sont les matrices apparaissant dans la décomposition $QR$ de $M$, est continue. En déduire un homéomorphisme entre $\mathrm{GL}_n(\mathbb{C})$ et

$$U_n(\mathbb{C}) \times \mathbb{C}^{\frac{n(n-1)}{2}} \times (\mathbb{R}_{>0})^n.$$

(2) Soit $M \in \mathrm{GL}_n(\mathbb{R})$ une matrice admettant une décomposition $LU$. Montrer que si $(M_j)_{j \geq 0}$ est une suite de matrices convergeant vers $M$, alors il existe $j_0$ tel que $M_j$ admet une décomposition $LU$ pour $j \geq j_0$, et que de plus si on note $M_j = L_j U_j$ la décomposition $LU$ de $M_j$ (pour $j \geq j_0$), alors les suites $(L_j)_{j \geq j_0}$ et $(U_j)_{j \geq j_0}$ convergent (vers des matrices qu'on notera $L$ et $U$) et que $M = LU$ est la décomposition $LU$ de $M$. En déduire qu'il existe un homéomorphisme entre $\mathbb{R}^{n(n-1)} \times (\mathbb{R} \smallsetminus \{0\})^n$ et un ouvert de $\mathrm{GL}_n(\mathbb{R})$ (qu'on explicitera).

Indications : pour (1) on pourra raisonner comme pour la décomposition polaire, c'est-à-dire utiliser le fait que dans un espace métrique compact, une suite converge si et seulement si elle n'admet qu'une seule valeur d'adhérence. Pour (2), on pourra raisonner par récurrence sur $n$ et utiliser les formules explicites pour le calcul de la décomposition $LU$, cf. par exemple [Se, §8.1].

### 4.3. **Opérations élémentaires sur les lignes et colonnes.**

**Exercice 10.** Soit $M \in \mathrm{M}_{m,n}(k)$. On appelle *pivot* d'une ligne non nulle de $M$ le coefficient non nul sur cette ligne situé dans la colonne la plus à gauche. La matrice $M$ est dite *échelonnée en lignes* si elle vérifie les conditions suivantes :

— si une ligne est nulle, toutes les suivantes sont nulles ;

— le pivot de chaque ligne non nulle est situé strictement plus à droite que les pivots de toutes les lignes précédentes.

Une telle matrice est dite *réduite* si chaque pivot vaut 1, et si les pivots sont les seuls coefficients non nuls de leur colonne.

Le *type* d'une matrice est la liste strictement croissante des indices de colonnes de pivot.

(1) Si $\mathbf{j} = (j_1, \cdots, j_r)$, montrer que l'ensemble des matrices échelonnées réduites en lignes de type $\mathbf{j}$ est un sous-espace affine de $\mathrm{M}_{m,n}(k)$, de dimension

$$\sum_{i=1}^{r}(n - r + i - j_i).$$

(2) Montrer que chaque orbite de $\mathrm{GL}_m(k)$ pour l'action sur $\mathrm{M}_{m,n}(k)$ donnée par $(P, M) \mapsto P \cdot M$ contient une et une seule matrice échelonnée réduite. (*Indication* : pour l'existence, on pourra utiliser le pivot de Gauss. Pour l'unicité, on pourra montrer par récurrence sur $r$ que pour tous $m, n \geq r$, si $E$ et $E'$ sont deux matrices échelonnées réduites de rang $r$ et $P$ une matrice inversible telle que $PE = E'$, alors $E = E'$ et $P$ est de la forme $\begin{pmatrix} I_r & A \\ 0 & B \end{pmatrix}$.)

(3) Si $K$ est une extension de $k$ et si $M_1, M_2 \in \mathrm{M}_{m,n}(k)$, montrer que $M_1$ et $M_2$ diffèrent par multiplication à gauche par une matrice de $\mathrm{GL}_n(k)$ si et seulement si elles diffèrent par multiplication à gauche par une matrice de $\mathrm{GL}_n(K)$.

Référence : [CG, Chap. IV, Théorème 2.3.1].

**Exercice 11.** (1) Montrer que deux matrices $A, B$ de $\mathrm{M}_{m,n}(k)$ se déduisent l'une l'autre par opérations élémentaires sur les lignes si et seulement si elles ont même noyau.

(2) Montrer que deux matrices $A, B$ de $\mathrm{M}_{m,n}(k)$ se déduisent l'une l'autre par opérations élémentaires sur les colonnes si et seulement si elles ont même image.

(3) Si $k$ est un corps fini, en déduire le nombre d'orbites de l'action de $\mathrm{GL}_m(k)$ sur l'ensemble des matrices de rang $r$ dans $\mathrm{M}_{m,n}(k)$ définie par $(P, M) \mapsto P \cdot M$.

(4) Même question pour l'action définie par $(P, M) \mapsto M \cdot P^{-1}$.

**Exercice 12.** Let but de cet exercice est de montrer (en utilisant des opérations élémentaires) que le groupe $\mathrm{SL}_2(\mathbb{Z})$ est engendré par les matrices

$$U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad V = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

On notera $G$ le sous-groupe de $\mathrm{SL}_2(\mathbb{Z})$ engendré par ces matrices.

(1) En calculant les puissances de $U$ et $V$, montrer que si une matrice $M$ appartient à $G$, alors pour tout $k \in \mathbb{Z}$ la matrice obtenue en effectuant l'opération $L_1 \leftarrow L_1 + kL_2$ appartient encore à $G$, et de même pour $L_2 \leftarrow L_2 + kL_1$.

(2) Montrer que si $M \in \mathrm{SL}_2(\mathbb{Z})$ il existe $g \in G$ telle que

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

avec $a \neq 0$ et $c \neq 0$, et qu'alors $a$ et $c$ sont premiers entre eux.

(3) Avec $M$ et $g$ comme dans la question précédente, et en considérant l'algorithme d'Euclide appliqué à $a$ et $c$, montrer qu'il existe $g_1, \cdots, g_r \in G$ telles que $g_1 \cdots g_r g M$ est de la forme

$$\begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} \quad \text{ou} \quad \begin{pmatrix} 0 & b \\ 1 & d \end{pmatrix}.$$

(4) Conclure.

Référence : [FGN2, Ex. 3.15].

4.4. **Méthodes itératives.**

**Exercice 13.** On dit qu'une matrice $A$ est *à diagonale strictement dominante* si pour tout $i$ on a

$$|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|.$$

Montrer que si $A$ est à diagonale strictement dominante, alors les méthodes de Jacobi et de Gauss–Seidel s'appliquent et convergent.

(*Indication* : Pour Jacobi on pourra consider la norme $\|\cdot\|_\infty$. Pour Gauss–Seidel on pourra étudier les valeurs propres de la matrice $M^{-1}N$.)

Référence : [Ro, p. 201–202].

## 5. Résolution effective de systèmes linéaires

Des références utiles pour cette leçon sont [Ci, Ro, Se]. La plupart des méthodes considérées ci-dessous ont des variantes "par blocs", qu'on ne considérera pas pour simplifier.

5.1. **Motivation.** Cette leçon est consacrée aux systèmes linéaires. Il est bien sûr nécessaire d'expliquer quelques résultats généraux sur ces systèmes, et de donner quelques "conséquences théoriques" ; cependant, le point central de la leçon est d'expliquer comment on résoud ces systèmes *en pratique*. Cette question est motivée notamment par des questions d'analyse numérique, comme la "méthode des différences finies", dans laquelle on remplace une équation différentielle par une succession de systèmes linéaires (faisant intervenir un grand nombre d'équations et d'inconnues). Pour mettre en œuvre cette méthode il faut donc savoir résoudre rapidement, souvent de façon approchée, des systèmes comportant un très grand nombre d'équations portant sur de nombreuses variables. Les matrices définissant ces équations ont souvent des formes assez spécifiques (notamment, elles sont "creuses" au sens où elles comportent beaucoup de 0), ce qui conduit parfois à faire des hypothèses permettant d'améliorer les techniques utilisées en les adaptant aux situations rencontrées en pratique.

Pour plus de détails sur ces choses-là, on pourra consulter [Ci, Chap. 3]. (Il est important de connaître ces motivations, et il peut être utile de le mentionner dans le plan si on a des choses intéressantes et précises à dire à ce sujet, mais ce n'est pas indispensable.)

Assez souvent, on se limite aux systèmes de la forme

$$Ax = b$$

avec $A \in \mathrm{GL}_n(k)$, $b \in k^n$ donnés, et $x$ l'inconnue cherchée. (C'est-à-dire qu'on suppose que la matrice définissant le système est carrée et inversible.) On peut donner plusieurs justifications pour cela :

(1) on peut toujours ramener un système quelconque à un système comme ci-dessus, quitte à "transformer" $b$ en un vecteur dépendant de certains paramètres qu'on peut choisir arbitrairement (c'est le contenu du Théorème de Rouché–Fontené) ;
(2) si $k = \mathbb{R}$ ou $\mathbb{C}$ (ce qui est le cas principalement étudié), $\mathrm{GL}_n(k)$ est dense dans $\mathrm{M}_n(k)$, ce qui peut se traduire concrètement en disant que "pratiquement toutes les matrices carrées sont inversibles".

Dans ce cadre l'équation a évidemment une solution unique, donnée par $x = A^{-1}b$. En fait, la question du calcul des solutions de $Ax = b$ est fortement liée à celle de $A^{-1}$. En effet, si on sait calculer $A^{-1}$ alors il suffit de poser $x = A^{-1}b$. Inversement, la $i$-ème colonne de $A^{-1}$ est la solution de l'équation $Ax = e_i$ où $e_i$ est le $i$-ème vecteur de la base canonique. Calculer $A^{-1}$ est donc équivalent à résoudre $n$ systèmes du type $Ax = b$.

5.2. **Méthodes directes.** Tout d'abord on considère des méthodes permettant de calculer la valeur exacte de la solution $x$.

5.2.1. *Méthode de Cramer.* Cette méthode est basée sur le résultat classique suivant.

**Théorème 1.** Soit $k$ un corps, et considérons $A \in \mathrm{GL}_n(k)$ et $b \in k^n$. Notons, pour tout $j \in \{1, \cdots, n\}$, $A_j$ la matrice obtenue à partir de $A$ en remplaçant sa $j$-ème colonne par $b$. Alors le vecteur $x = A^{-1}b$ a pour coordonnées :

$$x_j = \frac{\det(A_j)}{\det(A)}$$

pour $j \in \{1, \cdots, n\}$.

Ce théorème est équivalent à la formule exprimant l'inverse d'une matrice en termes de sa comatrice. Il peut avoir un intérêt théorique, mais est très peu efficace dans la pratique : si on utilise la formule habituelle pour le déterminant, le nombre d'opérations à effectuer pour calculer $x$ est de l'ordre de $n^2 n!$ (voir [Ro, p. 182]), ce qui n'est pas raisonnable dans les cas pratiques où $n$ est très grand.

5.2.2. *Méthode des pivots de Gauss.* Cette méthode est la plus classique des méthodes "raisonnables" pour résoudre un système. Elle s'adapte facilement au cas où $A$ n'est pas inversible, voire pas carrée. Il est indispensable de présenter cette méthode dans le plan, si possible de manière algorithmique (c'est-à-dire en donnant une "recette" qu'on peut suivre de façon mécanique quelle que soit la matrice donnée). Pour cela, on pourra consulter [Ro, §5.5] ou [Ci, §4.1]. Cette méthode nécessite de l'ordre de $n^3$ opérations pour résoudre un système carré de taille $n$.

Le principe de cette méthode utilise les opérations élémentaires sur les lignes (et éventuellement les colonnes), et est basée sur l'observation que si $M$ est une matrice inversible alors le système

$$Ax = b$$

est équivalent au système

$$(MA)x = Mb \ ;$$

on cherche donc à choisir $M$ de sorte que la matrice $MA$ soit "plus sympathique" que $A$.

Ces idées sont intimement liées au fait que le groupe $\mathrm{GL}_n(k)$ est engendré par les matrices de transvection et de dilatations (cf. par exemple [Ro, §5.4]). De façon plus

algébrique, on peut également voir cette méthode comme donnant une classification des orbites du groupe $\mathrm{GL}_n(k)$ pour l'action par multiplication à gauche sur $\mathrm{M}_n(k)$ (ou plus généralement sur $\mathrm{M}_{n,m}(k)$ si on veut). Pour ce point de vue, on pourra consulter [CG, §IV.2].

L'avantage principal de cette méthode est qu'elle est complètement générale (pas d'hypothèse sur le corps, ni sur la matrice). Mais pour cette raison elle ne prend en compte aucune des spécificités de la matrice donnée, et peut donc être améliorée si on accepte de faire des hypothèses sur $A$.

**Remarque.** La méthode du pivot permet de se ramener à la fin à un système triangulaire, qu'on résoud "par remontée" (cf. paragraphe suivant). Une variante, appelée "méthode de Gauss–Jordan" permet de se ramener à un système diagonal : voir [Ro, §5.10].

5.2.3. *Cas particuliers.* Les méthodes qui suivent sont basées sur l'observation qu'il est très facile de résoudre un système $Ax = b$ dans les cas suivants :

(1) si $A$ est triangulaire ;
(2) si $k = \mathbb{R}$ (resp. $\mathbb{C}$) et $A$ est orthogonale (resp. unitaire).

En effet, si $A$ est triangulaire supérieure la $i$-ème équation s'écrit

$$a_{i,i}x_i + a_{i,i+1}x_{i+1} + \cdots + a_{i,n}x_n = b_i,$$

avec $a_{i,i} \neq 0$. On peut alors résoudre le système "par remontée" en posant

$$x_n = \frac{b_n}{a_{n,n}}$$

puis, une fois que $x_{i+1}, \cdots, x_n$ ont été calculés, en posant

$$x_i = \frac{1}{a_{i,i}} \left( b_i - \sum_{j=i+1}^{n} a_{i,j}x_j \right).$$

Ce procédé nécessite de l'ordre de $n^2$ opérations (voir [Ci, p. 72] pour plus de détails).

Bien sûr, le cas d'une matrice triangulaire inférieure est similaire, en "descendant" plutôt qu'en remontant.

Enfin, si $A$ est orthogonale, resp. unitaire, alors on a $A^{-1} = {}^{\mathrm{t}}A$, resp. $A^{-1} = {}^{\mathrm{t}}\overline{A}$ ; on sait donc inverser $A$ sans effort, et ensuite résoudre tout système du type $Ax = b$.

5.2.4. *Décomposition LU.* L'idée de cette méthode est la suivante : supposons que $A = A_1A_2$ avec $A_1$ triangulaire inférieure et $A_2$ triangulaire supérieure. (Ces 2 matrices sont nécessairement inversibles si $A$ l'est.) Alors pour résoudre le système

$$Ax = b$$

il suffit de d'abord résoudre le système

$$A_1y = b$$

puis, une fois $y$ connu, de résoudre le système

$$A_2x = y.$$

On a vu ci-dessus que chacune de ces étapes est facile (et rapide) ; dans ce cas on peut donc aussi résoudre notre système aisément.

Malheureusement, toute matrice inversible ne peut pas nécessairement s'écrire sous la forme $A = A_1A_2$ comme ci-dessus. Mais on sait précisément lesquelles peuvent s'écrire de cette manière.

**Théorème 2.** Soit $A \in \mathrm{GL}_n(k)$. Alors il existe des matrices $A_1$ triangulaire inférieure et $A_2$ triangulaire supérieure telles que $A = A_1A_2$ si et seulement si chacune des matrices $(a_{ij})_{1\leq i,j\leq m}$ est inversible (pour $m \in \{1, \cdots, n-1\}$). De plus, si cette condition est vérifiée, l'écriture $A = A_1A_2$ est unique si on impose de plus que les coefficients diagonaux de $A_1$ sont tous égaux à 1.

Une décomposition comme dans ce théorème est appelée "décomposition $LU$". Il n'est pas difficile de voir que la condition du théorème est vérifiée pour "presque toute" matrice inversible (voir notamment l'exercice 2). Cette méthode peut donc souvent s'appliquer en pratique.

Le fait que si $A$ admet une décomposition $LU$ alors les matrices $(a_{ij})_{1\leq i,j\leq m}$ sont inversibles est très facile à voir, puisque chacune de ces matrices admet une décomposition en produit similaire. L'implication réciproque peut se démontrer (au moins) de 2 façons intéressantes :

(1) en remarquant que dans ce cas, quand on applique l'algorithme de Gauss à la résolution du système $Ax = b$, alors il n'y a aucune permutation de lignes à faire : voir [Ro, §5.7] ou [Ci, §4.3] ;

(2) par récurrence sur $n$, ce qui fournit une méthode algorithmique pratique pour ce calcul : voir [Se, §8.1] ou [Ro, p. 195] (où ce procédé est appelé "méthode de Crout").

En particulier, en suivant la méthode fournie par la 2ème preuve on voit que le calcul effectif de la décomposition $LU$ d'une matrice de taille $n$ peut se faire en $\frac{2}{3}n^3$ opérations environ. Ce nombre d'opérations est comparable à celui nécessaire à l'application du pivot de Gauss ; il est donc surtout utile quand on veut résoudre des systèmes $Ax = b$ pour une même matrice $A$ et de nombreuses valeurs pour $b$. (En effet, une fois la décomposition $LU$ trouvée, elle peut être utilisée pour résoudre le système pour tout choix de $b$.)

Pour une description explicite du calcul de la décomposition $LU$ dans le cas des matrices tridiagonales, on pourra consulter [Ci, Théorème 4.3.2]. Dans ce cas le nombre d'opérations à effectuer est linéaire en $n$, et donc très économique.

5.2.5. *Matrices symétriques et décomposition de Cholesky.* Supposons qu'une matrice symétrique inversible $A$ admette une décomposition $LU$. En factorisant les coefficients diagonaux de la deuxième matrice, on peut alors écrire

$$A = B \cdot D \cdot C$$

avec $B$ (resp. $C$) triangulaire inférieure (resp. supérieure) avec des 1 sur la diagonale, et $D$ diagonale (à coefficients non nuls). On a alors

$$A = {}^tA = {}^tC \cdot D \cdot {}^tB.$$

Par unicité de la décomposition $LU$, on doit alors avoir $C = {}^tB$, et donc

$$A = B \cdot D \cdot {}^tB.$$

En particulier, si $k = \mathbb{R}$ et si $A$ est symétrique définie positive, alors chaque $(a_{ij})_{1\leq i,j\leq m}$ l'est également, donc est inversible ; on en déduit que $A$ admet une décomposition $LU$. De plus, chacun des coefficients de $D$ est alors strictement

positif (voir par exemple l'exercice 3). On peut donc écrire $D = (D')^2$ avec $D'$ une matrice diagonale, ce qui permet d'écrire

$$A = B \cdot D' \cdot D' \cdot {}^{\mathrm{t}}B.$$

On a donc (essentiellement) démontré le théorème suivant.

**Théorème 3.** Si $A \in \mathrm{M}_n(\mathbb{R})$, alors $A$ est symétrique définie positive si et seulement si il existe une matrice $B$ triangulaire inférieure inversible telle que $A = B \cdot {}^{\mathrm{t}}B$. De plus, une telle écriture est unique si on impose que les coefficients diagonaux de $B$ soient strictement positifs.

La décomposition $A = B \cdot {}^{\mathrm{t}}B$ comme ci-dessus s'appelle *décomposition de Cholesky*. Pour plus de détails sur le calcul effectif de $B$, et le nombre d'opérations nécessaire, on pourra consulter [Ro, §5.9] ou [Ci, §4.4].

5.2.6. *Décomposition $QR$.* Les décompositions $LU$ et de Cholesky tirent leur utilité du fait que les systèmes d'équations triangulaires sont faciles à résoudre. Comme expliqué au §5.2.3, il y a une autre situation où le système est facile à résoudre : c'est celui où $A$ est unitaire (dans le cas $k = \mathbb{C}$). En suivant le même principe que pour la décomposition $LU$, on est donc amené à considérer des décompositions comme produit d'une matrice *unitaire* et d'une matrice triangulaire. Ceci est toujours possible, comme montré par l'énoncé suivant.

**Proposition 1.** Pour tout $M$ dans $\mathrm{GL}_n(\mathbb{C})$, il existe une matrice unitaire $Q$ et une matrice triangulaire supérieure $R$ telles que $M = QR$. De plus, cette factorisation est unique si on impose que les coefficients diagonaux de $R$ sont tous strictement positifs.

Comme d'habitude, l'unicité de la décomposition est facile à voir (en utilisant le fait qu'une matrice unitaire est diagonalisable). L'existence peut se prouver en utilisant la factorisation de Cholesky (dans sa variante hermitienne) : on écrit

$${}^{\mathrm{t}}\overline{M} \cdot M = {}^{\mathrm{t}}\overline{R} \cdot R$$

avec $R$ triangulaire supérieure à diagonale réelle positive, et on vérifie que $Q := MR^{-1}$ est alors unitaire. (Voir [Se, §8.3] pour les détails.)

Dans la pratique, il est préférable d'utiliser le procédé d'orthonormalisation de Gramm–Schmidt sur les colonnes de $M$ : voir [Se, p. 97–98]. Ce procédé admet une variante qui peut se décrire de façon matricielle en termes de "matrices de Householder" : voir [Ci, §4.5] pour ce point de vue. D'après [Ci, p. 94], cette variante est meilleure car elle évite des propagations d'erreurs d'arrondis.

5.3. **Méthodes itératives.** Dans la théorie, les méthodes directes de résolution de systèmes fournissent une solution exacte. Cependant, en pratique, les ordinateurs ne savent calculer qu'en faisant des arrondis ; cette précision est donc illusoire, et il est souvent suffisant de calculer une valeur approchée de la solution, ce qui peut se faire pour un coût moindre en utilisant une "méthode itérative".

5.3.1. *Principe.* On considère toujours le système $Ax = b$, avec $k = \mathbb{C}$ et $A$ inversible. Supposons qu'on puisse écrire $A = M - N$ avec $M, N \in \mathrm{M}_n(\mathbb{C})$ et $M$ inversible. Alors on a

$$Ax = b \quad \Leftrightarrow \quad Mx - Nx = b \quad \Leftrightarrow \quad x = M^{-1}Nx + M^{-1}b.$$

Ainsi, $x$ est l'unique point fixe de la fonction $x \mapsto M^{-1}Nx + M^{-1}b$. Suivant la méthode traditionnelle de calcul (approché) des points fixes, on est conduit à choisir $x^{(0)}$ dans $\mathbb{R}^n$, puis à poser

$$x^{(j+1)} = M^{-1}Nx^{(j)} + M^{-1}b.$$

Si la suite $(x^{(j)})_{j\geq 0}$ converge, alors sa limite sera un point fixe de la fonction considérée ci-dessus, et donc égale à $x$. Si c'est le cas pour tout choix de $x^{(0)}$, on dira que la méthode itérative (associée au choix de $M$ et $N$) est *convergente*.

5.3.2. *Étude théorique.* Considérons une suite $(x^{(j)})_{j\geq 0}$ comme ci-dessus ; alors il est facile de voir que pour tout $j \geq 0$ on a

$$x^{(j+1)} - x = M^{-1}N(x^{(j)} - x),$$

et donc

$$x^{(j)} - x = (M^{-1}N)^j(x^{(0)} - x).$$

On est donc amené à étudier la convergence vers 0 de suites du type $(B^j v)_{j\geq 0}$ pour une matrice $B$ et un vecteur $v$ donnés. Cette question est classique, et la réponse donnée par l'énoncé suivant (où on note $\rho(B)$ le rayon spectral de $B$, c'est-à-dire le maximum des modules des valeurs propres de $B$).

**Théorème 4.** Soit $B \in \mathrm{M}_n(\mathbb{C})$. Les conditions suivantes sont équivalentes :

(1) la suite $(B^j)_{j\geq 0}$ converge vers 0 dans $\mathrm{M}_n(\mathbb{C})$ ;

(2) pour tout $v$ dans $\mathbb{C}^n$, la suite $(B^j v)_{j\geq 0}$ converge vers 0 dans $\mathbb{C}^n$ ;

(3) il existe une norme matricielle [1] $\|\cdot\|$ sur $\mathrm{M}_n(\mathbb{C})$ telle que $\|B\| < 1$ ;

(4) $\rho(B) < 1$.

Pour une preuve, voir [Ci, Théorème 1.5.1] ou [Ro, Théorème 4.16, p. 140–141].

En appliquant ce théorème on voit que la méthode itérative est convergente si et seulement si $\rho(M^{-1}N) < 1$.

5.3.3. *Méthode de Jacobi.* Pour cette méthode on suppose que les coefficients diagonaux de $A$ sont non nuls. On choisit alors pour $M$ la matrice diagonale dont les coefficients sont $a_{1,1}, \cdots, a_{n,n}$, et on pose $N = M - A$. En appliquant le processus itératif, on est amenés à poser

$$x_i^{(j+1)} = \frac{1}{a_{i,i}} \left( b_i - \sum_{m\neq i} a_{i,m} x_m^{(j)} \right).$$

En ce qui concerne la convergence, on a le résultat suivant. Ici, on dit qu'une matrice $A$ est *à diagonale strictement dominante* si pour tout $i$ on a

$$|a_{i,i}| > \sum_{j\neq i} |a_{i,j}|.$$

**Proposition 2.** Si $A$ est à diagonale strictement dominante, alors la méthode de Jacobi s'applique et converge.

Pour une preuve, voir [Ro, Théorème 5.6, p. 201]. (Il suffit en fait de vérifier que $\|M^{-1}N\|_\infty < 1$.)

1. Par norme *matricielle*, on entend une norme $\|\cdot\|$ vérifiant $\|MN\| \leq \|M\| \cdot \|N\|$ pour tous $M, N$ dans $\mathrm{M}_n(\mathbb{C})$.

5.3.4. *Méthode de Gauss–Seidel.* Pour cette méthode on suppose encore que les coefficients diagonaux de $A$ sont non nuls, mais on prend cette fois-ci pour $M$ la matrice telle que

$$m_{i,j} = \begin{cases} a_{i,j} & \text{si } j \leq i\,; \\ 0 & \text{sinon} \end{cases}$$

(c'est-à-dire que $M$ est le "triangle inférieur" de $A$), et on pose $N = M - A$. Dans ce procédé, on est amené à poser

$$x_i^{(j+1)} = \frac{1}{a_{i,i}} \left( b_i - \sum_{m=1}^{i-1} a_{i,m} x_m^{(j+1)} - \sum_{m=i+1}^{n} a_{i,m} x_m^{(j)} \right).$$

Par rapport à la méthode de Jacobi, cela revient à "remplacer les $x_m^{(j)}$ par $x_m^{(j+1)}$ si $m < i$". En particulier, on peut oublier $x_m^{(j)}$ une fois que $x_m^{(j+1)}$ a été calculé ; cette méthode nécessite donc moins de mémoire que celle de Jacobi.

La convergence est garantie dans les cas suivants.

**Proposition 3.** Si $A$ est à diagonale strictement dominante, ou si $A$ est hermitienne définie positive, alors la méthode de Gauss–Seidel s'applique et converge.

La preuve du premier cas se trouve dans [Ro, Théorème 5.7, p. 202] ou [Se, Proposition 9.3.1] (dans un cadre plus général). Celle du deuxième cas se trouve dans [Ro, Théorème 5.8, p. 202–203], [Se, Théorème 9.3.1] ou [Ci, Théorème 5.3.2] (comme cas particulier d'un énoncé plus général).

5.3.5. *Méthode de relaxation.* Encore une fois on suppose que les coefficients diagonaux de $A$ sont non nuls, et on écrit $A = D + E + F$ avec $D$ la partie diagonale de $A$, $E$ sa partie triangulaire inférieure stricte, et $F$ sa partie triangulaire supérieure stricte. Pour un paramètre $\omega \in \mathbb{C}^\times$, on pose

$$M_\omega := \frac{1}{\omega} D + E, \qquad N_\omega = M_\omega - A = \left( \frac{1}{\omega} - 1 \right) D - F.$$

La méthode de Gauss–Seidel correspond au cas $\omega = 1$ ; ici on va chercher à faire un "meilleur" choix pour $\omega$.

Les formules nécessaires pour calculer $x^{(j+1)}$ en fonction de $x^{(j)}$ sont similaires aux cas précédents, cf. [Se, §9.2.3], [Ro, p. 204] ou [Ci, p. 101].

En ce qui concerne la convergence, on a déjà la condition nécessaire suivante (voir [Ro, Théorème 5.9, p. 204], [Se, Proposition 9.2.1] ou [Ci, Théorème 5.3.3]).

**Lemme 1.** Si la méthode de relaxation converge, alors $|\omega - 1| < 1$.

Pour des conditions suffisantes, il faut faire des hypothèses supplémentaires sur $A$.

**Proposition 4.** Si $A$ est à diagonale strictement dominante, alors la méthode de relaxation s'applique pour tout $\omega$, et converge pour tout $\omega \in ]0, 1]$.

Pour une preuve, voir [Se, Proposition 9.3.1].

**Proposition 5.** Si $A$ est hermitienne définie positive, alors la méthode de relaxation s'applique pour tout $\omega$, et converge dès que $|\omega - 1| < 1$.

Pour une preuve, voir [Se, Théorème 9.3.1], [Ro, Théorème 5.10, p. 205] ou [Ci, Théorème 5.3.2].

5.3.6. *Le cas des matrices tridiagonales.* On peut étudier les méthodes ci-dessus de façon très fine dans le cas où la matrice $A$ est tridiagonale (c'est-à-dire qu'elle vérifie $a_{i,j} = 0$ dès que $|i - j| \geq 2$), un cas qui se rencontre effectivement dans des problèmes concrets. Ceci est expliqué dans [Ro, p. 206–209], [Se, §9.4] et [Ci, p. 105–109]. On obtient notamment l'énoncé suivant.

**Théorème 5.** Supposons que $A$ est hermitienne définie positive, et tridiagonale. Alors la méthode de relaxation converge pour tout $\omega \in ]0, 2[$, et le paramètre optimal (c'est-à-dire tel que le rayon spectral de la matrice $M_\omega^{-1} N_\omega$ est minimal) est obtenu pour

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(J)^2}},$$

où $J = D^{-1}(-E - F)$.

5.4. **Méthodes de gradient.** Pour cette partie je recommande de consulter [Ro] (plutôt que [Se] ou [Ci], qui traitent ces choses de façon moins convaincante à mon goût).

5.4.1. *Reformulation du problème en terme de minimisation.* On s'intéresse à la résolution du système $Ax = b$ sous l'hypothèse où $A$ est symétrique définie positive. On notera $u$ l'unique solution ; le but est donc de calculer $u$. On note $\langle -, - \rangle$ le produit scalaire standard sur $\mathbb{R}^n$, et on définit $\varphi : \mathbb{R}^n \to \mathbb{R}$ en posant

$$\varphi(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle.$$

L'observation fondamentale est la suivante.

**Lemme 2.** La fonction $\varphi$ admet un unique minimum, obtenu pour $x = u$.

*Démonstration.* On calcule :

$$\varphi(x) - \varphi(u) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle - \frac{1}{2}\langle Au, u\rangle + \langle b, u\rangle.$$

Puisque $Au = b$, ceci se réécrit

$$\begin{aligned}
\varphi(x) - \varphi(u) &= \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + \frac{1}{2}\langle b, u\rangle \\
&= \frac{1}{2}\langle Ax, x\rangle - \frac{1}{2}\langle Au, x\rangle - \frac{1}{2}\langle Au, x\rangle + \frac{1}{2}\langle Au, u\rangle \\
&= \frac{1}{2}\langle A(x-u), x\rangle - \frac{1}{2}\langle Au, x - u\rangle \\
&= \frac{1}{2}\langle A(x-u), x\rangle - \frac{1}{2}\langle u, A(x-u)\rangle = \frac{1}{2}\langle A(x-u), x-u\rangle
\end{aligned}$$

où, dans la 4ème égalité, on a utilisé la symétrie de $A$. Maintenant, puisque $A$ est définie positive, on a $\langle A(x-u), x-u\rangle \geq 0$, avec égalité si et seulement si $x - u = 0$, ou en d'autres termes $x = u$. Ce qui prouve l'assertion voulue. □

Ce lemme montre que le problème de la résolution du système $Ax = b$ est équivalent à celui de la minimisation de la fonction $\varphi$. Il existe des techniques spécifiques pour ce genre de problème, notamment les algorithme de descentes, qui vont donc fournir de nouvelles méthodes pour résoudre les systèmes linéaires définis par une matrice symétrique définie positive.

5.4.2. *Principe.* Pour minimiser $\varphi$, on va procéder de la manière suivante : on va construire une suite $(x_j)_{j\geq 0}$ d'éléments de $\mathbb{R}^n$ de sorte que la suite $(\varphi(x_j))_{j\geq 0}$ soit décroissante. Cette construction utilisera une suite $(\delta_j)_{j\geq 0}$ de vecteurs non nuls, qui serviront de "directions de descente" successives. (Le choix de cette suite sera discuté plus tard ; il y en a plusieurs qui sont intéressants.) On fixe un choix arbitraire pour $x_0$. Puis, une fois $x_j$ fixé, on va minimiser la fonction $\varphi$ sur la droite $x_j + \mathbb{R}\delta_j$. Pour cela, on considère la fonction de $\mathbb{R}$ dans $\mathbb{R}$ donnée par

$$\begin{aligned} t \mapsto \varphi(x_j + t\delta_j) &= \frac{1}{2}\langle A(x_j + t\delta_j), x_j + t\delta_j\rangle - \langle b, x_j + t\delta_j\rangle \\ &= \frac{1}{2}\langle A\delta_j, \delta_j\rangle \cdot t^2 + \langle \delta_j, Ax_j - b\rangle \cdot t + \langle \frac{1}{2}Ax_j - b, x_j\rangle. \end{aligned}$$

Cette fonction est un polynôme du second degré, à coefficient dominant strictement positif (puisque $A$ est définie positive) ; elle atteint donc son minimum pour une unique valeur de $t$, donnée par

$$t_j := -\frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle}.$$

On est ainsi conduit à poser

$$x_{j+1} = x_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle}\delta_j.$$

5.4.3. *Un cas particulier.* La méthode de Gauss–Seidel peut être vue comme un cas particulier de la méthode décrite ci-dessus, quand on prend pour les vecteurs $(\delta_j)_{j\geq 0}$ les vecteurs de la base canonique, de façon cyclique ; voir [Ro, p. 215] pour plus de détails.

5.4.4. *Un critère général de convergence.* Il reste maintenant à comprendre si (et comment) on peut choisir les vecteur $(\delta_j)_{j\geq 0}$ de sorte que la suite $(x_j)_{j\geq 0}$ converge vers $u$ (pour n'importe quel choix de $x_0$).

**Théorème 6.** Supposons qu'il existe $\alpha > 0$ tel que

$$\langle Ax_j - b, \delta_j\rangle \geq \alpha \|Ax_j - b\| \cdot \|\delta_j\|$$

pour tout $j \geq 0$. Alors la suite $(x_j)_{j\geq 0}$ converge vers $u$.

La preuve utilisera le lemme suivant (pour lequel on remarque que toutes les valeurs propres de $A$ sont strictement positives, puisqu'elle est définie positive).

**Lemme 3.** Notons $\lambda_1$ la valeur propre minimale de $A$. Alors pour tout $x \in \mathbb{R}^n$ on a

$$\langle Ax, x\rangle \geq \lambda_1 \|x\|^2.$$

*Démonstration.* Puisque $A$ est symétrique, elle est diagonalisable en base orthonormée. Si on note $(f_1, \cdots, f_n)$ une telle base diagonalisant $A$, telle que $f_1$ a pour valeur propre associée $\lambda_1$, et si on note $\lambda_i$ la valeur propre associée à $f_i$ pour $i \in \{2, \cdots, n\}$, alors si $x = \sum_i \alpha_i f_i$ on a

$$\langle Ax, x\rangle = \left\langle \sum_i \alpha_i\lambda_i f_i, \sum_i \alpha_i f_i \right\rangle = \sum_i \alpha_i^2\lambda_i \geq \lambda_1 \sum_i \alpha_i^2 = \lambda_1\|x\|^2,$$

comme annoncé. □

*Preuve du Théorème 6.* Si on a $x_j = u$ pour un entier $j \geq 0$, alors il est facile de voir que $x_l = u$ pour tout $l \geq j$. Le résultat est donc clair dans ce cas. Dans la suite, on suppose que $x_j \neq u$ pour tout $j$. On pose alors

$$e_j := \langle A(x_j - u), x_j - u\rangle,$$

qui est un réel strictement positif pour tout $j \geq 0$ d'après notre hypothèse. D'après le Lemme 3 on a $e_j \geq \lambda_1 \|x_j - u\|^2$ pour tout $j$, de sorte que pour conclure il suffit de montrer que la suite réelle $(e_j)_{j\geq 0}$ tend vers 0.

Remarquons que pour tout $j \geq 0$ on a

$$\langle Ax_{j+1} - b, \delta_j\rangle = 0. \tag{1}$$

En effet $Ax_{j+1} = Ax_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle} A\delta_j$, de sorte que

$$\begin{aligned}\langle Ax_{j+1} - b, \delta_j\rangle = \langle Ax_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle} A\delta_j - b, \delta_j\rangle \\ = \langle Ax_j - b, \delta_j\rangle - \langle \delta_j, Ax_j - b\rangle = 0.\end{aligned}$$

Cette égalité implique que

$$e_{j+1} = \langle A(x_{j+1} - u), x_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle}\delta_j - u\rangle = \langle A(x_{j+1} - u), x_j - u\rangle.$$

En réutilisant l'égalité $Ax_{j+1} = Ax_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle} A\delta_j$, on trouve ensuite que

$$e_{j+1} = e_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle}\langle A\delta_j, x_j - u\rangle.$$

La symétrie de $A$ permet de réécrire cette égalité sous la forme suivante :

$$e_{j+1} = e_j - \frac{\langle \delta_j, Ax_j - b\rangle}{\langle A\delta_j, \delta_j\rangle}\langle \delta_j, Ax_j - b\rangle = e_j\left(1 - \frac{\langle \delta_j, Ax_j - b\rangle^2}{e_j \cdot \langle A\delta_j, \delta_j\rangle}\right).$$

Maintenant, par Cauchy–Schwarz on a

$$e_j = \langle A(x_j - u), A^{-1}(A(x_j - u))\rangle \leq \|A^{-1}\|_2 \|A(x_j - u)\|^2, \quad \langle A\delta_j, \delta_j\rangle \leq \|A\|_2 \|\delta_j\|^2,$$

et par hypothèse on a

$$\langle Ax_j - b, \delta_j\rangle \geq \alpha \|A(x_j - u)\| \cdot \|\delta_j\|.$$

Il s'ensuit que

$$e_{j+1} \leq e_j\left(1 - \frac{\alpha^2}{\|A^{-1}\|_2 \cdot \|A\|_2}\right). \tag{2}$$

Puisque $e_j$ et $e_{j+1}$ sont strictement positifs, le coefficient $1 - \frac{\alpha^2}{\|A^{-1}\|_2 \cdot \|A\|_2}$ est strictement positif. Par une récurrence immédiate on obtient donc que

$$e_j \leq \left(1 - \frac{\alpha^2}{\|A^{-1}\|_2 \cdot \|A\|_2}\right)^j e_0$$

pour tout $j \geq 0$. Maintenant par définition ce coefficient est strictement inférieur à 1 ; on en déduit comme désiré que $e_j \xrightarrow[j\to+\infty]{} 0$. □

**Remarque.** Le réel $\frac{1}{\|A^{-1}\|_2 \cdot \|A\|_2}$ est appelé le *conditionnement* de la matrice $A$ par rapport à la norme $\|\cdot\|_2$, et noté $\mathrm{cond}_2(A)$ ; voir [Ro, Définition 4.5, p. 148]. C'est ce coefficient qui "gouverne" la vitesse de convergence, puisqu'on a

$$e_j \leq (1 - \alpha^2 \mathrm{cond}_2(A))^j e_0.$$

Pour améliorer la convergence on a intérêt à faire en sorte que ce conditionnement soit aussi grand que possible ; c'est le point de départ des méthodes de "préconditionnement", voir [Se, p. 114]. (Notons que dans [Se] le conditionnement est défini comme étant égal à $\|A^{-1}\|_2 \cdot \|A\|_2$.)

5.4.5. *Méthode du gradient à pas optimal.* La méthode du *gradient à pas optimal* consiste à choisir $\delta_j = Ax_j - b$ tant que $Ax_j - b \neq 0$. L'hypothèse du Théorème 6 est alors vérifiée pour $\alpha = 1$, de sorte que la convergence est garantie, avec

$$\|x_j - u\| \leq \gamma(1 - \mathrm{cond}_2(A))^{j/2}$$

pour une certaine constante $\gamma$ (dépendant de la matrice et du choix de vecteur initial).

**Remarque.** Dans le nom "gradient à pas optimal", le terme "gradient" fait référence au fait que la direction de $\delta_j$ est celle que $Ax_j - b$, qui est égal au gradient $\nabla\varphi(x_j)$. (C'est la "direction de décroissance maximale locale" pour $\varphi$.) Le terme "pas optimal" fait référence au fait que le choix du coefficient $t_j$ est fait de sorte à minimiser la fonction sur la droite considérée. (Ce genre de méthode de minimisation s'applique à des fonctionnelles plus compliquées, pour lesquelles le problème de la minimisation sur une droite peut être plus difficile à résoudre ; dans ce cas il peut être judicieux de choisir d'autres "pas" plus facile à calculer.)

5.4.6. *Méthode du gradient conjugué.* Pour la méthode du *gradient conjugué*, on part d'un vecteur $\delta_0$ arbitraire, puis tant que $Ax_j - b \neq 0$ on choisit le vecteur $\delta_j$ de la forme

$$\delta_j = (Ax_j - b) + s_j \delta_{j-1},$$

en choisissant $s_j$ de sorte à minimiser le coefficient

$$1 - \frac{\langle \delta_j, Ax_j - b\rangle^2}{e_j \cdot \langle A\delta_j, \delta_j\rangle}$$

apparaissant dans la preuve du Théorème 6. En fait, comme expliqué dans cette preuve (voir (1)) on a

$$\langle Ax_j - b, \delta_{j-1}\rangle = 0,$$

et $e_j$ ne dépend pas de $\delta_j$ ; il s'agit donc de minimiser la quantité

$$\langle A\delta_j, \delta_j\rangle = \langle A\delta_{j-1}, \delta_{j-1}\rangle (s_j)^2 + 2\langle Ax_j - b, A\delta_{j-1}\rangle s_j + \langle A(Ax_j - b), Ax_j - b\rangle,$$

ce qui force à choisir

$$s_j = -\frac{\langle Ax_j - b, A\delta_{j-1}\rangle}{\langle A\delta_{j-1}, \delta_{j-1}\rangle}.$$

Ici, l'orthogonalité de $Ax_j - b$ et $\delta_{j-1}$ montre que le vecteur $\delta_j$ est nécessairement non nul ; il peut donc effectivement servir de direction de descente. D'autre part elle assure que

$$\langle Ax_j - b, \delta_j\rangle = \|Ax_j - b\|^2.$$

On a

$$\|\delta_j\|^2 = \|Ax_j - b\|^2 + s_j^2\|\delta_{j-1}\|^2 = \|Ax_j - b\|^2 + \frac{\langle Ax_j - b, A\delta_{j-1}\rangle^2}{\langle A\delta_{j-1}, \delta_{j-1}\rangle^2}\|\delta_{j-1}\|^2.$$

Par Cauchy–Schwarz on a

$$\langle Ax_j - b, A\delta_{j-1}\rangle \leq \|A\|_2 \cdot \|Ax_j - b\| \cdot \|\delta_{j-1}\|,$$

et d'après le Lemme 3 on a

$$\langle A\delta_{j-1}, \delta_{j-1}\rangle \geq \lambda_1\|\delta_{j-1}\|^2$$

où $\lambda_1$ est la valeur propre minimale de $A$. On en déduit que

$$\frac{\langle Ax_j - b, A\delta_{j-1}\rangle^2}{\langle A\delta_{j-1}, \delta_{j-1}\rangle^2} \leq \frac{\|A\|_2^2 \cdot \|Ax_j - b\|^2}{\lambda_1^2 \cdot \|\delta_{j-1}\|^2},$$

et donc que

$$\|\delta_j\|^2 \leq \|Ax_j - b\|^2 \cdot \left(1 + \frac{\|A\|_2^2}{\lambda_1^2}\right),$$

et finalement que

$$\langle Ax_j - b, \delta_j\rangle \geq \frac{1}{\sqrt{1 + \frac{\|A\|_2^2}{\lambda_1^2}}} \cdot \|Ax_j - b\| \cdot \|\delta_j\|.$$

Le Théorème 6 assure donc ici aussi la convergence de la méthode. Mais on peut en fait faire beaucoup mieux : comme expliqué dans [Ro, Lemme 5.12, p. 220] on a le résultat suivant.

**Proposition 6.** Supposons que $\delta_0 = Ax_0 - b$. Alors tant que la méthode s'applique (c'est-à-dire tant que $Ax_j - b$ est non nul), les vecteurs $(Ax_l - b : l \in \{0, \cdots, j\})$ sont deux à deux orthogonaux (pour le produit scalaire standard), et les vecteurs $(\delta_l : l \in \{0, \cdots, j\})$ sont deux à deux orthogonaux pour le produit scalaire $(x, y) \mapsto \langle Ax, y\rangle$.

Puisque toute famille de vecteurs deux à deux orthogonaux dans $\mathbb{R}^n$ est de cardinal au plus $n$, on en déduit que si on part de $\delta_0 = Ax_0 - b$ la méthode converge au plus tard pour $j = n - 1$. Ceci montre que la méthode du gradient conjugué est une méthode *exacte* de résolution de système linéaire. Cependant, en raison des erreurs d'arrondis (ou pour limiter le nombre d'itérations) elle peut également être utilisée comme méthode itérative.

## 6. Autres ressources sur cette leçon

6.1. **Fiches mises à disposition par des collègues.** Pour la décomposition de Bruhat, on pourra consulter le document suivant :

`https://perso.univ-rennes1.fr/matthieu.romagny/agreg/theme/decomposition_de_Bruhat.pdf`

6.2. **Sujets d'écrit en rapport avec la leçon.** La partie 5 du sujet MG 2014 porte sur une analyse fine de la décomposition de Bruhat.

## Références

[CG] P. Caldero et J. Germoni, *Nouvelles histoires hédonistes de groupes et de géométries, Tome I*, Calvage & Mounet, 2017.

[Ci] P. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, 1985.

[FGN2] S. Francinou, H. Gianella, S. Nicolas, *Oraux X-ENS, algèbre 2*, Cassini, 2009.

[Go] X. Gourdon, *Les maths en tête - Algèbre, 2ème édition*, Ellipses, 2009.

[Ro] J.-É. Rombaldi, *Analyse matricielle - Cours et exercices résolus*, EDP Sciences, 1999.

[Se] D. Serre, *Les matrices, théorie et pratique*, Dunod, 2001.