ISIMA Première Année

COURS D'ANALYSE NUMÉRIQUE

Analyse Numérique Matricielle et Optimisation

Vincent BARRA
Jonas KOKO
Philippe MAHEY

Institut Supérieur d'Informatique, de Modélisation et de leurs Applications Campus des Cézeaux
B.P. 1025 - 63173 AUBIERE CEDEX
http://www.isima.fr

Table des matières

Table des figures

Liste des algorithmes

Introduction au Calcul Matriciel

1.1 Représentations des systèmes d'équations linéaires

Considérons le système de 2 équations à 2 inconnues

$$\begin{aligned} 2x_1 - x_2 &= 1 \quad (1.1) \\ x_1 + x_2 &= 5 \quad (1.2) \end{aligned}$$

- *Première interprétation (lecture par ligne).* Chaque équation représente une droite de l'espace à 2 dimensions $\mathbb{R}^2$. La solution se trouve à l'intersection des deux droites (à moins qu'elles ne soient parallèles). On obtient ici le point $x_1 = 2$, $x_2 = 3$, c'est–à–dire le vecteur

$$x = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

de $\mathbb{R}^2$ (voir figure 1.1).

- *Deuxième interprétation (lecture par colonne).* On écrit le système sous la forme vectorielle

$$x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}.$$

Il faut donc trouver les coefficients d'une combinaison linéaire de deux vecteurs qui donnent un troisième (voir figure 1.2).

Exemple 1.1 Un exemple de système en dimension 3.

$$\begin{array}{rrrl} 2x & +y & +z & = 5 \\ 4x & -6y & & = -2 \\ -2x & +7y & +2z & = 9 \end{array}$$

Chaque équation représente un plan de l'espace $\mathbb{R}^3$, *i.e.* l'espace $0xyz$. Deux plans se coupent généralement suivant une droite. On dira que le plan a 2 dimensions et la droite une dimension. En général cette droite coupera le troisième plan en un point, solution unique du système.

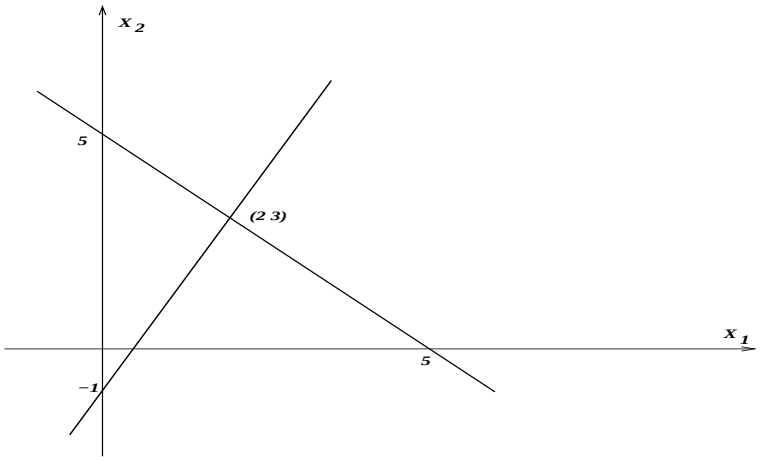

FIG. 1.1 – Équations linéaires

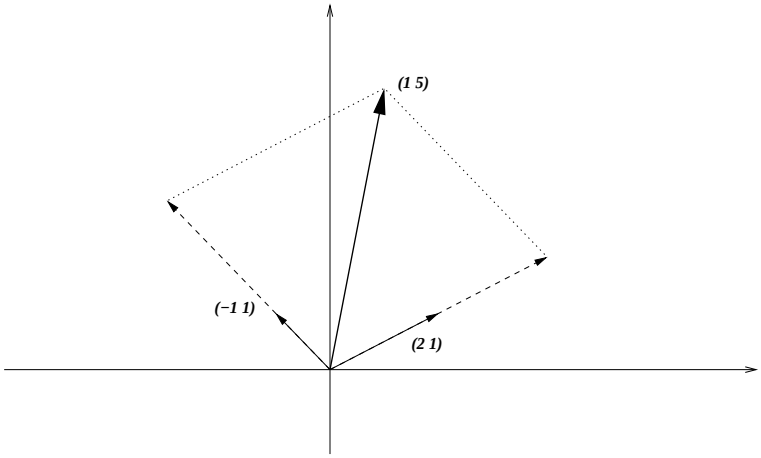

FIG. 1.2 – Combinaisons linéaires

En deuxième interprétation, on cherche les coefficients (x, y, z) d'une combinaison linéaire des vecteurs colonnes telle que :

$$x \begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix} + y \begin{pmatrix} 1 \\ -6 \\ 7 \end{pmatrix} + z \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ -2 \\ 9 \end{pmatrix}.$$

◇

Définition 1.1 (Système singulier) *On dira que le système est* ***singulier*** *s'il n'a pas de solution ou s'il a une infinité de solutions.*

1.2 Vecteurs et matrices dans $\mathbb{R}^n$

$\mathbb{R}^n$, $n \geq 1$, est l'espace des vecteurs réels à n composantes. On notera

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

le vecteur de composantes x_i, $i = 1, \ldots, n$; soit $x \in \mathbb{R}^n$.

Définition 1.2 (Base canonique) *On appelle* ***base canonique*** *de $\mathbb{R}^n$ l'ensemble des vecteurs e_i, $i = 1, \ldots, n$, dont les composantes sont toutes nulles sauf la i–ème qui vaut 1.*

Les éléments de la base canonique sont aussi appelés vecteurs unitaires. Tout vecteur de $\mathbb{R}^n$ peut donc s'écrire de manière unique comme une combinaison linéaire des vecteurs de la base canonique, *i.e.*

$$x = x_1 e_1 + \cdots + x_n e_n.$$

Remarque 1.1 *On ne distinguera pas par une notation spéciale (comme on le fait en Physique) les vecteurs des scalaires.*

1.2.1 Opérations sur les vecteurs

L'addition et la multiplication par un scalaire sont les deux opérations de base dans un espace vectoriel. Soit x et y deux vecteurs de $\mathbb{R}^n$, on a

$$x + y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}.$$

Si $\alpha \in \mathbb{R}$, on a

$$\alpha x = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix}$$

1.2.2 Produit scalaire

Le produit scalaire est une application bilinéaire de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$ notée $< \cdot, \cdot >$ ou plus communément $x^T y$:

$$x^T y = x_1 y_1 + \cdots + x_n y_n.$$

Le produit scalaire $x^T y$ est donc un réel. Le produit scalaire est symétrique

$$x^T y = y^T x.$$

Pour x, y et z des vecteurs de $\mathbb{R}^n$ et $\alpha \in \mathbb{R}$ on a :

$$x^T 0 = 0, \quad x^T e_i = x_i, \quad x^T(y+z) = x^T y + x^T z, \quad (\alpha x)^T y = \alpha(x^T y)$$

Remarque 1.2 *On utilise la même notation pour le vecteur nul $0 \in \mathbb{R}^n$ et pour le scalaire nul 0.*

Définition 1.3 (Orthogonalité) *On dit que deux vecteurs x et y de $\mathbb{R}^n$ sont orthogonaux si $x^T y = 0$.*

1.2.3 Matrices

Une matrice $m \times n$ est un tableau de scalaires rangés en m lignes et n colonnes. Les matrices servent principalement à représenter des transformations linéaires de $\mathbb{R}^n$ dans $\mathbb{R}^m$. On notera a_{ij}, les éléments de la matrice A, $i = 1, \ldots, m$, $j = 1, \ldots, n$. La matrice A représente la transformation linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$ suivante :

$$\forall x \in \mathbb{R}^n, \quad y = Ax \;\Leftrightarrow\; y \in \mathbb{R}^m \text{ et } y_i = a_{i1}x_1 + \cdots + a_{in}x_n, \quad i = 1, \ldots, m.$$

Si on note $A_{\bullet j}$, $j = 1, \ldots, n$, les colonnes de A et $A_{i\bullet}$, $i = 1, \ldots, m$ les vecteurs lignes de A, la transformation peut aussi s'écrire des deux manières suivantes :

$$y = \sum_{j=1}^{n} x_j A_{\bullet j} \qquad \text{ou} \qquad y_i = A_{i\bullet}^T x, \quad i = 1, \ldots, m.$$

Quand $m = n$, on parle de matrice carrée.

Exemple 1.2 Soit la matrice A donnée par

$$A = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}.$$

Le vecteur x dont l'image par A est $\begin{bmatrix} 1 \\ 5 \end{bmatrix}$ est solution du système

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ x_1 + x_2 &= 5 \end{aligned}$$

C'est donc le vecteur $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ trouvé plus haut. On retrouve sur la figure 1.3, les informations communes aux figures 1.1 et 1.2. ◇

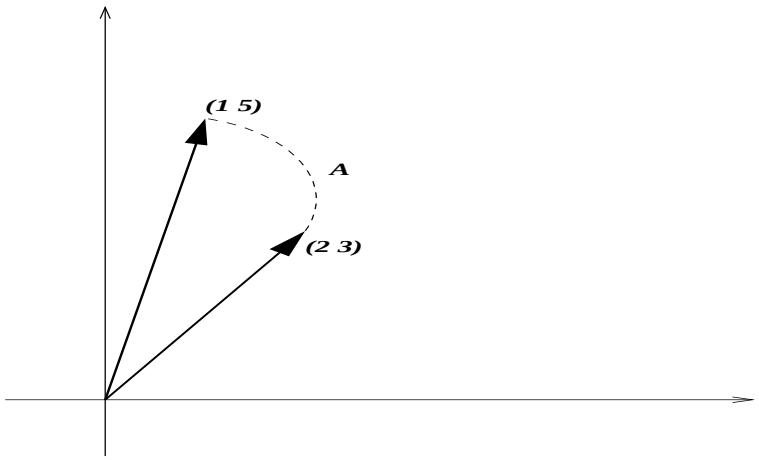

FIG. 1.3 – Autre représentation du système (1.1)-(1.2)

Soit A, B des matrices $m \times n$, soit x, y des vecteurs de $\mathbb{R}^n$, on a les propriétés élémentaires suivantes :

$$\begin{aligned} A(x+y) &= Ax + Ay \\ (A+B)x &= Ax + Bx \\ \alpha Ax &= A(\alpha x), \quad \forall \alpha \in \mathbb{R}. \end{aligned}$$

La matrice identité $\mathbb{I}_n$ est la matrice carrée dont les colonnes forment la base canonique de $\mathbb{R}^n$. On a donc

$$\mathbb{I}_n x = x, \quad \forall x \in \mathbb{R}^n.$$

Une matrice diagonale, notée $A = \text{diag}\{d_1, \ldots, d_n\}$, correspond à un changement d'échelle.

Définition 1.4 (Matrice transposée) *On appelle matrice transposée d'une matrice réelle A la matrice A^T telle que*

$$y^T(Ax) = (A^T y)^T x, \quad \forall x \in \mathbb{R}^n \ \forall y \in \mathbb{R}^m. \tag{1.3}$$

Il est bien entendu que le premier produit scalaire dans (1.3) est dans $\mathbb{R}^m$ et le second dans $\mathbb{R}^n$. On montre alors que la matrice A^T a pour lignes les colonnes de A (et donc pour colonnes les lignes de A). On remarque que la notation dans (1.3) est cohérente avec celle du produit scalaire.

Définition 1.5 (Matrice symétrique) *Une matrice carrée est dite symétrique si $A^T = A$ (on a alors $a_{ij} = a_{ji}$.*

1.2.4 Produit matriciel

Le produit matriciel correspond à la composition de deux applications linéaires. On trouve là le premier calcul sérieux du point de vue de la complexité :

$$C = AB \longrightarrow c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad \forall i, j \tag{1.4}$$

Définition 1.6 ("flop") *On appelle* ***flop*** *l'opération élémentaire d'un produit scalaire, une multiplication suivie d'une addition et d'une substitution.*

Un flop revient donc à mettre à jour une somme s en lui rajoutant le produit de deux nombres a et b, *i.e.*

$$s := s + a * b.$$

Le coût d'un produit scalaire de deux vecteurs de longueur n est donc de n flops.

Avec cette définition, le nombre d'opérations nécessaires pour effectuer le produit (1.4) est de l'ordre de n^3 flops. On verra plus loin que ce coût, très élevé, est en fait équivalent au coût de la résolution d'un système linéaire à n inconnues.

Pour le produit matriciel on la propriété suivante

$$(AB)^T = B^T A^T.$$

Définition 1.7 (Inverse) *L'inverse d'une matrice carrée, notée A^{-1} , est l'unique matrice (lorsqu'elle existe) telle que*

$$AA^{-1} = A^{-1}A = \mathbb{I}.$$

L'étude de l'existence de la matrice inverse est un problème clé de l'algèbre Linéaire qui sera traité au chapitre 2. On verra en particulier pourquoi les transformations singulières n'ont pas d'inverse.

1.3 Sous-espace engendré par un ensemble de vecteurs

Un sous-espace vectoriel de $\mathbb{R}^n$ est un ensemble fermé pour les opérations élémentaires d'un espace linéaire, l'addition et la multiplication par un scalaire.

L'ensemble de toutes les combinaisons linéaires d'un ensemble S de vecteurs de $\mathbb{R}^n$ est le sous-espace engendré par S, noté $\text{lin}\{S\}$:

$$\text{lin}\{S\} = \left\{x \in \mathbb{R}^n \mid x = \alpha_1 x^1 + \cdots + \alpha_n x^n,\ \alpha_i \in \mathbb{R},\ x^i \in S,\ \forall i = 1, \ldots, n\right\}.$$

On vérifie facilement que $\text{lin}\{S\}$ est bien un sous-espace et que $0 \in \text{lin}\{S\}$.

Prenons un vecteur u_1 de $\mathbb{R}^2$, par exemple $u_1 = (2\ 1)^T$. L'ensemble des vecteurs y de $\mathbb{R}^2$ tels que $y = \alpha u_1$, où α est un réel quelconque, est un sous-espace de $\mathbb{R}^2$, le sous-espace $\text{lin}\{u_1\}$. Tous ses éléments peuvent être décrits par le seul paramètre α, on dit que la *dimension* de $\text{lin}\{u_1\}$ est égale à 1. u_1 est le générateur de $\text{lin}\{u_1\}$.

Rajoutons maintenant un autre vecteur $u_2 = (-1\ 1)^T$. Le sous-espace engendré par $\{u_1, u_2\}$ est l'espace $\mathbb{R}^2$ tout entier, qui est de dimension 2.

Remarque 1.3 *Le nombre de générateurs peut être plus grand que la dimension : avec $u_3 = (4\ 2)^T$, le sous-espace engendré par $\{u_1, u_3\}$ est l'ensemble des y tels que*

$$y = \alpha u_1 + \beta u_3 = (\alpha + 2\beta)u_1.$$

D'où $\text{lin}\{u_1, u_3\} = \text{lin}\{u_1\}$ qui est de dimension 1.

On a vu que le nombre de générateurs est supérieur à la dimension si certains générateurs appartiennent au sous-espace engendré par les autres. On dit qu'ils sont *linéairement dépendants.* Dans le cas contraire, l'ensemble des générateurs est *linéairement indépendant* et se nomme *base* du sous-espace. Un sous-espace possède en général une infinité de bases, mais toutes ces bases ont la même cardinalité, la *dimension* du sous-espace.

Le résultat suivant fournit un moyen pratique de tester si k vecteurs sont linéairement indépendants.

Proposition 1.1 *k vecteurs de $\mathbb{R}^n$, $v_1, \ldots, v_k$ sont linéairement indépendants si, et seulement si,*

$$\alpha_1 v_1 + \cdots + \alpha_k v_k = 0 \iff \alpha_1 = \cdots = \alpha_k = 0.$$

Il est clair que tout ensemble de plus de n vecteurs de $\mathbb{R}^n$ est linéairement dépendant, ce qui montre que la dimension d'un sous-espace de $\mathbb{R}^n$ est un entier variant de 0 (sous-espace nul $\{0\}$) à n ($\mathbb{R}^n$ lui-même).

Une matrice A à m lignes et n colonnes représente une transformation linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$. Pour tout x de $\mathbb{R}^n$, on définit l'image de x par la transformation A comme le vecteur y de $\mathbb{R}^m$ tel que

$$y = Ax = \sum_{j=1}^{n} A_{\cdot j} x_j, \quad A_{\cdot j} \in \mathbb{R}^m, \ j = 1, \ldots n.$$

Le sous-espace de $\mathbb{R}^m$ engendré par les colonnes de A est appelé *sous-espace image* de A, noté $\mathrm{Im}(A)$. C'est donc l'ensemble des y de $\mathbb{R}^m$ qui s'écrivent sous la forme Ax pour un certain x de $\mathbb{R}^n$.

Remarque 1.4 (Rang d'une matrice) *Soit A une matrice $m \times n$. Le rang de A, noté* $\mathrm{rang}(A)$, *est le nombre maximal de colonnes de A linéairement indépendantes. C'est aussi le nombre maximal de lignes linéairement indépendantes. On a donc*

$$\mathrm{rang}(A) \leq \min\{m, n\}.$$

De la définition précédente, on tire le résultat fondamental suivant

$$\dim\left(\mathrm{Im}(A)\right) = \mathrm{rang}(A).$$

1.4 Systèmes d'équations linéaires homogènes

Définition 1.8 (Système homogène) *Un système d'équations linéaires est dit **homogène** si le second membre est nul.*

Soit A la matrice $m \times n$ du système (donc pas nécessairement carrée). Analysons l'ensemble des solutions du système d'équations linéaires homogènes

$$Ax = 0. \tag{1.5}$$

Il est clair que 0 est toujours solution.

Proposition 1.2 *L'ensemble des solutions du système linéaire (1.5) est un sous-espace vectoriel de $\mathbb{R}^n$. Ce sous-espace est appelé **noyau** de A, noté* $\ker(A)$.

En effet, si x et y sont solution de (1.5), $x + y$ est aussi solution, ainsi que αx, $\alpha \in \mathbb{R}$.

D'une manière générale, l'ensemble des vecteurs orthogonaux à un sous–ensemble S de vecteurs est un sous-espace vectoriel, noté $S^\perp$, *i.e.*

$$S^\perp = \left\{x \in \mathbb{R}^n \ \mid \ x^T s = 0, \forall s \in S\right\}.$$

On vérifiera que $S^\perp = (\mathrm{lin}\{S\})^\perp$.

On s'aperçoit donc que tout sous-espace vectoriel peut être représenté de deux manières distinctes :

- comme ensemble de combinaisons linéaires d'un nombre fini de vecteurs (espace des vecteurs colonnes ou image de la matrice formée par ces colonnes).
- comme ensemble des solutions d'un système linéaire homogène (noyau de la matrice dont les lignes contiennent les coefficients des équations).

Exemple 1.3 Considérons dans $\mathbb{R}^3$ le plan L des vecteurs dont la troisième composante x_3 est nul. L est donc l'ensemble des solutions de l'équation homogène $x_3 = 0$, ou bien

$$L = \{x \in \mathbb{R}^3 \mid Ax = 0\}, \quad A = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

Ce sous-espace vectoriel est clairement de dimension 2 et peut–être engendré par deux de ses éléments linéairement indépendants (non colinéaires dans ce cas). On prend par exemple les vecteurs $(1\ 0\ 0)^T$ et $(0\ 1\ 0)^T$. Donc

$$L = \{x \in \mathbb{R}^3 \mid x = Bz, z \in \mathbb{R}^2\}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

On remarque que le vecteur ligne de la matrice A est orthogonal à tout vecteur de L. Or ce vecteur de $\mathbb{R}^3$ engendre un sous-espace vectoriel de dimension 1, l'image de A^T. On dira alors que l'image de A^T (sous-espace vectoriel des lignes) est orthogonale au noyau. $\diamond$

Proposition 1.3 *Soit A une matrice $m \times n$. On a*

$$\begin{aligned} &\mathrm{Ker}(A) \perp \mathrm{Im}(A^T), \quad \textit{dans } \mathbb{R}^n \\ &\mathrm{Ker}(A^T) \perp \mathrm{Im}(A), \quad \textit{dans } \mathbb{R}^m. \end{aligned}$$

Ces propriétés géométriques concernant les paires de sous–espaces orthogonaux seront exploitées au chapitre 4.

Les dimensions respectives de ces quatre sous–espaces ne dépendent que du rang de A.

$$\begin{aligned} &\dim \mathrm{Ker}(A) = n - \mathrm{rang}(A),\ \dim \mathrm{Im}(A^T) = \mathrm{rang}(A), \quad \text{dans } \mathbb{R}^n \\ &\dim \mathrm{Ker}(A^T) = m - \mathrm{rang}(A),\ \dim \mathrm{Im}(A) = \mathrm{rang}(A), \quad \text{dans } \mathbb{R}^m \end{aligned}$$

Exemple 1.4 Soit la matrice 2×2 suivante

$$A = \begin{bmatrix} 1 & -3 \\ -2 & 6 \end{bmatrix}.$$

On a $\mathrm{rang}(A) = 1$.

- L'espace colonnes ou $\mathrm{Im}(A)$ contient tous les multiples du vecteur $(1\ -2)^T$.
- Le noyau de A contient les multiples de $(3\ 1)^T$.
- L'espace lignes ou $\mathrm{Im}(A^T)$ contient les multiples de $(1\ -3)^T$.
- Le noyau de A^T contient les multiples de $(2\ 1)^T$.

Ces quatre sous–espaces sont des droites de $\mathbb{R}^2$. Si on change la deuxième colonne en $(-3\ 7)^T$, les colonnes sont alors linéairement indépendantes et $\mathrm{rang}(A) = 2$. Dans ce cas $\mathrm{Im}(A) = \mathrm{Im}(A^T) = \mathbb{R}^2$ et $\mathrm{Ker}(A) = \mathrm{Ker}(A^T) = \{0\}$.

1.5 Transformations élémentaires sur les matrices

Définition 1.9 *On appelle transformation élémentaire des lignes (respectivement des colonnes) d'une matrice une combinaison des trois transformations suivantes :*
- *multiplication d'une ligne (resp. colonne) par un scalaire non nul ;*
- *addition de deux lignes (resp. colonnes) ;*
- *permutation de deux lignes (resp. colonnes).*

Principaux résultats

(i) Une transformation élémentaire ne change pas le rang d'une matrice (c'est une transformation non singulière).

(ii) Une transformation élémentaire sur les lignes (resp. colonnes) équivaut à multiplier la matrice à gauche (resp. à droite) par une matrice élémentaire obtenue en appliquant la même transformation à la matrice identité.

Exemple 1.5

$$A = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 3 & 4 \end{bmatrix}$$

Si on additionne les deux premières lignes et que l'on inscrive le résultat sur la première ligne sans changer les deux autres, on obtient

$$A' = \begin{bmatrix} 2 & 2 & 0 & 3 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 3 & 4 \end{bmatrix}$$

On peut vérifier que $A' = EA$, avec

$$E = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

◇

Pivotage d'une colonne

C'est une séquence de transformations élémentaires qui transforme une colonne d'une matrice en vecteur unitaire (c.a.d. de la base orthonormée de $\mathbb{R}^m$). L'élément qui doit se transformer en 1 est le *pivot*. Par exemple soit A_s la colonne à pivoter et a_{rs} le pivot ($a_{rs} \neq 0$). On effectue dans l'ordre
- multiplier la ligne r par $1/a_{rs}$
- pour tout $i \neq r$, remplacer la ligne i par la somme de la ligne i et de la ligne r multipliée par $-a_{is}$.

Exemple 1.6 Pivotons la première colonne de la matrice suivante

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ -2 & 1 & 3 \end{bmatrix}$$

avec $a_{11} = 1$ comme pivot. On doit donc multiplier A à gauche par la matrice élémentaire

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

et le résultat est

$$A' = E \cdot A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 3 & 5 \end{bmatrix}$$

◇

Il est clair que, si tous les pivots successifs sont non nuls, une séquence de n pivotages effectués sur les n colonnes de la matrice avec le pivot sur la diagonale transformera la matrice en l'identité. Si E_i est la matrice élémentaire associée au pivotage de la i–ème colonne, on a

$$\mathbb{I}_n = E_n E_{n-1} \cdots E_1 A$$

d'où une première méthode pour calculer l'inverse d'une matrice

$$A^{-1} = E_n E_{n-1} \cdots E_1.$$

Donc pour calculer l'inverse d'une matrice, il suffit d'effectuer les pivotages en parallèle sur la matrice identité. La matrice à inverser se transforme progressivement en la matrice identité alors que l'identité devient l'inverse. S'il est vrai que, quand aucun pivot nul n'est rencontré, cet algorithme a une complexité de $O(n^3)$, il peut atteindre $O(2n^3)$ dans le cas général et nécessite le stockage de deux matrices $n \times n$. On lui préférera les méthodes étudiées au chapitre suivant, basées elles aussi sur des pivotages successifs de la matrice, mais plus robustes et moins coûteuses.

1.6 Systèmes d'équations linéaires

Soit A une matrice $m \times n$ dont les lignes contiennent les coefficients du système linéaire et b un vecteur de $\mathbb{R}^m$ contenant le second membre. On cherche à décrire l'ensemble des x de $\mathbb{R}^n$ solutions du système

$$Ax = b. \tag{1.6}$$

Avant de nous poser le problème de l'existence d'une solution, montrons que l'ensemble des solutions, quand il n'est pas vide, est un objet de $\mathbb{R}^n$ d'aspect familier.

Soit x^0 une solution particulière du système (1.6). Alors, toute solution du système peut s'écrire sous la forme $x = x^0 + y$, $y \in \mathrm{Ker}(A)$. En effet comme x et x^0 sont deux solutions du système, on a $A(x - x^0) = 0$. Donc, l'ensemble des solutions s'obtient par une translation du sous–espace noyau de A. Il en a donc la forme géométrique et quand b varie, on construit une famille d'objets linéaires parallèles entre eux (voir figure 1.4). On les appelle des *variétés linéaires* ou *sous-espaces affines*. Si une variété linéaire passe par l'origine, c'est un sous-espace vectoriel.

Bien que la dimension ait été définie pour décrire la génération d'un sous–espace vectoriel, on l'emploie pour une variété linéaire par analogie pour exprimer le nombre de degrés de liberté. Si $V = \{x \in \mathbb{R}^n \mid Ax = b\}$ est une variété linéaire (donc que l'ensemble des solutions du système est non vide) telle que $V = \{x^0\} + \mathrm{Ker}(A)$, on définit sa dimension comme

$$\dim V = \dim \mathrm{Ker}(A) = n - \mathrm{rang}(A).$$

Si un système homogène a toujours une solution (l'origine), ce n'est pas vrai pour le cas général.

Théorème 1.1 *Le système linéaire (1.6) possède au moins une solution si, et seulement si,* $b \in \mathrm{Im}(A)$.

On remarque que cette condition s'exprime dans l'espace des colonnes $\mathbb{R}^m$. On sait déjà que le sous–espace $\mathrm{Im}(A)$ est de dimension $\mathrm{rang}(A)$. On en déduit que si $\mathrm{rang}(A) < m$, la probabilité pour que b soit dans $\mathrm{Im}(A)$ est nulle et le système sera presque toujours sans solution (voir figure 1.5).

Par contre si $\mathrm{rang}(A) = m$, on dit que A est de rang plein, le système a donc toujours des solutions. Si de plus, $m = n$, le noyau de A est de dimension nulle et le système a une solution unique. Cette solution est l'image de b par la transformation inverse de A, notée A^{-1}, *i.e.*

$$x = A^{-1}b.$$

Donc seules les matrices carrées de rang plein sont inversibles. On dit aussi que ces transformations sont *régulières*. Les matrices carrées non inversibles sont associées aux transformations *singulières* de $\mathbb{R}^n$.

Remarque 1.5 *Une condition équivalente, nécessaire et suffisante, pour que le système ait une solution est :*

$$\mathrm{rang}(A) = \mathrm{rang}\left([A \,|\, b]\right), \tag{1.7}$$

où $[A \,|\, b]$ *est la matrice* $m \times (n+1)$ *obtenue en rajoutant la colonne* b *à* A. *On verra de toute façon qu'il est rare qu'on ait à calculer explicitement l'inverse d'une matrice et que la résolution itérative d'un système linéaire permet de tester la condition (1.7) comme elle permet de calculer le rang de* A.

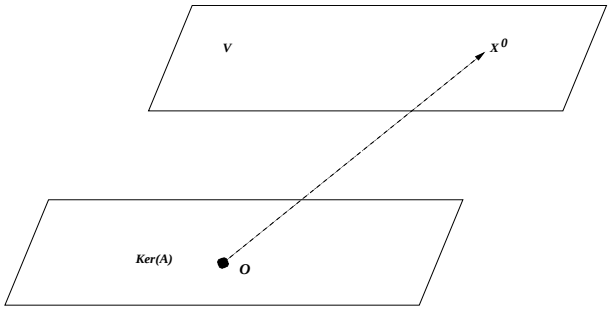

FIG. 1.4 – Variété linéaire : Dans $\mathbb{R}^n$, Ker(A) $= \{x \in \mathbb{R}^n \mid Ax = 0\}$ et $V = \{x \in \mathbb{R}^n \mid Ax = b\} =$ Ker(A) $+ \{x^0\}$

1.7 Exercices

Exercice 1.1 Soit A, une matrice $m \times n$. Montrer que

$$\begin{aligned} \ker A^T &= (\mathrm{Im}A)^{\perp} \\ \mathrm{Im}A^T &= (\ker A)^{\perp}. \end{aligned}$$

Préciser les espaces contenant ces ensembles.

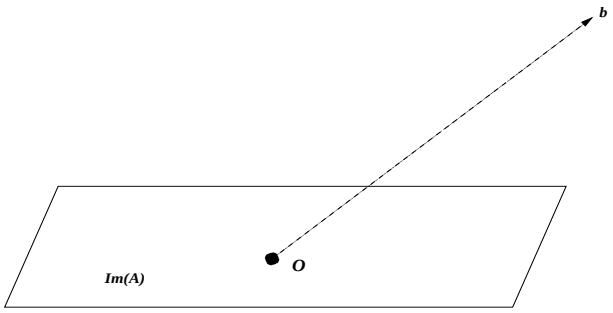

FIG. 1.5 – Système incompatible : Dans $\mathbb{R}^n$, $b \notin \mathrm{Im}(A)$. Donc on ne peut trouver $x \in \mathbb{R}^n$ tel que $b = Ax$

Soit maintenant

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

Quel est le rang de A. Donner les sous–espaces $\mathrm{Im}A$, $\ker A^T$, $\mathrm{Im}A^T$ et $\ker A^T$.

Exercice 1.2 La matrice A ci–dessous représente le passage de la base canonique $\{e_i\}_{1\leq i\leq 4}$ à la base $\mathcal{A} = \{a_i\}_{1\leq i\leq 4}$.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}.$$

1.— Pour changer l'ordre des lignes de A, on multiplie A par une matrice P. En notant a^i la ligne i de A, comment passe–t–on de

$$\begin{bmatrix} a^1 \\ a^2 \\ a^3 \\ a^4 \end{bmatrix} \quad \text{à} \quad \begin{bmatrix} a^1 \\ a^3 \\ a^2 \\ a^4 \end{bmatrix}$$

Exprimer la matrice A' correspondante et la matrice P qui réalise cette transformation.

Comment passe–t–on de

$$\begin{bmatrix} a^1 \\ a^2 \\ a^3 \\ a^4 \end{bmatrix} \quad \text{à} \quad \begin{bmatrix} a^1 - 2a^3 \\ a^3 + 3a^4 \\ a^2 \\ a^4 - a^1 \end{bmatrix}$$

Exprimer la matrice A'' correspondante et la matrice P qui réalise cette transformation.

2.— Déterminer les matrices P_1, P_2 et P_3 qui réalisent les transformations suivantes sur la matrice T ; en déduire l'inverse de T.

$$T = \begin{pmatrix} 1 & 3 & 3 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} P_1 \to \begin{pmatrix} 1 & 3 & 3 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} P_2 \to \begin{pmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} P_3 \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Exercice 1.3 Dans tout l'exercice, on note $\{e_1,\ e_2\}$ la base canonique de $\mathbb{R}^2$.
Soit $\{u_1,\ u_2\}$ un couple de vecteurs linéairement indépendants :

$$u_1 = \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}, \quad u_2 = \begin{pmatrix} u_{12} \\ u_{22} \end{pmatrix}.$$

Un vecteur x de $\mathbb{R}^2$ s'écrit

$$x = x_1 e_1 + x_2 e_2 = x'_1 u_1 + x'_2 u_2.$$

On illustrera les résultats des questions a) à d) avec les vecteurs

$$u_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

1. Montrer que l'on peut écrire $(u_1\ u_2) = M(e_1\ e_2)$, où M est une matrice que l'on explicitera.
2. En déduire x_1 et x_2 en fonction de x'_1 et x'_2 sous forme matricielle.
3. Inverser la relation précédente ; soit M^{-1} l'inverse de M.
4. Calculer $\| x' \|^2$ en fonction de x_1, x_2 et des éléments de M.
5. A quelle condition sur u_1 et u_2 a-t-on

$$\| x' \|^2 = \lambda^2 \| x \|^2 .$$

6. On pose $u_{11} = \cos\theta$ et $u_{12} = \sin\theta$. Déterminer les matrices M_+ et M_- telles que

$$\| x' \|^2 = \| x \|^2 .$$

7. Quels sont les vecteurs invariants par M_+ et M_- ?
8. Soit x le vecteur de coordonnées $x_1 = 1$ et $x_2 = 2$. Montrer que $M_+ x$ s'obtient par une rotation de θ et que $M_- x$ s'obtient par une symétrie par rapport à la droite d'angle $\theta/2$. On notera $M_+ = R_\theta$ et $M_- = S_{\theta/2}$. Calculer $R_{-\theta}$ et $(S_{\theta/2})^2$. En déduire R_θ^{-1} et $S_{\theta/2}^{-1}$.

Élimination de Gauss

2.1 Un exemple

Soit le système de 3 équations à 3 inconnues

$$\begin{aligned} 2x - 3y &= 3 && (2.1)\\ 4x - 5y + z &= 7 && (2.2)\\ 2x - y - 3z &= 5 && (2.3) \end{aligned}$$

La méthode d'élimination de Gauss (ou méthode du pivot) consiste à utiliser la première équation pour calculer x en fonction des autres variables puis de remplacer cette variable dans les équations suivantes. Cette élimination se poursuit avec y dans les nouvelles équations (sauf la première) jusqu'à l'obtention d'une équation à une seule inconnue. On remonte alors en remplaçant les variables calculées dans les équations ayant servi à l'élimination :
Tirons x de l'équation (2.1) :

$$x = \frac{3}{2}(1 + y) \qquad (2.4)$$

Remplaçons x dans les deux dernières équations (2.2)-(2.3) par son expression (2.4)

$$\begin{aligned} y + z &= 1 && (2.5)\\ 2y - 3z &= 2. && (2.6) \end{aligned}$$

Tirons y de l'équation (2.5) de ce nouveau système :

$$y = 1 - z. \qquad (2.7)$$

Remplaçons y dans l'équation (2.6) :

$$-5z = 0. \qquad (2.8)$$

La phase d'élimination est terminée. On effectue alors la substitution en sens inverse des variables grâce aux équations (2.8), (2.7) et (2.4) :

$$\begin{aligned} (2.8) &\Rightarrow z = 0\\ (2.7) &\Rightarrow y = 1\\ (2.4) &\Rightarrow x = 3 \end{aligned}$$

On remarque qu'à une étape donnée, l'élimination d'une variable peut se faire dans n'importe quelle équation, à condition, bien sûr, que cette équation contienne la variable en question.

2.2 Systèmes triangulaires

On constatera au paragraphe suivant que la phase d'élimination consiste à transformer le système original en un système triangulaire. Un système triangulaire est un système dont la matrice est triangulaire.

Une matrice *triangulaire supérieure* (respectivement *triangulaire inférieure*) est une matrice carrée dont les éléments sous la diagonale (resp. au dessus) sont nuls.

Un système triangulaire supérieur se résout par substitution arrière (on supposera tous les éléments diagonaux a_{ii} non nuls).

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{2n}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots & \cdots \\ a_{nn}x_n &= b_n \end{aligned}$$

On commence d'abord par calculer x_n dans l'équation n. Puis à l'aide de x_n, on peut calculer x_{n-1} dans l'équation $n-1$ et ainsi de suite jusqu'à x_1. Ce qui donne l'algorithme suivant.

$$x_n = \frac{b_n}{a_{nn}}, \tag{2.9}$$

$$x_k = \frac{1}{a_{kk}}\left[b_k - \sum_{j=k+1}^{n} a_{kj}x_j\right], \quad k = n-1, \ldots, 1. \tag{2.10}$$

On remarque que le calcul de x_k coûte $n-k$ flops et une division. Le coût total de l'algorithme est donc de

$$1 + 2 + \cdots + n - 1 = \frac{n(n-1)}{2},$$

soit (on ne garde que les termes de plus haut degré)

$$\boldsymbol{\frac{n^2}{2}} \textbf{ flops et } \boldsymbol{n} \textbf{ divisions.}$$

Dans le cas d'un système triangulaire inférieur, on effectue des substitutions directes et l'algorithme est

$$x_1 = \frac{b_1}{a_{11}}, \tag{2.11}$$

$$x_k = \frac{1}{a_{kk}}\left[b_k - \sum_{j=1}^{k-1} a_{kj}x_j\right], \quad k = 2, \ldots, n. \tag{2.12}$$

Le coût de l'algorithme de substitution (2.11)-(2.12) est le même que celui de substitution inverse (2.9)-(2.10).

2.3 Méthode de Gauss (ou du pivot) pour les systèmes linéaires

Montrons d'abord que la technique d'élimination décrite au §2.1 correspond à une opération de pivotage.

Éliminer x de l'équation (2.2) revient à faire la somme de l'équation (2.1) multipliée par -2 avec (2.2). De même éliminer x de (2.3) revient à faire la somme de l'équation (2.1) multipliée par -1 avec l'équation (2.3). On a donc effectué un pivotage de la première colonne (celle de x) en utilisant la première ligne comme ligne du pivot.

Remarque 2.1
1. *A chaque itération, l'élément de la colonne dans la ligne du pivot doit être non nul.*
2. *Les transformations élémentaires effectuées sur la matrice sont effectuées en parallèle sur le second membre.*
3. *A la fin de l'élimination (si tout se passe bien !), on obtient un système triangulaire avec les pivots sur la diagonale.*

Les transformations successives sont décrites ci-dessous pour l'exemple du §2.1.
Itération 1 :

$$\begin{bmatrix} 1 & & \\ -2 & 1 & \\ -1 & & 1 \end{bmatrix} \left[\begin{array}{ccc|c} (2) & -3 & 0 & 3 \\ 4 & -5 & 1 & 7 \\ 2 & -1 & -3 & 5 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & -3 & 0 & 3 \\ 0 & 1 & 1 & 1 \\ 0 & 2 & -3 & 2 \end{array}\right]$$

Itération 2 :

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & -2 & 1 \end{bmatrix} \left[\begin{array}{ccc|c} 2 & -3 & 0 & 3 \\ 0 & (1) & 1 & 1 \\ 0 & 2 & -3 & 2 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & -3 & 0 & 3 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -5 & 0 \end{array}\right]$$

Dans le cas général, on a donc à l'étape k de l'algorithme une matrice $A^{(k)}$ de n lignes et $n+1$ colonnes (la dernière colonne représente le second membre du système transformé) dont les $k-1$ premières colonnes sont triangulaires supérieures (cf. figure 2.1).

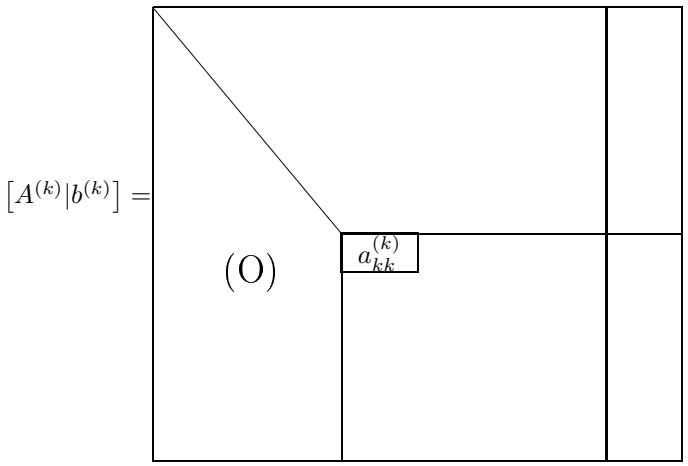

FIG. 2.1 – Étape k de la méthode de Gauss

On met donc à zéro chaque élément de la colonne k sous le pivot $a_{kk}^{(k)}$ (supposé non nul) en remplaçant la ligne i (pour $i = k+1, \ldots, n$) par

$$\text{ligne}\, i \;\leftarrow\; \text{ligne}\, i + \frac{-a_{ik}^{(k)}}{a_{kk}^{(k)}} \times \text{ligne}\, k.$$

L'élément $a_{ij}^{(k)}$ pour $i = k+1, \ldots, n$ et $j = k+1, \ldots, n+1$ devient

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}$$

La mise à jour de la ligne i requiert donc $n-k+1$ flops et 1 division. Donc l'étape k coûte $(n-k)(n-k+1)$ flops et $n-k$ divisions. On en déduit le coût total de l'algorithme de Gauss :

$$\begin{aligned} n(n-1) + (n-1)(n-2) + \cdots + 2 \cdot 1 &= n^2 + (n-1)^2 + \cdots + 2^2 + 1^2 - (n + n - 1 + \cdots + 2 + 1) \\ &= \frac{1}{6} n(n+1)(2n+1) - \frac{1}{2} n(n+1) \text{ flops} \end{aligned}$$

et

$$\frac{1}{2} n(n+1) \text{ divisions.}$$

On dira que la complexité de la méthode de Gauss est de $\mathbf{\frac{1}{3} n^3}$ **flops.**

L'algorithme d'élimination de Gauss est présenté dans l'algorithme 1. Le second membre est stocké dans la dernière colonne de A. L'élimination transforme la matrice A en une matrice triangulaire supérieure (*étape 1.*). Le système triangulaire supérieur est ensuite résolu par substitution inverse (*étape 2.*).

Algorithme 1 Méthode de Gauss

1. Élimination
Pour $k = 1, \ldots, n-1$ **Faire**
  Pour $i = k+1, \ldots, n$ **Faire**
    Pour $j = k+1, \ldots, n+1$ **Faire**
      $a_{ij} \leftarrow a_{ij} - \dfrac{a_{ik}}{a_{kk}} a_{kj}$
    Fin Pour
  Fin Pour
Fin Pour
2. Résolution du système triangulaire
$x_n \leftarrow \dfrac{a_{n,n+1}}{a_{nn}}$
Pour $k = n-1, \ldots, 1$ **Faire**
  $x_k \leftarrow \dfrac{1}{a_{kk}} \left[a_{k,n+1} - \displaystyle\sum_{j=k+1}^{n} a_{kj} x_j \right]$
Fin Pour

2.3.1 Stratégies pour les pivots nuls

Si le pivot est nul (en pratique, on évitera aussi les pivots de valeur absolue trop petite, cf. chapitre 3), on le remplace en effectuant une permutation avec un élément non nul parmi les éléments sous lui et/ou à sa droite. On distingue généralement deux stratégies :

Pivotage total On permute lignes et colonnes pour choisir le plus grand élément en valeur absolue dans la sous-matrice en bas à droite (voir Fig. 2.1) :

$$\max \left\{ \left| a_{ij}^{(k)} \right| \ ; \ i = k, \ldots, n; \ j = k, \ldots, n \right\}$$

Pivotage partiel On ne permute que les lignes sous le pivot en choisissant le plus grand élément en valeur absolue

$$\max\left\{\left|a_{ik}^{(k)}\right| \ ; \ i=k,\ldots,n\right\}$$

La stratégie du pivotage partiel est la plus utilisée car la plus économique (la recherche du plus grand élément sur une liste de p nombres coûte p comparaisons numériques, ce qui donne $n^3/3$ comparaisons pour le pivotage total). L'algorithme d'élimination de Gauss avec recherche du pivot partiel est présenté dans l'algorithme 2. Le second membre est stocké dans la dernière colonne de A.

Algorithme 2 Méthode de Gauss avec pivot partiel

1. Élimination
Pour $k=1,\ldots,n-1$ **Faire**
  Recherche du pivot
  $c_p \leftarrow |a_{kk}|$, $i_p \leftarrow k$
  Pour $i=k+1,\ldots,n$ **Faire**
    Si $|a_{ik}| > c_p$ **Alors**
      $c_p \leftarrow |a_{ik}|$, $i_p \leftarrow i$
    Fin Si
  Fin Pour
  Permutation
  Si $i_p \neq k$ **Alors**
    Permuter les lignes i_p et k de la matrice A
  Fin Si
  Pivotage
  Pour $i=k+1,\ldots,n$ **Faire**
    Pour $j=k+1,\ldots,n+1$ **Faire**
      $a_{ij} \leftarrow a_{ij} - \dfrac{a_{ik}}{a_{kk}} a_{kj}$
    Fin Pour
  Fin Pour
Fin Pour
2. Résolution du système triangulaire
$x_n \leftarrow \dfrac{a_{n,n+1}}{a_{nn}}$
Pour $k=n-1,\ldots,1$ **Faire**
  $x_k \leftarrow \dfrac{1}{a_{kk}}\left[a_{k,n+1} - \displaystyle\sum_{j=k+1}^{n} a_{kj}x_j\right]$
Fin Pour

2.3.2 Cas singulier et calcul du rang d'une matrice

Pivotage total

Si, à l'itération k de la méthode de Gauss avec pivot total,

$$\max\left\{\left|a_{ij}^{(k)}\right| \ ; \ i=k,\ldots,n; \ j=k,\ldots,n\right\} = 0$$

(donc tous les éléments de la sous-matrice sont nuls), on peut affirmer que le rang de la matrice A est égal à $k-1$. S'il existe un élément $b_i^{(k)}$, $k \leq i \leq n$, du second membre différent de zéro, alors le système n'a pas de solution.

Pivotage partiel

Si

$$\max\left\{\left|a_{ik}^{(k)}\right| \ ; \ i = k, \ldots, n\right\} = 0$$

on peut seulement affirmer que la colonne k est linéairement dépendante des $k-1$ premières. Cela implique que $\text{rang}(A) < n$ et que le système n'a probablement pas de solution. Toutefois, on ne peut en être sûr que sur le test du pivot total nul.

Remarque 2.2 *La méthode de Gauss et le test du pivot total nul apparaissent donc comme la meilleure stratégie pour calculer le rang d'une matrice.*

2.4 Facteurs LU d'une matrice non singulière

Quand on a plusieurs systèmes linéaires à résoudre avec la même matrice et des seconds membres différents, on a intérêt lors de la première résolution à garder les coefficients des pivotages successifs en mémoire. Cela correspond à garder la factorisation LU de la matrice. En effet, chaque pivotage peut être représenté par une matrice élémentaire qui ne diffère de l'identité que par une sous-colonne.

Reprenons le pivotage de la matrice $A^{(k)}$ à l'étape k. Soit η_k le vecteur de $\mathbb{R}^{n-k}$ dont les composantes sont $\eta_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$. On a donc (en supposant $a_{kk}^{(k)} \neq 0$)

$$A^{(k+1)} = E_k A^{(k)}$$

où E_k est la matrice élémentaire suivante

$$E_k = \begin{bmatrix} 1 & & \\ & \ddots & \\ & -\eta_k & 1 \end{bmatrix}$$

On a donc, après $n-1$ pivotages

$$A^{(n-1)} = E_{n-1} E_{n-2} \cdots E_1 A. \tag{2.13}$$

Notons U la matrice triangulaire supérieure $A^{(n-1)}$ et réécrivons la relation (2.13)

$$A = E_1^{-1} \cdots E_{n-1}^{-1} U$$

On vérifie que la matrice inverse E_k^{-1} a la même forme que E_k avec les éléments de la sous-colonne k changés de signe et que les produits $E_k^{-1} E_{k+1}^{-1}$ s'effectuent sans calcul en accolant les vecteurs η_k et η_{k+1} dans les colonnes k et $k+1$. On peut donc écrire $A = LU$ où L est une matrice triangulaire inférieure dont les éléments diagonaux sont égaux à 1 et les éléments sous la diagonale sont

$$l_{ij} = \eta_{ij}.$$

Observons que les éléments non diagonaux de L peuvent être rangés directement à la place des éléments de A correspondants. La matrice A est donc recouverte par sa factorisation LU et le coût de stockage est en n^2.

Grâce à cette factorisation (qui ne coûte donc pas plus cher que la triangularisation), tout nouveau système linéaire $Ax = b'$ peut être résolu par la résolution de deux systèmes triangulaires (donc en $O(n^2)$ flops). En effet, pour résoudre

$$LUx = b'$$

on résout d'abord

$$Ly = b'$$

puis

$$Ux = y.$$

Algorithme 3 Factorisation LU

$\sigma(i) = i$, $i = 1, \ldots, n$ (initialisation du vecteur des permutations)
Pour $k = 1, \ldots, n-1$ **Faire**
  Recherche du pivot
  $c_p \leftarrow |a_{kk}|$, $i_p \leftarrow k$
  Pour $i = k+1, \ldots, n$ **Faire**
    Si $|a_{ik}| > c_p$ **Alors**
      $c_p \leftarrow |a_{ik}|$, $i_p \leftarrow i$
    Fin Si
  Fin Pour
  Permutation
  Si $i_p \neq k$ **Alors**
    Permuter les lignes i_p et k de la matrice A
    $\sigma(k) = i_p$, $\sigma(i_p) = k$
  Fin Si
  Pivotage
  Pour $i = k+1, \ldots, n$ **Faire**
    Remplissage de la colonne k par les coefficients η_{ik}
    $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$
    Modification des lignes qui n'ont pas encore été ligne-pivot
    Pour $j = k+1, \ldots, n+1$ **Faire**
      $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$
    Fin Pour
  Fin Pour
Fin Pour

Quand aucun pivot nul n'est rencontré, A peut se mettre sous la forme LU et cette factorisation est unique. En effet, s'il existe deux factorisations L_1U_1 et L_2U_2 de A, on a alors $L_1U_1 = L_2U_2$. Ce qui implique que $L_2^{-1}L_1 = U_2U_1^{-1}$ et le produit de deux matrices inférieures (resp. supérieures) étant une matrice triangulaire inférieure (resp. supérieure), ces produits sont nécessairement une matrice diagonale. C'est l'identité car $(l_1)_{ii} = (l_2)_{ii} = 1$ pour tout i.

Dans le cas d'une stratégie de pivot partiel, si P_k est la matrice de permutation des lignes à l'itération k, on peut écrire

$$A^{(k+1)} = E_k P_k A^{(k)}.$$

En fait, les différentes permutations peuvent être résumées dans la matrice

$$P = P_{n-1}P_{n-2}\cdots P_1$$

et on obtient la décomposition générale suivante.

Théorème 2.1 *Pour toute matrice A non singulière d'ordre n, il existe une matrice de permutation P, une matrice triangulaire inférieure L telle que $l_{ii} = 1$, pour tout i, et une matrice triangulaire supérieure U, telles que*

$$PA = LU.$$

DÉMONSTRATION : On a $E_{n-1}P_{n-1}\cdots E_1P_1A = U$. On montre alors que la matrice $L = P(E_{n-1}P_{n-1}\cdots E_1P_1)^{-1}$ est bien triangulaire inférieure. On remarque que dans ce cas $|l_{ij}| \leq 1$. □

L'algorithme 3 montre les différentes étapes de la factorisation LU avec recherche du pivot partiel. En entrée on a la matrice A et le vecteur des permutations σ. En sortie, les A contient les facteurs L et U de la matrice et σ les permutations de lignes éventuelles. Si $\sigma_i = j$ alors les lignes i et j ont été permutées. Les permutations doivent être repercutées sur le second membre lors de la résolution de $Ly = b$, *conf.* algorithme 4.

Algorithme 4 Substitutions directes/inverses

1. Substitutions directes $Ly = b$

$x_1 \leftarrow b_{\sigma_1}$

Pour $k = 2, \ldots, n$ **Faire**

$$x_k \leftarrow b_{\sigma_k} - \sum_{j=1}^{k-1} a_{kj}x_j$$

Fin Pour

2. Substitutions inverses $Ux = y$

$$x_n \leftarrow \frac{x_n}{a_{nn}}$$

Pour $k = n-1, \ldots, 1$ **Faire**

$$x_k \leftarrow \frac{1}{a_{kk}}\left[x_k - \sum_{j=k+1}^{n} a_{kj}x_j\right]$$

Fin Pour

2.4.1 Cas particuliers

Matrices symétriques

Dans ce cas, U peut s'écrire $U = DL^T$ où D est la matrice diagonale contenant les pivots successifs. On a donc la factorisation $A = LDL^T$. La complexité de l'algorithme est alors de $n^3/6$ flops (cf. exercice).

Matrices bandes

Ce sont des matrices symétriques telles que $a_{ij} = 0$ pour $|i-j| > p$ (p est la largeur de bande de la matrice, $p < n$). Ces matrices interviennent couramment dans la discrétisation d'équations différentielles. Il est alors facile de montrer que les facteurs LU respectent la bande. On a alors intérêt de stocker la matrice (et ses facteurs LU) sous la forme de tableau à n lignes et p colonnes et la complexité est en $np^2/2$ flops.

Matrices symétriques définies positives

Ces matrices seront étudiées au chapitre 6. Elles possèdent une factorisation unique LDL^T avec des pivots successifs strictement positifs. La factorisation peut s'effectuer directement sans

pivotage par identification terme à terme en $n^3/6$ flops par l'algorithme de Cholesky (cf. chapitre 6).

2.5 Autres applications

2.5.1 Calcul de l'inverse d'une matrice

On a vu précédemment que la résolution d'un système linéaire ne nécessite pas le calcul explicite de l'inverse d'une matrice. Quand on a besoin néanmoins de la calculer, on peut procéder de la manière suivante, basée sur la factorisation LU de la matrice :

- Calculer les facteurs LU de la matrice : $PA = LU$
- Résoudre les n systèmes linéaires $LUx^i = Pe_i$, où e_i, $1 \leq i \leq, n$, est le i-ème vecteur de la base canonique de $\mathbb{R}^n$. La solution x^i est la i-ème colonne de A^{-1}.

Le coût total apparent est de $n^3/3 + n^3 = 4n^3/3$ flops. Mais on peut montrer que, grâce à la structure particulière des seconds membres des systèmes linéaires successifs, le coût réel n'est que de n^3 flops (cf. exercice).

Une approche équivalente couramment utilisée, mais qui ne passe pas par le calcul des facteurs LU, est la méthode dite de Gauss-Jordan qui consiste à pivoter complètement le système paramétré

$$Ax - y = 0.$$

On pivote cette fois sur la colonne entière de façon à transformer le système en un système diagonal $x - A^{-1}y = 0$. On peut observer que cette technique consiste à effectuer en parallèle à partir de la matrice identité les pivotages nécessaires à la transformation de A en la matrice identité.

2.5.2 Calcul du déterminant

Bien que très important par ses nombreuses utilisations en Analyse et en Géométrie différentielle et bien que remarquable par ses propriétés étonnantes dans la théorie des matrices, le déterminant est difficile à définir et à exploiter dans le cadre de ce cours. Le déterminant est la valeur d'une forme multilinéaire sur un ensemble de n vecteurs qui change de signe à chaque permutation des vecteurs et qui vaut 1 pour les n vecteurs d'une base orthonormée. C'est aussi le volume du n-parallélépipède engendré par n vecteurs dans $\mathbb{R}^n$. On se bornera ici à rappeler les principales propriétés du déterminant et à donner le lien essentiel avec la méthode de Gauss, c'est-à-dire, le déterminant est égal (au signe près) au produit des pivots.

Propriétés du déterminant

- Le déterminant dépend linéairement de chaque ligne séparément. Il s'ensuit que $\det(A+B) \neq \det(A) + \det(B)$ et $\det(aA) \neq a\det(A)$. En effet, à chaque fois qu'on multiplie une ligne par a, le déterminant est multiplié par a ; donc $\det(aA) = a^n \det(A)$.
- Le déterminant change de signe si on permute deux lignes.
- $\det(\mathbb{I}) = 1$
- $\det(A) = 0 \Leftrightarrow A$ est singulière.
- $\det(AB) = \det(A)\det(B)$ (donc $\det(A^{-1}) = (\det(A))^{-1}$.
- $\det(A^T) = \det(A)$.
- Si $T = (t_{ij})$ est triangulaire,

$$\det(T) = \prod_{i=1}^{n} t_{ii}.$$

Déterminant d'une matrice d'ordre 2

Le calcul du déterminant d'une matrice 2×2 est bien connu :

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc \tag{2.14}$$

ce qui permet de déterminer explicitement l'inverse d'une matrice 2×2 :

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{2.15}$$

La généralisation des formules (2.14)-(2.15) au cas des matrices $n \times n$ conduit aux fameuses formules de Cramer que nous ne reproduirons pas ici car elles ont une complexité exponentielle, ce qui les rend impraticables pour des dimensions très petites. A titre d'exemple, pour calculer le déterminant d'une matrice 20×20 par la formule de Cramer il faut à peu près 15400 ans de calcul sur une machine de 100 Mips (soit 10^8 instructions par seconde). Avec la méthode des pivots le coût n'est que de $3 \cdot 10^{-5}$ secondes !

En pratique on calculera le déterminant après pivotage :

$$\det(A) = (-1)^p \prod_{i=1}^{n} u_{ii}$$

où les u_{ii} $(1 \leq i \leq n)$ sont les pivots et p le nombre de permutations effectuées au cours de la factorisation.

2.6 Exercice

Exercice 2.1 (Assemblage de ressorts) 1.— Calculer les facteurs LU de la matrice A suivante

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

En déduire les facteurs LU d'une matrice symétrique A, $n \times n$,

$$\begin{aligned} a_{ii} &= 2 \\ a_{i,i+1} &= -1 \\ a_{ij} &= 0, \quad \forall j > i+1. \end{aligned}$$

2.— On considère $n+1$ ressorts de même longueur L et de constante de raideur k, assemblés bout à bout et fixés aux deux extrémités de la chaîne. Chaque noeud i est associé à une force f_i dirigée selon l'axe des ressorts et on ne considère que les déplacements longitudinaux du système. On note u_i le déplacement du noeud i. De la relation mécanique

$$f = \frac{k}{L} \delta u,$$

on tire le bilan des efforts au noeud i

$$\frac{k}{L}(u_i - u_{i-1}) - \frac{k}{L}(u_{i+1} - u_i) + f_i = 0.$$

On a aux extrémités, $u_0 = u_{n+1} = 0$.

Mettre ce système d'équations sous la forme $Ax = b$, où A est une matrice $n \times n$ et b un vecteur de $\mathbb{R}^n$ que l'on explicitera.

En tenant compte du profil de A, donner le nombre de flops nécessaires à la résolution du système des ressorts.

Exercice 2.2 (Matrice de Hilbert) Trianguler par la méthode de Gauss, la matrice de Hilbert 3×3 suivante

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}$$

Vérifier que le déterminant de H est égal au produit des pivots successifs. Calculer l'inverse de H.

Exercice 2.3 (Produit matrice–vecteur dans $\mathbb{R}^n$) Soit a et b, deux vecteurs de $\mathbb{R}^n$. On définit la matrice A $(n \times n)$ par

$$\begin{aligned} a_{ii} &= 1 + a_i^2 + b_i^2 \\ a_{ij} &= a_i a_j + b_i b_j, \quad \text{si } i \neq j. \end{aligned}$$

a) Exprimer la matrice A en fonction de $\mathbb{I}_n$ et de a, b, a^T, b^T.

b) On désire effectuer le produit Ax, $x \in \mathbb{R}^n$. Écrire un algorithme réalisant ce produit en $O(kn)$ flops, où k est un facteur à déterminer.

STABILITÉ NUMÉRIQUE

3.1 Introduction

Les algorithmes décrits dans ce cours et en particulier ceux du chapitre 2 font intervenir un certain nombre d'opérations élémentaires destinées à être traitées par un ordinateur. Chaque ordinateur a une manière propre de représenter les nombres réels et l'ensemble des réels qu'il peut représenter est fini, la clé pour comprendre la structure de ce sous-ensemble de la droite des réels étant la précision machine.

L'arithmétique en précision finie et la réalité des données inexactes nous obligent à considérer les questions suivantes reliées à la résolution d'un système linéaire $Ax = b$:

- si A et b sont perturbés par une 'petite quantité', comment les solutions exactes x et calculées $x_c$sont elles affectées ?
- que signifie numériquement le fait que A est presque singulière ?
- si $b \notin \mathrm{Im}(A)$, comment déterminer x pour que Ax soit suffisamment 'proche' de b ?

On s'attardera ici à donner quelques éléments de réponse aux deux premières questions, la troisième étant plus particulièrement traitée dans le chapitre suivant.

3.2 Normes matricielles et condition d'une matrice

3.2.1 Rappels sur les normes vectorielles

Une norme vectorielle est une fonction notée $\|.\|$ définie sur un espace vectoriel et satisfaisant aux trois axiomes suivants :

1. $\| x \| \geq 0$ pour tout x et $\| x \| = 0 \Leftrightarrow x = 0$
2. $\| ax \| = |a| \, \| x \|$ pour tout x et tout scalaire a
3. $\| x + y \| \leq \| x \| \| y \|$ pour tous x, y

Exemple 3.1 *1. dans $\mathbb{R}^n$ muni du produit scalaire $x^T y$,*

$$\| x \|_2 = \left(x^T x\right)^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

est une norme. C'est la norme euclidienne, la plus couramment employée.

2. $\| x \|_\infty = \max\limits_{1 \leq i \leq n} \{|x_i|\}$ est la norme du max (ou norme de Tchebychev, ou norme l_∞)

3. $\| x \|_1 = \sum_{i=1}^{n} |x_i|$ *est la norme* l_1

4. *les trois normes précédentes sont des cas particuliers des normes* l_p *:* $\| x \|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$

L'inégalité triangulaire correspondante est un résultat important de l'analyse fonctionnelle dû à Minkowski et sa démonstration découle de l'inégalité fondamentale suivante dite inégalité de **Hölder** :

$$(\forall p, q > 1 \ / \ \frac{1}{p} + \frac{1}{q} = 1) \quad \sum_{i=1}^{n} |x_i y_i| \leq \| x \|_p \| y \|_q \ .$$

On observera également que dans $\mathbb{R}^n$ toutes les normes sont équivalentes dans le sens où, pour deux normes $\| \ . \ \|_a$ et $\| \ . \ \|_b$, il existe deux constantes positives α et β satisfaisant

$$\alpha \| x \|_a \leq \| x \|_b \leq \beta \| x \|_a, \quad \forall x \in \mathbb{R}^n.$$

3.2.2 Normes matricielles

On définit maintenant des normes sur l'espace des matrices carrées $(n \times n)$. On exige de plus que ces normes vérifient une condition supplémentaire, dite condition de norme **sous-multiplicative** :

$$\| AB \| \leq \| A \| \cdot \| B \|$$

Pour évaluer les effets d'une transformation linéaire sur la norme d'un vecteur, on s'arrangera pour utiliser des normes de matrices **subordonnées** aux normes vectorielles, dans le sens suivant :

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

où $\| A \|$ définit une norme subordonnée à la norme vectorielle utilisée dans le calcul des normes vectorielles $\| x \|$ et $\| Ax \|$. On retrouve la définition usuelle d'une norme d'application linéaire. On a alors la relation $\| Ax \| \leq \| A \| \| x \|$ pour tout x.

Observations :

1. le sup ci-dessus doit être pris dans $\mathbb{C}^n$, mais coïncide avec le sup dans $\mathbb{R}^n$ pour les normes l_1, l_2 et l_∞.
2. par la compacité de la sphère unité, le sup est atteint pour un x non nul

Propriétés des normes subordonnées

1. $\| A \| = \inf \{ \rho / \ \| Ax \| \leq \rho \| x \| \}$
2. $\| I \| = 1$

Exemple 3.2 1. *la norme matricielle subordonnée à la norme euclidienne est* $\| A \|_2 = [\lambda_{max}]^{\frac{1}{2}}$, *où* λ_{max} *est la plus grande valeur propre de* $A^T A$, *cf. chap. 5.*

2. *on vérifie que les normes matricielles subordonnées aux normes* l_1 *et* l_∞ *sont données par* $\| A \|_1 = \max_j \sum_i |a_{ij}|$ *et* $\| A \|_\infty = \max_i \sum_j |a_{ij}|$

3. *la norme euclidienne étant difficile à calculer, on lui préfère souvent la norme de la trace ou norme de Frobenius :*

$$\| A \|_F = \left(Tr(A^T A)\right)^{\frac{1}{2}} = \left(\sum_i \sum_j a_{ij}^2 \right)^{\frac{1}{2}}.$$

Cette norme n'est subordonnée à aucune norme vectorielle sur $\mathbb{R}^n$.

On peut vérifier que, pour une matrice orthogonale H (*i.e.* telle que $H^T H = H H^T = I$), on a

$$\begin{aligned} \| H \|_2 &= 1 \\ \| H \|_F &= \sqrt{n}. \end{aligned}$$

Remarque 3.1 *Le lien entre* $\| A \|_2$ *et les valeurs propres de* A *sera précisé au chapitre 5.*

Le résultat suivant montre l'utilité pratique des normes matricielles :

Théorème 3.1 *Soit* $\| . \|$ *une norme matricielle subordonnée et* E *une matrice telle que* $\| E \| < 1$. *Alors la matrice* $A = I + E$ *est inversible.*

DÉMONSTRATION : considérons le système homogène $Ax = 0$. Il s'écrit $x + Ex = 0$, et ainsi $\| x \| = \| Ex \| \leq \| E \| \| x \|$, car la norme est subordonnée. Comme $\| E \| < 1$, on a nécessairement $x = 0$ et A est donc non singulière. □

3.2.3 Condition d'une matrice

Étudions les variations de la solution d'un système d'équations linéaires quand le second membre subit une perturbation :

$$Ax = b \Rightarrow A(x + \delta x) = b + \delta b$$

On peut donc écrire $\delta x = A^{-1} \delta b$ d'où les relations écrites avec des normes subordonnées appropriées :
$\| \delta x \| \leq \| A^{-1} \| \| \delta b \|$
$\| A \| \| x \| \geq \| b \|$
et on peut écrire l'estimation de l'erreur relative en norme sur x en fonction de la perturbation relative sur b :

$$\frac{\| \delta x \|}{\| x \|} \leq \| A \| \| A^{-1} \| \frac{\| \delta b \|}{\| b \|}$$

La quantité $\| A \| \| A^{-1} \|$ est appelée la **condition** de la matrice est notée $\sigma(A)$.

Propriété 3.1

1. $\sigma(A) \geq 1$. *Plus la condition est grande, plus la matrice est dite mal conditionnée, i.e. plus le système* $Ax = b$ *est instable.*
2. $\sigma(A) = \sigma(A^{-1})$
   $\sigma(aA) = \sigma(A)$ *pour tout scalaire* $a \neq 0$.
3. *Si* H *est une matrice orthogonale* $\sigma(H) = 1$

3.2.4 Conditionnement pour la norme euclidienne

La condition est très liée aux valeurs propres extrêmes de A, et si la norme utilisée est la norme euclidienne, on vérifie que :

$$\sigma(A) = \left(\frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} \right)^{\frac{1}{2}}$$

de sorte que si A est symétrique

$$\sigma(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

Le calcul de la valeur exacte de $\sigma(A)$ est coûteux et il est possible d'essayer *a priori* de réduire le risque d'erreur.

Équilibrage *(scaling)*

C'est une méthode pragmatique qui n'est pas évidente à mettre en oeuvre. Son principe est d'essayer d'équilibrer le poids de tous les coefficients de la matrice. Il faut trouver des matrices diagonales $D^{(1)} = (d_{ii}^{(1)}, 1 \leq i \leq n)$ et $D^{(2)}$ telles que $\sigma(D^{(1)} A {D^{(2)}}^{-1}) << \sigma(A)$, la matrice $D^{(1)} A {D^{(2)}}^{-1}$ s'écrivant :

$$D^{(1)} A {D^{(2)}}^{-1} = \left(a_{ij} \frac{d_{ii}^{(1)}}{d_{jj}^{(2)}} \right)$$

On résout alors

$$\begin{cases} D^{(1)} A {D^{(2)}}^{-1} y & = D^{(1)} b \\ D^{(2)} x & = y \end{cases}$$

Ce n'est pas toujours facile, et pas toujours efficace

Raffinement itératif

On peut supposer que la perturbation due aux erreurs d'arrondis ne porte que sur le second membre. On a par exemple réalisé en machine une factorisation

$$L_c U_c = A + E$$

et on a résolu exactement le système

$$(A + E) x_c = b$$

On vérifie généralement dans ce cas que le résidu calculé $r_c = b - A x_c$ n'est pas nul.
Posons alors $x^{(0)} = x_c$ et $r^{(0)} = r_c$. Ces valeurs initialisent un processus récursif. On calcule pour tout $k \geq 0$

$$\begin{cases} (A + E) e^{(k+1)} & = r^{(k)} \\ x^{(k+1)} & = \left(x^{(k)} + e^{(k+1)} \right)_c \\ r^{(k+1)} & = \left(b - A x^{(k+1)} \right)_c \end{cases}$$

En supposant que l'erreur d'arrondi porte essentiellement sur la résolution des systèmes linéaires, on pose

$$\begin{aligned} x^{(k+1)} - x^{(k)} &= (A + E)^{(-1)} r^{(k)} \\ e^{(k+1)} &= (A + E)^{(-1)} A (x - x^{(k)}) \end{aligned}$$

d'où l'on tire

$$x - x^{(k+1)} = (I - (A + E)^{(-1)} A)(x - x^{(k)})$$

Posons $G = I - (A+E)^{(-1)}A = I - (I + A^{-1}E)^{-1}$, on a donc obtenu, pour tout $k \geq 1$

$$x - x^{(k+1)} = G(x - x^{(k)}) = G^{k+1}(x - x^{(0)})$$

Supposons alors que pour une norme subordonnée on ait $\|A^{-1}E\| = \alpha < 1$, alors $G = \sum_{k=1}^{\infty}(-A^{-1}E)^k$ est tel que $\|G\| < \frac{\alpha}{1-\alpha}$ et

$$x - x^{(k)} \leq \left(\frac{\alpha}{1-\alpha}\right)^k (x - x^{(0)})$$

L'expérience prouve qu'une ou deux itérations de ce processus peuvent être intéressantes, mais pas plus.

Exercice d'application :
On considère le système suivant :

$$(S) : \begin{cases} 10x + 7y + 8z + 7w = 32 \\ 7x + 5y + 6z + 5w = 23 \\ 8x + 6y + 10z + 9w = 33 \\ 7x + 5y + 9z + 10w = 31 \end{cases}$$

dont la solution est apparente : $x = y = z = w = 1$. Recalculer la solution avec un second membre perturbé b=(32.1 22.9 33.1 30.9)^T.
Que remarquez-vous ? Calculer la condition de la matrice du système (on utilisera la norme de Frobenius et on justifiera ce choix).

3.3 Vitesse de convergence des suites

Une des difficultés de la construction d'algorithmes est le contrôle de la convergence, c'est-à-dire, pouvoir identifier si la suite des solutions calculées à chaque itération converge ou pas, et si oui, avec quelle vitesse s'approche-t-elle de la solution du problème. On dira qu'une suite $\{x_k\}$ de $\mathbb{R}^n$ converge vers $\overline{x} \in \mathbb{R}^n$ si la suite de réels $r_k = \| x_k - \overline{x} \|$ converge, dans $\mathbb{R}$, vers 0. L'équivalence des norme montre que la notion de convergence est indépendante du choix de la norme dans $\mathbb{R}^n$.

Définition 3.1 (Ordre de convergence) *Considérons une suite de réels $\{r_k\}$ qui converge vers une valeur r^*. On appelle* **ordre** *de convergence de la suite $\{r_k\}$ le plus grand entier $p > 0$ tel que :*

$$\lim_{k\to\infty} sup \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p} = \beta < +\infty$$

La situation de loin la plus courante est $p = 1$, appelée convergence linéaire. C'est le cas typiquement pour une suite géométrique $r_k = a^k, 0 < a < 1$. La limite β est le rayon de convergence, et dans le cas géométrique, on trouve justement a. Donc, plus le rayon de convergence est proche de 0, plus la convergence est rapide (on parle de convergence superlinéaire quand $\beta = 0$, ce qui est le cas par exemple pour la suite $r_k = (1/k)^k$). D'autre part, plus le taux est proche de 1, plus la convergence est lente (par exemple $r_k = (1/k)$).
Quand $p = 2$, on parle de convergence quadratique (par exemple $r_k = a^{2^k}$).

3.4 Stabilité de la méthode de Gauss

On a vu précédemment l'influence du mauvais conditionnement sur la stabilité de la solution d'un système d'équations linéaires. Il faut observer qu'un mauvais conditionnement ne signifie pas nécessairement que la matrice est quasi singulière. C'est le déterminant qui mesure cette quasi singularité quand il s'approche de 0. Par exemple la matrice $(n \times n)$ aI, avec $a = 10^{-1}$ voit son déterminant tendre vers 0 quand n tend vers l'infini, alors que sa condition reste égale à 1. La situation inverse est encore plus flagrante avec une matrice non symétrique dans l'exemple ci-dessous :

$$A = \begin{bmatrix} 1 & 100 \\ 0 & 1 \end{bmatrix}$$

Si

$$b = \begin{bmatrix} 100 \\ 1 \end{bmatrix}, \qquad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

mais si

$$b = \begin{bmatrix} 100 \\ 0 \end{bmatrix}, \qquad x = \begin{bmatrix} 100 \\ 0 \end{bmatrix}$$

Le déterminant de A est bien sûr égal à 1, pourtant la condition de cette matrice est très mauvaise ($> 10^4$).

Le mauvais conditionnement d'une matrice est souvent très difficile à évaluer et surtout, la question du remède à y apporter reste délicate. Toutefois, certaines situations peuvent être facilement contournées : c'est le cas par exemple où les hétérogénéités numériques du système apparaissent suivant les lignes ou les colonnes. On peut alors effectuer un changement d'échelle (**scaling**) et améliorer la condition de la matrice. On se contentera ici d'illustrer ce fait sur un exemple très simple :

$$(S) : \left\{ \begin{array}{rcrcr} 10x_1 & + & 100000x_2 & = & 100000 \\ x_1 & + & x_2 & = & 2 \end{array} \right.$$

Une résolution directe du système en arithmétique tronquée à trois chiffres significatifs fournira une solution x=(0.00 1.00). Si on multiplie la première équation par un facteur d'échelle de 10^{-5} on obtient l'approximation x'=(1.00 1.00) beaucoup plus satisfaisante.

Il faut pour terminer mettre l'accent sur le fait que même une matrice bien conditionnée est de déterminant très supérieur à 0 peut provoquer des difficultés numériques si l'algorithme d'élimination est mal contrôlé. C'est le cas sur des pivots trop petits sont acceptés au cours des calculs.

Prenons un exemple simple :

$$(S) : \left\{ \begin{array}{rcrcr} 0.0000x_1 & + & x_2 & = & 1 \\ x_1 & + & x_2 & = & 2 \end{array} \right.$$

Si on maintient 0.0001 comme premier pivot et si les calculs sont effectués avec trois chiffres significatifs, on obtiendra x_2=1, et la substitution arrière donnera $0.0001x_1 + 1 = 1$, soit $x_1 = 0$, et l'amplification de l'erreur d'arrondi est catastrophique.

Le remède à cette difficulté est connu : c'est la stratégie du **pivot partiel** qui revient ici à permuter les deux lignes et commencer l'élimination par celle qui présente le plus grand pivot.

3.5 Exercices

Exercice 3.1 Considérons le système linéaire $Ax = b$ dans $\mathbb{R}^4$ avec

$$A = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 1 & 3 & 1 & 2 \\ -1 & -2 & 2 & -1 \\ -2 & -1 & 3 & 11 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \alpha \end{pmatrix}$$

1. Calculer le rang de la matrice A.
2. Pour quelles valeurs de α le système linéaire est–il compatible ? Caractériser alors l'ensemble des solutions du système.

Exercice 3.2 Considérons le système linéaire

$$2x + 6y = 8 \tag{3.1}$$
$$2x + 6.00001y = 8.00001 \tag{3.2}$$

Ce problème est-il bien conditionné ? Pourquoi ? Résoudre le système (3.1)-(3.2).
Considérons maintenant le système

$$2x + 6y = 8 \tag{3.3}$$
$$2x + 5.99999y = 8.00002 \tag{3.4}$$

Résoudre le système (3.3)-(3.4). Conclusion.

Exercice 3.3 Soit A une matrice non singulière d'ordre n. Donner un algorithme de résolution de

$$A^2 x = b$$

qui évite le calcul explicite de A^2. Évaluer le gain en nombre de flops si A est pleine.

Exercice 3.4 Considérons les matrices

$$A = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}$$

Appliquer la méthode d'élimination de Gauss avec une précision de 10^{-3}. Vérifier que les facteurs LU des deux matrices sont obtenus avec une bonne précision et que le nombre de 0 de B est maintenu.

Exercice 3.5 Soit A la matrice carrée d'ordre n suivante

$$A = \begin{pmatrix} 1 & 2 & 0 & \cdots & 0 \\ 0 & 1 & 2 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Calculer $\det A$, $\| A \|_1$ et $\| A \|_\infty$. Calculer A^{-1} puis $\sigma(A)$.

MOINDRES CARRÉS ET TRANSFORMATIONS ORTHOGONALES

4.1 Projections orthogonales

On s'intéresse ici aux mécanismes de calcul de la projection orthogonale d'un vecteur de $\mathbb{R}^n$ sur un sous-espace. Les résultats fondamentaux d'existence, d'unicité et de caractérisation de la projection orthogonale d'un point sur un sous-espace, valables dans des espaces plus généraux que $\mathbb{R}^n$, sont rappelés ci-après sans démonstration.

Théorème 4.1 *Soit L un sous-espace vectoriel de $\mathbb{R}^n$. Étant donné un point $y \in \mathbb{R}^n$, il existe un unique point p de L, appelé la projection orthogonale de y sur L, tel que $\| y-p \| \leq \| y-x \|, \forall x \in L$. Une condition nécessaire et suffisante pour que $p \in L$ soit la projection orthogonale de y sur L est $y - p \in L^\perp$*

4.1.1 Projection sur une droite passant par l'origine

Une droite qui passe par l'origine est un sous-espace de dimension 1. Soit $y \in \mathbb{R}^n$ et D le sous-espace de vecteurs engendrés par un vecteur v non nul : $D = \{z \in \mathbb{R}^n / z = xv, x \in \mathbb{R}\}$.
Si $p \in D$ est la projection orthogonale de y sur la droite, on peut écrire : $y = p + u$, avec $p = xv$ et $u^T v = 0$.
On remarque que la dernière relation signifie que $u \in D^\perp$ et donc que u est la projection orthogonale de y sur $D^\perp$. (Figure 4.1). Le calcul de p est alors

$$v^T y = x v^T v \Rightarrow x = \frac{1}{v^T v}.v^T y$$

soit

$$p = \frac{v^T y}{v^T v}.v$$

Ainsi, cette expression qui exprime le fait que p a la même direction que v peut s'écrire différemment pour représenter la transformation qui transforme y en p : on vérifie que $(v^T y).v = (vv^T)y$ où l'on observe que $v^T y$ est un scalaire alors que $vv^\perp$ est une matrice, de rang 1 (puisque toutes les colonnes sont des multiples de v) qui **projette** l'espace $\mathbb{R}^n$ sur la droite D. La projection orthogonale sur une droite qui passe par l'origine est donc une transformation linéaire, de matrice :

$$P = \frac{1}{\| v \|^2} vv^T$$

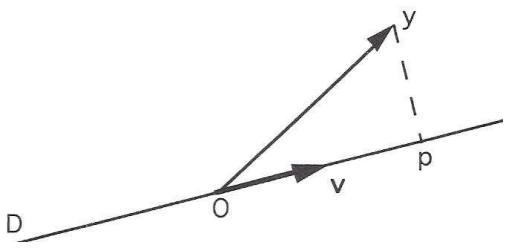

FIG. 4.1 – Projection sur une droite

Remarques :

1. Comme $p = Py$, la projection orthogonale sur le sous-espace orthogonal $D^{\perp}$ est $u = y - p = (I - P)y$, donc $I - P$ est la matrice de projection orthogonale sur $D^{\perp}$.
2. Soit θ l'angle entre les directions des vecteurs v et y. On a alors :

$$cos(\theta) = \frac{v^T y}{\| v \| \| y \|}$$

et on en déduit l'inégalité de Schwarz :

$$(\forall a, b) |a^T b| \leq \| a \| \| b \|$$

Exemple :
$v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
Comme $v^T v$=1, on a bien $p = Py = \begin{bmatrix} y_1 \\ 0 \end{bmatrix}$

4.1.2 Projection sur une droite ne passant pas par l'origine

On utilise la représentation d'une droite comme un sous-espace affine parallèle à un sous-espace :

$$D = \{z \in \mathbb{R}^n / z = z^0 + xv, x \in \mathbb{R}\}$$

Soit $y \in \mathbb{R}^n$ et p sa projection orthogonale sur la droite D. Alors :
$y = p + u$ avec $p = z^0 + xv$ et $u^T v = 0$, d'où $p = z^0 + P(y - z^0) = (I - P)z^0 + Py$,
où P est la matrice de projection sur le sous-espace engendré par v obtenu dans le paragraphe 4.1.1.

4.1.3 Projection sur un sous-espace

Soit A une matrice $(n \times r)$ de rang r (donc $r \leq n$). Considérons la projection p d'un vecteur de $\mathbb{R}^n$ sur le sous-espace image de A (Fig 4.2).

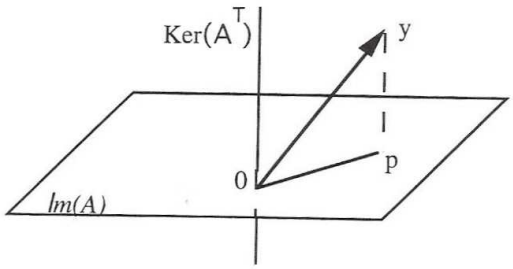

FIG. 4.2 – Projection sur un sous-espace

Alors $y = p + u$ avec $p = Ax, x \in \mathbb{R}^r$ et $A^T u = 0$ car $u \in Ker(A^T)$.
On obtient alors : $A^T y = A^T Ax$, et comme A est de rang plein, la matrice $(r \times r)$ $A^T A$ est inversible et :
$p = Py$ avec $P = A(A^T A)^{-1} A^T$.
De plus, on retrouve la matrice de projection orthogonale sur le noyau de A^T : $P' = I - P$.

4.1.4 Matrices de projection

Les matrices de projection sont des matrices qui possèdent les deux propriétés suivantes :

- $P = P^T$
- $P^2 = P$

On vérifiera ces propriétés sur les matrices précédentes.
Les matrices de projection sont en général singulières, puisqu'elles ramènent l'espace à un sous-espace de dimension plus petite. De plus, elles contractent les normes : $\| Py \| \leq \| y \|$.

4.2 Moindres carrés linéaires

4.2.1 Identification des paramètres

Un système est observé à partir des mesures de sa sortie y en fonction de différentes entrées t. On veut valider un modèle théorique de ce système qui dépend de n paramètres $x_1 \cdots x_n$. Soit $x = (x_1 \cdots x_n)^T$ le vecteur de paramètres à déterminer. On a donc pour chaque entrée $t_i, 1 \leq i \leq m$ une sortie mesurée Y_i et une sortie théorique $y_i = f(t_i; x)$ (Fig 4.3)

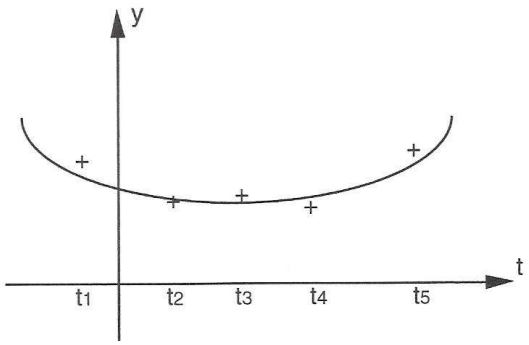

FIG. 4.3 – $f(t) = x_1 + x_2 t + x_3 t^2$ mesurée pour $t = t_1 \cdots t_5$

L'erreur entre le modèle théorique et la sortie mesurée est $e_i = y_i - Y_i$ et la solution aux moindres carrés consiste à choisir x qui minimise la comme des carrés des erreurs :

Trouver $x \in \mathbb{R}^n$ tel que $\sum_{i=1}^{m} e_i^2$ soit minimale.

Si la fonction f est **linéaire par rapport aux paramètres** x, on parle de problème au moindres carrés linéaires. Limitons nous à ce cas et interprétons le problème dans l'espace des mesures : $e_i = y_i - Y_i = a_i^T x - Y_i$.

Soit e le vecteur de $\mathbb{R}^m$ dont les composantes sont les erreurs e_i et soit A la matrice $(m \times n)$ dont les lignes sont les a_i^T. Le problème consiste alors à

Trouver $x \in \mathbb{R}^n$ qui **minimise** $\| Ax - Y \|^2$.

Dans l'espace $\mathbb{R}^m$, il s'agit donc de trouver le point de l'image de A le plus proche au sens de la norme euclidienne du vecteur Y (qui a peu de chances d'appartenir à $Im(A)$, car $m >> n$). L'unique solution dans l'espace des mesures est donc la **projection orthogonale** du vecteur Y sur le sous-espace $Im(A)$. La solution a déjà été calculée au paragraphe 4.1.3, c'est la solution du système linéaire suivant, dit aux **équations normales** :

$$A^T A x = A^T Y$$

Si $\text{rang}A = n$ (hypothèse raisonnable car $m >> n$ et les a_i dépendent des entrées t_i), ce système a une solution unique $x = (A^T A)^{-1} A^T Y$, dite solution aux moindres carrés. La matrice $A^+ = (A^T A)^{-1} A^T$ (Attention !! ce n'est pas la matrice de projection) est souvent appelée la **pseudo inverse** de la matrice rectangulaire A. Elle satisfait $A^+ A = I$ et $AA^+ = P$.

4.2.2 Systèmes incompatibles

Soit un système linéaire incompatible $Ax = b$, où $A(m \times n)$ est telle que $\text{rang}A < m$ et $b \notin Im(A)$. On supposera par exemple (comme dans le cas des moindres carrés) que $m > n$ et $\text{rang}A = n$. Le système n'a donc pas de solution, et on le remplace par le système aux équations normales obtenu en le multipliant par A^T :

$$A^T A x = A^T b$$

Ce système est en général mal conditionné car la condition de A^TA est la condition de A au carré dans le cas d'une matrice carrée A. La méthode de Gauss risque d'être inefficace et on lui préférera des méthodes basées sur des transformations orthogonales qui ont l'avantage d'être numériquement stables.

Finalement, si les colonnes de A sont orthonormées (*i.e.* orthogonales deux à deux et de norme 1) $A^TA = I$ (l'identité dans la 'petite' dimension n) et la solution des équations normales est simplement $x^* = A^Tb$. La 'bonne' stratégie pour résoudre le problème des moindres carrés est donc de construire une base orthonormée de $Im(A)$ pour calculer explicitement la projection. On verra au paragraphe 4.3 que cette construction revient à trianguliariser la matrice par des transformations orthogonales.

4.2.3 Exemple : régression linéaire

Soit un modèle affine d'un système à une entrée t : $y = a + bt$, où a et b sont deux paramètres à déterminer pour minimiser au sens des moindres carrés l'erreur des trois mesures (t, Y) suivantes : (-1,4),(0,5),(1,9).
Le système

$$(S) : \begin{cases} a & - & b & = & 4 \\ a & & & = & 5 \\ a & + & b & = & 9 \end{cases}$$

n'a bien sûr pas de solution. On construit donc les équations normales en multipliant par $A^T = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix}$, et donc $A^TA = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$, donc $(A^TA)^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix}$
d'où la solution aux moindres carrés $a = 6 =$ et $b = 5/2$.
On remarque que le vecteur des erreurs $Y - Ax = [1/2\ -1\ 1/2]^T$ est bien orthogonal dans R^3 aux colonnes de A. Chaque erreur peut être également représentée par la distance verticale entre la mesure Y_i et la droite $f(t) = 6 + \frac{5}{2}t$ pour $t = t_i$ (Fig 4.4)

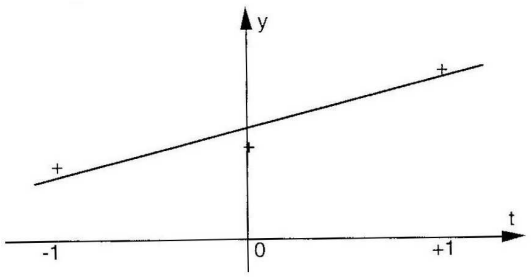

FIG. 4.4 – régression linéaire

4.3 Transformations orthogonales

4.3.1 Matrices orthogonales

Définition 4.1 *Une matrice carrée H est dite* **orthogonale** *si* $H^TH = HH^T = I$

Une matrice orthogonale est donc une matrice carrée dont les colonnes sont orthonormées. Les matrices de rotation, de symétrie, de permutation et l'identité sont des exemples de matrices or-

thogonales.
Une matrice orthogonale H est naturellement inversible par définition, et l'inverse est $H^{-1} = H^T$.

Proposition 4.1 *Propriété fondamentale*
Les transformation orthogonales sont des isométries, les normes (euclidiennes), les produits scalaires et les angles sont conservés :

$$H \text{ orthogonale} \Leftrightarrow (\forall x \in \mathbb{R}^n) \parallel Hx \parallel = \parallel x \parallel$$

En effet, $\parallel Hx \parallel^2 = (Hx)^T(Hx) = x^T H^T Hx = x^T x = \parallel x \parallel^2$.

Cette propriété entraîne une stabilité numérique des méthodes utilisant ces transformations. On les utilise principalement pour :
- orthonormaliser un système de générateurs
- résoudre un systèmes aux équations normales
- trianguliariser un système mal conditionné
- calculer les valeurs propres d'une matrice (chapitre 5)

4.3.2 Orthogonalisation de Gram-Schmidt

A partir d'une famille de vecteurs indépendants de $\mathbb{R}^p$, ou d'une matrice A, on peut construire une famille $\{q_1 \cdots q_p\}$, base orthonormée de $Im(A)$. C'est un outil fondamental pour la résolution de systèmes surdéterminés. L'idée générale est donc de construire une base orthonormée du sous-espace image d'un ensemble de vecteurs. L'intérêt numérique est que cette construction équivaut à trianguliariser la matrice formée par ces vecteurs. On présentera ici la construction de cette base orthonormée, connue sous le nom de procédure de Gram-Schmidt, dont l'intérêt est purement académique. En pratique, Gram-Schmidt coûte trop cher et on utilise la méthode de factorisation **QR** qui permet d'atteindre le même objectif grâce à des transformations orthogonales élémentaires (**rotations de Givens** ou transformations de **Householder**) numériquement stables et seulement deux fois plus chères que la méthode de Gauss.
Considérons tout d'abord une matrice carrée $A \in \mathcal{M}_{n,n}(\mathbb{R})$ dont les colonnes a_i sont linéairement indépendantes. Montrons que l'on peut construire itérativement une matrice orthogonale Q, dont les colonnes sont notées q_i telle que $Q^T A = R$, où R est triangulaire supérieure.
Comme Q est orthogonale, on a : $A = QR$. Or, les éléments de R sont de la forme :

$$\begin{array}{llll} r_{11} = q_1^t a_1 & & & \\ r_{12} = q_1^t a_2, & r_{22} = q_2^t a_2 & & \\ \vdots & & & \\ r_{1n} = q_1^t a_n, & \cdots & r_{nn} = q_n^t a_n & \end{array}$$

On identifie alors le produit QR par colonne :

$$\begin{array}{ll} a_1 = (q_1^t a_1) q_1 & \Rightarrow q_1 = \frac{a_1}{\parallel a_1 \parallel} \text{ et } r_{11} = \parallel a_1 \parallel \\ a_2 = (q_1^t a_2) q_1 + (q_2^t a_2) q_2, & \Rightarrow q_2 = \frac{a_2 - (q_1^T a_2) q_1}{\parallel a_2 - (q_1^T a_2) q_1 \parallel} \\ \vdots & \end{array}$$

On remarque que chaque colonne q_i est obtenue en soustrayant de q_i ses projections orthogonales

sur les $i-1$ premiers vecteurs de la base orthonormée déjà calculés, puis en normant le résultat (Fig 4.5)

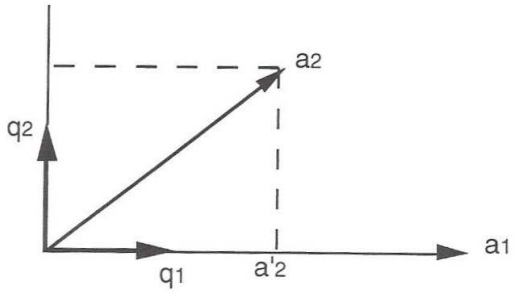

FIG. 4.5 – Gram-Schmidt

On peut réaliser la même construction si le nombre de vecteurs a_i est inférieur à n (matrice rectangulaire $(n \times p)$). On construit p colonnes q_i orthonormées comme précédemment de manière à obtenir $A = Q_1 R_1$, où R_1 est une matrice carrée triangulaire $(p \times p)$. On complète alors la base orthonormée en continuant la procédure de Gram-Schmidt avec $n-p$ vecteurs arbitraires, mais tels que les n colonnes formées avec les a_i soient linéairement indépendantes. Soit Q_2 la matrice des $n-p$ derniers vecteurs orthonormés. On a alors bien :
$A^T Q_2 = R_1^T Q_1^T Q_2 = 0$ ce qui montre que : $A = QR = [Q_1 Q_2] = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$.

Les colonnes de Q_1 forment une base orthonormée de $Im(A)$, et les colonnes de Q_2 forment une base orthonormée de $Ker(A^T)$

4.3.3 Transformations de Householder

On peut interpréter la méthode de Gram-Schmidt comme une méthode de triangularisation de la matrice A, au même titre que la méthode de Gauss. Il est possible de réorganiser les calculs en construisant des transformations élémentaires orthogonales qui effectuent cette triangularisation colonne par colonne (ou élément par élément). Les symétries de Householder et les rotations de Givens sont des exemples simples et intéressants de telles transformations, car elles conduisent à des algorithmes numériquement plus stables que la méthode de Gram-Schmidt.

Définition 4.2 *Une matrice de Householder est une matrice carrée H qui s'écrit $H = I - 2P$, où P est la matrice de projection sur la droite engendrée par un vecteur v non nul.*

On vérifie (Fig. 4.6) que H représente une symétrie par rapport au sous-espace $v^\perp$.

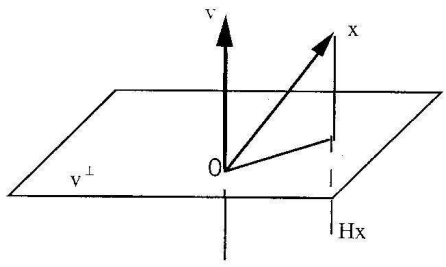

FIG. 4.6 – Transformation de Householder

Pour trianguliser la colonne k d'une matrice A, on effectuera la transformation sous la diagonale, pour conserver les éléments nuls construits aux itérations précédentes. Soit a_k le sous-vecteur associé aux $n-k+1$ dernières composantes de la colonne k de A. On détermine la symétrie H_k dans $\mathbb{R}^{n-k+1}$ telle que $H_k a_k = z_k$, où :

$$\begin{cases} z_{1k} = \| a_k \| \\ z_{ik} = 0, 2 \leq i \leq n-k+1 \end{cases}$$

En effet, H_k est orthogonale, ce qui implique que $\| z_k \| = \| a_k \|$. On obtient alors :

$$H_k = I - 2\frac{v_k v_k^T}{v_k^T v_k}$$

avec $v_k = a_k \pm \| a_k \| e_1$, et e_1 est le premier vecteur de la base canonique de $\mathbb{R}^{n-k+1}$.
En pratique, on choisira v_k de norme maximale.

Plus généralement, le théorème suivant montre qu'il est toujours possible de trouver une matrice de Householder permettant de transformer a_k en z_k, et plus généralement un vecteur quelconque en un vecteur colinéaire à un vecteur donné.

Théorème 4.2 *Soient f et e deux vecteurs non colinéaires de $\mathbb{R}^n$; avec $\|e\|_2 = 1$. Il est alors possible de trouver $u \in \mathbb{R}^n$ tel que*

1. $\|u\|_2 = 1$
2. $H(u)f = \alpha e$

DÉMONSTRATION :Remarquons tout d'abord que si $H(u)$ est une matrice de Householder, alors $H(u)f = f - 2u(u^T f)$ et $\|H(u)f\|_2 = \|f\|_2$.
Posons alors $| \alpha | = \|f\|_2$. On cherche alors u tel que $H(u)f = \alpha e$, soit

$$f - 2u(u^T f) = \alpha e$$
$$u = \frac{1}{2u^T f}(f - \alpha e)$$

Si $\beta = u^T f$, en multipliant à gauche par f^T :

$$2\beta^2 = \alpha^2 - \alpha f^T e$$

et β existe si $\alpha^2 - \alpha f^T e > 0$. Or l'inégalité de Cauchy-Schwarz nous donne

$$| f^T e | \leq \|f\|_2 \|e\|_2 = \|\alpha\|$$

et l'inégalité est de plus stricte par hypothèse (f et e non colinéaires). Ainsi :

$$u = \frac{1}{2\beta}(f - \alpha e)$$

répond à la question.

Remarque 4.1 *Si f et e sont colinéaires, $H = I$ ou $H = I - 2ee^T$ répondent à la question.*

La figure 4.7 présente un exemple de construction dans $\mathbb{R}^2$.

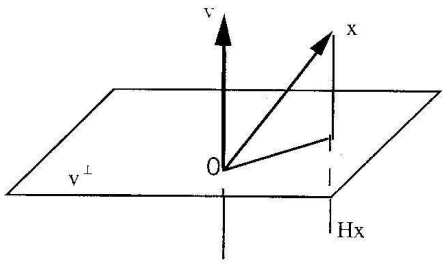

FIG. 4.7 – Transformation de Householder

Observons que la transformation orthogonale associée à la triangularisation de la colonne k dans $\mathbb{R}^n$ est la matrice Q_k représentée ci-dessous, *i.e.* la matrice identité de $\mathbb{R}^n$ où le bloc diagonal des $n - k + 1$ dernières colonnes a été remplacé par H_k.

$$Q = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & H_k \end{pmatrix}$$

4.4 Exercices

Exercice 4.1 Trouver la meilleure approximation, $\bar{x}$, au sens des moindres carrés, du système

$$3x = 10, \qquad 4x = 5.$$

Déterminer le carré de l'erreur e^2 et montrer que le vecteur erreur $(10 - 3\bar{x}\ 5 - 4\bar{x})^T$ est orthogonal à $(3\ 4)^T$

Exercice 4.2 Trouver la droite $f(t) = at + b$ qui approche le mieux (au sens des moindres carrés) les mesures :

$$f(0) = 0, \quad f(1) = 1, \quad f(3) = 2, \quad f(4) = 5.$$

Exercice 4.3 Soit

$$P = \frac{1}{\| v \|^2} v v^T,$$

la matrice de projection sur la droite de direction v.

Interpréter géométriquement la matrice $Q = I - 2P$. Montrer que $Q^2 = I$ et $Q^T = Q$ (*i.e.* Q est orthogonale). Soit $y = (y_1, y_2) \in \mathbb{R}^2$. Déterminer v pour que la seconde coordonnée de $z = Qy$ soit nulle.

Exercice 4.4 Projeter le vecteur $b = (0\ 3\ 0)^T$ sur les droites de directions respectives

$$a^1 = \begin{pmatrix} 2/3 \\ 2/3 \\ -1/3 \end{pmatrix}, \quad a^2 = \begin{pmatrix} -1/3 \\ 2/3 \\ 2/3 \end{pmatrix}.$$

Trouver la projection de b sur le plan engendré par a^1 et a^2.

Exercice 4.5 Soit V et W deux sous–espaces vectoriels de $\mathbb{R}^n$. Montrer que

$$\dim(V + W) = \dim V + \dim W - \dim V \cap W.$$

Illustrer ce résultat avec V et W respectivement ensembles des matrices triangulaires inférieures et supérieures.

Analyse spectrale

5.1 Introduction

L'analyse spectrale est l'étude des valeurs propres et des vecteurs propres d'une matrice carrée. Les valeurs propres d'une matrice carrée $A(n \times n)$ sont les n solutions dans $\mathbb{C}$ de l'équation caractéristique

$$det(\lambda I - A) = 0$$

Du point de vue de l'algèbre linéaire, cela signifie que le noyau de $\lambda I - A$ contient des vecteurs non nuls, appelés vecteurs propres associés à λ. Donc, si $x \neq 0$ est un vecteur propre de A associé à la valeur propre λ, on a $Ax = \lambda x$.

Le calcul des n solutions de l'équation caractéristique est très coûteux dès que $n > 2$ et le théorème d'Abel montre qu'on ne peut espérer le résoudre par des radicaux dès que $n > 4$. On recherchera donc des méthodes itératives qui permettent d'approcher ces racines et non de les calculer explicitement car, à la différence des méthodes de résolution de systèmes linéaires vues dans le chapitre 2, la convergence sera ici asymptotique. En fait, les méthodes qui seront présentées pour le calcul des valeurs propres sont utilisées pour extraire les racines d'un polynôme en passant par la matrice compagne :

$$\sum_{i=0}^{n-1} a_i t^i + t^n = 0$$

est le polynôme caractéristique de la matrice compagne

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

5.2 Intérêts de l'analyse spectrale

Considérons par exemple un système d'équations différentielles ordinaires :
Trouver les fonctions $v(t)$ et $w(t)$, pour $t \in \mathbb{R}^+$, telles que :

$$\begin{aligned} \frac{dv}{dt} &= 4v - 5w \\ \frac{dw}{dt} &= 2v - 3w \end{aligned}$$

avec comme conditions initiales $v(0) = 8, w(0) = 5$. Ce système peut être mis sous la forme vectorielle

$$\frac{du}{dt} = Au \tag{5.1}$$

avec :

$$u(t) = \begin{pmatrix} v(t) \\ w(t) \end{pmatrix}, u(0) = \begin{pmatrix} 8 \\ 5 \end{pmatrix}, A = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}$$

Sachant que la solution de l'équation $x' = ax$ en dimension 1 est $x(t) = x_0 e^{at}$, qui diverge si $a > 0$ et se stabilise asymptotiquement à zéro si $a < 0$, on cherchera des solutions particulières de la forme

$$A = \begin{cases} v(t) = & \tilde{v}e^{\lambda t} \\ w(t) = & \tilde{w}e^{\lambda t} \end{cases}$$

En remplaçant dans (5.1), on voit que

$$u(t) = e^{\lambda t}\begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix} \Rightarrow \frac{du}{dt} = \lambda e^{\lambda t}\begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix} = Au(t)$$

d'où

$$A\begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix} = \lambda\begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix}$$

et λ doit donc être une valeur propre de A associée au vecteur propre $\begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix}$. Dans le cas présent, on calcule aisément les éléments propres de A :

$$\lambda_1 = -1 \Rightarrow \begin{pmatrix} \tilde{v_1} \\ \tilde{w_1} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow u^1(t) = \begin{pmatrix} e^{-t} \\ e^{-t} \end{pmatrix}$$
$$\lambda_1 = 2 \Rightarrow \begin{pmatrix} \tilde{v_2} \\ \tilde{w_2} \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \end{pmatrix} \Rightarrow u^2(t) = \begin{pmatrix} 5e^{2t} \\ 2e^{2t} \end{pmatrix}$$

La linéarité du système implique que toute combinaison linéaire des solutions particulières u^1, u^2 est solution de l'équation différentielle. La solution générale s'écrit donc :

$$u(t) = \alpha u^1(t) + \beta u^2(t)$$

où $\alpha = 3$ et $\beta = 1$ sont déterminées par les conditions initiales.

5.3 Résultats généraux

On donnera ici quelques propriétés d'intérêt surtout pratique concernant les valeurs propres et les vecteurs propres de certaines matrices. Les démonstrations sont omises, et on se référera à l'ouvrage de G. Strang [10], pour plus de détails.

Une matrice carrée A de dimension n à coefficients complexes ou réels possède n valeurs propres non nécessairement distinctes dans $\mathbb{C}$. L'ensemble de ces valeurs propres est le **spectre** de A, et le nombre de fois où apparaît une valeur propre λ dans ce spectre est appelé sa **multiplicité**. Le **rayon spectral**, noté $\rho(A)$, est le plus grand module des valeurs propres de A. La somme des valeurs propres est égale à la **trace** de la matrice :

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii}$$

et on en déduit donc que si une matrice réelle possède une valeur propre complexe, son conjugué est aussi valeur propre. De même, le produit des valeurs propres est égal au **déterminant** de A :

$$\prod_{i=1}^{n} \lambda_i = det(A)$$

Attention ! Il n'est pas possible en général d'obtenir valeurs et vecteurs propres de la somme ou du produit de deux matrices dont on connaît le spectre, mais certains cas particuliers où les vecteurs propres sont les mêmes permettent de conclure :

- si λ est valeur propre de A, λ^k est valeur propre de A^k (en particulier, si A est inversible, λ^{-1} est valeur propre de l'inverse)
- si λ est valeur propre de A, $\lambda + \alpha$ est valeur propre de $A + \alpha I$.

Si λ est une valeur propre de A, le système $(A - \lambda I)x = 0$ possède des solutions non nulles appelées **vecteurs propres** de A associés à λ. On appelle **sous-espace propre** associé à une valeur propre λ le noyau de $A - \lambda I$. Sa dimension est donc au moins égale à 1.

Lorsque les valeurs propres sont distinctes, les vecteurs propres sont linéairement indépendants. L'implication réciproque est fausse, on pensera au cas de la matrice identité pour s'en convaincre.

Si les n vecteurs propres $x_i, 1 \leq i \leq n$ peuvent être choisis linéairement indépendants, ils forment une matrice $X = [x_i]$ non singulière, et on a :

$$X^{-1}AX = \Lambda = \text{diag}\{\lambda_1 \cdots \lambda_n\}$$

En effet, AX est la matrice dont les colonnes sont les vecteurs $Ax_i = \lambda x_i$ qui est bien égale à $X\Lambda$. On dit dans ce cas que A **est diagonalisable** et cette propriété ne peut s'écrire qu'avec la matrice des vecteurs propres. Elle caractérise le fait que X a pour colonnes des vecteurs propres et que ces vecteurs propres sont linéairement indépendants.

Observation : il existe des matrices qui ne sont pas diagonalisables (on les appelle matrices **défectives**). Elles satisfont aux deux conditions suivantes :

- il existe des valeurs propres multiples
- la dimension du sous-espace propre associé est strictement inférieure à la multiplicité de la valeur propre.

Attention, le seul fait de l'existence de valeurs propres multiples ne suffit pas à impliquer que la matrice soit défective (*cf.* la matrice identité par exemple). Un exemple typique de matrice défective est $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, qui possède la valeur propre double 0. Les vecteurs propres sont de la forme $x = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$, ils engendrent donc un sous-espace de dimension 1. A n'est donc pas diagonalisable.

Dans certains cas, il est relativement facile d'inférer sur la valeur ou la nature des valeurs propres d'une matrice A :

- si A est diagonale, les valeurs propres sont les éléments diagonaux
- si A est triangulaire, les valeurs propres sont également les éléments diagonaux
- si A est non singulière, les valeurs propres sont toutes différentes de 0
- si A est orthogonale, les valeurs propres ont pour module 1 et il est possible de choisir une base de vecteurs propres orthonormés
- si A est symétrique, ses valeurs propres sont réelles. Les vecteurs propres associés à des valeurs propres distinctes sont alors orthogonaux.

Le dernier point implique en particulier que toute matrice réelle symétrique est diagonalisable par une matrice orthogonale : $\Lambda = Q^T AQ$,où Q est formée par n vecteurs propres orthonormés. On a alors la **factorisation spectrale** d'une matrice symétrique réelle :

$$A = Q\Lambda Q^T$$

En résumé de ce qui précède :
- **l'inversibilité est liée aux valeurs propres**
- **la diagonalisabilité est liée aux vecteurs propres**

5.4 Similitudes

L'objectif est encore une fois de transformer une matrice par des transformations simples en une matrice dont on connaît les valeurs propres, c'est-à-dire, une matrice triangulaire ou diagonale. Les transformations qui maintiennent le spectre d'une matrice sont des similitudes.

5.4.1 Définition et propriétés

Définition 5.1 *Deux matrices carrées A et B sont dites* ***semblables*** *s'il existe une matrice S non singulière telle que*

$$B = S^{-1}AS$$

La transformation de A vers B est une **similitude**. En l'écrivant sous la forme $AS = SB$, on retrouve une généralisation de la définition des valeurs propres et des vecteurs propres. On a d'ailleurs le résultat fondamental :

Proposition 5.1 *Deux matrices semblables ont les mêmes valeurs propres*

DÉMONSTRATION : soit x un vecteur propre de A associé à la valeur propre λ. On a donc $Ax = \lambda x$, qui s'écrit $SBS^{-1}x = \lambda x$, ce qui veut dire que λ est valeur propre de B associé au vecteur propre $S^{-1}x$. □

L'intérêt de ces transformations est double :
- les valeurs propres sont inchangées
- en supposant les vecteurs propres linéairement indépendants, la similitude associée à la matrice X dont les colonnes sont les vecteurs propres transforme A en une matrice diagonale dont les éléments diagonaux sont les valeurs propres de A : $X^{-1}AX = \Lambda$.

Montrons maintenant sur un exemple que deux matrices semblables représentent la même transformation linéaire sur deux bases différentes : soit P la matrice de projection dans $\mathbb{R}^2$ sur la droite L d'angle θ :

$$P = \begin{bmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{bmatrix} = uu^T \quad \text{avec} \quad u = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}$$

Si on fait maintenant tourner la base canonique orthonormée d'un angle θ, la projection devient maintenant une projection sur l'axe horizontal et s'écrit

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Pour passer de P à Q, on utilise la matrice de rotation d'angle θ

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

Le changement de base se traduit par : si x est le vecteur de coordonnées dans la base de départ, et u le vecteur de coordonnées dans la base d'arrivée, on a

$$x = Ru$$

Soit y la projection de x sur L dans la base de départ, et v ses coordonnées dans la base d'arrivée. Alors

$$\begin{aligned} y &= Px \\ Rv &= PRu \\ \Rightarrow v &= R^{-1}PRu = Qu \end{aligned}$$

et donc

$$Q = R^{-1}PR$$

et donc P et Q ont les mêmes valeurs propres. Comme Q est diagonale, on retrouve ici le résultat connu : toute matrice de projection sur une droite dans $\mathbb{R}^2$ a pour valeurs propres 1 et 0, et pour vecteurs propres associés les colonnes de la matrice de rotation R, c'est-à-dire les directions de L et $L^\perp$ respectivement.

5.4.2 Théorème de Gershgorin

Une alternative pratique aux algorithmes de calcul approché des valeurs propres que nous décrirons plus loin est le théorème suivant qui permet de localiser les valeurs propres dans des disques, dits disques de Gershgorin, du plan complexe.

Théorème 5.1 *Si on représente une matrice A (ou toute matrice semblable à A) sous la forme $A = diag\{d_1 \cdots d_n\} + F$, où F est une matrice de diagonale nulle, alors le spectre de A est contenu dans l'union des disques $D_i, 1 \leq i \leq n$ du plan complexe, tels que*

$$D_i = \left\{ z \in \mathbb{C}, |z - d_i| \leq \sum_{j=1}^{n} |f_{ij}| \right\}$$

DÉMONSTRATION : soit λ une valeur propre supposée différente des $d_i, 1 \leq i \leq n$. Alors $(D - \lambda I) + F$ est singulière et d'après le théorème 3.1, on a la majoration $\|(D - \lambda I)^{-1} + F\| \geq 1$. Le choix de la norme $\|.\|_\infty$ fournit le résultat cherché. □

Une application intéressante de ce résultat est l'estimation des valeurs propres d'une matrice obtenue en perturbant une matrice dont on connaît le spectre.

Exemple 5.1

$$A = \begin{bmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{bmatrix}$$

dont les valeurs propres sont situées dans les disques suivants (Figure 5.1)

$$\begin{aligned} D_1 &= \{z \in \mathbb{C}, |z - 1| \leq 0.2\} \\ D_2 &= \{z \in \mathbb{C}, |z - 2| \leq 0.4\} \\ D_3 &= \{z \in \mathbb{C}, |z - 3| \leq 0.2\} \end{aligned}$$

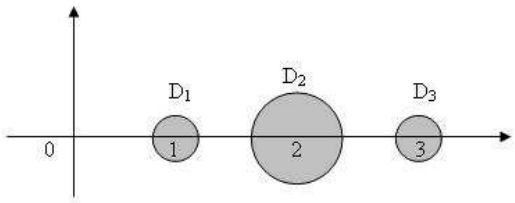

FIG. 5.1 – disques de Gershgorin

5.5 Calcul des valeurs propres d'une matrice symétrique : méthode de Jacobi

L'intérêt principal des matrices réelles symétriques est qu'il existe une base de vecteurs propres orthonormés. On peut donc la diagonaliser par une transformation orthogonale : soient A une telle matrice et Q la matrice orthogonale dont les colonnes sont les vecteurs propres de A, alors $Q^T AQ = diag\{\lambda_1 \cdots \lambda_n\}$.

La méthode de Jacobi est une méthode d'élimination symétrique itérative utilisant des similitudes orthogonales :

- la matrice transformée tend vers une matrice diagonale
- le produit des transformations orthogonales tend vers la matrice des vecteurs propres.

5.5.1 Principe d'élimination symétrique

Ce principe, basé sur des rotations successives, sera illustré tout d'abord sur une matrice (2×2) : soit A la sous-matrice (2×2) symétrique

$$A = \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \quad a_{pq} \neq 0.$$

Soit R la matrice de rotation d'angle $-\theta$:

$$R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

Pour éliminer l'élément a_{pq} non diagonal, on détermine θ tel que

$$R^T AR = \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}$$

Le calcul donne

$$\operatorname{cotg}(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}}$$

Considérons maintenant une matrice symétrique $n \times n$ A telle que l'élément a_{pq} soit non nul. La transformation orthogonale Ω telle que $A' = \Omega^T A \Omega$, avec $a'_{pq} = a'_{qp} = 0$ est :

$$\Omega = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & \cos\theta & \sin\theta & & \\ & & -\sin\theta & \cos\theta & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ p \\ q \\ \\ \\ \end{matrix}$$

On vérifie que seules les lignes et colonnes p et q de A sont modifiées. En posant $c = \cos\theta$, $s = \sin\theta$, $t = \tan\theta$, la mise à jour s'écrit :

$$\begin{aligned} a'_{pj} &= ca_{pj} - sa_{qj}, j \neq p, q \\ a'_{qj} &= ca_{qj} + sa_{pj}, j \neq p, q \\ a'_{pp} &= a_{pp} - ta_{pq} \\ a'_{qq} &= a_{qq} + ta_{pq} \end{aligned}$$

Un choix classique pour p et q est celui qui permet d'éliminer l'élément non diagonal de plus grand module : $|a_{pq}| = \max_{i \neq j} |a_{ij}|$

5.5.2 Convergence

Soit $\{A_k\}$ la suite de matrices engendrée par l'algorithme en éliminant à chaque itération un élément non diagonal non nul ($A_0 = A$) et son symétrique. Comme la transformation est orthogonale, la norme de Frobenius de la matrice est conservée (*cf* chapitre 3) :

$$\sum_{i,j} a'^2_{ij} = \sum_{i,j} a^2_{ij}$$

Mais ce même résultat est vrai pour la matrice (2×2) transformée ci-dessus. Donc :

$$a'^2_{pp} + a'^2_{qq} = a^2_{pp} + a^2_{qq} + 2a^2_{pq}$$

Comme seules les lignes p et q sont modifiées, ces résultats impliquent que la somme des carrés des éléments diagonaux augmente strictement de la valeur $2a^2_{pq}$, et que la somme des carrés des éléments non diagonaux diminue strictement de la même quantité.
On démontre alors le théorème suivant :

Théorème 5.2 *La suite de matrices $\{A_k\}$ engendrée par la méthode de Jacobi classique converge vers une matrice diagonale contenant toutes les valeurs propres de A sur la diagonale*

DÉMONSTRATION : voir par exemple Ciarlet [2].

On peut également démontrer que, si les valeurs propres sont toutes distinctes, la suite des matrices $Q_k = \Omega_1 \cdots \Omega_k$ converge vers la matrice orthogonale contenant les vecteurs propres.

Attention ! : un élément annulé peut redevenir non nul aux itérations suivantes.

5.5.3 Observations

- Le tri du plus grand élément parmi $n(n-1)/2$ coûtant relativement cher,on lui préfère d'autres stratégies plus économiques (balayage cyclique ou choix avec seuil)

– La méthode de Jacobi présente d'excellentes performances pour des matrices pleines de faible dimension (typiquement inférieure à 100). Dans le cas général, on lui préférera la méthode QR (*cf.* ci-après).

5.6 Calcul de certains vecteurs propres : les puissances itérées

5.6.1 Quotient de Rayleigh

Pour une matrice symétrique A, le quotient de Rayleigh est le rapport défini pour tout vecteur $x \neq 0$ par :

$$\rho_A(x) = \frac{x^T A x}{x^T x}$$

On vérifie immédiatement que si x est vecteur propre, le quotient de Rayleigh fournit la valeur propre associée : en effet $Ax = \lambda x \Rightarrow x^T Ax = \lambda x^T x$.

Si λ_1 et λ_n sont respectivement la plus petite et la plus grande valeur propre de A, et x^1,x^n les vecteurs propres associés, on a également les résultats suivants :

$$\lambda_1 = \rho_A(x^1) = \min_{x \in \mathbb{R}^n} \{\rho_A(x)\}$$
$$\lambda_n = \rho_A(x^n) = \max_{x \in \mathbb{R}^n} \{\rho_A(x)\}$$

De plus, si les valeurs propres sont rangées dans l'ordre croissant, on a

$$\lambda_1 = \min_{S_i} \{\max_{x \in S_i} \{\rho_A(x)\}\}$$
$$\lambda_n = \max_{S_{i-1}} \{ \min_{x \in S_{i-1}^{\perp}} \{\rho_A(x)\}\}$$

où S_i est un sous-espace quelconque de dimension i.

Le sous-espace S_i pour lequel le quotient de Rayleigh est maximum est le sous-espace propre associé aux i premières valeurs propres. Le sous-espace pour lequel il est minimum est orthogonal au sous-espace propre associé aux $i-1$ premières valeurs propres. C'est donc le sous-espace engendré par les $n - i + 1$ vecteurs propres associés à $\{\lambda_i \cdots \lambda_n\}$.

5.6.2 Méthode des puissances itérées

La méthode des puissances itérées permet de calculer le vecteur propre associé à la plus grande valeur propre.
Supposons A symétrique de valeurs propres ordonnées selon

$$|\lambda_1| \leq \cdots |\lambda_{n-1}| < |\lambda_n|$$

On considère l'itération suivante définie à partir d'un vecteur initial q_0 donné, tel que $\|q_0\| = 1$, et q_0 n'est pas orthogonal à v^n, le vecteur propre associé à la plus grande valeur propre isolée λ_n :

$$x_{k+1} = Aq_k$$
$$q_{k+1} = \frac{x_{k+1}}{\|x_{k+1}\|}$$

Par récurrence, on montre que

$$q_k = \frac{A^k q_0}{\|A^k q_0\|}$$

et comme les vecteurs propres $\{v^1 \cdots v^n\}$ forment une base de $\mathbb{R}^n$, on peut écrire

$$q_0 = \sum_{i=1}^{n} \alpha_i v^i, \quad \alpha_n \neq 0$$

et

$$A^k q_0 = \alpha_n \lambda_n^k \left(v^n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n} \left(\frac{\lambda_i}{\lambda_n} \right)^k v^i \right)$$

Lorsque $k \to \infty$, les rapports $\left(\frac{\lambda_i}{\lambda_n}\right)^k$ tendent vers 0 pour $i \neq n$, ce qui signifie que la suite des itérés $\{q_k\}$ converge vers le vecteur propre v^n. On peut montrer de plus que $\|Aq_k\|$ tend vers λ_n et que la convergence est linéaire de taux $\left|\frac{\lambda_{n-1}}{\lambda_n}\right|$ si $\alpha_{n-1} \neq 0$.

5.6.3 Méthode des puissances inverses

Pour les mêmes raisons, l'itération

$$\begin{aligned} Ax_{k+1} &= q_k \\ q_{k+1} &= \frac{x_{k+1}}{\|x_{k+1}\|} \end{aligned}$$

avec $\|q_0\| = 1$, et q_0 n'est pas orthogonal à v^1, converge vers la direction du vecteur propre associé à la plus petite valeur propre.

5.6.4 Remarques

1. Accélération par **décalage** : la matrice $A + \alpha I$ a les mêmes vecteurs propres que A et ses valeurs propres sont décalées de la quantité α. La méthode des puissances itérées inverses converge d'autant plus vite que les rapports $\left(\frac{\lambda_1}{\lambda_2}\right)^k$ tendent rapidement vers 0. On a donc intérêt à ce que $\|\lambda_1\|$ soit le plus proche possible de 0, et de plus, la méthode sera d'autant plus rapide que l'écart entre les deux plus petites valeurs propres se creuse. La technique du décalage consiste donc à remplacer A par $A + \alpha I$, avec $\alpha \approx -\lambda_1$. Plus l'estimation de λ_1 sera précise, plus la convergence sera rapide. Toutefois, il faut que $\alpha \neq -\lambda_1$ pour éviter que la matrice ne devienne singulière
2. Technique de **déflation** : la méthode des puissances itérées peut être étendue pour permettre le calcul de toutes les valeurs propres d'une matrice symétrique. Supposons en effet calculée la plus grande valeur propre λ_n ainsi qu'un vecteur propre associé v^n. Soit P_n la la matrice de projection orthogonale sur l'hyperplan $(v^n)^\perp$. La matrice $P_n A$ possède les mêmes vecteurs propres que A et les mêmes valeurs propres à l'exception de λ_n qui est remplacée par 0. L'application de la méthode des puissances itérées à $P_n A$ permettra donc de calculer la deuxième plus grande valeur propre de A. Cette technique, dite de déflation, permet théoriquement de calculer toutes les valeurs propres de A. Elle est toutefois numériquement instable sans précautions, et on lui préférera généralement la méthode des puissances groupées présentée ci-après.

5.7 Puissances groupées et méthode QR

On peut généraliser la méthode des puissances itérées à des itérations dites groupées, permettant d'identifier les vecteurs propres associés aux p plus grandes valeurs propres. A chaque itération,

on applique la transformation A à chacun des vecteurs orthonormés, et on orthogonalise (par Gram-Schmidt ou Householder, *cf.* chapitre 4) le système résultant.

5.7.1 Itération des puissances itérées

Soit $\{q_1^{(k)} \cdots q_p^{(k)}\}$ un système de p vecteurs orthonormés obtenu à l'itération k
- construire le système $\{Aq_1^{(k)} \cdots Aq_p^{(k)}\}$
- calculer les quotients de Rayleigh : $\lambda_i = {q_i^{(k)}}^T Aq_i^{(k)}, \quad 1 \leq i \leq p$
- tester la convergence : $\max_{1 \leq i \leq p} \|Aq_i^{(k)} - \lambda_i q_i^{(k)}\| < Tol$
- orthogonaliser $\{Aq_1^{(k)} \cdots Aq_p^{(k)}\} \rightarrow \{q_1^{(k+1)} \cdots q_p^{(k+1)}\}$

5.7.2 Méthode QR

La méthode QR correspond au cas $p = n$: soit Q_k la matrice $(n \times n)$ dont les colonnes sont les vecteurs orthogonaux $q_i^{(k)}$ et soit $A_k = Q_k^T A Q_k$ la représentation de A sur la base des $\{q_i^{(k)}\}$. On peut écrire alors (Gram-Schmidt) :

$$AQ_k = Q_{k+1} R_{k+1}$$

où R_{k+1} est triangulaire supérieure. Et ainsi

$$A_k = Q_k^T Q_{k+1} R_{k+1} = Q R_{k+1}$$
$$A_{k+1} = Q_{k+1}^T A Q_{k+1} = R_{k+1} Q_k^T Q_{k+1} = R_{k+1} Q$$

En résumé, Q est la matrice orthogonale qui permet de trianguläriser A_k et on écrit Q_{k+1} en inversant le produit QR. La méthode QR prend alors la forme particulièrement simple suivante :
- $A_0 = A$
- $A_k = QR$ (par Gram-Schmidt par exemple)
- $A_{k+1} = RQ$
- test sur le plus grand élément non diagonal

Observons que $A_{k+1} = Q^T A_k Q$, ce qui implique que les matrices sont toutes semblables et que la suite des matrices Q converge vers la matrice des vecteurs propres. La matrice A_k converge vers une matrice diagonale où les valeurs propres sont rangées dans l'ordre décroissant (Figure 5.2).

Les performances sont considérablement améliorées si on intègre deux modifications :
- **décalage** de la matrice en prenant comme approximation de la plus petite valeur propre l'élément $(n \times n)$ de A_k
- transformation préalable de A en une matrice **tridiagonale**. Les matrices A_k restent tridiagonales et l'orthogonalisation s'effectue en $O(n)$ flops.

5.8 Exercices

Exercice 5.1 1.— Soit a et b deux réels et D une matrice carrée. Montrer que si λ est valeur propre de D, alors $a\lambda + b$ est une valeur propre de $aD + b\mathbb{I}$, où $\mathbb{I}$ est la matrice identité.

2.— Soit D la matrice carrée tridiagonale définie par

$$d_{ii} = 0, \quad i = 1, \ldots, n$$
$$d_{i,i+1} = d_{i+1,i} = 1, \quad i = 1, \ldots, n-1$$
$$d_{ij} = 0, \quad \text{ailleurs.}$$

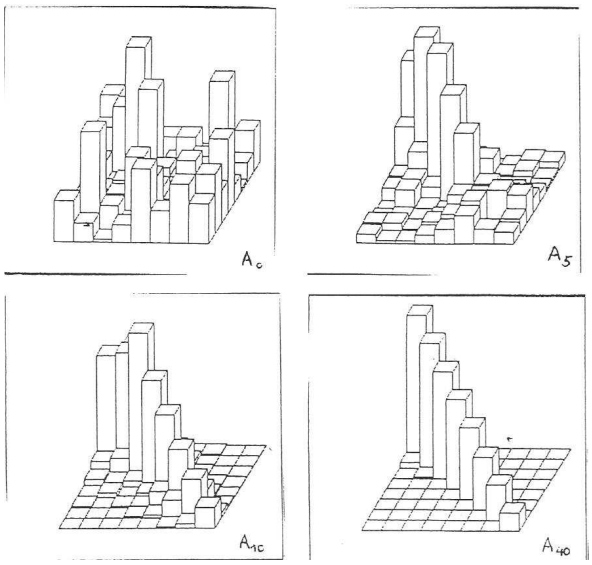

FIG. 5.2 – diagonalisation d'une matrice 8×8 par QR d'après [6]

Pour $j = 1, \ldots, n$, soit $\mathbf{x}^{(j)}$ le vecteur de $\mathbb{R}^n$ dont la i–ème composante est

$$x_i^{(j)} = \sin 2i\frac{j\pi}{n+1}.$$

Montrer que $\mathbf{x}^{(j)}$ est un vecteur propre de D associée à la valeur propre

$$\lambda_j = 2\cos 2\frac{j\pi}{n+1}.$$

3.— En déduire les vecteurs propres et les valeurs propres de la matrice 4×4 suivante

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

Exercice 5.2 (Suite de Fibonacci) On considère la suite 1, 1, 2, 3, 5, 8, 13, ... définie par

$$\begin{aligned} &u_0 = 1, \quad u_1 = 1, \\ &u_k = u_{k-1} + u_{k-2}, \quad k \geq 2. \end{aligned}$$

Calculer une valeur approchée de u_k, en utilisant les puissances itérées.

Exercice 5.3 Soit a et b deux vecteurs non colinéaires de $\mathbb{R}^n$. Déterminer les vecteurs propres et les valeurs propres de la matrice $n \times n$

$$A = aa^T + bb^T.$$

Exercice 5.4 (Convergence de la méthode de Jacobi) Soit $A' = A + E$, la somme de deux matrices symétriques. Les valeurs propres de ces trois matrices sont notées

$$\lambda'_1 \geq \lambda'_2 \geq \cdots \geq \lambda'_n, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \quad \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n.$$

1.— Montrer les propriétés suivantes pour tout $i = 1, \ldots, n$

(i) $\lambda_i + \mu_n \leq \lambda'_i \leq \lambda_i + \mu_1$

(ii) $|\lambda'_i - \lambda_i| \leq \| E \|$, quelle que soit la norme matricielle $\| \cdot \|$.

2.— Soit $E^{(k)} = A^{(k)} - \mathrm{diag}(a_{ii}^{(k)})$, où les $A^{(k)}$ sont les matrices engendrées par la méthode de Jacobi. En utilisant la norme $\| E \|_F = (\mathrm{trace}(E^T E))^{1/2}$, montrer que $E^{(k)}$ tend vers la matrice nulle, lorsque $k \longmapsto \infty$.

3.— En déduire le théorème de convergence de la méthode de Jacobi, i.e.

(i) $a_{ij}^{(k)} \longrightarrow 0$, pour $i \neq j$

(ii) chaque $a_{ii}^{(k)} \longrightarrow \lambda_i$, où λ_i est une valeur propre de A.

Exercice 5.5 Supposons calculée la plus grande valeur propre λ_1 d'une matrice symétrique A et le vecteur propre associé v_1.

1.— Quelle est la matrice P_1 de projection orthogonale sur $v_1^{\perp}$.

2.— Montrer que $P_1 A$ a les mêmes vecteurs propres que A, à l'exception de v_1. Quelles sont les valeurs propres de $P_1 A$. En déduire une première méthode de calcul des plus grandes valeurs propres de A.

Matrices définies positives

6.1 Introduction

6.1.1 Définition

Bien que les notions développées dans ce chapitre soient très générales, on ne considérera ici que le cas des **matrices symétriques**.

Définition 6.1 (Matrice définie positive) *Une matrice carrée A de dimension n est* ***définie positive*** *si et seulement si :*

$$(\forall x \in \mathbb{R}^n, x \neq 0) \quad x^T A x > 0$$

Lorsque l'inégalité est large, on parle de matrice **semi-définie positive**. L'inégalité inverse ($<$ ou $\leq$) permet de définir les notions de matrices **définies négatives** et **semi-définies négatives**.

6.1.2 Exemples

1. *Formes quadratiques :* soit $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ une forme **bilinéaire** définie par $f(x, y) = x^T A y$, où A est une matrice $(n \times n)$. Lorsque A est symétrique, on lui associe la fonction q, appelée **forme quadratique**, de $\mathbb{R}^n$ dans $\mathbb{R}$, définie par :

   $$(\forall x \in \mathbb{R}^n) \quad q(x) = x^T A x$$

   Si A est définie positive, q est à valeurs positives et exhibe donc un unique point de minimum en $x = 0$, où elle s'annule. Par exemple, c'est le cas de la fonction carré de la norme

   $$q(x) = \|x\|^2$$

   donc la matrice identité I est définie positive. On verra dans la suite que q est une fonction **convexe**.
   Soit dans $\mathbb{R}^2$ la fonction

   $$q(x) = x_1^2 + 2x_2^2 + 2x_1 x_2 = x^T A x, \quad \text{avec} \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

   On observe sur cet exemple que toute forme quadratique $q(x) = \sum_i \sum_j q_{ij} x_i x_j$ peut s'écrire de manière unique sous la forme $x^T A x$, avec A symétrique. Il suffit de partager également le

terme croisé $q_{ij}x_ix_j$ avec $a_{ij} = a_{ji} = \frac{q_{ij}}{2}, \quad i \neq j$.
On peut mettre q sous la forme :

$$q(x) = (x_1 + x_2)^2 + x_2^2 \geq 0$$

qui ne s'annule qu'en (0,0). Donc la forme q est définie positive ainsi que la matrice A associée.

2. *Matrice tridiagonale de l'exemple des ressorts du TP2 :*

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & & \\ & & & -1 \\ & & -1 & 2 \end{pmatrix} \Rightarrow q(x) = \sum_{i=1}^{n} 2x_i^2 - 2\sum_{i=1}^{n-1} x_i x_{i+1}.$$

Or, q peut être mise sous la forme d'une somme de carrés :

$$q(x) = x_1^2 + (x_1 - x_2)^2 + \cdots + (x_{n-1} - x_n)^2 + x_n^2$$

la matrice étant non singulière, elle est définie positive. On retrouve ici au coefficient près l'expression de l'énergie mécanique du système des ressorts.

3. *Moindres carrés* : soit B une matrice $(p \times n)$ telle que $rang(B) = n$. Alors la matrice $A = B^T B$ est définie positive. On retrouve la matrice du système des équations normales (*cf.* chapitre 4). Si B n'est pas de rang plein, A est seulement semi-définie positive.

6.2 Caractérisation des matrices définies positives

6.2.1 Mise sous forme de carrés

Exemple :

$$\begin{aligned} q(x) &= 10x_1^2 + 5x_2^2 + 2x_3^2 + 4x_1x_2 - 2x_2x_3 + 6x_3x_1 \\ &= (x_1 + 2x_2)^2 + (x_2 - x_3)^2 + (3x_1 + x_3)^2 \\ &= x^T A x \quad \text{avec} \quad A = \begin{pmatrix} 10 & 2 & 3 \\ 2 & 5 & -1 \\ 3 & -1 & 2 \end{pmatrix} \end{aligned}$$

Donc A est semi-définie positive.

Attention ! Pour montrer que la matrice est définie positive, il faut montrer qu'elle est de plus non singulière (ce qui est le cas ci-dessus).

6.2.2 Valeurs propres positives

Théorème 6.1 *Une matrice symétrique définie positive possède des valeurs propres strictement positives.*

DÉMONSTRATION :

1. Supposons A symétrique définie positive. Soit λ_i une valeur propre associée au vecteur propre x_i. On a donc $Ax_i = \lambda_i x_i$, et $x_i^T A x_i = \lambda_i x_i^T x_i = \lambda_i \|x_i\|^2 > 0$

2. Réciproquement, si toutes les valeurs propres sont positives, on a pour tout vecteur propre x_i $x_i^T A x_i > 0$. Ces derniers formant une base orthonormée (*cf.* chapitre 5), on écrit pour tout $x \neq 0$

$$x^T A x = \left(\sum_{i=1}^{n} \alpha_i x_i\right)^T A \left(\sum_{i=1}^{n} \alpha_i x_i\right) = \sum_{i=1}^{n} \alpha_i^2 \lambda_i > 0$$

Par exemple, les valeurs propres de la matrice du système des ressorts sont :

$$(\forall j \in \{1 \cdots n\}) \quad \lambda_i = 2 - 2cos\frac{2j\pi}{n+1} > 0$$

6.2.3 Par les sous-matrices carrés symétriques par rapport à la diagonale

Toutes les sous-matrices carrés symétriques par rapport à la diagonale d'une matrice définie positive sont définies positives. En effet, soit A_k une telle sous-matrice, constituée des k premières lignes et colonnes d'une matrice A définie positive. Choisissons un vecteur x dont les $n-k$ dernières composantes sont nulles :

$$x^T A x = \begin{pmatrix} x_k^T & 0 \end{pmatrix} \begin{pmatrix} A_k & * \\ * & * \end{pmatrix} \begin{pmatrix} x_k \\ 0 \end{pmatrix} = x_k^T A_k x_k$$

ce qui démontre la propriété.

En particulier, une condition **nécessaire** pour qu'une matrice soit définie positive est que les éléments diagonaux soient positifs.

6.2.4 Matrice à dominance diagonale

Définition 6.2 (Matrice à diagonale dominante) *Une matrice A est dite à **diagonale strictement dominante** si*

$$(\forall i \in \{1 \cdots n\}) \quad a_{ii} > \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

Théorème 6.2 *Les matrices à diagonale strictement dominante sont définies positives*

DÉMONSTRATION :On utilise d'abord la symétrie de A et le fait que

$$2|x_i||x_j| \leq x_i^2 + x_j^2$$

pour écrire :

$$\begin{aligned} -\sum_{i\neq j} a_{ij} x_i x_j &\leq \sum_{i<j} |a_{ij}||x_i||x_j| \\ &\leq \sum_{i<j} |a_{ij}| x_i^2 + \sum_{i<j} |a_{ij}| x_j^2 \end{aligned}$$

Le dernier second membre peut s'écrire $\sum_i \left(\sum_{j\neq i} |a_{ij}| x_i^2\right)$, et la propriété de dominance diagonale implique que ce terme est strictement inférieur à $\sum_i a_{ii} x_i^2$, d'où

$$-\sum_{i\neq j} a_{ij} x_i x_j < \sum_i a_{ii} x_i^2$$

□

6.2.5 Pivots positifs

Théorème 6.3 *Une matrice définie positive sera triangularisée sans permutations avec des pivots positifs.*

DÉMONSTRATION :Soit A une matrice définie positive. Toutes les sous-matrices principales sont donc définies positives (donc non singulières). Cela implique qu'il existe une factorisation $A = LDL^T$, avec D diagonale et L triangulaire inférieure. Donc $D = L^{-1}A(L^T)^{-1}$ et chaque élément de D s'écrit x^TAx, où x^T est une ligne de L^{-1}. Les pivots sont donc positifs. □

Ce résultat, certainement d'un grand intérêt pratique, vient du résultat plus général suivant : si on écrit une forme quadratique par rapport à une nouvelle base, on réalise une **congruence**. Prenons le changement de variable $x = Cy$, C non singulière. La forme quadratique $q(x) = x^TAx$ devient $q(y) = y^TC^TACy$. On a alors, sans démonstration :

Théorème 6.4 (Loi d'inertie de Sylvester) *C^TAC a le même nombre de valeurs propres positives, négatives ou nulles que A.*

Quand on pivote une matrice A symétrique sans permutation, on la met sous la forme $A = LDL$, avec D diagonale contenant tous les pivots. Donc, les **signes des pivots** coïncident avec ceux des valeurs propres (mais pas leurs valeurs).

6.2.6 Racine carrée d'une matrice

Si A est une matrice définie positive, il existe une matrice non singulière R telle que

$$A = R^TR.$$

On appelle parfois R la racine carrée de A. Le choix de R n'est pas unique :
- avec $A = LDL^T$, on obtient $R = D^{1/2}L^T$
- soit Q la matrice des vecteurs propres, $Q^TAQ = L$, donc $R = L^{1/2}Q^T$
- $R = Q'L^{1/2}Q^T$ avec Q' orthogonale

Remarque 6.1 *Une fonction quadratique définie positive peut être considérée comme une extension de la norme euclidienne après changement de variable, c'est donc aussi une norme :*

$$Ry = x \Rightarrow \|x\|^2 = y^TR^TRy.$$

6.3 Méthode de Cholesky

L'adaptation de la méthode de Gauss aux matrices symétriques définies positives conduit à la factorisation de Cholesky.

Comme il a été observé au paragraphe précédent, la factorisation $A = LDL^T = R^TR$ s'écrit avec R triangulaire supérieure. L'algorithme ci-dessous remplace les éléments a_{ij} par les éléments r_{ij} :

Cet algorithme demande $\mathcal{O}\left(\frac{n^3}{6}\right)$ flops. Les éléments de R satisfont de plus

$$r_{ij}^2 \leq \sum_{k=1}^{i} r_{ik}^2 = a_{ii}, \quad \forall i, j = 1, \ldots, n.$$

Donc tous les éléments de R sont bornés en module par les éléments diagonaux de A, ce qui contribue à la stabilité numérique de la méthode (sans la rendre totalement insensible au mauvais conditionnement).

Algorithme 5 Méthode de Cholesky

Pour $k \in \{1 \cdots n\}$ **Faire**

$$a_{kk} \leftarrow \left[a_{kk} - \sum_{p=1}^{k-1} a_{kp}^2 \right]^{1/2}$$

Pour $i \in \{k+1 \cdots n\}$ **Faire**

$$a_{ik} \leftarrow \frac{1}{a_{kk}} \left[a_{ik} - \sum_{p=1}^{k-1} a_{ip} a_{kp} \right]$$

Fin Pour

Fin Pour

6.4 Fonctions quadratiques convexes

Une **fonction quadratique** de $\mathbb{R}^n$ dans $\mathbb{R}$ peut s'écrire sous la forme générale

$$f(x) = x^T A x + a^T x + b$$

où A est une matrice symétrique, $a \in \mathbb{R}^n, b \in \mathbb{R}$. Ainsi, une fonction quadratique est l'extension à $\mathbb{R}^n$ d'un polynôme du second degré, et on peut la représenter comme la somme d'une forme quadratique et d'une fonction affine (cf. figure 6.1)

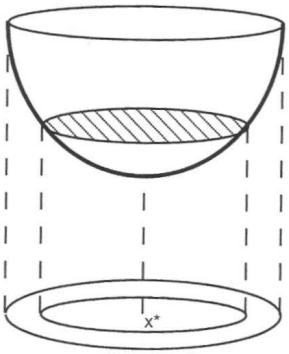

FIG. 6.1 – fonction quadratique de $\mathbb{R}^2$

6.4.1 Fonctions convexes

Définition 6.3 (Fonction convexe) *On dit qu'une fonction à n variables est* ***convexe*** *si*

$$\forall x, x' \in \mathbb{R}^n,\ \forall \lambda \in [0,1]) \quad f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x'). \tag{6.1}$$

Si l'inégalité (6.1) est stricte, on parle de fonction **strictement convexe**.
De plus, si pour un $\alpha > 0$, qui ne dépend que de f on a

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x') - \frac{1}{2}\lambda(1-\lambda)\alpha \|x - x'\|^2$$

on dit que f est **fortement convexe**.

Si f est convexe, la fonction $g = -f$ est **concave**.

Proposition 6.1 *Soient f_1, f_2 deux fonctions convexes :*

1. *$f_1 + f_2$ est convexe*
2. *αf_1 est convexe pour $\alpha \geq 0$*
3. *$sup\{f_1, f_2\}$ est convexe*

On va retrouver géométriquement par l'étude des courbes de niveau de ces fonctions le résultat fondamental suivant :

Théorème 6.5 *Une fonction quadratique est convexe si et seulement si elle est associée à une matrice semi-définie positive. Si la matrice est définie positive, alors la fonction est fortement convexe*

DÉMONSTRATION :Il suffit de montrer que la forme quadratique $f(x) = x^T Ax$ avec A définie positive est une fonction convexe. A étant symétrique, elle est diagonalisable sur la base de ses vecteurs propres orthonormés, soit $A = XDX^T$. Après changement de base $x = Xy$, f s'écrit $g(y) = y^T Dy$, qui est convexe car somme de fonctions carrées à coefficients positifs. Donc f est convexe.
De plus, pour une matrice définie positive, on a :

$$\lambda_1 \|x\|^2 \leq x^T Ax \leq \lambda_n \|x\|^2$$

où λ_1 (resp. λ_n) est la plus petite (resp. grande) valeur propre de A.On en déduit (calculs intermédiaires laissés en exercice) que f est fortement convexe avec $\alpha = \lambda_1$. □

6.4.2 Normes elliptiques

Etudions l'effet d'un changement de variables sur la norme euclidienne. Soit B une matrice non singulière et $x = By$ un changement de variable. Le carré de la norme euclidienne est une forme quadratique définie positive dans $\mathbb{R}^n$:

$$\|x\|^2 = x^T x = y^T B^T By = \|y\|_B^2$$

où $\|.\|_B^2$ est une norme elliptique associée à B ($B^T B$ est définie positive et on démontre facilement que l'inégalité du triangle est conservée).

Exemple 6.1 $B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ *d'où* $A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$ *et* $\|y\|_B^2 = 4y_1^2 + y_2^2$

Le changement de variable est ici un changement d'échelle. L'ensemble des points tels que $\|x\|^2 = 1$ (la boule unité), devient l'ellipse $4y_1^2 + y_2^2 = 1$. L'applatissement de la sphère initiale suivant l'axe horizontal est proportionnel à $\sqrt{a_{11}}$ (cf. figure 6.2)

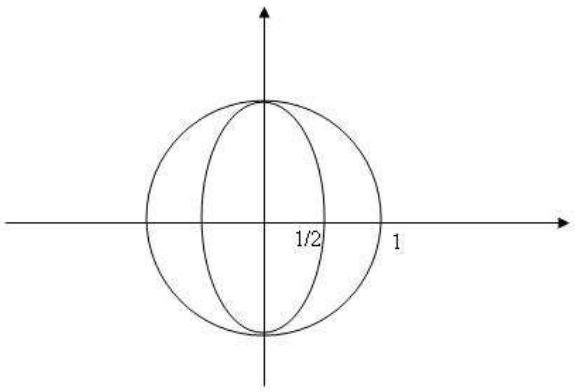

FIG. 6.2 – Boule unité elliptique

6.4.3 Fonctions elliptiques dans $\mathbb{R}^2$

Soit $A = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ et la fonction quadratique de $\mathbb{R}^2$ associée :

$$q(x) = x^T A x$$

Pour diagonaliser A, on calcule ses valeurs propres et ses vecteurs propres :

$$\lambda_1 = 1 \quad \text{avec } q_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$
$$\lambda_1 = 9 \quad \text{avec } q_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

Soit Q la matrice orthogonale dont les colonnes sont les vecteurs propres. Effectuons alors le changement de variables $x = Qy$. la forme quadratique q devient alors :

$$y^T Q^T A Q y = y^T \Lambda y = y_1^2 + 9y_2^2$$

et on remarque de plus que

$$y = Q^T x \Rightarrow \begin{cases} y_1 = \frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{2}} \\ y_2 = \frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}} \end{cases}$$

La forme quadratique initiale une fois diagonalisée s'écrit donc sous la forme d'une somme de carrés :

$$q(x) = \left(\frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{2}} \right)^2 + 9 \left(\frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}} \right)^2$$

La Figure 6.3 représente la boule unité elliptique $q(x) = 1$ qui s'obtient à partir de la boule euclidienne par transformation orthogonale Q (rotation de $-\pi/4$), puis applatissement suivant les axes proportionnel à $\frac{1}{\sqrt{\lambda_1}}$.

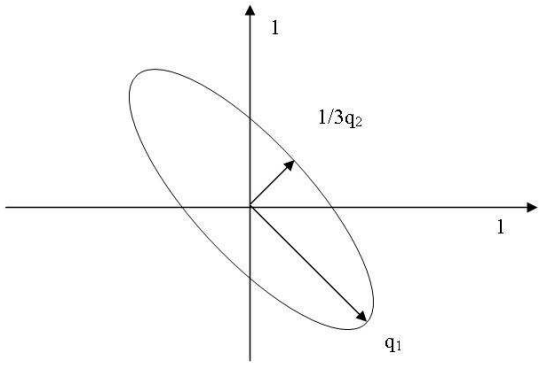

FIG. 6.3 – Fonction elliptique de $\mathbb{R}^2$

INTRODUCTION À L'OPTIMISATION EN DIMENSION FINIE

7.1 Différentiabilité

On considère dans ce chapitre des fonctions continues définies sur tout ou partie de l'espace vectoriel $\mathbb{R}^n$. On notera $x^T y$ le produit scalaire usuel de deux vecteurs x et y et $\| x \|$ la norme euclidienne du vecteur x ($\| x \| = (x^T x)^{1/2}$).

Définition 7.1 (Dérivée directionnelle) *La dérivée directionnelle d'une fonction f en $x \in \mathbb{R}^n$ dans la direction $d \in \mathbb{R}^n$ est la limite, quand elle existe, de l'expression*

$$\frac{1}{t}(f(x+td) - f(x))$$

quand t tend vers 0. On notera

$$f'(x;d) = \lim_{t \to +0} \frac{f(x+td) - f(x)}{t}.$$

On a donc pour t suffisamment petit l'approximation suivante de la fonction f dans la direction d dite *approximation du premier ordre*

$$f(x+td) = f(x) + tf'(x;d) + t\varepsilon(t) \tag{7.1}$$

où $\varepsilon(t)$ tend vers 0 avec t.

Définition 7.2 (Gâteaux-différentiabilité) *Une fonction f est dite Gâteaux-différentiable (on dira simplement différentiable par la suite) en x si elle possède une dérivée directionnelle en x dans toutes les directions et s'il existe un vecteur $\nabla f(x)$ de $\mathbb{R}^n$ tel que*

$$f'(x;d) = \nabla f(x)^T d.$$

Le vecteur $\nabla f(x)$ est appelé *gradient* de f en x défini par

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Observation : Si on choisit la direction $d = e_i$ de la base canonique, on voit que $f'(x; e_i) = \dfrac{\partial f(x)}{\partial x_i}$.

Définition 7.3 (Courbes de niveau) *Les courbes de niveau d'une fonction f sont des points de $\mathbb{R}^n$ tels que $f(x) = \alpha$, où α est une constante.*

Si f est différentiable en x, le gradient de f en x est orthogonal à la courbe de niveau en x et pointe vers la région où la valeur de f est plus grande qu'en x.

L'ensemble

$$S_\alpha = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$$

est une **section** de f.

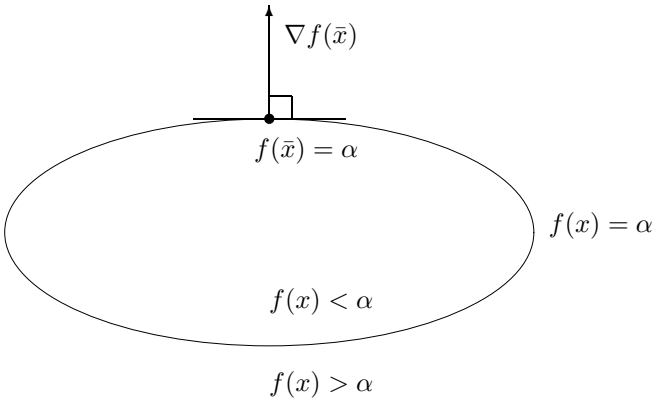

FIG. 7.1 – Courbes de niveau et gradient

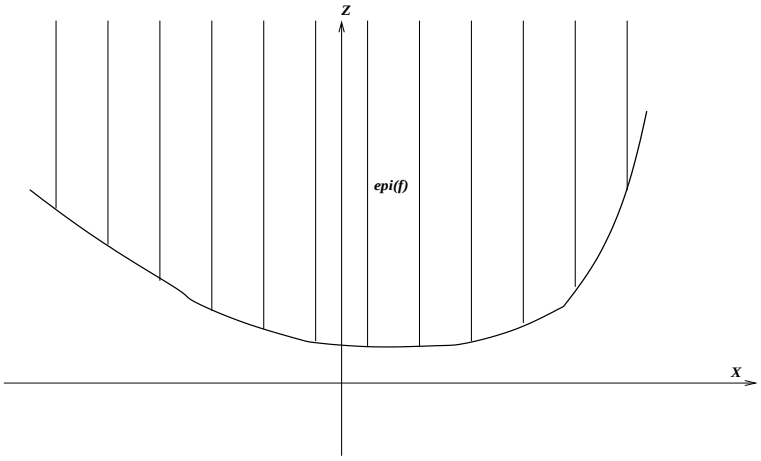

FIG. 7.2 – Épigraphe et section

Définition 7.4 (Epigraphe) *L'épigraphe d'une fonction f est la région de $\mathbb{R}^{n+1}$ située au dessus du graphe de f*

$$\text{epi}(f) = \left\{(x, z) \in \mathbb{R}^{n+1} \mid f(x) \leq z\right\}.$$

La région située au dessous du graphe est appelée hypographe.

Les courbes de niveau sont dans $\mathbb{R}^n$ et l'épigraphe est dans $\mathbb{R}^{n+1}$! Dans $\mathbb{R}^{n+1}$, le vecteur $(\nabla f(x), -1)$ définit le **plan tangent** à l'épigraphe au point $(x, f(x))$.

7.1.1 Différentiabilité et optimalité

Définition 7.5 (Minimum global - minimum local) *On appelle minimum global de la fonction f un point x^* tel que*

$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n. \tag{7.2}$$

Si on restreint la condition (7.2) à un certain voisinage de x^*, on a un minimum local. Si l'inégalité précédente est stricte, on parlera de minimum strict. Pour un maximum, il suffit bien sûr de changer le sens des inégalités dans les définitions correspondantes.

Théorème 7.1 (Condition nécessaire d'optimalité du premier ordre) *Si x^* est un minimum local de f, différentiable, on a*

$$\nabla f(x^*) = 0. \tag{7.3}$$

DÉMONSTRATION : De la définition d'un minimum local, on tire $\nabla f(x^*)^T d \geq 0$, $\forall d \in \mathbb{R}^n$. La linéarité du produit scalaire implique que $\nabla f(x^*) = 0$. □

La condition (7.3) n'est généralement pas suffisante car elle est vérifiée également par un maximum local (e.g. $f(x) = -x^2$ au point 0) ou un point d'inflexion (e.g. $f(x, y) = xy$ au point $(0, 0)$).

Les points en lesquels le gradient d'une fonction s'annule sont appelés *points stationnaires* de la fonction. Il peut s'agir de minima, de maxima ou de point-selles.

7.1.2 Fonctions deux fois différentiables

Une multiapplication est un vecteur de fonction noté $f(x) = (f_1(x), \ldots, f_p(x))^T$. Supposons les f_i différentiables ; la matrice $p \times n$ dont les lignes sont les gradients des fonctions f_i est le Jacobien de f

$$\nabla f(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_p(x)^T \end{bmatrix}.$$

Une fonction f est dite deux fois différentiable en x si elle est différentiable en x et si chaque composante du gradient est une fonction différentiable en x. On peut alors définir le Jacobien de la multiapplication $F(x) = \nabla f(x)$. La matrice formée par les vecteurs gradients $\nabla F_i(x)$ est appelée *Hessien* ou matrice Hessienne de f, notée $\nabla^2 f(x)$. Le Hessien est une matrice carrée, symétrique dont l'élément h_{ij} est la dérivée seconde $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$. On a donc

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i,j=1,\ldots,n}.$$

On peut définir comme précédemment l'approximation du second ordre dans la direction d d'une fonction f en x où f est deux fois différentiable

$$f(x + td) = f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + t^2 \varepsilon(t). \tag{7.4}$$

Théorème 7.2 *Si x^* est un minimum local de f, deux fois différentiable, on a $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est semi-définie positive.*

DÉMONSTRATION : Immédiat à partir de l'approximation du deuxième ordre (7.4) car l'inégalité $d^T \nabla^2 f(x) d \geq 0$, $\forall d$, implique que $\nabla^2 f(x^*)$ est semi-définie positive. □

La condition précédente n'est pas suffisante comme le montre l'exemple de la fonction $f(x) = x^3$ au point $x = 0$. Par contre, si $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est définie positive, alors x^* est un minimum local strict de f (**condition suffisante**).

7.1.3 Quelques règles de différentiation

Dans les formules ci-dessous, $a \in \mathbb{R}^n$, A matrice carrée $(n \times n)$, B matrice $(p \times n)$; f est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$.

$$\begin{aligned} f(x) = a^T x + \alpha &\implies \nabla f(x) = a \\ f(x) = x^T A x + a^T x &\implies \nabla f(x) = (A + A^T)x + a \\ F(x) = Bx - b &\implies \nabla F(x) = B. \end{aligned}$$

Soit g une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ et h une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^p$. La fonction composée $f(x) = (g \circ h)(x) = g(h(x))$ a pour gradient

$$\nabla f(x) = \nabla h(x)^T \nabla g(h(x)).$$

Comme application de la formule ci-dessus, pour $t \in \mathbb{R}$, $d \in \mathbb{R}^n$, on a

$$\frac{df(x + td)}{dt} = d^T \nabla f(x + td).$$

Une autre application de la dérivation d'une fonction composée. Si $f(x) = \| Bx - b \|^2$ on a

$$\nabla f(x) = 2B^T(Bx - b).$$

7.2 Convexité

Définition 7.6 (Ensemble convexe) *Un sous-ensemble C de $\mathbb{R}^n$ est un ensemble convexe si $\forall x_1, x_2 \in C$, le segment $[x_1,\ x_2]$ est dans C, i.e. $\lambda x_1 + (1 - \lambda)x_2 \in C$, $\forall \lambda \in (0, 1)$.*

Exemple 7.1 *Les objets linéaires (sous-espaces vectoriels, variétés linéaires, polyèdres) de $\mathbb{R}^n$; les boules unités associées à toute norme de $\mathbb{R}^n$.*

Par convention, on supposera que l'ensemble vide est convexe. Si C_1 et C_2 sont deux convexes de $\mathbb{R}^n$, alors on a

- $C_1 + C_2$ est convexe
- αC_1 est convexe, α un réel ;
- $C_1 \cap C_2$ est convexe.

Définition 7.7 (Fonction convexe) *Une fonction f est convexe sur un ensemble convexe C de $\mathbb{R}^n$ si*

$$\forall x_1, x_2 \in C, \quad f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2), \quad \forall \lambda \in (0, 1).$$

Si f est convexe, $g = -f$ est concave.

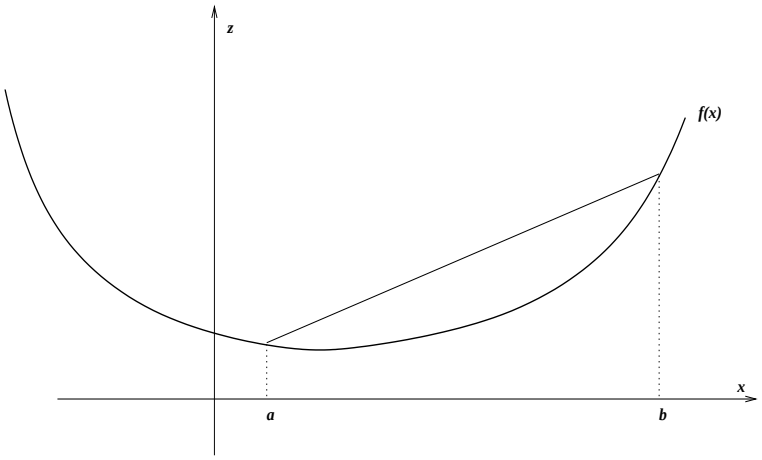

FIG. 7.3 – Fonction convexe

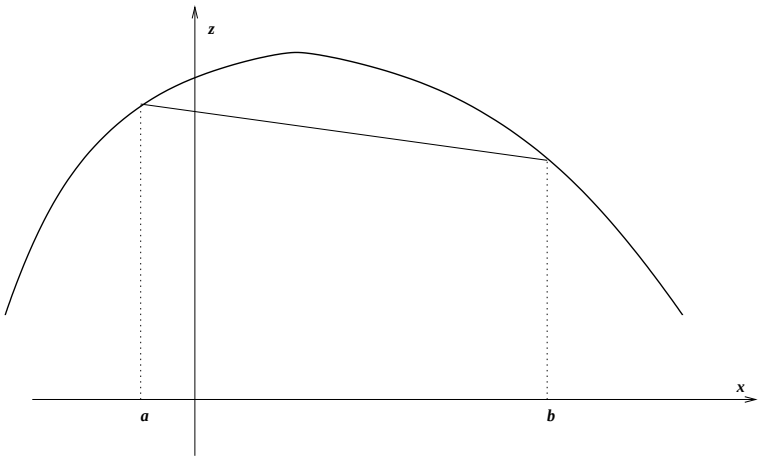

FIG. 7.4 – Fonction concave

Exemple 7.2 (Exemples de fonctions convexes) *1. $f(x) = a^T x + \alpha$ (fonction affine)*

2. *$f(x) = x^T Hx$ (fonction quadratique) avec H matrice carrée semi-définie positive*
3. *$f(x) = \mathrm{e}^{c(x)}$, où c est une fonction convexe*
4. *$f(x) = \| x \|$, où $\| \cdot \|$ est une norme vectorielle quelconque*
5. *$f(x) = af_1(x)$, où f_1 est convexe et $a \geq 0$*
6. *$f(x) = f_1(x) + f_2(x)$, où f_1 et f_2 sont deux fonctions convexes*
7. *$f(x) = \max\{f_1(x), f_2(x)\}$, où f_1 et f_2 sont deux fonctions convexes.*

Définition 7.8 (Domaine d'une fonction convexe) *Le domaine d'une fonction convexe f, noté $\mathrm{dom}(f)$, est l'ensemble des x tels que $f(x) < +\infty$.*

Propriété 7.1 *Soit f convexe, $x_0 \in \mathrm{dom}(f)$ et d une direction de $\mathbb{R}^n$ telle que $x_0 + d \in \mathrm{dom}(f)$. La fonction*

$$q(t) = \frac{1}{t}\left[f(x_0 + td) - f(x_0)\right]$$

est monotone croissante sur $(0,\ 1)$.

Démonstration : Soit $0 \leq h \leq k \leq 1$; appliquons le définition sur le segment $[x_0,\ x_0 + kd]$:

$$f(x_0 + hd) \leq \lambda f(x_0) + (1 - \lambda) f(x_0 + kd)$$

avec $h = (1 - \lambda)k$. On en déduit aisément que $q(h) \leq q(k)$. □

Comme conséquences on a que :

- En tout point de l'intérieur relatif de $\mathrm{dom}(f)$, f est continue et la dérivée directionnelle existe. De plus f est localement Lipschitzienne, i.e.

$$|f(y_1) - f(y_2)| \leq L \| y_1 - y_2 \|,$$

  où y_1, y_2 sont quelconques dans un voisinage du point x_0 et $L > 0$ dépend de x_0.
- Si f est convexe et différentiable en x_0, on a pour $z \in \mathrm{dom}(f)$

$$f(z) \geq f(x_0) + \nabla f(x_0)^T (z - x_0).$$

- Si f est convexe et deux fois différentiable, le Hessien $\nabla^2 f(x_0)$ est une matrice semi-définie positive.

Théorème 7.3 (Condition d'optimalité pour une fonction convexe différentiable) *Un point $x^* \in \mathbb{R}^n$ minimise f convexe différentiable si, et seulement si, $\nabla f(x^*) = 0$.*

Observez que dans le cas convexe, la condition nécessaire du premier ordre devient nécessaire et suffisante. On retrouve ici le fait que tout minimum local d'une fonction convexe est un minimum global (propriété qui n'exige pas la différentiabilité).

7.3 Méthodes de descente

Définition 7.9 (Direction de descente) *Un vecteur d de $\mathbb{R}^n$ est une direction de descente pour la fonction f au point x si*

$$f'(x; d) = \nabla f(x)^T d < 0.$$

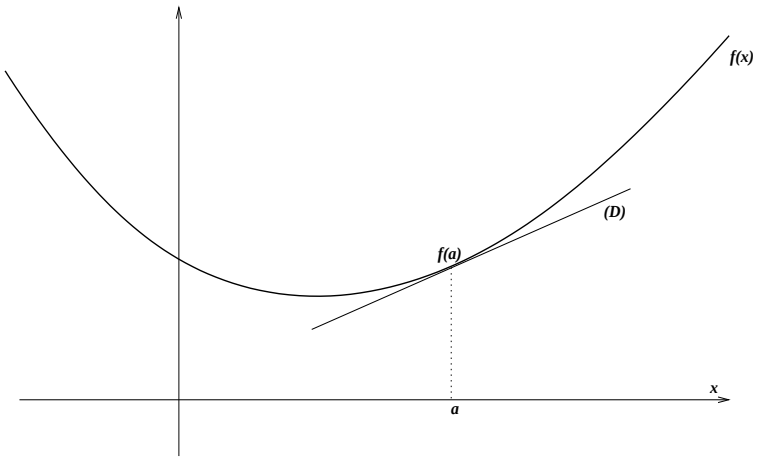

FIG. 7.5 – Fonction convexe différentiable. La courbe de f est toujours au dessus de la tangente (D)

Soit d est une direction de descente et t un réel positif. On suppose t suffisamment petit pour que l'approximation du premier ordre suivante soit valable

$$f(x+td) \approx f(x) + t\nabla f(x)^T d.$$

Donc on a $f(x+td) - f(x) = t\nabla f(x)^T d < 0$, i.e. $f(x+td) < f(x)$.

Les méthodes de descente sont des méthodes itératives pour la minimisation des fonctions différentiables sur $\mathbb{R}^n$ dans lesquelles une direction de descente est choisie à chaque itération à partir des informations généralement locales. On minimise alors la fonction dans cette direction (minimisation unidirectionnelle ou recherche linéaire). Le schéma général est donné ci-dessous (l'indice représente ici l'itération) :

Initialisation. Choisir x_0.

Itération k. Choisir une direction de descente d_k. S'il n'en existe pas, arrêt de l'algorithme.
Sinon

$$x_{k+1} = x_k + t_k d_k,$$

avec $t_k > 0$ tel que $f(x_{k+1}) \leq f(x_k + td_k)$, $\forall t \in (0,\ \delta)$.

Pour la suite, on posera $g_k = \nabla f(x_k)$.

Proposition 7.1 *Dans la recherche linéaire, si t_k minimise exactement la fonction $\theta(t) = f(x_k + td_k)$, le nouveau gradient g_{k+1} au point $x_{k+1} = x_k + t_k d_k$ est orthogonal à la direction d_k.*

DÉMONSTRATION : Si t_k minimise la fonction $\theta(t)$, alors $\theta'(t_k) = 0$. Or, les règles de calcul de la dérivée θ' (chaînage) donnent :

$$\theta'(t) = \nabla f(x_k + td_k)^T d_k.$$

En particulier on a

$$\begin{aligned} \theta'(0) &= g_k^T d_k \\ \theta'(t_k) &= g_{k+1}^T d_k. \end{aligned}$$

La **direction de la plus grande pente** est le vecteur d normé qui minimise $f'(x;d)$. C'est donc la direction opposée au gradient :

$$d_k = \frac{-g_k}{\| g_k \|}.$$

En pratique, on utilise aussi la version non normée, *i.e.*

$$d_k = -g_k.$$

La **méthode du gradient** (ou de plus grande pente) choisit la direction de plus grande descente à chaque itération. On remarque que, dans ce cas, les directions successives sont orthogonales.

Il est clair que la direction de plus grande pente peut être arbitrairement mauvaise en ce qui concerne la direction du minimum. Le cas des fonctions quadratiques est particulièrement instructif et on verra dans la section suivante comment améliorer à peu de frais cette direction pour suivre la vallée et non la pente vers la solution.

7.4 Direction de Newton et directions conjuguées

Dans le cas d'une fonction quadratique fortement convexe (donc associée à une matrice définie positive), on peut montrer (cf. Luenberger [7]) que la méthode du gradient avec recherches linéaires exactes présente une convergence linéaire avec un taux égal

$$\left(\frac{r-1}{r+1}\right)^2$$

où r est le rapport entre la plus grande et la plus petite valeur propre de la matrice associée à la forme quadratique. Elle devient sous-linéaire quand la fonction est mal conditionnée, c'est-à-dire quand r devient grand.

7.4.1 Méthode de Newton

Soit, dans $\mathbb{R}^n$,

$$q(x) = \frac{1}{2}x^T Ax,$$

avec A symétrique définie positive. En un point $x \in \mathbb{R}^n$ différent de l'origine (l'unique minimum global de q). Le gradient de q est $\nabla q(x) = Ax$.

On se place en un point x_0. On fait une approximation d'ordre 2 de q en x_0. On obtient

$$q(x) = \frac{1}{2}x_0^T Ax_0 + (x - x_0)^T Ax_0 + \frac{1}{2}(x - x_0)^T A(x - x_0).$$

On cherche maintenant, à partir de x_0, le point x^* tel que $\nabla q(x^*) = 0$, ce qui donne $x^* = x_0 + d_N$, où $d_N = -A^{-1}\nabla q(x_0)$ est la direction de Newton. On observe que cette direction fournit instantanément (sans recherche linéaire) le minimum de la fonction car $x_0 + d_N = x^* = 0$ (cf. figure 7.6). L'algorithme converge donc en une seule itération.

Pour une fonction f deux fois différentiable, on écrit l'approximation d'ordre 2

$$\tilde{f}(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k).$$

On cherche ensuite x_{k+1} tel que $\nabla \tilde{f}(x_{k+1}) = 0$. On trouve

$$x_{k+1} = x_k + d_N, \quad d_N = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k).$$

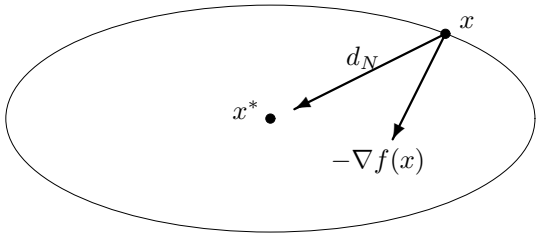

FIG. 7.6 – Directions de Newton et du gradient

Comme pour les fonctions quadratiques, la direction et le pas de déplacement sont fixés. Si l'ordre de convergence devient quadratique, l'application de cette idée dans le cas général est limitée par le coût du calcul de l'inverse du Hessien et par les sévères hypothèses qui garantissent cette convergence. En pratique, la direction est approchée itérativement (cf. cours 2ème année F4) et nous allons illustrer cette possibilité avec la méthode du gradient conjugué.

7.4.2 Méthode des directions conjuguées

Les méthodes des directions conjuguées sont des méthodes itératives qui, appliquées à une fonction quadratique convexe à n variables, conduisent à l'optimum en n étapes au plus.

Définition 7.10 *A étant une matrice symétrique définie positive, deux vecteurs non nuls x et y de $\mathbb{R}^n$ sont dits **conjugués** par rapport à A si $x^T Ay = 0$.*

Proposition 7.2 *Si p vecteurs non nuls de $\mathbb{R}^n$ ($p \leq n$) sont mutuellement conjugués, ils sont linéairement indépendants.*

DÉMONSTRATION : Soit α_i tels que

$$\sum_{i=1}^{p} \alpha_i d_i = 0. \tag{7.5}$$

On multiplie (7.5) à gauche par A, on obtient

$$\sum_{i=1}^{p} \alpha_i A d_i = 0. \tag{7.6}$$

En multipliant (7.6) par d_1^T, on trouve $\alpha_1 d_1^T A d_1 = 0$, puisque les vecteurs sont mutuellement conjugués. On déduit que $\alpha_1 = 0$. En répétant le processus avec $d_2, \ldots, d_p$, on arrive à $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$. Donc les vecteurs $d_1, \ldots, d_p$ sont linéairement indépendants. □

Soit la fonction quadratique convexe

$$q(x) = \frac{1}{2} x^T A x + b^T x + c.$$

Principe de la méthode : On part d'un point x_0, on minimise q successivement suivant n directions $d_0, d_1, \ldots, d_{n-1}$ conjuguées par rapport à la matrice A, i.e. $d_i^T A d_j = 0$, pour $i \neq j$. La séquence $\{x_k\}$ est définie par $x_{k+1} = x_k + t_k d_k$, où t_k minimise $\theta(t) = q(x_k + t d_k)$. Le point obtenu à la n-ème itération, i.e.

$$x_n = x_{n-1} + t_{n-1} d_{n-1} = x_0 + \sum_{i=0}^{n-1} t_j d_j$$

est l'optimum du problème ($\nabla q(x_n) = A x_n + b = 0$).

Théorème 7.4 *Les gradients successifs calculés aux points x_k générés par la méthode des directions conjuguées satisfont*

$$g_{k+1}^T d_i = 0, \quad \forall i \leq k.$$

DÉMONSTRATION : Par hypothèse sur le choix optimal de t_k, on a $g_{k+1}^T d_k = 0$. Comme q est quadratique, on a les relations

$$\begin{aligned} g_{k+1} &= g_k + t_k A d_k, \\ g_{k+1} &= g_{i+1} + \sum_{j=i+1}^{k} t_j A d_j, \quad i < k. \end{aligned}$$

D'où le résultat. □

Corollaire 7.1 *x_{k+1}, calculé à l'itération k de la méthode des directions conjuguées, minimise q sur le sous-espace affine $V_k = \{x_0\} + \mathrm{lin}\{d_0, \ldots, d_k\}$.*

DÉMONSTRATION : A l'itération $k+1$, le gradient g_{k+1} est orthogonal à $d_0, \ldots, d_k$, donc orthogonal à V_k. C'est la condition d'optimalité du problème de minimisation de q sur V_k. □

Corollaire 7.2 *La méthode des directions conjuguées, appliquée à une fonction quadratique fortement convexe de $\mathbb{R}^n$, converge en au plus n itérations.*

DÉMONSTRATION : Soit $g_k = 0$ pour $k < n$, soit $g_n = 0$ car $V_n = \mathbb{R}^n$ d'après le corollaire 7.1.□
Une manière simple de construire les directions conjuguées itérativement est

1. $d_0 = -g_0$
2. Si $g_{k+1} \neq 0$, alors

$$\begin{aligned} \beta_k &= \frac{g_{k+1}^T A d_k}{d_k^T A d_k}, \\ d_{k+1} &= -g_{k+1} + \beta_k d_k. \end{aligned}$$

Le coefficient β_k est calculé de sorte que $d_k^T A d_{k+1} = 0$. La méthode s'appelle alors la **méthode du gradient conjugué**. Elle peut être utilisée pour minimiser une fonction quadratique fortement convexe, c'est-à-dire pour résoudre un système d'équations linéaires dont la matrice est définie positive.

Soit à résoudre le système

$$Ax = b \tag{7.7}$$

avec A définie positive. On peut résoudre ce système à travers la minimisation de la fonction quadratique fortement convexe

$$q(x) = \frac{1}{2} x^T A x - b^T x$$

dont le gradient est $\nabla q(x) = Ax - b$. De sorte que si $\nabla q(x^*) = 0$ alors x^* est solution de (7.7).

Algorithme du gradient conjugué pour les fonctions quadratiques

Initialisation $k = 0$. x_0 donné, $g_0 = \nabla q(x_0) = Ax_0 - b$, $d_0 = -g_0$

Itération $k \geq 0$. x_k, g_k et d_k connus

- Calcul (explicite) du pas de déplacement

$$t_k = -\frac{g_k^T d_k}{d_k^T A d_k}$$

- Mise à jour

$$x_{k+1} = x_k + t_k d_k$$

- Nouvelle direction du gradient conjugué

$$\beta_k = \frac{g_{k+1}^T A d_k}{d_k^T A d_k}$$
$$d_{k+1} = -g_{k+1} + \beta_k d_k.$$

En pratique on arrête les itérations dès que $\| g_k \|$ est "suffisamment" petit. Cette méthode est très populaire et peut concurrencer la méthode de Cholesky (*cf.* Chap. 6) quand la dimension de l'espace n est très élevée.

L'algorithme ci-dessus peut également être adapté aux fonctions non quadratiques différentiables. On évite alors le calcul du Hessien et les performances sont toujours supérieures à celles de la méthode du gradient.

Algorithmes du gradient conjugué de Fletcher-Reeves et Polak-Ribière

Initialisation $k = 0$. x_0 donné, $g_0 = \nabla q(x_0) = Ax_0 - b$, $d_0 = -g_0$

Itération $k \geq 0$. x_k, g_k et d_k connus

- Calcul du pas de déplacement par recherche linéaire

$$t_k = \arg\min_{t>0} f(x_k + t d_k)$$

- Mise à jour

$$x_{k+1} = x_k + t_k d_k$$

- Nouvelle direction du gradient conjugué

$$\beta_k = \frac{\| g_{k+1} \|^2}{\| g_k \|^2}, \quad \text{(Fletcher-Reeves)}$$

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{\| g_k \|^2}, \quad \text{(Polak-Ribière)}$$

$$d_{k+1} = -g_{k+1} + \beta_k d_k.$$

Remarque 7.1 *Après n itérations, on a généré $\mathbb{R}^n$. Il faut donc réinitialiser l'algorithme en posant $d_k = -g_k$ (au lieu des formules du gradient conjugué) toutes les n itérations.*

7.5 Optimisation sous contraintes

Un problème d'optimisation dans $\mathbb{R}^n$ prend la forme générale suivante

$$\begin{array}{ll} \text{Minimiser} & f(x) \\ \text{sous} & g_i(x) \leq 0, \quad i = 1, \ldots, p \\ & h_j(x) = 0, \quad j = 1, \ldots, q \\ & x \in S \end{array}$$

où f, g_i, h_j sont des fonctions de $\mathbb{R}^n$ dans $\mathbb{R}$ et S un ensemble de $\mathbb{R}^n$.

On a représenté ici trois types de contraintes : contraintes décrites par des inégalités, contraintes décrites par des égalités et contraintes décrites par un ensemble (non nécessairement explicite). Le problème consiste donc à trouver parmi tous les x qui satisfont conjointement toutes les contraintes un x qui rend f minimum. On fera comme précédemment la distinction entre minimum local et minimum global.

7.5.1 Contraintes égalités

On considère le problème

$$(PE) \qquad \begin{array}{ll} \text{Minimiser} & f(x) \\ \text{sous} & h_j(x) = 0, \quad \forall j = 1, \ldots, q. \end{array}$$

On supposera les fonctions f, h_1,...,h_q continument différentiables.

Théorème 7.5 *Si x^* est minimum local pour le problème (PE) et sous l'hypothèse que les gradients $\nabla h_j(x^*)$, $j = 1, \ldots, q$, sont linéairement indépendants, alors il existe q multiplicateurs de Lagrange λ_j, $1 \leq j \leq q$, tels que*

$$\nabla f(x^*) + \sum_{j=1}^{q} \lambda_j \nabla h_j(x^*) = 0. \tag{7.8}$$

DÉMONSTRATION : La démonstration est basée sur le théorème de la fonction implicite et on ne donnera ici qu'une description sommaire. L'ensemble des points dits réalisables, c'est-à-dire qui satisfont simultanément toutes les contraintes, est une surface S de $\mathbb{R}^n$. L'hypothèse sur les gradient ∇h_j (hypothèse de régularité) implique que

(i) La surface S est localement une variété différentiable de dimension $n - q$.

(ii) Le plan tangent à S en x^* est un sous-espace affine de dimension $n - q$ et peut être défini en fonction des gradients ∇h_j.

Comme on ne s'intéresse qu'aux directions à partir de x^*, on appellera plan tangent à S en x^* le sous-espace

$$M = \{y \in \mathbb{R}^n \ : \ Jy = 0\}$$

où J est le Jacobien des contraintes en x^* ($\nabla h_j(x^*)^T y = 0, \forall j$). Une condition nécessaire pour que x^* soit un minimum local de f sur S est qu'il n'existe pas de directions de descente dans le plan tangent M. Plus précisément

$$d^T \nabla f(x^*) \geq 0, \quad \forall d \in M$$

ce qui implique (car $-d \in M$) que $d^T \nabla f(x^*) = 0$ ou encore que $\nabla f(x^*)$ est orthogonal à M (cf. figure 7.7). Or une représentation du sous-espace $M^\perp$ est (cf. chap. 1)

$$M^\perp = \left\{y \in \mathbb{R}^n \ \mid \ y = J^T \lambda\right\}$$

d'où le résultat. □

Exemple 7.3 *Soit le problème*

$$\begin{array}{cc} \textit{Minimiser} & f(x) = x_1^2 + x_2^2 \\ \textit{Sous} & x_1 + x_2 = 1. \end{array}$$

On écrit les conditions d'optimalités (7.8), ce qui donne

$$\begin{aligned} 2x_1 + \lambda &= 0 \\ 2x_2 + \lambda &= 0 \\ x_1 + x_2 &= 1. \end{aligned}$$

Un calcul direct donne la solution $x_1^* = 1/2$, $x_2^* = 1/2$, $\lambda = -1$.

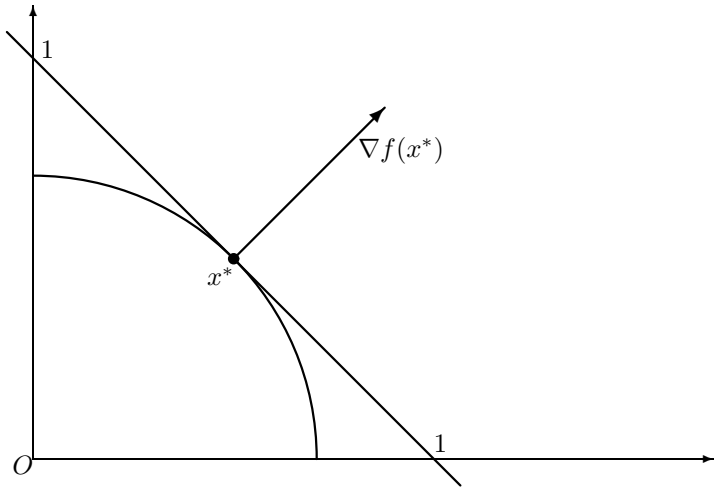

FIG. 7.7 – Plan tangent M et gradient

On remarque donc que les conditions d'optimalité (7.8) forment avec les contraintes un système de $n+q$ équations à $n+q$ inconnues. Dans le cas général ce système est non linéaire et doit être résolu par une méthode itérative. En fait ces méthodes itératives, regroupées sous le nom de Programmation Non Linéaire, progressent le long de directions de descente jusqu'à ce que les conditions d'optimalité soient approximativement satisfaites. Chaque multiplicateur de Lagrange peut-être interprété comme un prix marginal de la ressource fixée au second membre ($\lambda_j = -\partial f/\partial h_j$).

Définition 7.11 (Le Lagrangien) *C'est la fonction* $\mathcal{L}$ *de* $\mathbb{R}^n \times \mathbb{R}^q$ *dans* $\mathbb{R}$ *définie par*

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{j=1}^{q} \lambda_j h_j(x). \tag{7.9}$$

La condition d'optimalité du premier ordre exprime donc le fait que le couple (x^*, λ^*) est un point stationnaire de $\mathcal{L}$ dans $\mathbb{R}^n \times \mathbb{R}^q$.

Quand les fonctions en jeu sont deux fois différentiables, on peut écrire une condition nécessaire du deuxième ordre. La sous-matrice du Hessien du Lagrangien $\mathcal{L}$ associée aux dérivées secondes par rapport à x est

$$L^* = \nabla^2 f(x^*) + \sum_{j=1}^{q} \lambda_j \nabla^2 h_j(x^*).$$

Si x^* est minimum local, alors la matrice L^* est telle que

$$d^T L^* d \geq 0, \quad \forall d \in M.$$

C'est la restriction de L^* au sous-espace M qui doit être semi-définie positive. Il n'est pas nécessaire que L^* le soit.

7.5.2 Contraintes inégalités

On considère le problème

$$(PI) \qquad \begin{array}{ll} \text{Minimiser} & f(x) \\ \text{sous} & g_i(x) \leq 0, \quad \forall i = 1, \ldots, p. \end{array}$$

Définition 7.12 (Contrainte active) *On appelle contrainte active (ou saturée) en x les contraintes dont les indices sont dans*

$$I(x) = \{i \in \{1, \ldots, p\} \mid g_i(x) = 0\}.$$

Proposition 7.3 *Si x^* est solution optimale locale de (PI), x^* est solution optimale locale de*

$$\begin{array}{ll} \textit{Minimiser} & f(x) \\ \textit{sous} & g_i(x) = 0, \quad \forall i \in I(x^*). \end{array}$$

Donc les contraintes inactives n'ont aucune influence sur l'optimalité du problème (PI).

Théorème 7.6 (de Kuhn-Tucker) *Sous l'hypothèse de différentiabilité et de régularité des contraintes actives (les gradients $\nabla g_i(x^*)$, $i \in I(x^*)$ sont linéairement indépendants), si x^* est un minimum local pour le problème (PI), alors il existe p multiplicateurs de Kuhn-Tucker $\mu_i \geq 0$, $i = 1, \ldots, p$, tels que*

$$\nabla f(x^*) + \sum_{i \in I(x^*)} \mu_i \nabla g_i(x^*) = 0, \tag{7.10}$$

$$\mu_i \geq 0, \quad i = 1, \ldots, p \tag{7.11}$$

$$\mu_i g_i(x^*) = 0, \quad i = 1, \ldots, p. \tag{7.12}$$

Observons que les équations (7.11)-(7.12) dites de complémentarité signifient que le multiplicateur μ_i associé à une contrainte inactive est nul. On renvoie à Minoux [9] pour la démonstration du théorème que l'on illustrera par un exemple.

Exemple 7.4 *Soit le problème*

$$\begin{array}{ll} \textit{Minimiser} & f(x) = (x_1 - 2)^2 + (x_2 - 1)^2 \\ \textit{sous :} & x_1^2 - x_2 \leq 0 \\ & x_1 + x_2 - 2 \leq 0 \\ & -x_1 \leq 0. \end{array}$$

L'ensemble Ω des solutions réalisables est représenté sur la figure 7.8. La solution optimale est la projection du point $(2,1)^T$ sur cet ensemble. C'est donc le point $x^* = (1,1)^T$. Écrivons les conditions de Kuhn-Tucker en x^*. Seules les deux premières contraintes sont actives :

$$\begin{aligned} 2(x_1 - 2) + 2\mu_1 x_1 + \mu_2 &= 0 \\ 2(x_2 - 1) - \mu_1 + \mu_2 &= 0. \end{aligned}$$

On obtient donc $\mu_1 = \mu_2 = 2/3$.

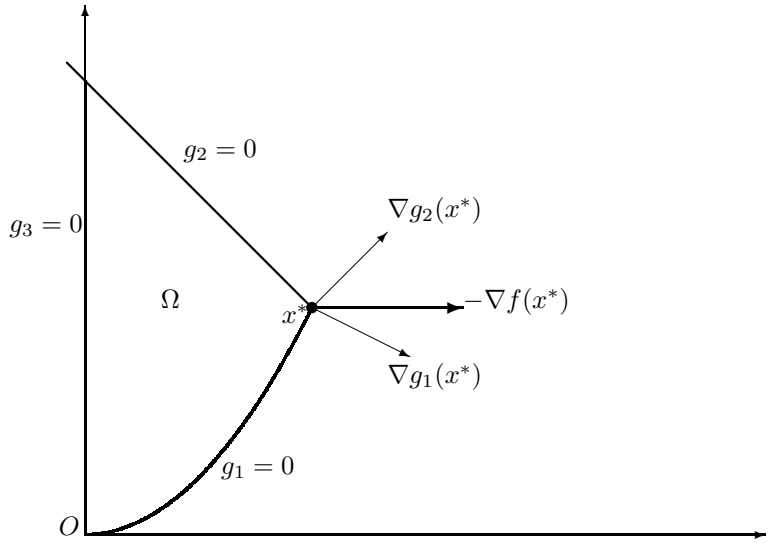

FIG. 7.8 – Solutions réalisables et solution optimale : le gradient $\nabla g_2(x^*)$ et orthogonal au plan tangent à la contrainte $g_2(x) = 0$ en x^* de même pour $\nabla g_3(x^*)$.

7.6 Exercice

Exercice 7.1 Soit f la fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par

$$f(x, y) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Montrer que $x^* = (1, 1)$ est un point stationnaire de f. Calculer $\nabla^2 f(x^*)$ et donner la nature de x^*.

Exercice 7.2 Soit f la norme euclidienne sur $\mathbb{R}^n$ définie, pour $x \in \mathbb{R}^n$, par

$$f(x) = \| x \|_2 = \sqrt{(x_1)^2 + \cdots (x_n)^2}.$$

1.— Montrer que f admet une dérivée directionnelle en tout point de $\mathbb{R}^n \setminus \{0\}$ et que

$$f'(x; d) = \frac{1}{\| x \|_2} x^T d.$$

2.— On se place dans $\mathbb{R}$. Alors $f(x) = |x|$ n'est pas dérivable en 0. On définit

$$f_\varepsilon(x) = \begin{cases} |x| - \frac{\varepsilon}{2} & \text{si } |x| > \varepsilon \\ \frac{1}{2\varepsilon} x^2 & \text{sinon} \end{cases}$$

Étudier la dérivabilité de f_ε. Quelle est l'expression de f_ε pour $f = \| \cdot \|_2$ sur $\mathbb{R}^n$.

Exercice 7.3 On considère la fonction quadratique définie dans $\mathbb{R}^2$ par

$$f(x, y) = 5x^2 + 5y^2 + 8xy - 10x - 8y.$$

a) Montrer que f est une fonction convexe et déterminer son minimum global dans $\mathbb{R}^2$.

b) Représenter les courbes de niveau de f.

c) A partir du point initial $(-4, 4)$, appliquer 2 itérations des méthodes du gradient et du gradient conjugué. Comparer les résultats obtenus.

Exercice 7.4 Soit la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ donnée par

$$f(x) = x_1^2 + 2x_2^2 + 4x_1 + 4x_2.$$

Montrer par récurrence que la méthode de la plus forte pente, avec $d_k = -\nabla f(x_k)$, appliquée à f en partant de $(0, 0)$ génère une suite vectorielle $\{x_k\}$ de terme général

$$x_k = \begin{pmatrix} \frac{2}{3^k} - 2 \\ \left(-\frac{1}{3}\right)^k - 1 \end{pmatrix}$$

En déduire le minimum de f.

Exercice 7.5 Soit $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ deux fois différentiable et A une matrice $p \times n$ de rang p. Soit $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ définie par

$$q(x) = f(Ax).$$

a) Calculer le gradient et le Hessien de q en fonction du gradient et du Hessien de f.

b) Application : Problème des moindres carrés.

$$q(x) = \| Ax - b \|^2,$$

où b est un vecteur de $\mathbb{R}^p$ donné. Calculer x^* qui minimise q et mettre x^* sous la forme $x^* = A^* b$. Quelles sont les propriétés de A^*.

Exercice 7.6 Montrer que dans la méthode du gradient conjugué appliquée à une fonction quadratique convexe, avec minimisation exacte, les notations suivantes sont vérifiées.

(a) $\mathcal{V}_k = \text{lin}\{d_0, d_1, \ldots, d_k\} = \text{lin}\{g_0, g_1, \ldots, g_k\}$

(b) $d_{k+1}^T A d_i = 0, \ \forall i = 0, 1, \ldots, k.$

(c) $t_k = \dfrac{\| g_k \|^2}{d_k^T A d_k}$

Exercice 7.7 Donner les courbes de niveau de la fonction

$$f(x) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2.$$

En déduire le minimum de f.

a) Montrer que la méthode de la plus forte pente appliquée à f, en partant de $(0, 0)$ ne peut converger en un nombre fini d'itérations.

b) Trouver un point initial pour que la méthode du gradient conjugué de Fletcher–Reeves appliquée à f converge en une seule itération.

Exercice 7.8 Soit une fonction quadratique $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ définie par

$$f(x) = \frac{1}{2} x^T Q x - b^T x,$$

où Q est une matrice carrée symétrique d'ordre n.

1.— Discuter de l'existence et de la nature des points stationnaires de f en fonction du rang de la matrice Q.

2.— Soit $\bar{x}$ un point stationnaire, λ_j une valeur propre de Q et u_j le vecteur propre associé à λ_j. Montrer que

$$f(\bar{x} + tu_j) = f(\bar{x}) + \frac{1}{2}t^2\lambda_j, \quad \forall t \in \mathbb{R}.$$

3.— Discuter de la géométrie des courbes de niveau et de l'épigraphe de f dans les cas suivants :

(C1) La matrice Q est définie positive.

(C2) La matrice Q est sémi–définie positive et il existe des points stationnaires.

(C3) La matrice Q est sémi–définie positive et il n'existe pas de points stationnaires.

(C4) La matrice Q est indéfinie et non singulière.

Comme application, on pourra utiliser les matrices Q et les vecteurs b suivants, dans $\mathbb{R}^2$

$$Q = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \qquad b = \begin{pmatrix} 9 \\ 18 \end{pmatrix} \qquad \text{cas } (\mathbf{C1})$$

$$Q = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{cas } (\mathbf{C2})$$

$$Q = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{cas } (\mathbf{C3})$$

$$Q = \begin{pmatrix} -2 & 2 \\ 2 & 6 \end{pmatrix}, \qquad b = \begin{pmatrix} 6 \\ -6 \end{pmatrix} \qquad \text{cas } (\mathbf{C4})$$

MÉTHODES ITÉRATIVES

On étudie dans ce chapitre des méthodes itératives pour résoudre un système d'équations linéaires. Il s'agit de méthodes de type point fixe qui engendrent une suite de solutions qui converge vers la solution du système. On verra que chaque itération peut s'effectuer en $O(n^2)$ (en $O(n)$ pour des matrices bandes) et que ces méthodes peuvent être compétitives avec les méthodes de pivotage si n est très grand.

8.1 La méthode de Jacobi

Soit

$$Ax = b$$

un système linéaire de matrice A $(n \times n)$ non singulière dont l'unique solution est x^*. On note $x_i^{(k)}$ la i-ème composante du vecteur x calculé à l'itération k. On supposera les éléments a_{ii} diagonaux de A tous non nuls. La méthode consiste à calculer chaque variable séparément, les autres étant fixées à leur valeur de l'itération précédente.

Algorithme de Jacobi

$k = 0$ $x^{(0)} \in \mathbb{R}^n$ donné

$k \geq 1$ Pour $i = 1, \ldots, n$, calculer

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \Big[b_i - \sum_{j=1, j \neq i}^{n} a_{ij} x_j^{(k)} \Big]$$

La convergence de la méthode nécessite des hypothèses supplémentaires sur A comme le montre l'exemple suivant :

$$\begin{bmatrix} 1 & 10 \\ 10 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 12 \end{bmatrix}$$

dont la solution est $x^* = (1\ 1)^T$. Choisissons $x^{(0)} = (0\ 0)^T$; les itérés de Jacobi sont

$$x^{(1)} = \begin{bmatrix} 11 \\ 6 \end{bmatrix},\ x^{(2)} = \begin{bmatrix} -49 \\ -49 \end{bmatrix},\ x^{(3)} = \begin{bmatrix} 501 \\ 251 \end{bmatrix},\ x^{(4)} = \begin{bmatrix} -2499 \\ -2499 \end{bmatrix}$$

et la méthode diverge.

Permutons alors les deux colonnes de A et on obtient

$$x^{(1)} = \begin{bmatrix} 1.2 \\ 1.1 \end{bmatrix},\ x^{(2)} = \begin{bmatrix} .98 \\ .98 \end{bmatrix},\ x^{(3)} = \begin{bmatrix} 1.004 \\ 1.002 \end{bmatrix},\ x^{(4)} = \begin{bmatrix} .9992 \\ .9996 \end{bmatrix}$$

et la méthode converge.

8.2 La méthode de Gauss-Seidel

Les mises à jour sont effectuées séquentiellement en prenant pour x_j, $j < i$, la dernière valeur calculée soit $x_j^{(k+1)}$.

Algorithme de Gauss-Seidel

$k = 0$ $x^{(0)} \in \mathbb{R}^n$ donné

$k \geq 1$ Pour $i = 1, \ldots, n$, calculer

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right]$$

Remarque 8.1 *Chaque itération (mise à jour des n composantes) s'effectue en $O(n^2)$ flops pour les deux méthodes. L'implémentation de l'itération de Gauss-Seidel ne nécessite que le stockage d'un seul tableau dans lequel les x_i sont mis à jour séquentiellement alors que deux tableaux sont nécessaires pour Jacobi. L'intérêt de cette dernière méthode réside plutôt dans le fait que les mises à jour des composantes de x peuvent s'effectuer en parallèle sur des processeurs différents.*

Remarque 8.2 *On peut écrire des versions des algorithmes de Jacobi ou Gauss-Seidel par blocs : chaque x_i est un bloc de composantes tel que la sous-matrice diagonale a_{ii} est non singulière. Une mise à jour de type Jacobi revient alors à résoudre un système linéaire de "petite taille"*

$$[a_{ii}] x_i^{(k+1)} = b_i - \sum_{j=1, j \neq i}^{n} [a_{ij}] x_j^{(k)}.$$

8.3 Convergence des méthodes itératives

8.3.1 Résultat général

On donne tout d'abord un résultat de convergence sur les puissances de matrices qui précise le rôle des valeurs propres. On rappelle que $\rho(A)$ est le rayon spectral de la matrice A, i.e. le plus grand module des valeurs propres de A, voir chap. 5.

Théorème 8.1 *Soit une matrice carrée B ; les conditions suivantes sont équivalentes*

(i) $\lim_{k \to +\infty} B^k = 0$

(ii) $\lim_{k \to +\infty} B^k v = 0$, $\forall v \in \mathbb{R}^n$

(iii) $\rho(B) < 1$

Pour la démonstration, voir P.G. Ciarlet [2, p.21].

Théorème 8.2 *Soit une matrice B $(n \times n)$ non singulière, un processus itératif dans $\mathbb{R}^n$:*

$$u^{(k+1)} = Bu^{(k)} + b$$

défini à partir d'un partir d'un point u^0 et supposons qu'il possède un point fixe, c'est-à-dire, qui satisfait $u^* = Bu^* + b$. Le processus converge si, et seulement si, $\rho(B) < 1$.

DÉMONSTRATION : $e^{(k)} = u^{(k)} - u^* = B^k e_0$ tend vers 0 si $\rho(B) < 1$ d'après le théorème 8.1. □

Remarque 8.3 *Comme le rayon spectral est souvent difficile à évaluer, la condition nécessaire et suffisante du théorème 2 peut-être remplacée par la condition suffisante*

$$\| B \| < 1$$

où $\| \cdot \|$ *est une norme subordonnée quelconque.*

8.3.2 Convergence des méthodes et relaxation

Les méthodes itératives étudiées dans ce chapitre correspondent à une décomposition de la matrice A du système linéaire sous la forme $A = M - N$ avec M non singulière. Le système linéaire peut alors s'écrire comme un point fixe

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

D'après le théorème 2, la méthode itérative sera convergente si le rayon spectral de la matrice $M^{-1}N$ est strictement inférieur à 1.

Pour expliciter la décomposition correspondant aux méthodes de Jacobi et Gauss-Seidel, on note $A = A - E - F$ où $D = \text{diag}\{a_{11}, \ldots, a_{22}\}$ et $-E$ et $-F$ sont les parties de A triangulaires respectivement au dessous et au dessus de la diagonale (cf. figure 8.1). On obtient alors :

Méthode de Jacobi : $M = D$, $N = E + F$ (on notera $J = D^{-1}(E + F)$

Méthode de Gauss-Seidel : $M = D - E$, $N = F$.

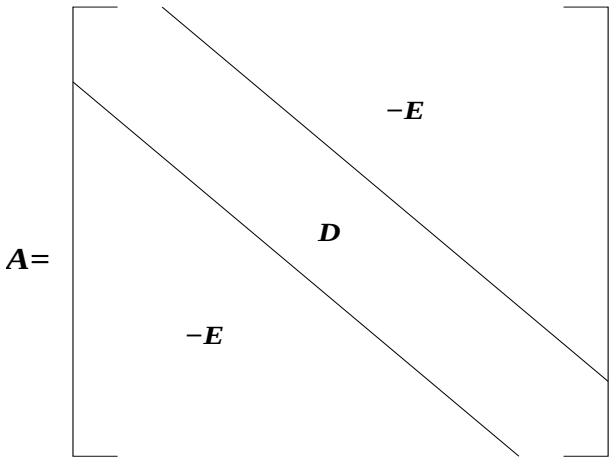

FIG. 8.1 – Décomposition de la matrice pour les méthodes de Jacobi et de Gauss-Seidel

On peut accélérer ces deux méthodes en introduisant un *coefficient de relaxation* ω. Si $\tilde{x}^{(k+1)}$ est l'itéré calculé par une des deux méthodes, on calculera

$$x^{(k+1)} = \omega \tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}.$$

En fait, on utilisera cette idée surtout avec Gauss-Seidel. On obtient la méthode dite de relaxation. La décomposition correspondante est :

$$M = \frac{1}{\omega}D - E, \quad N = \frac{1 - \omega}{\omega}D + F.$$

L'opérateur de point fixe est noté $L_\omega = (D-\omega E)^{-1}[(1-\omega)D+\omega F]$. L_1 est donc l'opérateur de point fixe de la méthode de Gauss-Seidel. On parlera de *sous-relaxation* si $\omega < 1$ et de *sur-relaxation* si $\omega > 1$.

Théorème 8.3 *Supposons A symétrique et la décomposition $A = M - N$ telle que $M^T + N$ soit définie positive. Alors $\rho(M^{-1}N) < 1$ si, et seulement si, A est définie positive.*

DÉMONSTRATION : *(i)* Supposons que A est définie positive. On observe d'abord que :

$$M^T + N = (A^T + N^T) + N = (A + N) + N^T = M + N^T.$$

Puisque le rayon spectral est borné supérieurement par toute norme matricielle, on va évaluer $\| M^{-1}N \|$ avec la norme subordonnée à la norme vectorielle $\| v \|_A = (v^T Av)^{1/2}$:

$$\| M^{-1}N \| = \| I - M^{-1}A \| = \sup_{\|v\|_A=1} \| v - M^{-1}Av \|_A .$$

Posons $w = M^{-1}Av$. Alors

$$\begin{aligned} \| v - w \|_A^2 &= (v-w)^T A(v-w) \\ &= 1 - w^T Av - v^T Aw + w^T Aw \\ &= 1 - w^T Mw - w^T M^T w + w^T Aw \\ &= 1 w^T (M^T + N) w. \end{aligned}$$

On déduit alors que $\| v - w \|_A^2 < 1$ grâce à l'hypothèse sur $M^T + N$ (pour $w \neq 0$).

(ii) Supposons que $\rho(M^{-1}N) < 1$. Si A n'est pas définie positive, il existe un x_0 tel que $\alpha_0 = x_0^T Ax_0 \leq 0$. Étudions la suite $\{\alpha_k\}$ définie par α_0 et par $\alpha_k = x_k^T Ax_k$ et $x_k = Bx_{k-1}$. Cette suite tend vers 0 car $\rho(B) < 1$. Comme $x_{k-1} - x_k = M^{-1}Ax_{k-1} = N^{-1}Ax_k$, on a :

$$\begin{aligned} \alpha_{k-1} - \alpha_k &= (x_{k-1} - x_k)^T M^T A^{-1} M (x_{k-1} - x_k) - (x_{k-1} - x_k)^T N^T A^{-1} N (x_{k-1} - x_k) \\ &= (x_{k-1} - x_k)^T (M^T + N)(x_{k-1} - x_k) \end{aligned}$$

car $A^{-1}M - I = A^{-1}N$ et $N^T A^{-1} + I = M^T A^{-1}$. Or $x_{k-1} - x_k \neq 0$ sinon on aurait $Bx_{k-1} = x_k$ ce qui impliquerait que B a une valeur propre égale à 1, ce qui est impossible puisque $\rho(B) < 1$. Donc l'hypothèse sur $M^T + N$ implique que la suite $\{\alpha_k\}$ est décroissante et comme $\alpha_0 \leq 0$, il y a contradiction avec le fait que cette suite converge vers 0. Donc A est définie positive. □.

Corollaire 8.1 *A étant symétrique définie positive, la méthode de relaxation converge si, et seulement si, $0 < \omega < 2$.*

DÉMONSTRATION : On évalue simplement $M^T + N$

$$M^T + N = \frac{1}{\omega} D - E^T + \frac{1-\omega}{\omega} D + F = \frac{2-\omega}{\omega} D.$$

Les matrices non symétriques définies positives donnent lieu également à des résultats de convergence si leur diagonale est strictement dominante.

Théorème 8.4 *Si A est une matrice à diagonale strictement dominante, la méthode de Jacobi est convergente.*

DÉMONSTRATION : Une matrice à diagonale strictement dominante satisfait bien $a_{ii} > 0$, $\forall i$. La matrice $J = D^{-1}(E + F)$ satisfait

$$J_{ii} = 0, \quad J_{ij} = -\frac{a_{ij}}{a_{ii}} \text{ pour } i \neq j.$$

Si A est à diagonale strictement dominante, $a_{ii} > \sum_{j \neq i} |a_{ij}|$, donc $\sum_j |J_{ij}| < 1$, d'où $\rho(J) < 1$. □

8.3.3 Choix du coefficient de relaxation

C'est une question difficile à laquelle on peut répondre dans certains cas particuliers comme le cas des matrices tridiagonales par blocs.

Théorème 8.5 *Soit A une matrice tridiagonale par blocs définie positive. Alors $\rho(L_1) = \rho(J)^2 < 1$. La valeur optimale du coefficient de relaxation :*

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(J)^2}}.$$

DÉMONSTRATION : Voir Ciarlet [2, p. 105].

Exemple 8.1 *Illustrons ces derniers résultats par un exemple :*

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

1.– Méthode de Jacobi.

$$M = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},\ N = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ J = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}.$$

Les valeurs propres de J sont $1/2$ et $-1/2$ (la méthode converge)

$$\begin{aligned} 2u_1^{(k+1)} &= u_2^{(k)} + b_1 \\ 2u_2^{(k+1)} &= u_1^{(k)} + b_2. \end{aligned}$$

L'erreur est divisée par 2 à chaque itération.

2.– Méthode de Gauss-Seidel.

$$M = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix},\ N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ L_1 = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}.$$

Les valeurs propres de L_1 sont 0 et $1/4$ (la méthode converge)

$$\begin{aligned} 2u_1^{(k+1)} &= u_2^{(k)} + b_1 \\ 2u_2^{(k+1)} &= u_1^{(k+1)} + b_2 \\ &= \frac{1}{2}(u_2^{(k)} + b_1) + b_2. \end{aligned}$$

L'erreur est divisée par 4 à chaque itération.

3.– Méthode de Relaxation.

$$M = \begin{bmatrix} \frac{2}{\omega} & 0 \\ -1 & \frac{2}{\omega} \end{bmatrix},\ N = \begin{bmatrix} -2 + \frac{2}{\omega} & 1 \\ 0 & -2 + \frac{2}{\omega} \end{bmatrix},\ L_\omega = \begin{bmatrix} 1 - \omega & \frac{\omega}{2} \\ \frac{\omega(1-\omega)}{2} & 1 - \omega + \frac{\omega^2}{4} \end{bmatrix}.$$

Calculons le produit des 2 valeurs propres égal au déterminant :

$$\det(L_\omega) = (1 - \omega)^2.$$

La valeur optimale ω_{opt} est supérieure à 1 et elle correspond au rayon spectral minimal ; donc les 2 valeurs propres doivent être égales à $\omega - 1$. Or leur somme vaut $2 - 2\omega + \omega^2/4 = 2\omega - 2$. D'où $\omega_{opt} \approx 1.07$. Le taux de convergence est alors égal à 0.07 ($\approx (1/4)^2$).

Bibliographie

[1] Bonnans J. F., Gilbert J. C., Lemaréchal C. and Sagastizábal. *Optimisation Numérique*, volume 36 of *Mathématiques et Applications*. Springer, 1997.

[2] Ciarlet P.G. *Introduction à Analyse Numérique Matricielle et à l'Optimisation*. Masson, 1982.

[3] Lascaux P. et Théodor R. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, volume 1. Masson, Paris, 1994.

[4] Lascaux P. et Théodor R. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, volume 2. Masson, Paris, 1994.

[5] Golub G.H. and Van Loan C.F. *Matrix Computations*. The John Hopkins University Press, Baltimore, 1989.

[6] Laurent-Gengoux P. and Trystram D. *Comprendre l'informatique numérique*. Lavoisier, 1989.

[7] Luenberger D. *Linear and Nonlinear Programming*. Addison Wesley, Reading, MA, 1989.

[8] Meurant G. *Computer Solution of Large Systems*. Studies in Mathematics and its Applications. North Holland, 1999.

[9] Minoux M. *Programmation Mathématique Tome 1*. Dunond, Paris, 1983.

[10] Strang G. *Linear Algebra and its Applications*. Brooks Cole, 1988.

Index