Faculté des Sciences
Aix*Marseille Université

CTES
Centre de Télé-Enseignement Sciences

LICENCE 3 MATHEMATIQUES

Expédition dans la semaine n°	Etape	Code UE	N° d'envoi de l'UE
46	**2L3MAT**	SMI6U01T	**1**

Nom de l'UE : Analyse numérique et optimisation

Le cours contient 3 chapitres (systèmes linéaires, systèmes non linéaires, optimisation). Pour chaque semaine, il est proposé d'étudier une partie du cours, de faire des exercices (corrigés) et, éventuellement, de réaliser un TP en python. Les TP sont conseillés mais non obligatoires. Deux devoirs sont à rendre afin de bénéficier d'une note de contrôle continu.
note finale=max(note-examen, 1/3(2 note-examen + note-contrôle-continu)).

- Contenu de l'envoi : Polycopié, chapitre 1, paragraphe 1 à 4. TP 1 et 2

- Guide du travail à effectuer

Semaine 1 :
Etudier les paragraphes 1.1 (Objectifs), 1.2.1 (rappels d'algèbre linéaire) et 1.2.2 (discrétisation d'une équation) proposés (avec corrigés) : 3 (Théorème du rang), 4, 6 (Vrai ou faux) , 9 (La matrice K_3).
L'exercice 11 (Résolution d'un système sous forme particulière) fait partie du premier devoir (à rendre ultérieurement)
Semaine 2 :
Etudier le paragraphe 1.3 (méthodes directes) jusqu'au théorème 1.22 (décomposition de Choleski) sans la démonstration
Exercices proposés (avec corrigés) : 19 (Vrai ou faux), 21 (LU), 27 (Sur la méthode LLt), 28 (Décomposition LU d'une matrice à paramètres). Faire le TP 1
Semaine 3 :
Etudier la démonstration du théorème 1.22, terminer le paragraphe 1.3.
Exercices proposés (avec corrigés) : 28 (Décomposition LU d'une matrice à paramètres) et 29 (Echelonnement et factorisation LU et LDU).
Semaine 4 :
Etudier le paragraphe 1.4 (Normes et conditionnement d'une matrice) Exercices proposés (avec corrigés) : 39 (Normes induites particulières), 42 (Matrice diagonalisable et rayon spectral), 44 (Série de Neumann). Faire le TP2
L'exercice 48 (Conditionnement de la matrice transposée) fait partie du premier devoir (à rendre ultérieurement)

-Coordonnées de l'enseignant responsable de l'envoi
R. Herbin, I2M, 39 rue Joliot Curie, 13453 marseille cedex 13
email : raphaele.herbin@univ-amu.fr
Vous pouvez aussi consulter la page web: http://www.i2m.univ-amu.fr/~herbin
et me poser des questions par email

Introduction

L'objet de l'analyse numérique est de concevoir et d'étudier des méthodes de résolution de certains problèmes mathématiques, en général issus de la modélisation de problèmes "réels", et dont on cherche à calculer la solution à l'aide d'un ordinateur.
Les méthodes numériques pour la résolution des équations différentielles sont abordées dans le cours d'équations différentielles. Dans le cadre de ce cours, nous aborderons les thèmes suivants, qui font l'objet de trois grands chapitres :

- Systèmes linéaires
- Systèmes non linéaires
- Optimisation

On pourra consulter les ouvrages suivants pour ces différentes parties (ceci est une liste non exhaustive !) :

- A. Quarteroni, R. Sacco et F. Saleri, Méthodes Numériques : Algorithmes, Analyse et Applications, Springer 2006.
- P.G. Ciarlet, Introduction à l'analyse numérique et à l'optimisation, Masson, 1982, (pour les chapitre 1 à 3 de ce polycopié).
- L. Dumas, Modélisation à l'oral de l'agrégation, calcul scientifique, Collection CAPES/Agrégation, Ellipses, 1999.
- E. Hairer, polycopié du cours "Analyse Numérique", http ://www.unige.ch/ hairer/polycop.html
- J. Hubbard et F. Hubert, Calcul Scientifique, Vuibert.
- P. Lascaux et R. Théodor, Analyse numérique matricielle appliquée à l'art de l'ingénieur, tomes 1 et 2, Masson, 1987
- L. Sainsaulieu, Calcul scientifique cours et exercices corrigés pour le 2ème cycle et les éécoles d'ingénieurs, Enseignement des mathématiques, Masson, 1996.
- M. Schatzman, Analyse numérique, cours et exercices, (chapitres 1,2 et 4).
- D. Serre, Les matrices, Masson, (2000). (chapitres 1,2 et 4).
- P. Lascaux et R. Theodor, Analyse numérique sappliquée aux sciences de l'ingénieur, Paris, (1994)
- R. Temam, Analyse numérique, Collection SUP le mathématicien, Presses Universitaires de France, 1970.

Et pour les anglophiles...

- G. Dahlquist and A. Björck, Numerical Methods, Prentice Hall, Series in Automatic Computation, 1974, Englewood Cliffs, NJ.
- R. Fletcher, Practical methods of optimization, J. Wiley, New York, 1980 (chapitre 3).
- G. Golub and C. Van Loan, Matrix computations, The John Hopkins University Press, Baltimore (chapitre 1).
- R.S. Varga, Matrix iterative analysis, Prentice Hall, Englewood Cliffs, NJ 1962.

Pour des rappels d'algègre linéaire :

— Poly d'algèbre linéaire de première année, P. Bousquet, R. Herbin et F. Hubert, http ://www.cmi.univ-mrs.fr/ herbin/PUBLI/L1alg.pdf

— Introduction to linear algebra, Gilbert Strang, Wellesley Cambridge Press, 2008

Ce cours a été rédigé pour la licence de mathématiques à distance (téléenseignement) du CTES de l'université d'Aix-Marseille. Chaque section est suivie d'un certain nombre d'exercices. On donne ensuite des suggestions pour effectuer les exercices, puis des corrigés détaillés. Il est fortement conseillé d'essayer de faire les exercices d'abord sans ces indications, et de ne regarder les corrigés détaillés qu'une fois l'exercice achevé (même si certaines questions n'ont pas pu être effectuées), ceci pour se préparer aux conditions d'examen. N'hésitez pas à me contacter pour toute question sur le contenu du cours ou des exercices.

Chapitre 1

Systèmes linéaires

1.1 Objectifs

On note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées d'ordre n. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$, l'objectif est de résoudre le système linéaire $A\boldsymbol{x} = b$, c'est-à-dire de trouver $\boldsymbol{x}$ solution de :

$$\left\{ \begin{array}{l} \boldsymbol{x} \in \mathbb{R}^n \\ A\boldsymbol{x} = b \end{array} \right. \tag{1.1}$$

Comme A est inversible, il existe un unique vecteur $\boldsymbol{x} \in \mathbb{R}^n$ solution de (1.1). Nous allons étudier dans les deux paragraphes suivants des méthodes de calcul de ce vecteur $\boldsymbol{x}$: la première partie de ce chapitre sera consacrée aux méthodes "directes" et la deuxième aux méthodes "itératives". Nous aborderons ensuite en troisième partie les méthodes de résolution de problèmes aux valeurs propres.
Un des points essentiels dans l'efficacité des méthodes envisagées concerne la taille des systèmes à résoudre. La taille de la mémoire des ordinateurs a augmenté de façon drastique de 1980 à nos jours.
Le développement des méthodes de résolution de systèmes linéaires est liée à l'évolution des machines informatiques. C'est un domaine de recherche très actif que de concevoir des méthodes qui permettent de profiter au mieux de l'architecture des machines (méthodes de décomposition en sous domaines pour profiter des architectures parallèles, par exemple).
Dans la suite de ce chapitre, nous verrons deux types de méthodes pour résoudre les systèmes linéaires : les méthodes directes et les méthodes itératives. Pour faciliter la compréhension de leur étude, nous commençons par quelques rappels d'algèbre linéaire.

1.2 Pourquoi et comment ?

Nous donnons dans ce paragraphe un exemple de problème dont la résolution numérique recquiert la résolution d'un système linéaire, et qui nous permet d'introduire des matrices que nous allons beaucoup étudier par la suite. Nous commençons par donner ci-après après quelques rappels succincts d'algèbre linéaire, outil fondamental pour la résolution de ces systèmes linéaires.

1.2.1 Quelques rappels d'algèbre linéaire

Quelques notions de base

Ce paragraphe rappelle des notions fondamentales que vous devriez connaître à l'issue du cours d'algèbre linéaire de première année. On va commencer par revisiter le **produit matriciel**, dont la vision combinaison linéaire de lignes est fondamentale pour bien comprendre la forme matricielle de la procédure d'élimination de Gauss.

Soient A et B deux matrices carrées d'ordre n, et $M = AB$. Prenons comme exemple d'illustration

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} -1 & 0 \\ 3 & 2 \end{bmatrix} \text{ et } M = \begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix}$$

On note $a_{i,j}$, $b_{i,j}$ et $m_{i,j}$, $i, j = 1, \dots n$ les coefficients respectifs de A, B et M. Vous savez bien sûr que

$$m_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}. \tag{1.2}$$

On peut écrire les matrices A et B sous forme de lignes (notées $\boldsymbol{\ell}_i$) et colonnes (notées $\boldsymbol{c}_j$) :

$$A = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \dots \\ \boldsymbol{\ell}_n(A) \end{bmatrix} \text{ et } B = \begin{bmatrix} \boldsymbol{c}_1(B) & \dots & \boldsymbol{c}_n(B) \end{bmatrix}$$

Dans nos exemples, on a donc

$$\boldsymbol{\ell}_1(A) = \begin{bmatrix} 1 & 2 \end{bmatrix}, \ \boldsymbol{\ell}_2(A) = \begin{bmatrix} 0 & 1 \end{bmatrix}, \ \boldsymbol{c}_1(B) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \ \boldsymbol{c}_2(B) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

L'expression (1.2) s'écrit encore

$$m_{i,j} = \boldsymbol{\ell}_i(A) \boldsymbol{c}_j(B),$$

qui est le produit d'une matrice $1 \times n$ par une matrice $n \times 1$, qu'on peut aussi écrire sous forme d'un produit scalaire :

$$m_{i,j} = (\boldsymbol{\ell}_i(A))^t \cdot \boldsymbol{c}_j(B)$$

où $(\boldsymbol{\ell}_i(A))^t$ désigne la matrice transposée, qui est donc maintenant une matrice $n \times 1$ qu'on peut identifier à un vecteur de $\mathbb{R}^n$. C'est la technique "habituelle" de calcul du produit de deux matrices. On a dans notre exemple :

$$\begin{aligned} m_{1,2} &= \boldsymbol{\ell}_1(A)\, \boldsymbol{c}_2(B) = \boldsymbol{\ell}_1(A)\, \boldsymbol{c}_2(B) = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\ &= (\boldsymbol{\ell}_i(A))^t \cdot \boldsymbol{c}_j(B) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\ &= 4. \end{aligned}$$

Mais de l'expression (1.2), on peut aussi avoir l'expression des lignes et des colonnes de $M = AB$ en fonction des lignes de B ou des colonnes de A :

$$\boldsymbol{\ell}_i(AB) = \sum_{k=1}^{n} a_{i,k} \boldsymbol{\ell}_k(B) \tag{1.3}$$

$$\boldsymbol{c}_j(AB) = \sum_{k=1}^{n} b_{k,j} \boldsymbol{c}_k(A) \tag{1.4}$$

Dans notre exemple, on a donc :

$$\boldsymbol{\ell}_1(AB) = \begin{bmatrix} -1 & 0 \end{bmatrix} + 2 \begin{bmatrix} 3 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 4 \end{bmatrix}$$

ce qui montre que la ligne 1 de AB est une combinaison linéaire des lignes de B. Le colonnes de AB, par contre, sont des combinaisons linéaires de colonnes de A. Par exemple :

$$\boldsymbol{c}_2(AB) = 0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

Il faut donc retenir que dans un produit matriciel AB,

les colonnes de AB sont des combinaisons linéaires des colonnes de A
les lignes de AB sont des combinaisons linéaires des lignes de B.

Cette remarque est très importante pour la représentation matricielle de l'élimination de Gauss : lorqu'on calcule des systèmes équivalents, on effectue des combinaisons linéaires de lignes, et donc on multiplie à gauche par une matrice d'élimination.
Il est intéressant pour la suite de ce cours de voir ce que donne la multiplication d'une matrice par une matrice de permutation.
Commençons par une exemple. Soit P et A des matrices carrées d'ordre 2 définies par

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},\ PA = \begin{bmatrix} c & d \\ a & b \end{bmatrix},\ AP = \begin{bmatrix} b & a \\ d & c \end{bmatrix}.$$

La multiplication de A par la matrice P échange les lignes de A lorqu'on multiplie A par P à gauche, et elle échange les colonnes de A lorqu'on multiplie A par P à droite. Noter que ceci montre d'ailleurs bien que le produit matriciel n'est pas commutatif...La matrice P s'appelle matrice de permutation. Les matrices de permutation auront un fort rôle à jouer dans l'élaboration d'algorithmes de résolution des systèmes linéaires (voir l'algorithme de Gauss avec pivot partiel).
De manière plus générale, on peut définir une matrice de permutation de la façon suivante :

Définition 1.1 (Matrice de permutation)**.** *Soit $n \in \mathbb{N}$ et soient $i, j \in \{1, \ldots, n\}$. On notera $P^{(i \leftrightarrow j)} \in \mathcal{M}_n(\mathbb{R})$ la matrice telle que :*

1. *Si $i = j$, $P^{(i \leftrightarrow j)} = \mathrm{Id}_n$,*
2. *Si $i \neq j$, $p_{i,i}^{(i \leftrightarrow j)} = p_{j,j}^{(i \leftrightarrow j)} = 0$, $p_{i,j}^{(i \leftrightarrow j)} = p_{j,i}^{(i \leftrightarrow j)} = 1$, et pour tout $k, l \in \{1, \ldots, n\}$ tel que $(k,l) \notin \{(i,i),(i,j),(j,i),(j,j)\}$, si $k = l$, $p_{k,l}^{(i \leftrightarrow j)} = 1$ sinon $p_{k,l}^{(i \leftrightarrow j)} = 0$.*

La matrice $P^{(i \leftrightarrow j)}$ est alors appelée matrice de permutation élémentaire. Une matrice de permutation est définie comme le produit d'un nombre fini de permutations élémentaires.

Remarquons qu'une matrice de permutation possède alors n termes égaux à 1, et tous les autres égaux à 0, tels que chaque ligne et chaque colonne comprenne exactement l'un des termes égaux à 1 (pour les amateurs de jeu d'échecs, ces termes sont disposés comme n tours sur un échiquier de taille $n \times n$ telles qu'aucune tour ne peut en prendre une autre).
Pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ et toute matrice de permutation P, la matrice PA est obtenue à partir de A par permutation des lignes de A, et la matrice AP est obtenue à partir de A par permutation des colonnes de A. Dans un système linéaire $Ax = b$, on remarque qu'on ne change pas la solution x si on permute des lignes, c'est à dire si l'on résout $PAx = Pb$. Notons que le produit de matrices de permutation est évidemment une matrice de permutation, et que toute matrice de permutation P est inversible et $P^{-1} = P^t$ (voir exercice 2).

Le tableau ci-dessous est la traduction littérale de "Linear algebra in a nutshell", par Gilbert Strang [1] Pour une matrice carrée A, on donne les caractérisations du fait qu'elle est inversible ou non.
On rappelle pour une bonne lecture de ce tableau les quelques définitions suivantes (s'il y a des notions que vous avez oubliées ou que vous ne maîsez :

Définition 1.2 (Pivot)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n. On appelle pivot de A le premier élément non nul de chaque ligne dans la forme échelonnée de A obtenue par élimination de Gauss. Si la matrice est inversible, elle a donc n pivots (non nuls).*

1. Voir la page web de Strang `www.mit.edu/~gs` pour une foule d'informations et de cours sur l'algèbre linéaire.

A inversible	A non inversible
Les vecteurs colonne sont indépendants	Les vecteurs colonne sont liés
Les vecteurs ligne sont indépendants	Les vecteurs ligne sont liés
Le déterminant est non nul	Le déterminant est nul
$A\boldsymbol{x} = 0$ a une unique solution $\boldsymbol{x} = \mathbf{0}$	$A\boldsymbol{x} = \mathbf{0}$ a une infinité de solutions
Le noyau de A est réduit à $\{\mathbf{0}\}$	Le noyau de A contient au moins un vecteur non nul
$A\boldsymbol{x} = \boldsymbol{b}$ a une solution unique $x = A^{-1}\boldsymbol{b}$	$A\boldsymbol{x} = \boldsymbol{b}$ a soit aucune solution, soit une infinité
A a n pivots (non nuls)	A a $r < n$ pivots
A est de rang maximal : $\text{rang}(A) = n$.	$\text{rang}(A) = r < n$
La forme totatement échelonnée R de A est la matrice identité	R a au moins une ligne de zéros
L'image de A est tout $\mathbb{R}^n$	L'image de A est strictement incluse dans $\mathbb{R}^n$
L'espace $L(A)$ engendré par les lignes de A est tout $\mathbb{R}^n$	$L(A)$ est de dimension $r < n$
Toutes les valeurs propres de A sont non nulles	Zéro est valeur propre de A
A^tA est symétrique définie positive [2]	A^tA n'est que semi-définie

TABLE 1.1: Extrait de "Linear algebra in a nutshell", G. Strang

Définition 1.3 (Valeurs propres)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $\boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{x} \neq 0$ tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. L'élément $\boldsymbol{x}$ est appelé vecteur propre de A associé à λ.*

Définition 1.4 (Déterminant)**.** *Il existe une unique application, notée* $\det$ *de $\mathcal{M}_n(\mathbb{R})$ dans $\mathbb{R}$ qui vérifie les propriétés suivantes*

(D1) Le déterminant de la matrice identité est égal à 1.

(D2) Si la matrice $\tilde{A}$ est obtenue à partir de A par échange de deux lignes, alors $\det\tilde{A} = -\det A$.

(D3) Le déterminant est une fonction linéaire de chacune des lignes de la matrice A.

(D3a) (multiplication par un scalaire) si $\tilde{A}$ est obtenue à partir de A en multipliant tous les coefficients d'une ligne par $\lambda \in \mathbb{R}$, alors $\det(\tilde{A}) = \lambda\det(A)$.

(D3b) (addition) si $A = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \boldsymbol{\ell}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$, $\tilde{A} = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \tilde{\boldsymbol{\ell}}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$ *et* $B = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \boldsymbol{\ell}_k(A) + \tilde{\boldsymbol{\ell}}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$, *alors*

$$\det(B) = \det(A) + \det(\tilde{A}).$$

On peut déduire de ces trois propriétés fondamentales un grand nombre de propriétés importantes, en particulier le fait que $\det(AB) = \det A \, \det B$ *et que le déterminant d'une matrice inversible est le produit des pivots : c'est de cette manière qu'on le calcule sur les ordinateurs. En particulier on n'utilise jamais la formule de Cramer, beaucoup trop coûteuse en termes de nombre d'opérations.*

On rappelle que si $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, les valeurs propres sont les racines du **polynôme caractéristique** P_A de degré n, qui s'écrit :

$$P_A(\lambda) = \det(A - \lambda I).$$

Matrices diagonalisables

Un point important de l'algèbre linéaire, appelé "réduction des endomorphismes" dans les programmes français, consiste à se demander s'il existe une base de l'espace dans laquelle la matrice de l'application linéaire est diagonale ou tout au moins triangulaire (on dit aussi trigonale).

Définition 1.5 (Matrice diagonalisable dans $\mathbb{R}$)**.** *Soit A une matrice réelle carrée d'ordre n. On dit que A est diagonalisable dans $\mathbb{R}$ s'il existe une base $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ de $\mathbb{R}^n$ et des réels $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots, n$. Les réels $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A, et les vecteurs $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ sont des vecteurs propres associés.*

Vous connaissez sûrement aussi la diagonalisation dans $\mathbb{C}$: une matrice réelle carrée d'ordre n admet toujours n valeurs propres dans $\mathbb{C}$, qui ne sont pas forcément distinctes. Une matrice est diagonalisable dans $\mathbb{C}$ s'il existe une base $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ de $\mathbb{C}^n$ et des nombres complexes $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots, n$. Ceci est vérifié si la dimension de chaque sous espace propre $E_i = \mathrm{Ker}(A - \lambda_i \mathrm{Id})$ (appelée multiplicité géométrique) est égale a la multiplicité algébrique de λ_i, c'est-à-dire son ordre de multiplicité en tant que racine du polynôme caractéristique.
Par exemple la matrice $A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ n'est pas diagonalisable dans $\mathbb{C}$ (ni évidemment, dans $\mathbb{R}$). Le polynôme caractéristique de A est $P_A(\lambda) = \lambda^2$, l'unique valeur propre est donc 0, qui est de multiplicité algébrique 2, et de multiplicité géométrique 1, car le sous espace propre associé à la valeur propre nulle est $F = \{\boldsymbol{x} \in \mathbb{R}^2 \, ; A\boldsymbol{x} = 0\} = \{\boldsymbol{x} = (0, t), t \in \mathbb{R}\}$, qui est de dimension 1.
Ici et dans toute la suite, comme on résout des systèmes linéaires réels, on préfère travailler avec la diagonalisation dans $\mathbb{R}$; cependant il y a des cas où la diagonalisation dans $\mathbb{C}$ est utile et même nécessaire (étude de stabilité des systèmes diférentiels, par exemple). Par souci de clarté, nous préciserons toujours si la diagonalisation considérée est dans $\mathbb{R}$ ou dans $\mathbb{C}$.

Lemme 1.6. *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

$$A = P\mathrm{diag}(\lambda_1, \ldots, \lambda_n)P^{-1},$$

où P est la matrice dont les vecteurs colonnes sont égaux à des vecteurs propres $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ associées aux valeurs propres $\lambda_1, \ldots, \lambda_n$.

DÉMONSTRATION – Par définition d'un vecteur propre, on a $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots n$, et donc, en notant P la matrice dont les colonnes sont les vecteurs propres $\boldsymbol{u}_i$,

$$\begin{bmatrix} A\boldsymbol{u}_1 & \ldots & A\boldsymbol{u}_n \end{bmatrix} = A \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \end{bmatrix} = AP$$

et donc

$$AP = \begin{bmatrix} \lambda_1 \boldsymbol{u}_1 & \ldots & \lambda_n \boldsymbol{u}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} = P\mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Notons que dans ce calcul, on a fortement utilisé la multiplication des matrices par colonnes, c.à.d.

$$\boldsymbol{c}_i(AB) = \sum_{j=1}^{n} a_{i,j} \boldsymbol{c}_j(B).$$

Remarquons que P lest aussi la matrice définie (de manière unique) par $P\boldsymbol{e}_i = \boldsymbol{u}_i$, où $(e_i)_{i=1,\dots,n}$ est la base canonique de $\mathbb{R}^n$, c'est-à-dire que $(\boldsymbol{e}_i)_j = \delta_{i,j}$. La matrice P est appelée matrice de passage de la base $(\boldsymbol{e}_i)_{i=1,\dots,n}$ à la base $(\boldsymbol{u}_i)_{i=1,\dots,n}$; (il est bien clair que la i-ème colonne de P est constituée des composantes de $\boldsymbol{u}_i$ dans la base canonique $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_n)$.
La matrice P est inversible car les vecteurs propres forment une base, et on peut donc aussi écrire :

$$P^{-1}AP = \mathrm{diag}(\lambda_1,\dots,\lambda_n) \text{ ou } A = P\mathrm{diag}(\lambda_1,\dots,\lambda_n)P^{-1}.$$

■

La diagonalisation des matrices réelles symétriques est un outil qu'on utilisera souvent dans la suite, en particulier dans les exercices. Il s'agit d'un résultat extrêmement important.

Lemme 1.7 (Une matrice symétrique est diagonalisable dans $\mathbb{R}$)**.** *Soit E un espace vectoriel sur $\mathbb{R}$ de dimension finie : $dimE = n$, $n \in \mathbb{N}^*$, muni d'un produit scalaire i.e. d'une application*

$$E \times E \to \mathbb{R},$$
$$(\boldsymbol{x},\boldsymbol{y}) \to (\boldsymbol{x} \mid \boldsymbol{y})_E,$$

qui vérifie :

$$\forall \boldsymbol{x} \in E, (\boldsymbol{x} \mid \boldsymbol{x})_E \geq 0 \text{ et } (\boldsymbol{x} \mid \boldsymbol{x})_E = 0 \Leftrightarrow x = 0,$$
$$\forall (\boldsymbol{x},\boldsymbol{y}) \in E^2, (\boldsymbol{x} \mid \boldsymbol{y})_E = (\boldsymbol{y} \mid \boldsymbol{x})_E,$$
$$\forall y \in E, \text{ l'application de } E \text{ dans } \mathbb{R}, \text{ définie par } \boldsymbol{x} \to (\boldsymbol{x} \mid \boldsymbol{y})_E \text{ est linéaire.}$$

Ce produit scalaire induit une norme sur E définie par $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x} \mid \boldsymbol{x})_E}$.
Soit T une application linéaire de E dans E. On suppose que T est symétrique, c.à.d. que $(T(\boldsymbol{x}) \mid \boldsymbol{y})_E = (\boldsymbol{x} \mid T(\boldsymbol{y}))_E$, $\forall (\boldsymbol{x},\boldsymbol{y}) \in E^2$. Alors il existe une base orthonormée $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ de E (c.à.d. telle que $(\boldsymbol{f}_i \mid \boldsymbol{f}_j)_E = \delta_{i,j}$) et $\lambda_1,\dots,\lambda_n$ dans $\mathbb{R}$ tels que $T(f_i) = \lambda_i \boldsymbol{f}_i$ pour tout $i \in \{1 \dots n\}$.

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^n$, le produit scalaire canonique de $\boldsymbol{x} = (x_1,\dots,x_n)^t$ et $\boldsymbol{y} = (y_1,\dots,y_n)^t$ est défini par $(\boldsymbol{x} \mid \boldsymbol{y})_E = \boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^n x_i y_i$. Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(\boldsymbol{x}) = A\boldsymbol{x}$ est linéaire, et :

$$(T\boldsymbol{x}|\boldsymbol{y}) = A\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x} \cdot A^t\boldsymbol{y} = \boldsymbol{x} \cdot A\boldsymbol{y} = (\boldsymbol{x} \mid T\boldsymbol{y}).$$

Donc T est linéaire symétrique. Par le lemme précédent, il existe $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ et $(\lambda_1 \dots \lambda_n) \in \mathbb{R}$ tels que $T\boldsymbol{f}_i = A\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i \, \forall\, i \in \{1,\dots,n\}$ et $f_i \cdot \boldsymbol{f}_j = \delta_{i,j}, \forall\, (i,j) \in \{1,\dots,n\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_n)$ base canonique de $\mathbb{R}^n$ dans la base $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ dont la i-ème colonne de P est constituée des coordonnées de $\boldsymbol{f}_i$ dans la base $(\boldsymbol{e}_1 \dots \boldsymbol{e}_n)$. On a : $P\boldsymbol{e}_i = \boldsymbol{f}_i$. On a alors $P^{-1}AP\boldsymbol{e}_i = P^{-1}A\boldsymbol{f}_i = P^{-1}(\lambda_i \boldsymbol{f}_i) = \lambda_i \boldsymbol{e}_i = \mathrm{diag}(\lambda_1,\dots,\lambda_n)\boldsymbol{e}_i$, où $\mathrm{diag}(\lambda_1,\dots,\lambda_n)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1,\dots,\lambda_n$. On a donc :

$$P^{-1}AP = \begin{bmatrix} \lambda_i & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = D.$$

De plus P est orthogonale, *i.e.* $P^{-1} = P^t$. En effet,

$$P^tP\boldsymbol{e}_i \cdot \boldsymbol{e}_j = P\boldsymbol{e}_i \cdot P\boldsymbol{e}_j = (\boldsymbol{f}_i|\boldsymbol{f}_j) = \delta_{i,j}, \ \forall i,j \in \{1 \dots n\},$$

et donc $(P^tP\boldsymbol{e}_i - \boldsymbol{e}_i) \cdot \boldsymbol{e}_j = 0, \forall j \in \{1 \dots n\}, \forall i \in \{1,\dots n\}$. On en déduit que $P^tP\boldsymbol{e}_i = \boldsymbol{e}_i$ pour tout $i = 1,\dots n$, *i.e.* $P^tP = PP^t = \mathrm{Id}$.

DÉMONSTRATION *du lemme 1.7* Cette démonstration se fait par récurrence sur la dimension de E. On note $(\cdot|\cdot)$ le produit scalaire dans E et $\|\cdot\|$ la norme associée.

1ère étape. On suppose $dimE = 1$. Soit $\boldsymbol{e} \in E$, $e \neq 0$, alors $E = \mathbb{R}\boldsymbol{e} = \mathbb{R}\boldsymbol{f}_1$ avec $\boldsymbol{f}_1 = \frac{1}{\|e\|}\boldsymbol{e}$. Soit $T : E \to E$ linéaire. On a : $T\boldsymbol{f}_1 \in \mathbb{R}\boldsymbol{f}_1$ donc il existe $\lambda_1 \in \mathbb{R}$ tel que $T\boldsymbol{f}_1 = \lambda_1\boldsymbol{f}_1$.

2ème étape. On suppose le lemme vrai si dim $E < n$. On montre alors le lemme si $dimE = n$. Soit E un espace vectoriel normé sur $\mathbb{R}$ tel que $dimE = n$ et $T : E \to E$ linéaire symétrique. Soit φ l'application définie par :

$$\begin{aligned} \varphi : \quad & E \to \mathbb{R} \\ & \boldsymbol{x} \to (T\boldsymbol{x}|\boldsymbol{x}). \end{aligned}$$

L'application φ est continue sur la sphère unité $S_1 = \{\boldsymbol{x} \in E|\ \|\boldsymbol{x}\| = 1\}$ qui est compacte car dim $E < +\infty$; il existe donc $\boldsymbol{e} \in S_1$ tel que $\varphi(x) \leq \varphi(\boldsymbol{e}) = (T\boldsymbol{e} \mid \boldsymbol{e}) = \lambda$ pour tout $\boldsymbol{x} \in E$. Soit $\boldsymbol{y} \in E \setminus \{0\}$ et soit $t \in]0, \frac{1}{\|y\|}[$ alors $\boldsymbol{e} + t\boldsymbol{y} \neq 0$. On en déduit que :

$$\frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y}) \in S_1 \text{ et donc } \varphi(\boldsymbol{e}) = \lambda \geq \left(T\left(\frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y})\right) | \frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y}))\right)_E$$

donc $\lambda(\boldsymbol{e}+t\boldsymbol{y} \mid \boldsymbol{e}+t\boldsymbol{y})_E \geq (T(\boldsymbol{e}+t\boldsymbol{y}) \mid \boldsymbol{e}+t\boldsymbol{y})$. En développant on obtient :

$$\lambda[2t(\boldsymbol{e} \mid \boldsymbol{y}) + t^2(\boldsymbol{y} \mid \boldsymbol{y})_E] \geq 2t(T(\boldsymbol{e}) \mid \boldsymbol{y}) + t^2(T(\boldsymbol{y}) \mid \boldsymbol{y})_E.$$

Comme $t > 0$, ceci donne :

$$\lambda[2(\boldsymbol{e} \mid \boldsymbol{y}) + t(\boldsymbol{y} \mid \boldsymbol{y})_E] \geq 2(T(\boldsymbol{e}) \mid \boldsymbol{y}) + t(T(\boldsymbol{y}) \mid \boldsymbol{y})_E.$$

En faisant tendre t vers 0^+, on obtient $2\lambda(\boldsymbol{e} \mid \boldsymbol{y})_E \geq 2(T(\boldsymbol{e}) \mid \boldsymbol{y})$, soit encore $0 \geq (T(\boldsymbol{e}) - \lambda\boldsymbol{e} \mid \boldsymbol{y})$ pour tout $\boldsymbol{y} \in E \setminus \{0\}$. De même pour $\boldsymbol{z} = -\boldsymbol{y}$ on a $0 \geq (T(\boldsymbol{e}) - \lambda\boldsymbol{e}|\boldsymbol{z})$ donc $(T(\boldsymbol{e}) - \lambda\boldsymbol{e} \mid \boldsymbol{y}) \geq 0$. D'où $(T(\boldsymbol{e}) - \lambda\boldsymbol{e} \mid \boldsymbol{y}) = 0$ pour tout $\boldsymbol{y} \in E$. On en déduit que $T(\boldsymbol{e}) = \lambda\boldsymbol{e}$. On pose $\boldsymbol{f}_n = \boldsymbol{e}$ et $\lambda_n = \lambda$.

Soit $F = \{\boldsymbol{x} \in E; (\boldsymbol{x} \mid \boldsymbol{e}) = 0\}$, on a donc $F \neq E$, et $E = F \bigoplus \mathbb{R}\boldsymbol{e}$: On peut décomposer $\boldsymbol{x} \in E$ comme $\boldsymbol{x} = \boldsymbol{x} - (\boldsymbol{x} \mid \boldsymbol{e})\boldsymbol{e} + (\boldsymbol{x} \mid \boldsymbol{e})\boldsymbol{e}$. Si $\boldsymbol{x} \in F$, on a aussi $T(\boldsymbol{x}) \in F$ (car T est symétrique). L'application $S = T|_F$ est alors une application linéaire symétrique de F dans F et on a $dimF = n - 1$. On peut donc utiliser l'hypothèse de récurrence : $\exists\lambda_1 \dots \lambda_{n-1}$ dans $\mathbb{R}$ et $\exists\boldsymbol{f}_1 \dots \boldsymbol{f}_{n-1}$ dans E tels que $\forall\, i \in \{1 \dots n-1\}$, $S\boldsymbol{f}_i = T\boldsymbol{f}_i = \lambda_i\boldsymbol{f}_i$, et $\forall i, j \in \{1 \dots n-1\}, \boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$. Et donc $(\lambda_1 \dots \lambda_n)$ et $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ conviennent. ∎

1.2.2 Discrétisation de l'équation de la chaleur

Dans ce paragraphe, nous prenons un exemple très simple pour obtenir un système linéaire à partir de la discrétisation d'un problème continu.

L'équation de la chaleur unidimensionnelle

Discrétisation par différences finies de $-u'' = f$ Soit $f \in C([0,1], \mathbb{R})$. On cherche u tel que

$$-u''(x) = f(x) \tag{1.5a}$$

$$u(0) = u(1) = 0. \tag{1.5b}$$

Remarque 1.8 (Problèmes aux limites, problèmes à conditions initiales)**.** *L'équation différentielle $-u'' = f$ admet une infinité de solutions. Pour avoir existence et unicité, il est nécessaire d'avoir des conditions supplémentaires. Si l'on considère deux conditions en 0 (ou en 1, l'origine importe peu) on a ce qu'on appelle un problème de Cauchy, ou problème à conditions initiales. Le problème* (1.5) *est lui un problème aux limites : il y a une condition pour chaque bord du domaine. En dimension supérieure, le problème $-\Delta u = f$ nécessite une condition sur au moins "un bout" de frontière pour être bien posé : voir le cours d'équations aux dérivées partielles de master pour plus de détails à ce propos.*

On peut montrer (on l'admettra ici) qu'il existe une unique solution $u \in C^2([0,1], \mathbb{R})$. On cherche à calculer u de manière approchée. On va pour cela introduire la méthode de discrétisation dite *par différences finies*. Soit $n \in \mathbb{N}^*$, on définit $h = 1/(n+1)$ le *pas de discrétisation*, c.à.d. la distance entre deux points de discrétisation,

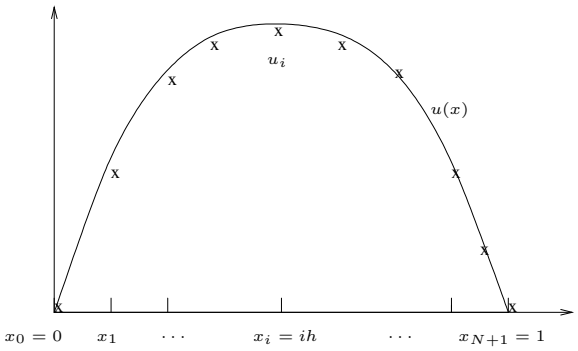

FIGURE 1.1: Solution exacte et approchée de $-u'' = f$

et pour $i = 0, \ldots, n+1$ on définit les points de discrétisation $x_i = ih$ (voir Figure 1.1), qui sont les points où l'on va écrire l'équation $-u'' = f$ en vue de se ramener à un système discret, c.à.d. à un système avec un nombre fini d'inconnues $u_1, \ldots, u_n$. Remarquons que $x_0 = 0$ et $x_{n+1} = 1$, et qu'en ces points, u est spécifiée par les conditions limites (1.5b). Soit $u(x_i)$ la valeur exacte de u en x_i. On écrit la première équation de (1.5a) en chaque point x_i, pour $i = 1 \ldots n$.

$$-u''(x_i) = f(x_i) = b_i \, \forall i \in \{1 \ldots n\}. \tag{1.6}$$

Supposons que $u \in C^4([0,1], \mathbb{R})$ (ce qui est vrai si $f \in C^2$). Par développement de Taylor, on a :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i),$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i),$$

avec $\xi_i \in]x_i, x_{i+1}[$ et $\eta_i \in]x_i, x_{i+1}[$. En sommant ces deux égalités, on en déduit que :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i).$$

On définit l'erreur de consistance, qui mesure la manière dont on a approché $-u''(x_i)$; l'erreur de consistance R_i au point x_i est définie par

$$R_i = u''(x_i) - \frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2}. \tag{1.7}$$

On a donc :

$$\begin{aligned} |R_i| &= \left| -\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} + u''(x_i) \right| \\ &\leq \left| \frac{h^2}{24}u^{(4)}(\xi_i) + \frac{h^2}{24}u^{(4)}(\eta_i) \right| \\ &\leq \frac{h^2}{12}\|u^{(4)}\|_\infty. \end{aligned} \tag{1.8}$$

où $\|u^{(4)}\|_\infty = \sup_{x \in]0,1[} |u^{(4)}(x)|$. Cette majoration nous montre que l'erreur de consistance tend vers 0 comme h^2 : on dit que le schéma est *consistant d'ordre 2*.
On introduit alors les inconnues $(u_i)_{i=1,\ldots,n}$ qu'on espère être des valeurs approchées de u aux points x_i et qui sont les composantes de la solution (si elle existe) du système suivant, avec $b_i = f(x_i)$,

$$\begin{cases} -\dfrac{u_{i+1} + u_{i-1} - 2u_i}{h^2} = b_i, \quad \forall i \in [\![1, n]\!], \\ u_0 = u_{n+1} = 0. \end{cases} \tag{1.9}$$

On cherche donc $\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \in \mathbb{R}^n$ solution de (1.9). Ce système peut s'écrire sous forme matricielle :$K_n \boldsymbol{u} = \boldsymbol{b}$ où $\boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ et K_n est la matrice carrée d'ordre n de coefficients $(k_{i,j})_{i,j=1,n}$ définis par :

$$\left\{ \begin{array}{ll} k_{i,i} & = \dfrac{2}{h^2}, \, \forall \, i = 1, \ldots, n, \\ k_{i,j} & = -\dfrac{1}{h^2}, \, \forall \, i = 1, \ldots, n, \;\; j = i \pm 1, \\ k_{i,j} & = 0, \, \forall \, i = 1, \ldots, n, \; |i - j| > 1. \end{array} \right. \tag{1.10}$$

On remarque immédiatement que K_n est tridiagonale.
On peut montrer que K_n est symétrique définie positive (voir exercice 14 page 20), et elle est donc inversible Le système $K_n \boldsymbol{u} = \boldsymbol{b}$ admet donc une unique solution. C'est bien, mais encore faut il que cette solution soit ce qu'on espérait, c.à.d. que chaque valeur u_i soit une approximation pas trop mauvaise de $u(x_i)$. On appelle erreur de discrétisation en x_i la différence de ces deux valeurs :

$$e_i = u(x_i) - u_i, \;\; i = 1, \ldots, n. \tag{1.11}$$

Si on appelle $\boldsymbol{e}$ le vecteur de composantes e_i et R le vecteur de composantes R_i on déduit de la définition (1.7) de l'erreur de consistance et des équations (exactes) (1.6) que

$$K_n \boldsymbol{e} = R \text{ et donc } \boldsymbol{e} = K_n^{-1} R. \tag{1.12}$$

Le fait que le schéma soit consistant est une bonne chose, mais cela ne suffit pas à montrer que le schéma est convergent, c.à.d. que l'erreur entre $\max_{i=1,\ldots,n} e_i$ tend vers 0 lorsque h tend vers 0, parce que K_n dépend de n (c'est-à-dire de h). Pour cela, il faut de plus que le schéma soit *stable*, au sens où l'on puisse montrer que $\|K_n^{-1}\|$ est borné indépendamment de h, ce qui revient à trouver une estimation sur les valeurs approchées u_i indépendante de h. La stabilité et la convergence font l'objet de l'exercice 57, où l'on montre que le schéma est convergent, et qu'on a l'estimation d'erreur suivante :

$$\max_{i=1\ldots n} \{|u_i - u(x_i)|\} \leq \frac{h^2}{96} \|u^{(4)}\|_\infty.$$

Cette inégalité donne la précision de la méthode (c'est une méthode dite d'ordre 2). On remarque en particulier que si on raffine la discrétisation, c'est–à–dire si on augmente le nombre de points n ou, ce qui revient au même, si on diminue le pas de discrétisation h, on augmente la précision avec laquelle on calcule la solution approchée.

L'équation de la chaleur bidimensionnelle

Prenons maintenant le cas d'une discrétisation du Laplacien sur un carré par différences finies. Si u est une fonction de deux variables x et y à valeurs dans $\mathbb{R}$, et si u admet des dérivées partielles d'ordre 2 en x et y, l'opérateur laplacien est défini par $\Delta u = \partial_{xx} u + \partial_{yy} u$. L'équation de la chaleur bidimensionnelle s'écrit avec cet opérateur. On cherche à résoudre le problème :

$$\begin{array}{l} -\Delta u = f \text{ sur } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ sur } \partial\Omega, \end{array} \tag{1.13}$$

On rappelle que l'opérateur Laplacien est défini pour $u \in C^2(\Omega)$, où Ω est un ouvert de $\mathbb{R}^2$, par

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Définissons une discrétisation uniforme du carré par les points (x_i, y_j), pour $i = 1, \ldots, M$ et $j = 1, \ldots, M$ avec $x_i = ih$, $y_j = jh$ et $h = 1/(M+1)$, representée en figure 1.2 pour $M = 6$. On peut alors approcher les dérivées secondes par des quotients différentiels comme dans le cas unidimensionnel (voir page 12), pour obtenir un système linéaire : $A\boldsymbol{u} = \boldsymbol{b}$ où $A \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{b} \in \mathbb{R}^n$ avec $n = M^2$. Utilisons l'ordre"*lexicographique*" pour numéroter les inconnues, c.à.d. de bas en haut et de gauche à droite : les inconnues sont alors numérotées de 1 à $n = M^2$ et le second membre s'écrit $b = (b_1, \ldots, b_n)^t$. Les composantes $b_1, \ldots, b_n$ sont définies par :pour $i, j = 1, \ldots, M$, on pose $k = j + (i-1)M$ et $b_k = f(x_i, y_j)$.

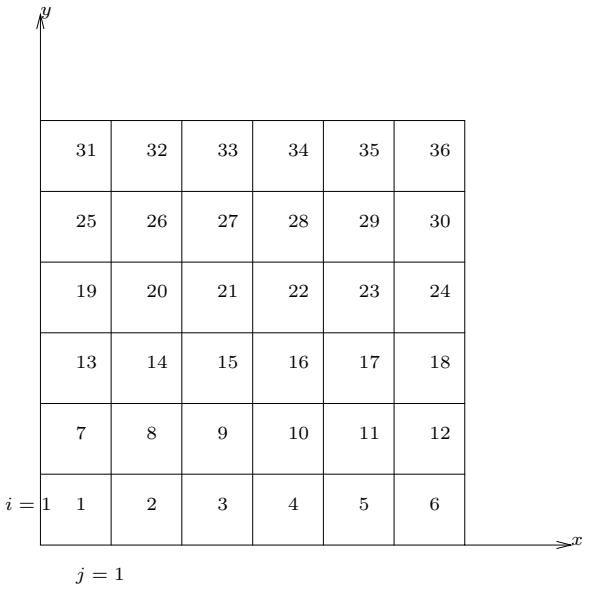

FIGURE 1.2: Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

Les coefficients de $A = (a_{k,\ell})_{k,l=1,n}$ peuvent être calculés de la manière suivante :

$$
\begin{cases}
\text{Pour} & i, j = 1, \ldots, M, \text{ on pose } k = j + (i-1)M, \\
a_{k,k} & = \dfrac{4}{h^2}, \\
a_{k,k+1} = & \begin{cases} -\dfrac{1}{h^2} & \text{si } j \neq M, \\ 0 & \text{sinon,} \end{cases} \\
a_{k,k-1} = & \begin{cases} -\dfrac{1}{h^2} \text{ si } j \neq 1, \\ 0 \qquad\qquad \text{sinon,} \end{cases} \\
a_{k,k+M} = & \begin{cases} -\dfrac{1}{h^2} \text{ si } i < M, \\ 0 \qquad\qquad \text{sinon,} \end{cases} \\
a_{k,k-M} = & \begin{cases} -\dfrac{1}{h^2} \text{ si } i > 1, \\ 0 \qquad\qquad \text{sinon,} \end{cases} \\
\text{Pour} & k = 1, \ldots, n, \text{ et } \ell = 1, \ldots, n; \\
a_{k,\ell} & = 0, \forall\, k = 1, \ldots, n,\ 1 < |k - \ell| < n \text{ ou } |k - \ell| > n.
\end{cases}
$$

La matrice est donc tridiagonale par blocs, plus précisément si on note

$$D = \begin{bmatrix} 4 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \\ 0 & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & & 0 & -1 & 4 \end{bmatrix},$$

les blocs diagonaux (qui sont des matrices de dimension $M \times M$), on a :

$$A = \begin{bmatrix} D & -\mathrm{Id} & 0 & \dots & \dots & 0 \\ -\mathrm{Id} & D & -\mathrm{Id} & 0 & \dots & 0 \\ 0 & -\mathrm{Id} & D & -\mathrm{Id} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & -\mathrm{Id} & D & -\mathrm{Id} \\ 0 & \dots & & 0 & -\mathrm{Id} & D \end{bmatrix}, \tag{1.14}$$

où Id désigne la matrice identité d'ordre M, et 0 la matrice nulle d'ordre M.

Matrices monotones, ou à inverse positive Une propriété qui revient souvent dans l'étude des matrices issues de la discrétisation d'équations différentielles est le fait que si leur action sur un vecteur $\boldsymbol{u}$ donne un vecteur positif $\boldsymbol{v}$ (composante par composante) alors le vecteur $\boldsymbol{u}$ de départ doit être positif (composante par composante) ; on dit souvent que la matrice est "monotone", ce qui n'est pas un terme très évocateur... Dans ce cours, on lui préfèrera le terme "à inverse positive" ; en effet, on montre à la proposition 1.10 qu'une matrice A est monotone si et seulement si elle est inversible et à inverse positive.

Définition 1.9 (IP-matrice ou matrice monotone)**.** *Si $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $\boldsymbol{x} \geq 0$ [resp. $\boldsymbol{x} > 0$] si toutes les composantes de $\boldsymbol{x}$ sont positives [resp. strictement positives].*
Soit $A \in \mathcal{M}_n(\mathbb{R})$, on dit que A est une matrice monotone si elle vérifie la propriété suivante :

$$\text{Si } \boldsymbol{x} \in \mathbb{R}^n \text{ est tel que } A\boldsymbol{x} \geq 0, \text{ alors } \boldsymbol{x} \geq 0,$$

ce qui peut encore s'écrire : $\{\boldsymbol{x} \in \mathbb{R}^n \text{ t.q. } A\boldsymbol{x} \geq 0\} \subset \{\boldsymbol{x} \in \mathbb{R}^n \text{ t.q. } \boldsymbol{x} \geq 0\}$.

Proposition 1.10 (Caractérisation des matrices monotones)**.** *Une matrice A est monotone si et seulement si elle inversible et à inverse positive (c.à.d. dont tous les coefficients sont positifs).*

La démonstration de ce résultat est l'objet de l'exercice 13. Retenez que toute matrice monotone est inversible et d'inverse positive. Cette propriété de monotonie peut être utilisée pour établir une borne de $\|A^{-1}\|$ pour la matrice de discrétisation du Laplacien, dont on a besoin pour montrer la convergence du schéma. C'est donc une propriété qui est importante au niveau de l'analyse numérique.

1.2.3 Exercices (matrices, exemples)

Exercice 1 (A faire sans calcul !)**.** Effectuer le produit matriciel

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Exercice 2 (Permutations et matrices)**.** Pour $n \geq 1$, on note Σ_n l'ensemble des bijections de $\{1, \ldots, n\}$ dans lui-même (ces bijections s'appellent des permutations), et pour $i = 1, \ldots,$ on note $E_i \in \mathcal{M}_{n,1}(\mathbb{R})$ la matrice colonne dont tous les éléments sont nuls sauf le i-ème, qui est égal à 1. A tout élément $\sigma \in \Sigma_n$, on associe la matrice $P_\sigma \in \mathcal{M}_n(\mathbb{R})$ dont les colonnes sont $E_{\sigma(1)}, \ldots, E_{\sigma(n)}$.

1. Dans cette question seulement, on suppose $n = 2$. Ecrire toutes les matrices de la forme P_σ.
2. Même question avec $n = 3$.
3. Montrer que pour tout $\sigma \in \Sigma_n$, P_σ est une matrice de permutation.
4. Montrer que si P est une matrice de permutation, alors il existe $\sigma \in \Sigma_n$ tel que $P = P_\sigma$.
5. Montrer que
$$P_\sigma \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{\sigma^{-1}(1)} \\ \vdots \\ x_{\sigma^{-1}(n)} \end{bmatrix}.$$
6. Montrer que si $\sigma_1, \sigma_2 \in \Sigma_n$, alors $P_{\sigma_1} P_{\sigma_2} = P_{\sigma_2 \circ \sigma_1}$. En déduire que le produit de 2 matrices de permutation est une matrice de permutation.
7. Montrer que $P_{\sigma^{-1}} = (P_\sigma)^t$. En déduire que toute matrice de permutation est inversible, d'inverse sa transposée.

Exercice 3 (Théorème du rang)**.** *Corrigé en page 22.*
Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$ $(n, p \geq 1)$. On rappelle que $\mathrm{Ker}(A) = \{\boldsymbol{x} \in \mathbb{R}^p;\ A\boldsymbol{x} = 0\}$, $\mathrm{Im}(A) = \{A\boldsymbol{x},\ \boldsymbol{x} \in \mathbb{R}^p\}$ et $\mathrm{rang}(A) = \dim(\mathrm{Im}(A))$. Noter que $\mathrm{Ker}(A) \subset \mathbb{R}^p$ et $\mathrm{Im}(A) \subset \mathbb{R}^n$.
Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ une base de $\mathrm{Im}(A)$ (donc $r \leq n$) et, pour $i \in \{1, \ldots, r\}$, $\boldsymbol{a}_i$ tel que $A\boldsymbol{a}_i = f_i$.

1. Montrer que la famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ est une famille libre de $\mathbb{R}^p$ (et donc $r \leq p$).
2. On note G le sous espace vectoriel de $\mathbb{R}^p$ engendré par $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$. Montrer que $\mathbb{R}^p = G \oplus \mathrm{Ker}(A)$. En déduire que (théorème du rang)
$$p = \dim(\mathrm{Ker}(A)) + \dim(\mathrm{Im}(A)).$$
3. On suppose ici que $n = p$. Montrer que l'application $\boldsymbol{x} \mapsto A\boldsymbol{x}$ (de $\mathbb{R}^n$ dans $\mathbb{R}^n$) est injective si et seulement si elle est surjective.

Exercice 4 ($\mathrm{rang}(A)$=$\mathrm{rang}(A^t)$)**.** *Corrigé en page 22.*
Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$ $(n, p \geq 1)$.

1. Soient P une matrice inversible de $\mathcal{M}_n(\mathbb{R})$ et Q une matrice inversible de $\mathcal{M}_p(\mathbb{R})$. Montrer que $\dim(\mathrm{Im}(PA)) = \dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$. Montrer aussi que les matrices P^t et Q^t sont inversibles.

Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ une base de $\mathrm{Im}(A)$ (donc $r \leq p$) et, pour $i \in \{1, \ldots, r\}$, $\boldsymbol{a}_i$ tel que $A\boldsymbol{a}_i = \boldsymbol{f}_i$. Soit $\boldsymbol{a}_{r+1}, \ldots, \boldsymbol{a}_p$ une base de $\mathrm{Ker}(A)$ (si $\mathrm{Ker}(A) \neq \{0\}$). La famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ est une base de $\mathbb{R}^p$ (voir question 1. de l'exercice 3). De même, on complète (si $r < n$) $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ par $\boldsymbol{f}_{r+1}, \ldots, \boldsymbol{f}_n$ de manière à avoir une base $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ de $\mathbb{R}^n$.

2. Montrer qu'il existe deux matrices $P \in \mathcal{M}_p(\mathbb{R})$ et $Q \in \mathcal{M}_n(\mathbb{R})$ telles que $P\boldsymbol{e}_i = \boldsymbol{a}_i$ (pour tout $i = 1, \ldots, p$) et $Q\boldsymbol{f}_j = \bar{\boldsymbol{e}}_j$ (pour tout $j = 1, \ldots, n$) ou $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ est la base canonique de $\mathbb{R}^p$ et $\bar{\boldsymbol{e}}_1, \ldots, \bar{\boldsymbol{e}}_n$ est la base canonique de $\mathbb{R}^n$. Montrer que P et Q sont inversibles.

On pose $J = QAP$.

3. calculer les colonnes de J et de J^t et en déduire que les matrices J et J^t sont de même rang.
4. Montrer que A et A^t sont de même rang.
5. On suppose maintenant que $n = p$. Montrer que les vecteurs colonnes de A sont liés si et seulement si les vecteurs lignes de A sont liés.

Exercice 5 (Décomposition de $\mathbb{R}^n$ à partir d'une matrice)**.** Soit $n \geq 1$ et $A \in \mathcal{M}_n(\mathbb{R})$.

1. On suppose que la matrice A est diagonalisable. Montrer que $\mathbb{R}^n = \mathrm{Ker}(A) \oplus \mathrm{Im}(A)$.
2. Donner un exemple pour lequel $\mathbb{R}^n \neq \mathrm{Ker}(A) \oplus \mathrm{Im}(A)$ (on pourra se limiter au cas $n = 2$).

Exercice 6 (Vrai ou faux ? Motiver les réponses...). *Suggestions en page 21, corrigé en page 23*

On suppose dans toutes les questions suivantes que $n \geq 2$.

1. Soit $Z \in \mathbb{R}^n$ un vecteur non nul. La matrice ZZ^t est inversible.
2. La matrice inverse d'une matrice triangulaire inférieure est triangulaire supérieure.
3. Les valeurs propres sont les racines du polynôme caractéristique.
4. Toute matrice inversible est diagonalisable dans $\mathbb{R}$.
5. Toute matrice inversible est diagonalisable dans $\mathbb{C}$.
6. Le déterminant d'une matrice A est égal au produit de ses valeurs propres (comptées avec leur multiplicité et éventuellement complexes).
7. Soit A une matrice carrée telle que $A\boldsymbol{x} = \mathbf{0} \Longrightarrow \boldsymbol{x} = \mathbf{0}$, alors A est inversible.
8. Soit A une matrice carrée telle que $A\boldsymbol{x} \geq \mathbf{0} \Longrightarrow \boldsymbol{x} \geq \mathbf{0}$, alors A est inversible.
9. Une matrice symétrique est inversible.
10. Une matrice symétrique définie positive est inversible.
11. Le système linéaire
$$\sum_{j=1}^{n+1} a_{i,j} x_j = 0 \text{ pour tout } i = 1, \ldots, n$$
admet toujours une solution non nulle.
12. La fonction $A \mapsto A^{-1}$ est continue de $GL_n(\mathbb{R})(\mathbb{R})$ dans $GL_n(\mathbb{R})(\mathbb{R})$ ($GL_n(\mathbb{R})$ désigne l'ensemble des matrices carrées inversibles d'ordre n).

Exercice 7 (Sur quelques notions connues). *Corrigé en page 23*

1. Soit A une matrice carrée d'ordre n et $\boldsymbol{b} \in \mathbb{R}^n$. Peut il exister exactement deux solutions distinctes au système $A\boldsymbol{x} = \boldsymbol{b}$?
2. Soient A, B et C de dimensions telles que AB et BC existent. Montrer que si $AB = \mathrm{Id}$ et $BC = \mathrm{Id}$, alors $A = C$.
3. Combien y a -t-il de matrices carrées d'ordre 2 ne comportant que des 1 ou des 0 comme coefficients ? Combien d'entre elles sont inversibles ?
4. Soit $B = \begin{bmatrix} 3 & 2 \\ -5 & -3 \end{bmatrix}$. Montrer que $B^{1024} = \mathrm{Id}$.

Exercice 8 (A propos de $BB^t = I$).
Pour $n \geq 1$, on note I_n la matrice identité d'ordre n.

1. Existe-t-il $B \in \mathcal{M}_{2,1}(\mathbb{R})$ telle que $BB^t = I_2$ (justifier la réponse) ?
2. Soit $n > 2$, Existe-t-il $B \in \mathcal{M}_{n,1}(\mathbb{R})$ telle que $BB^t = I_n$ (justifier la réponse) ?

Exercice 9 (La matrice K_3). *Suggestions en page 21. Corrigé en page 24*
Soit $f \in C([0,1], \mathbb{R})$. On cherche u tel que

$$-u''(x) = f(x), \ \forall x \in (0,1), \tag{1.15a}$$
$$u(0) = u(1) = 0. \tag{1.15b}$$

1. Calculer la solution exacte $u(x)$ du problèmes lorsque f est la fonction identiquement égale à 1 (on admettra que cette solution est unique), et vérifier que $u(x) \geq 0$ pour tout $x \in [0,1]$.

On discrétise le problème suivant par différences finies, avec un pas $h = \frac{1}{4}$ avec la technique vue en cours.

2. On suppose que u est de classe C^4 (et donc f est de calsse C^2). A l'aide de dévloppements de Taylor, écrire l'approximation de $u''(x_i)$ au deuxième ordre en fonction de $u(x_i)$, $u(x_{i-1})$ et $u(x_{i+1})$. En déduire le schéma aux différences finies pour l'approximation de (1.15), qu'on écrira sous la forme :

$$K_3 \boldsymbol{u} = \boldsymbol{b}, \tag{1.16}$$

où K_3 est la matrice de discrétisation qu'on explicitera, $\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$ et $\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix}$.

3. Résoudre le système linéaire (1.16) par la méthode de Gauss. Lorsque f est la fonction identiquement égale à 1, comparer u_i et $u(x_i)$ pour $i = 1, 2, 3$, et expliquer pourquoi l'erreur de discrétisation $u(x_i) - u_i$ est nulle.
4. Reprendre les questions précédentes en remplaçant les conditions limites (1.15b) par :

$$u(0) = 0, \quad u'(1) = 0. \tag{1.17}$$

5. Soit $c \in \mathbb{R}$. On considère maintenant le problème suivant :

$$-u''(x) = c, \ \forall x \in (0,1), \tag{1.18a}$$
$$u'(0) = u'(1) = 0, \tag{1.18b}$$

   (a) Montrer que le problème (1.18) admet soit une infinité de solutions, soit pas de solution.

   (b) Ecrire la discrétisation du problème (1.18), toujours avec $h = \frac{1}{4}$, sous la forme $\widetilde{K}\boldsymbol{u} = \widetilde{\boldsymbol{b}}$ en explicitant $\widetilde{K}$ et $\widetilde{\boldsymbol{b}}$.

   (c) Montrer que la matrice $\widetilde{K}$ n'est pas inversible : on part d'un problème continu mal posé, et on obtient par discrétisation un problème discret mal posé...

Exercice 10 (Matrices symétriques définies positives)**.** *Suggestions en page 21, corrigé en page 25.*
On rappelle que toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ symétrique est diagonalisable dans $\mathbb{R}$ (cf. lemme 1.7 page 10). Plus précisément, on a montré en cours que, si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, il existe une base de $\mathbb{R}^n$, notée $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$, et il existe $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ t.q. $A\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$, pour tout $i \in \{1, \dots, n\}$, et $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, n\}$ ($x \cdot y$ désigne le produit scalaire de x avec y dans $\mathbb{R}^n$).

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est symétrique définie positive, montrer que les éléments diagonaux de A sont strictements positifs.
2. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Montrer que A est symétrique définie positive si et seulement si toutes les valeurs propres de A sont strictement positives.
3. Soit $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est symétrique définie positive. Montrer qu'on peut définir une unique matrice $B \in \mathcal{M}_n(\mathbb{R})$, symétrique définie positive t.q. $B^2 = A$ (on note $B = A^{\frac{1}{2}}$).

Exercice 11 (Résolution d'un système sous forme particulière)**.** *Suggestions en page 21.*
Soit $n \geq 1$, $p \geq 1$, $A \in \mathcal{M}_n(\mathbb{R})$ et $B \in \mathcal{M}_{n,p}(\mathbb{R})$. On suppose que A est une matrice symétrique définie positive et que $\text{rang}(B) = p$ (justifier que ceci implique que $p \leq n$).
Pour $i \in \{1, \dots, p\}$, on pose $\boldsymbol{z}_i = A^{-1}B\boldsymbol{e}_i$ où $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ désigne la base canonique de $\mathbb{R}^p$ ($B\boldsymbol{e}_i$ est donc la i-ieme colonne de B).

1. Montrer que $\{B\boldsymbol{e}_i, i \in \{1, \dots, p\}\}$ est une base de $\text{Im}(B)$.
2. Montrer que A^{-1} est une matrice symétrique définie positive et que $\text{Ker}(B^t A^{-1} B) = \text{Ker}(B) = \{0\}$. En déduire que $\{B^t \boldsymbol{z}_i, i \in \{1, \dots, p\}\}$ est une base de $\mathbb{R}^p$.

Soient $\boldsymbol{b} \in \mathbb{R}^n$ et $\boldsymbol{c} \in \mathbb{R}^p$. On cherche le couple $(\boldsymbol{x}, \boldsymbol{y})$, avec $\boldsymbol{x} \in \mathbb{R}^n$ et $\boldsymbol{y} \in \mathbb{R}^p$, solution du système suivant (écrit sous forme de blocs) :

$$\begin{bmatrix} A & B \\ B^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}. \tag{1.19}$$

On pose $\boldsymbol{u} = A^{-1}\boldsymbol{b}$ et on note $y_1, \dots, y_p$ les composantes de $\boldsymbol{y}$.

3. Montrer que $(\boldsymbol{x}, \boldsymbol{y})$ est solution de (1.19) si et seulement si

$$\sum_{i=1}^{p} y_i B^t \boldsymbol{z}_i = B^t \boldsymbol{u} - \boldsymbol{c}, \tag{1.20}$$

$$\boldsymbol{x} = \boldsymbol{u} - \sum_{i=1}^{p} y_i \boldsymbol{z}_i. \tag{1.21}$$

   En déduire que le système (1.19) a une unique solution.

4. Montrer que la matrice (symétrique) $\begin{bmatrix} A & B \\ B^t & 0 \end{bmatrix}$ est inversible mais n'est pas symétrique définie positive.

Exercice 12 (Diagonalisation dans $\mathbb{R}$)**.**

Soit E un espace vectoriel réel de dimension $n \in \mathbb{N}$ muni d'un produit scalaire, noté $(\cdot, \cdot)$. Soient T et S deux applications linéaires symétriques de E dans E (T symétrique signifie $(Tx, y) = (x, Ty)$ pour tous $x, y \in E$). On suppose que T est définie positive (c'est-à-dire $(Tx, x) > 0$ pour tout $x \in E \setminus \{0\}$).

1. Montrer que T est inversible. Pour $x, y \in E$, on pose $(x, y)_T = (Tx, y)$. Montrer que l'application $(x, y) \mapsto (x, y)_T$ définit un nouveau produit scalaire sur E.
2. Montrer que $T^{-1}S$ est symétrique pour le produit scalaire défini à la question précédente. En déduire, avec le lemme 1.7 page 10, qu'il existe une base de E, notée $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$ et une famille $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{R}$ telles que $T^{-1}S\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$ pour tout $i \in \{1, \dots, n\}$ et t.q. $(T\boldsymbol{f}_i, \boldsymbol{f}_j) = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, n\}$.

Exercice 13 (IP-matrice)**.** *Corrigé en page 26*

Soit $n \in \mathbb{N}^\star$, on note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices de n lignes et n colonnes et à coefficients réels.
Si $x \in \mathbb{R}^n$, on dit que $x \geq 0$ [resp. $x > 0$] si toutes les composantes de x sont positives [resp. strictement positives].
Soit $A \in \mathcal{M}_n(\mathbb{R})$, on dit que A est une IP-matrice si elle vérifie la propriété suivante :

$$\text{Si } x \in \mathbb{R}^n \text{ est tel que } Ax \geq 0, \text{ alors } x \geq 0,$$

ce qui peut encore s'écrire : $\{x \in \mathbb{R}^n \text{ t.q. } Ax \geq 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x \geq 0\}$.

1. Soit $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$. Montrer que A est une IP-matrice si et seulement si A est inversible et $A^{-1} \geq 0$ (c'est-à-dire que tous les coefficients de A^{-1} sont positifs).
2. Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice réelle d'ordre 2. Montrer que A est une IP-matrice si et seulement si :

$$\begin{cases} ad < bc, \\ a \leq 0,\ d \leq 0 \\ b > 0,\ c > 0 \end{cases} \quad \text{ou} \quad \begin{cases} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{cases} \tag{1.22}$$

   En déduire que les matrices $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ et $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ sont des IP-matrices.
3. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une IP-matrice alors A^t (la transposée de A) est une IP-matrice.

4. Montrer que si A est telle que

$$a_{i,j} \leq 0, \text{ pour tout } i,j = 1,\dots,n,\ i \neq j, \text{ et } a_{i,i} > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}|, \text{ pour tout } i = 1,\dots,n, \tag{1.23}$$

alors A est une IP-matrice ; en déduire que si A^t satisfait (1.23), alors A est une IP-matrice.

5. Soit A une matrice **inversible** telle que

$$a_{i,j} \leq 0, \text{ pour tout } i,j = 1,\dots,n,\ i \neq j, \text{ et } a_{i,i} \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}|, \text{ pour tout } i = 1,\dots,n. \tag{1.24}$$

Pour tout $\varepsilon \geq 0$, on définit la matrice $A_\varepsilon = A + \varepsilon \mathrm{Id}$, où Id désigne la matrice identité.

   (a) Prouver que, pour $\varepsilon > 0$, la matrice A_ε est une IP-matrice.

   (b) Prouver que la matrice A_ε est inversible pour tout $\varepsilon \geq 0$, et que les coefficients de A_ε^{-1} sont des fonctions continues de ε.

   (c) En déduire que A est une IP-matrice.

6. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une IP-matrice et si $x \in \mathbb{R}^n$ alors :

$$Ax > 0 \Rightarrow x > 0.$$

c'est-à-dire que $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$.

7. Montrer, en donnant un exemple, qu'une matrice A de $\mathcal{M}_n(\mathbb{R})$ peut vérifier $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$ et ne pas être une IP-matrice.

8. On suppose dans cette question que $A \in \mathcal{M}_n(\mathbb{R})$ est inversible et que $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$. Montrer que A est une IP-matrice.

9. (Question plus difficile) Soit E l'espace des fonctions continues sur $\mathbb{R}$ et admettant la même limite finie en $+\infty$ et $-\infty$. Soit $\mathcal{L}(E)$ l'ensemble des applications linéaires continues de E dans E. Pour $f \in E$, on dit que $f > 0$ (resp. $f \geq 0$) si $f(x) > 0$ (resp. $f(x) \geq 0$) pour tout $x \in \mathbb{R}$. Montrer qu'il existe $T \in \mathcal{L}(E)$ tel que $Tf \geq 0 \Longrightarrow f \geq 0$, et $g \in E$ tel que $Tg > 0$ et $g \not> 0$ (ceci démontre que le raisonnement utilisé en 2 (b) ne marche pas en dimension infinie).

Exercice 14 (Matrice du Laplacien discret 1D)**.** *Corrigé détaillé en page 27.*
Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit K_n la matrice définie par (1.10) page 13, issue d'une discrétisation par différences finies avec pas constant du problème (1.5a) page 11.
Montrer que K_n est symétrique définie positive.

Exercice 15 (Pas non constant)**.**
Reprendre la discrétisation vue en cours avec un pas $h_i = x_{i+1} - x_i$ non constant, et montrer que dans ce cas,le schéma est consistant d'ordre 1 seulement.

Exercice 16 (Réaction diffusion 1d.)**.** *Corrigé détaillé en page 28.*
On s'intéresse à la discrétisation par Différences Finies du problème aux limites suivant :

$$\begin{aligned} &-u''(x) + u(x) = f(x),\ x \in]0,1[, \\ &u(0) = u(1) = 0. \end{aligned} \tag{1.25}$$

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\dots,n}$ une "valeur approchée" de la solution u du problème (1.25) aux points $\left(\frac{j}{n+1}\right)_{j=1,\dots,n}$. Donner la discrétisation par différences finies de ce problème sous la forme $AU = b$.

Exercice 17 (Discrétisation)**.** On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points du problème (1.5a) page 11, avec $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. On note u est la solution exacte, $x_i = ih$, pour $i = 1, \ldots, n$ les points de discrétisation, et $(u_i)_{i=1,\ldots,n}$ la solution du système discrétisé (1.9).

1. Montrer que si $u \in C^4([0,1]$, alors la propriété (1.7) est vérifiée, c.à.d. :
$$-\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} = -u''(x_i) + R_i \text{ avec } |R_i| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty.$$
2. Montrer que si f est constante, alors
$$\max_{1\leq i\leq n} |u_i - u(x_i)| = 0.$$
3. Soit n fixé, et $\max_{1\leq i\leq n} |u_i - u(x_i)| = 0$. A-t-on forcément que f est constante sur $[0,1]$?

Exercice 18 (Déterminant d'une matrice sous forme de blocs)**.**
Soient $A \in \mathcal{M}_n(\mathbb{R})$ $(n > 1)$, $b, c \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$. On s'intéresse à la matrice $\bar{A} \in \mathcal{M}_{n+1}(\mathbb{R})$ définie sous forme de blocs de la manière suivante :
$$\bar{A} = \begin{bmatrix} A & b \\ c^t & \lambda \end{bmatrix} \tag{1.26}$$
On montre dans cet exercice que les deux assertions suivantes sont, sauf cas particuliers, fausses :

A1 $\det(\bar{A}) = \lambda\det(A) - \det(bc^t)$,

A2 $\det(\bar{A}) = \lambda\det(A) - c^t\, b$,

1. Dans cette question, on prend $n \geq 2$, $A = 0$, $b = c$ et on suppose que $b \neq 0$.
   (a) Montrer que rang$(\bar{A}) \leq 2$ et en déduire que $\bar{A}$ n'est pas inversible.
   (b) En déduire que l'assertion A2 est fausse pour cet exemple.
2. Dans cette question, on suppose que A est symétrique définie positive, $\lambda = 0$, $b = c$ et que $b \neq 0$.
   (a) Montrer que $\bar{A}$ est inversible et que rang$(bb^t) = 1$.
   (b) En déduire que l'assertion A1 est fausse pour cet exemple.

1.2.4 Suggestions pour les exercices

Exercice 6 page 17 (Vrai ou faux ?)

1. Considérer la matrice ZZ^t.
12. Ecrire que $A^{-1} = \frac{1}{\det}(A)\ \mathrm{com}(A)^t$ où $\det(A)$ est le déterminant (non nul) de A et $\mathrm{com}(A)$ la comatrice de A.

Exercice 9 page 17 (La matrice K_3)

2. Ecrire le développement de Taylor de $u(x_i + h)$ et $u(x_i - h)$.
3. Pour l'erreur de discrétisation, se souvenir qu'elle dépend de l'erreur de consistance, et regarder sa majoration.
4. Pour tenir compte de la condition limite en 1, écrire un développement limité de $u(1 - h)$.
5.1 Distinguer les cas $c = 0$ et $c \neq 0$.

Exercice 10 page 18 (Matrices symétriques définies positives)

3. Utiliser la diagonalisation sur les opérateurs linéaires associés.

Exercice 9 page 17 (Résolution d'un système sous forme particulière)

1. Utiliser le fait que $Im(B)$ est l'ensemble des combinaisons linéaires des colonnes de B.
2. Utiliser le caractère s.d.p. de A puis le théorème du rang.

1.2.5 Corrigés des exercices

Exercice 3 page 16 (Théorème du rang)

1. Soit $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ dans $\mathbb{R}$ tel que $\sum_{i=1}^r \alpha_i \boldsymbol{a}_i = 0$. On a donc

$$0 = A(\sum_{i=1}^r \alpha_i \boldsymbol{a}_i) = \sum_{i=1}^r \alpha_i A\boldsymbol{a}_i = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i.$$

   Comme la famille $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ est une famille libre, on en déduit que $\alpha_i = 0$ pour tout $i \in \{1, \ldots, r\}$ et donc que la famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ est libre.

2. Soit $\boldsymbol{x} \in \mathbb{R}^p$. Comme $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ est une base de $\mathrm{Im}(A)$, il existe $\alpha_1, \ldots, \alpha_r$ tel que $A\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i$. On pose $\boldsymbol{y} = \sum_{i=1}^r \alpha_i \boldsymbol{a}_i$. On a $A\boldsymbol{y} = A\boldsymbol{x}$ et $\boldsymbol{x} = (\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{y}$. Comme $\boldsymbol{y} \in G$ et $A(\boldsymbol{x} - \boldsymbol{y}) = 0$, on en déduit que $\mathbb{R}^p = G + \mathrm{Ker}A$.

   Soit maintenant $\boldsymbol{x} \in \mathrm{Ker}A \cap G$. Comme $\boldsymbol{x} \in G$, il existe $\alpha_1, \ldots, \alpha_r$ tel que $\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{a}_i$. On a donc $A\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i$. Comme $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ est une famille libre et que $A\boldsymbol{x} = 0$, on en déduit que $\alpha_i = 0$ pour tout $i \in \{1, \ldots, r\}$ et donc $\boldsymbol{x} = 0$. Ceci montre que $\mathbb{R}^p = G \oplus \mathrm{Ker}(A)$. Enfin, comme $\dim G = r = \dim(\mathrm{Im}A)$, on en déduit bien que $p = \dim(\mathrm{Ker}(A)) + \dim(\mathrm{Im}(A))$.

3. On suppose ici $p = n$. Comme $n = \dim(\mathrm{Ker}(A)) + \dim(\mathrm{Im}(A))$, on a $\dim(\mathrm{Ker}(A)) = 0$ si et seulement si $\dim(\mathrm{Im}(A)) = n$. Ceci montre que l'application $\boldsymbol{x} \mapsto A\boldsymbol{x}$ (de $\mathbb{R}^n$ dans $\mathbb{R}^n$) est injective si et seulement si elle est surjective.

Exercice 4 page 16 (rang(A)=rang(A^t))

1. On remarque tout d'abord que le noyau de PA est égal au noyau de A. En effet, soit $\boldsymbol{x} \in \mathbb{R}^p$. Il est clair que $A\boldsymbol{x} = 0$ implique $PA\boldsymbol{x} = 0$. D'autre part, comme P est inversible, $PA\boldsymbol{x} = 0$ implique $A\boldsymbol{x} = 0$. On a donc bien $\mathrm{Ker}(PA) = \mathrm{Ker}(A)$. On en déduit que $\dim(\mathrm{Ker}(PA)) = \dim(\mathrm{Ker}(A))$ et donc, avec le théorème du rang (exercice 3), que $\dim(\mathrm{Im}(PA)) = \dim(\mathrm{Im}(A))$.

   Pour montrer que $\dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$, on remarque directement que $\mathrm{Im}(AQ) = \mathrm{Im}(A)$. En effet, on a, bien sûr, $\mathrm{Im}(AQ) \subset \mathrm{Im}(A)$ (l'inversibilité de Q est inutile pour cette inclusion). D'autre part, si $\boldsymbol{z} \in \mathrm{Im}(A)$, il existe $\boldsymbol{x} \in \mathbb{R}^p$ tel que $A\boldsymbol{x} = \boldsymbol{z}$. Comme Q est inversible, il existe $\boldsymbol{y} \in \mathbb{R}^p$ tel que $\boldsymbol{x} = Q\boldsymbol{y}$. On a donc $\boldsymbol{z} = AQ\boldsymbol{y}$, ce qui prouve que $\mathrm{Im}(A) \subset \mathrm{Im}(AQ)$. Finalement, on a bien $\mathrm{Im}(AQ) = \mathrm{Im}(A)$ et donc $\dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$.

   Pour montrer que P^t est inversible, il suffit de remarquer que $(P^{-1})^t P^t = (PP^{-1})^t = I_n$ (où I_n désigne la matrice Identité de $\mathbb{R}^n$). Ceci montre que P^t est inversible (et que $(P^t)^{-1} = (P^{-1})^t$). Bien sûr, un raisonnement analogue donne l'inversibilité de Q^t.

2. Par définition du produit matrice vecteur, $P\boldsymbol{e}_i = \boldsymbol{c}_i(P)$, i-ème colonne de P ; l suffit de prendre pour P la matrice dont les colonnes sont les vecteurs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$; l'image de P est égale à $\mathbb{R}^p$ car la famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ est une base de $\mathbb{R}^p$, ce qui prouve que P est inversible (on a $\mathrm{Im}(P) = \mathbb{R}^p$ et $\mathrm{Ker}P = \{0\}$ par le théorème du rang).

   Soit maintenant $R \in \mathcal{M}_n(\mathbb{R})$ dont les colonnes sont les vecteurs $\boldsymbol{f}_j$; la matrice R est bien inversible car la famille $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ est une base $\mathbb{R}^n$. On a donc, toujours par définition du produit matrice vecteur, $R\bar{\boldsymbol{e}}_j = \boldsymbol{c}_j(R) = \boldsymbol{f}_j$ pour $j = 1, n$. Posons $Q = R^{-1}$; on a alors $QR\bar{\boldsymbol{e}}_j = \bar{\boldsymbol{e}}_j = Q\boldsymbol{f}_j$, et la matrice Q est évidemment inversible.

3. Pour $i \in \{1, \ldots, p\}$, la i-ème colonne de J est donnée par $\boldsymbol{c}_i(J) = QAPe_i = QA\boldsymbol{a}_i$. Si $i \in \{1, \ldots, r\}$, on a donc $\boldsymbol{c}_i(J) = Q\boldsymbol{f}_i = \bar{\boldsymbol{e}}_i$. Si $i \in \{r+1, \ldots, p\}$, on a $\boldsymbol{c}_i(J) = 0$ (car $\boldsymbol{a}_i \in \mathrm{Ker}A$). Ceci montre que $\mathrm{Im}(J)$ est l'espace vectoriel engendré par $\bar{\boldsymbol{e}}_1, \ldots, \bar{\boldsymbol{e}}_r$ et donc que le rang de J est r.

   La matrice J appartient à $\mathcal{M}_{n,p}(\mathbb{R})$, sa transposée appartient donc à $\mathcal{M}_{p,n}(\mathbb{R})$. En transposant la matrice J, on a, pour tout $i \in \{1, \ldots, r\}$, $\boldsymbol{c}_i(J^t) = e_i$ et, pour tout $i \in \{r+1, \ldots, n\}$, $\boldsymbol{c}_i(J^t) = 0$. Ceci montre que $\mathrm{Im}(J^t)$ est l'espace vectoriel engendré par $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r$ et donc que le rang de J^t est aussi r.

4. Il suffit maintenant d'appliquer la première question, elle donne que le rang que A est le même que le rang de J et, comme $J^t = P^t A^t Q^t$, que le rang que A^t est le même que le rang de J^t. Finalement le rang de A et de A^t est r.
5. Les vecteurs colonnes de A sont liés si et seulement si le rang de A est strictement inférieur à n. Les vecteurs colonnes de A^t sont liés si et seulement si le rang de A^t est strictement inférieur à n. Comme les vecteurs colonnes de A^t sont les vecteurs lignes de A, on obtient le résultat désiré grâce au fait que A et A^t ont même rang.

Exercice 6 page 17 (Vrai ou faux ?)

1. Faux : La matrice ZZ^t est de rang 1 et donc non inversible.
2. Faux : La matrice inverse d'une matrice triangulaire inférieure est triangulaire inférieure.
3. Vrai : le polynôme caractéristique d'une matrice A est le déterminant de $A - \lambda \mathrm{Id}$.
4. Faux : la matrice $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ est inversible et non diagonalisable dans $\mathbb{R}$.
5. Faux : la matrice $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ est inversible et non diagonalisable dans $\mathbb{C}$.
6. Vrai : c'est le terme de degré 0 du polynôme caractéristique.
7. Vrai : si $\mathrm{Ker}(A) = \{0\}$, alors A est inversible.
8. Vrai : on va montrer que $\mathrm{Ker}(A) = \{0\}$, Supposons que $A\boldsymbol{x} = \mathbf{0}$, alors $A\boldsymbol{x} \geq \mathbf{0}$ et $A\boldsymbol{x} \leq \mathbf{0}$, ou encore $A(-\boldsymbol{x}) \geq \mathbf{0}$ Donc par hypothèse, $\boldsymbol{x} \geq \mathbf{0}$ et $-\boldsymbol{x} \geq \mathbf{0}$, et donc $\boldsymbol{x} = 0$, ce qui montre que $\mathrm{Ker}(A) = \{0\}$.
9. Faux : la matrice nulle est symétrique.
10. Vrai : Si A est s.d.p.alors $A\boldsymbol{x} = \mathbf{0}$ entraîne $A\boldsymbol{x} \cdot \boldsymbol{x} = 0$ et donc $\boldsymbol{x} = \mathbf{0}$, ce qui montre que $\mathrm{Ker}(A) = \{0\}$ et donc que A est inversible.
11. Vrai : l'ensemble des solutions est le noyau de la matrice $A \in \mathcal{M}_{n,n+1}(\mathbb{R})$ qui est de dimension au moins un par le théorème du rang.
12. Vrai : on peut écrire que $A^{-1} = \frac{1}{\det}(A)\mathrm{com}(A)^t$ où $\det(A)$ est le déterminant (non nul) de A et $\mathrm{com}(A)$ la comatrice de A, c.à.d. la matrice des cofacteurs des coefficients de A ; on rappelle que le cofacteur $c_{i,j}$ de l'élément $a_{i,j}$ est défini par $c_{i,j} = (-1)^{i+j}\Delta_{i,j}$ où $\Delta_{i,j}$ est le mineur relatif à (i,j), *i.e.* le déterminant de la sous matrice carrée d'ordre $n-1$ obtenue à partir de A en lui retirant sa i-ème ligne et sa j-ème colonne). On peut vérifier facilement que les applications $A \mapsto \det(A)$ et $A \mapsto c_{i,j}$ sont continues de $GL_n(\mathbb{R})(\mathbb{R})$ dans $\mathbb{R}^*$ et $\mathbb{R}$ respectivement (comme polynôme en les éléments de la matrice A), et que donc $A \mapsto A^{-1}$ est continue.

Exercice 7 page 17 (Sur quelques notions connues)

1. Supposons qu'il existe deux solutions distinctes $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ au système $A\boldsymbol{x} = \boldsymbol{b}$. Soit $\boldsymbol{z} = \boldsymbol{x}_1 - \boldsymbol{x}_2$. On a donc $A\boldsymbol{z} = \mathbf{0}$ et $z \neq 0$.
   — Si A est inversible, on a donc $\boldsymbol{z} = \mathbf{0}$ en contradiction avec $x_1 \neq x_2$.
   — Si A est non inversible, alors $A(t\boldsymbol{z}) = \mathbf{0}$ pour tout $t \in \mathbb{R}$, et donc il y a une infinité de solutions au système $A\boldsymbol{x} = \boldsymbol{b}$.
2. $C = (AB)C = A(BC) = A$.
3. Les matrices carrées d'ordre 2 ont quatre coefficients, et donc il y a $2^4 = 16$ matrices ne comportant que des 1 ou des 0 comme coefficients. Une matrice $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ est inversible si $ad - bc \neq 0$. Dans le cas de matrices ne comportant que des 1 ou des 0 comme coefficients, les valeurs non nulles possibles de $ad - bc$ sont 1 et -1, obtenues respectivement pour ($ad = 1$, $bc = 0$) et ($ad = 0$, $bc = 1$), c.à.d pour les matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

et

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

4. Les valeurs propres de B sont i et $-i$ (car la trace de B est nulle et son déterminant est égal à 1). Donc $B^{1024} = \mathrm{Id}$

Exercice 9 page 17 (La matrice K_3)

1. La solution est $-\frac{1}{2}x(x-1)$, qui est effectivement positive.
2. Avec les développements limités vus en cours, on obtient :

$$K_3 = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix},\ \boldsymbol{b} = \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \end{bmatrix},\ \text{où } h = \frac{1}{4}$$

3. L'échelonnement du système $K_3\boldsymbol{x} = \boldsymbol{b}$ sur la matrice augmentée (ou la méthode de Gauss) donne :

$$\frac{1}{h^2}\left[\begin{array}{ccc|c} 2 & -1 & 0 & b_1 \\ -0 & \frac{3}{2} & -1 & b_2 + \frac{1}{2}b_1 \\ 0 & 0 & \frac{4}{3} & b_3 + \frac{2}{3}b_2 + \frac{1}{3}b_1 \end{array}\right]$$

Donc pour $h = \frac{1}{4}$ et $b_1 = b_2 = b_3 = 1$ on obtient

$$u_1 = \frac{3}{32}, u_2 = \frac{1}{8} \text{ et } u_3 = \frac{3}{32}.$$

On a $u_i = u(x_i)$, ce qui veut dire que l'erreur de discrétisation est nulle. On a vu en cours (formule (1.8)) que l'erreur de consistance R peut être majorée par $\frac{h^2}{12}\|u^{(4)}\|_\infty$.. Ici u est un polynôme de degré 2, et donc $R = 0$. Or par l'inégalité (1.12), l'erreur de discrétisation $\boldsymbol{e} = (u(x_1) - u_1, u(x_2) - u_2, u(x_3) - u_3)^t$ satisfait $\boldsymbol{e} = K_3^{-1}R$. On en déduit que cette erreur de discrétisation est nulle.
Notons qu'il s'agit là d'un cas tout à fait particulier dû au fait que la solution exacte est un polynôme de degré inférieur ou égal à 3.

4. Avec la condition limite (1.17), la solution exacte du problème pour $f \equiv 1$ est maintenant $u(x) = -\frac{1}{2}x(x-2)$.
Pour prendre en compte la condition limite (1.17), on effectue un développement limité de u à l'ordre 2 en $x = 1$

$$u(1-h) = u(1) - hu'(1) + \frac{1}{2}h^2u''(\zeta) \text{ avec } \zeta \in [1-h, 1].$$

Les inconnues discrètes sont maintenant les valeurs approchées recherchées aux points x_i, $i \in \{1,2,3,4\}$, notées u_i, $i \in \{1,2,3,4\}$. Comme $u'(1) = 0$, l'égalité précédente suggère de prendre comme équation discrète $u_3 = u_4 - (1/2)f(1)$ (on rappelle que $x_4 = 1$).
Le système discret à reśoudre est donc :

$$\begin{aligned} 2u_1 - u_2 &= h^2 f(x_1), \\ -u_1 + 2u_2 - u_3 &= h^2 f(x_2) \\ -u_2 + 2u_3 - u_4 &= h^2 f(x_3) \\ -u_3 + u_4 &= \frac{1}{2}h^2 f(x_4) \end{aligned}$$

Le système linéaire à résoudre est donc $K\boldsymbol{u} = \boldsymbol{b}$, avec

$$K = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix},\ \boldsymbol{b} = \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \\ \frac{1}{2}f(4h) \end{bmatrix}.$$

En notant $b_i = f(x_i)$, l'échelonnement du système $h^2 K\boldsymbol{x} = h^2\boldsymbol{b}$ sur la matrice augmentée donne :

$$\left[\begin{array}{cccc|c} 2 & -1 & 0 & 0 & h^2 b_1 \\ 0 & \frac{3}{2} & -1 & 0 & h^2(b_2 + \frac{1}{2}b_1) \\ 0 & 0 & \frac{4}{3} & -1 & h^2(b_3 + \frac{2}{3}b_2 + \frac{1}{3}b_1) \\ 0 & 0 & 0 & \frac{1}{4} & h^2(\frac{1}{2}b_4 + \frac{1}{2}b_2 + \frac{1}{4}b_1 + \frac{3}{4}b_3) \end{array}\right]$$

Donc pour $h = \frac{1}{4}$ et $b_1 = b_2 = b_3 = b_4 = 1$ on obtient

$$u_1 = \frac{7}{32}, u_2 = \frac{3}{8}, u_3 = \frac{15}{32} \text{ et } u_4 = \frac{1}{2}.$$

La solution exacte aux points de discrétisation est :

$$u(x_1) = \frac{1}{2}\frac{1}{4}(2 - \frac{1}{4}) = \frac{7}{32}, u(x_2) = \frac{1}{2}\frac{1}{2}(2 - \frac{1}{2}) = \frac{3}{8}, u(x_3) = \frac{1}{2}\frac{3}{4}(2 - \frac{3}{4}) = \frac{15}{32}, u(x_4) = \frac{1}{2}.$$

On a donc $u(x_i) = u_i$ pour tout $i \in \{1, 2, 3, 4\}$, ce qu'on aurait pu deviner sans calculs car ici aussi l'erreur de discrétisation est nulle car l'erreur de consistance est nulle en raison du traitement que nous avons fait de la condition aux limites de Neumann ($u'(1) = 0$) et du fait que la solution exacte est un polynôme de degré au plus égal à 2.

5.

(a) Il est facile de voir que si $c \neq 0$, aucune fonction ne peut satisfaire le problème (1.18), alors que si $c = 0$, toutes les fonctions constantes conviennent.

(b) On a maintenant une condition de Neumann en 0 et en 1.
Un raisonnement similaire aux questions précédentes nous conduit à introduire 5 inconnues discrètes u_i, $i \in \{1, \ldots, 5\}$. Le système à résoudre est maintenant :

$$\widetilde{K} = \frac{1}{h^2}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix},\ \widetilde{\boldsymbol{b}} = \begin{bmatrix} \frac{1}{2}f(0) \\ f(h) \\ f(2h) \\ f(3h) \\ \frac{1}{2}f(4h) \end{bmatrix}.$$

(c) La matrice $\widetilde{K}$ n'est pas inversible car la somme de ses colonnes est égale au vecteur nul : on part d'un problème continu mal posé, et on obtient effectivement par discrétisation un problème discret mal posé.

Exercice 10 page 18 (Matrices symétriques définies positives)

1. On note $e_1, \ldots, e_n$ la base canonique de $\mathbb{R}^n$. Pour tout $i \in \{1, \ldots, n\}$, on a $a_{i,i} = Ae_i \cdot e_i$ et donc, comme A est définie positive, on en déduit $a_{i,i} > 0$.

2. On utilise le rappel donné dans l'énoncé. Les λ_i sont les valeurs propres de A. Soit $x \in \mathbb{R}^n$, décomposons x sur la base orthonormée $(\boldsymbol{f}_i)_{i=1,n}$: $x = \sum_{i=1}^n \alpha_i \boldsymbol{f}_i$. On a donc :

$$Ax \cdot x = \sum_{i=1}^n \lambda_i \alpha_i^2. \tag{1.27}$$

Montrons d'abord que si les valeurs propres sont strictement positives alors A est définie positive :
Supposons que $\lambda_i \geq 0$, $\forall i = 1, \ldots, n$. Alors pour $\forall x \in \mathbb{R}^n$, d'après (1.27), $Ax \cdot x \geq 0$ et la matrice A est positive. Supposons maintenant que $\lambda_i > 0$, $\forall i = 1, \ldots, n$. Alors pour $\forall x \in \mathbb{R}^n$, toujours d'après (1.27), $(Ax \cdot x = 0) \Rightarrow (x = 0)$, et la matrice A est donc bien définie.

Montrons maintenant la réciproque : si A est définie positive, alors $A\boldsymbol{f}_i \cdot \boldsymbol{f}_i > 0$, $\forall i = 1, \ldots, n$ et donc $\lambda_i > 0$, $\forall i = 1, \ldots, n$.

3. On note T l'application (linéaire) de $\mathbb{R}^n$ dans $\mathbb{R}^n$ définie par $T(x) = Ax$. On prouve tout d'abord l'existence de B. Comme A est s.d.p., toutes ses valeurs propres sont strictement positives, et on peut donc définir l'application linéaire S dans la base orthonormée $(f_i)_{i=1,n}$ par : $S(\boldsymbol{f}_i) = \sqrt{\lambda_i}\boldsymbol{f}_i$, $\forall i = 1, \ldots, n$. On a évidemment $S \circ S = T$, et donc si on désigne par B la matrice représentative de l'application S dans la base canonique, on a bien $B^2 = A$.
Pour montrer l'unicité de B, on peut remarquer que, si B^2=A, on a, pour tout $i \in \{1, \ldots, n\}$,

$$(B + \sqrt{\lambda_i}I)(B - \sqrt{\lambda_i}I)f_i = (B^2 - \lambda_i I)f_i = (A - \lambda_i I)f_i = 0,$$

où I désigne la matrice identité. On a donc $(B - \sqrt{\lambda_i}I)f_i \in \mathrm{Ker}(B + \sqrt{\lambda_i}I)$. Mais, comme B est s.d.p., les valeurs propres de B sont des réels strictement positifs, on a donc $\mathrm{Ker}(B + \sqrt{\lambda_i}I) = \{0\}$ et donc $Bf_i = \sqrt{\lambda_i}f_i$. Ce qui détermine complètement B.

Exercice 13 page 19 (IP-matrice)

1. Supposons d'abord que A est inversible et que $A^{-1} \geq 0$; soit $x \in \mathbb{R}^n$ tel que $b = Ax \geq 0$. On a donc $x = A^{-1}b$, et comme tous les coefficients de A^{-1} et de b sont positifs ou nuls, on a bien $x \geq 0$.
   Réciproquement, si A est une IP-matrice, alors $Ax = 0$ entraine $x = 0$ ce qui montre que A est inversible. Soit e_i le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on a : $AA^{-1}e_i = e_i \geq 0$, et donc par la propriété de IP-matrice, $A^{-1}e_i \geq 0$, ce qui montre que tous les coefficients de A^{-1} sont positifs.
2. La matrice inverse de A est $A^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ avec $\Delta = ad - bc$. Les coefficients de A^{-1} sont donc positifs ou nuls si et seulement si

$$\begin{cases} ad < bc, \\ a \leq 0,\ d \leq 0 \\ b \geq 0,\ c \geq 0 \end{cases} \quad \text{ou} \quad \begin{cases} ad > bc, \\ a \geq 0,\ d \geq 0, \\ b \leq 0,\ c \leq 0. \end{cases}$$

   Dans le premier cas, on a forcément $bc \neq 0$: en effet sinon on aurait $ad < 0$, or $a \leq 0$ et $d \leq 0$ donc $ad \geq 0$.
   Dans le second cas, on a forcément $ad \neq 0$: en effet sinon on aurait $bc < 0$, or $b \leq 0$ et $c \leq 0$ donc $bc \geq 0$.
   Les conditions précédentes sont donc équivalentes aux conditions (1.22).
3. La matrice A^t est une IP-matrice si et seulement A^t est inversible et $(A^t)^{-1} \geq 0$. Or $(A^t)^{-1} = (A^{-1})^t$. D'où l'équivalence.
4. Supposons que A vérifie (1.23), et soit $x \in \mathbb{R}^n$ tel que $Ax \geq 0$. Soit $k \in 1, \ldots, n$ tel que $x_k = \min\{x_i, i = 1, \ldots, n\}$. Alors

$$(Ax)_k = a_{k,k}x_k + \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{k,j}x_j \geq 0.$$

   Par hypothèse, $a_{k,j} \leq 0$ pour $k \neq j$, et donc $a_{k,j} = -|a_{k,j}|$. On peut donc écrire :

$$a_{k,k}x_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}|x_j \geq 0,$$

et donc :

$$(a_{k,k} - \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k,j}|)x_k \geq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k,j}|(x_j - x_k).$$

Comme $x_k = \min\{x_i, i = 1, \ldots, n\}$, on en déduit que le second membre de cette inégalité est positif ou nul, et donc que $x_k \geq 0$. On a donc $x \geq 0$.

5. (a) Puisque la matrice A vérifie l'hypothèse (1.24) et puisque $\varepsilon > 0$, la matrice A_ε vérifie l'hypothèse (1.23), et c'est donc une IP-matrice par la question précédente.

   (b) Pour $\varepsilon > 0$, la matrice A_ε est une IP-matrice donc inversible, et pour $\varepsilon = 0$, $A_\varepsilon = A$ et A est inversible par hypothèse. La fonction $\varepsilon \mapsto A + \varepsilon \mathrm{Id}$ est continue de $\mathbb{R}$ dans $\mathcal{M}_n(\mathbb{R})$, et la fonction $M \mapsto M^{-1}$ est continue de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$. Par composition, les coefficients de A_ε^{-1} sont donc des fonctions continues de ε.

   (c) Comme la matrice A_ε est une IP-matrice, les coefficients de A_ε^{-1} sont tous positifs ou nuls. Par continuité, les coefficients de A^{-1} sont donc aussi tous positifs ou nuls, et donc A est une IP-matrice.

6. Soit $\mathbf{1}$ le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1. Si $Ax > 0$, comme l'espace $\mathbb{R}^n$ est de dimension finie, il existe $\epsilon > 0$ tel que $Ax \geq \epsilon \mathbf{1}$. Soit $z = \epsilon A^{-1}\mathbf{1} \geq 0$; on a alors $A(x - z) \geq 0$ et donc $x \geq z$, car A est une IP-matrice.

   Montrons maintenant que $z > 0$: tous les coefficients de A^{-1} sont positifs ou nuls et au moins l'un d'entre eux est non nul par ligne (puisque la matrice A^{-1} est inversible). On en déduit que $z_i = \epsilon \sum_{i=1}^{n} (A^{-1})_{i,j} > 0$ pour tout $i = 1, \ldots, n$. On a donc bien $x \geq z > 0$.

7. Soit A la matrice nulle, on a alors $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} = \emptyset$, et donc $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$. Pourtant A n'est pas inversible, et n'est donc pas une IP-matrice.

8. Soit x tel que $Ax \geq 0$, alors il existe $\varepsilon \geq 0$ tel que $Ax + \varepsilon \mathbf{1} \geq 0$. Soit maintenant $b = A^{-1}\mathbf{1}$; on a $A(x + \varepsilon b) > 0$ et donc $x + \varepsilon b > 0$. En faisant tendre ε vers 0, on en déduit que $x \geq 0$.

9. Soit $T \in \mathcal{L}(E)$ défini par $f \in E \mapsto Tf$, avec $Tf(x) = f(\frac{1}{x})$ si $x \neq 0$ et $f(0) = \ell$, avec $\ell = \lim_{\pm\infty} f$. On vérifie facilement que $Tf \in E$. Si $Tf \geq 0$, alors $f(\frac{1}{x}) \geq 0$ pour tout $x \in \mathbb{R}$; donc $f(x) \geq 0$ pour tout $x \in \mathbb{R} \setminus \{0\}$; on en déduit que $f(0) \geq 0$ par continuité. On a donc bien $f \geq 0$.

   Soit maintenant g définie de $\mathbb{R}$ dans $\mathbb{R}$ par $g(x) = |\arctan x|$. On a $g(0) = 0$, donc $g \not> 0$. Or $Tg(0) = \frac{\pi}{2}$ et $Tg(x) = |\arctan \frac{1}{x}| > 0$ si $x > 0$, donc $Tg > 0$.

Exercice 14 page 20 (Matrice du laplacien discret 1D.)

Il est clair que la matrice A est symétrique.
Pour montrer que A est définie positive (car A est évidemment symétrique), on peut procéder de plusieurs façons :

1. *Par échelonnement* :
2. *Par les valeurs propres* :Les valeurs propres sont calculées à l'exercice 55 ; elles sont de la forme :

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos \frac{k\pi}{n+1}), k = 1, \ldots, n,$$

   et elles sont donc toutes strictement positives ; de ce fait, la matrice est symétrique définie positive (voir exercice 10).

3. *Par la forme quadratique associée* : on montre que $Ax \cdot x > 0$ si $x \neq 0$ et $Ax \cdot x = 0$ ssi $x = 0$. En effet, on a

$$Ax \cdot x = \frac{1}{h^2}\left[x_1(2x_1 - x_2) + \sum_{i=2}^{n-1} x_i(-x_{i-1} + 2x_i - x_{i+1}) + 2x_n^2 - x_{n-1}x_n\right]$$

On a donc

$$\begin{aligned} h^2 Ax \cdot x &= 2x_1^2 - x_1x_2 - \sum_{i=2}^{n-1} \left(x_i x_{i-1} + 2x_i^2\right) - \sum_{i=3}^{n} x_i x_{i-1} + 2x_n^2 - x_{n-1}x_n \\ &= \sum_{i=1}^{n} x_i^2 + \sum_{i=2}^{n} x_{1-i}^2 + x_n^2 - 2\sum_{i=1}^{n} x_i x_{i-1} \\ &= \sum_{i=2}^{n} (x_i - x_{i-1})^2 + x_1^2 + x_n^2 \geq 0. \end{aligned}$$

De plus, $Ax \cdot x = 0 \Rightarrow x_1^2 = x_n = 0$ et $x_i = x_{i-1}$ pour $i = 2$ à n, donc $x = 0$.

Exercice 16 page 20 (Réaction diffusion 1D.)

La discrétisation du probllème consiste à chercher U comme solution du système linéaire

$$AU = \left(f(\frac{j}{N+1}) \right)_{j=1,\ldots,n}$$

où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est définie par $A = (N+1)^2 K_n + \mathrm{Id}$, Id désigne la matrice identité et

$$K_n = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

1.3 Les méthodes directes

1.3.1 Définition

Définition 1.11 (Méthode directe)**.** *On appelle méthode directe de résolution de (1.1) une méthode qui donne exactement x (A et b étant connus) solution de (1.1) après un nombre fini d'opérations élémentaires : addition, soustraction, mutiplication, division, et extraction de racine carrée pour la methode de choleski.*

Parmi les méthodes de résolution du système (1.1), la plus connue est la *méthode de Gauss* (avec pivot), encore appelée *méthode d'échelonnement* ou *méthode LU* dans sa forme matricielle.
Nous rappelons la méthode de Gauss et sa réécriture matricielle qui donne la méthode LU et nous étudierons plus en détails la méthode de Choleski, qui est adaptée aux matrices symétriques.

1.3.2 Méthode de Gauss, méthode LU

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$. On cherche à calculer $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = \boldsymbol{b}$. Le principe de la méthode de Gauss est de se ramener, par des opérations simples (combinaisons linéaires), à un système triangulaire équivalent, qui sera donc facile à inverser.
Commençons par un exemple pour une matrice 3×3. Nous donnerons ensuite la méthode pour une matrice $n \times n$.

Un exemple 3×3

On considère le système $A\boldsymbol{x} = \boldsymbol{b}$, avec

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \\ -1 & 1 & -2 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}.$$

On écrit la **matrice augmentée**, constituée de la matrice A et du second membre $\boldsymbol{b}$.

$$\tilde{A} = \begin{bmatrix} A & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & 2 & -1 & 1 \\ -1 & 1 & -2 & -2 \end{bmatrix}.$$

Gauss et opérations matricielles Allons y pour Gauss :
La première ligne a un 1 en première position (en gras dans la matrice), ce coefficient est non nul, et c'est un **pivot**. On va pouvoir diviser toute la première ligne par ce nombre pour en soustraire un multiple à toutes les lignes d'après, dans le but de faire apparaître des 0 dans tout le bas de la colonne.
La deuxième équation a déjà un 0 dessous, donc on n'a rien besoin de faire. On veut ensuite annuler le premier coefficient de la troisième ligne. On retranche donc (-1) fois la première ligne à la troisième [3] :

$$\begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & 2 & -1 & 1 \\ -1 & 1 & -2 & -2 \end{bmatrix} \overset{\ell_3 \leftarrow \ell_3 + \ell_1}{\longrightarrow} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix}$$

Ceci revient à multiplier $\tilde{A}$ à gauche par la matrice $E_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$.
La deuxième ligne a un terme non nul en deuxième position (2) : c'est un pivot. On va maintenant annuler le deuxième terme de la troisième ligne ; pour cela, on retranche 1/2 fois la ligne 2 à la ligne 3 :

$$\begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \overset{\ell_3 \leftarrow \ell_3 - 1/2\ell_2}{\longrightarrow} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 0 & -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{1}{2} \end{bmatrix}.$$

Ceci revient à multiplier la matrice précédente à gauche par la matrice $E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$. On a ici obtenu une matrice sous forme triangulaire supérieure à trois pivots : on peut donc faire la remontée pour obtenir la solution du système, et on obtient (en notant x_i les composantes de $\boldsymbol{x}$) : $x_3 = 1$ puis $x_2 = 1$ et enfin $x_1 = 1$.
On a ainsi résolu le système linéaire.
Le fait de travailler sur la matrice augmentée est extrêmement pratique car il permet de travailler simultanément sur les coefficients du système linéaire et sur le second membre.
Finalement, au moyen des opérations décrites ci-dessus, on a transformé le système linéaire

$$A\boldsymbol{x} = \boldsymbol{b} \text{ en } U\boldsymbol{x} = E_2E_1b, \text{ où } U = E_2E_1A$$

est une matrice triangulaire supérieure.

3. Bien sûr, ceci revient à ajouter la première ligne ! Il est cependant préférable de parler systématiquement de "retrancher" quitte à utiliser un coefficient négatif, car c'est ce qu'on fait conceptuellement : pour l'élimination on enlève un multiple de la ligne du pivot à la ligne courante.

Factorisation LU Tout va donc très bien pour ce système, mais supposons maintenant qu'on ait à résoudre 3089 systèmes, avec la même matrice A mais 3089 seconds membres b différents[4]. Il serait un peu dommage de recommencer les opérations ci-dessus 3089 fois, alors qu'on peut en éviter une bonne partie. Comment faire ? L'idée est de "factoriser" la matrice A, c.à.d de l'écrire comme un produit $A = LU$, où L est triangulaire inférieure (lower triangular) et U triangulaire supérieure (upper triangular). On reformule alors le système $A\boldsymbol{x} = \boldsymbol{b}$ sous la forme $LU\boldsymbol{x} = \boldsymbol{b}$ et on résout maintenant deux systèmes faciles à résoudre car triangulaires : $L\boldsymbol{y} = \boldsymbol{b}$ et $U\boldsymbol{x} = \boldsymbol{y}$. La factorisation LU de la matrice découle immédiatement de l'algorithme de Gauss. Voyons comment sur l'exemple précédent.

1/ On remarque que $U = E_2E_1A$ peut aussi s'écrire $A = LU$, avec $L = (E_2E_1)^{-1}$.

2/ On sait que $(E_2E_1)^{-1} = (E_1)^{-1}(E_2)^{-1}$.

3/ Les matrices inverses E_1^{-1} et E_2^{-1} sont faciles à déterminer : comme E_2 consiste à retrancher 1/2 fois la ligne 2 à la ligne 3, l'opération inverse consiste à ajouter 1/2 fois la ligne 2 à la ligne 3, et donc

$$E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}.$$

Il est facile de voir que $E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ et donc $L = E_1^{-1}E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & \frac{1}{2} & 1 \end{bmatrix}$.
La matrice L est une matrice triangulaire inférieure (et c'est d'ailleurs pour cela qu'on l'appelle L, pour "lower" in English...) dont les coefficients sont particulièrement simples à trouver : les termes diagonaux sont tous égaux à un, et **chaque terme non nul sous-diagonal $\ell_{i,j}$ est égal au coefficient par lequel on a multiplié la ligne pivot i avant de la retrancher à la ligne j**.

4/ On a bien donc $A = LU$ avec L triangulaire inférieure (lower triangular) et U triangulaire supérieure (upper triangular).

La procédure qu'on vient d'expliquer s'appelle **méthode LU** pour la résolution des systèmes linéaires, et elle est d'une importance considérable dans les sciences de l'ingénieur, puisqu'elle est utilisée dans les programmes informatiques pour la résolution des systèmes linéaires.
Dans l'exemple que nous avons étudié, tout se passait très bien car nous n'avons pas eu de zéro en position pivotale. Si on a un zéro en position pivotale, la factorisation peut quand même se faire, mais au prix d'une permutation. Le résultat général que l'on peut démontrer est que si la matrice A est inversible, alors il existe une matrice de permutation P, une matrice triangulaire inférieure L et une matrice triangulaire supérieure U telles que $PA = LU$: voir le théorème 1.20.

Le cas général d'une matrice $n \times n$

De manière plus générale, pour une matrice A carrée d'ordre n, la méthode de Gauss s'écrit :
On pose $A^{(1)} = A$ et $\boldsymbol{b}^{(1)} = \boldsymbol{b}$. Pour $i = 1, \ldots, n-1$, on cherche à calculer $A^{(i+1)}$ et $\boldsymbol{b}^{(i+1)}$ tels que les systèmes $A^{(i)}\boldsymbol{x} = \boldsymbol{b}^{(i)}$ et $A^{(i+1)}\boldsymbol{x} = \boldsymbol{b}^{(i+1)}$ soient équivalents, où $A^{(i+1)}$ est une matrice dont les coefficients sous-diagonaux des colonnes 1 à i sont tous nuls, voir figure 1.3. Une fois la matrice $A^{(n)}$ (triangulaire supérieure) et le vecteur $\boldsymbol{b}^{(n)}$ calculés, il sera facile de résoudre le système $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$. Le calcul de $A^{(n)}$ est l'étape de "factorisation", le calcul de $\boldsymbol{b}^{(n)}$ l'étape de "descente", et le calcul de $\boldsymbol{x}$ l'étape de "remontée". Donnons les détails de ces trois étapes.

Etape de factorisation et descente Pour passer de la matrice $A^{(i)}$ à la matrice $A^{(i+1)}$, on va effectuer des combinaisons linéaires entre lignes qui permettront d'annuler les coefficients de la i-ème colonne situés en dessous de la ligne i (dans le but de se rapprocher d'une matrice triangulaire supérieure). Evidemment, lorsqu'on fait ceci, il faut également modifier le second membre $\boldsymbol{b}$ en conséquence. L'étape de factorisation et descente s'écrit donc :

4. Ceci est courant dans les applications. Par exemple on peut vouloir calculer la réponse d'une structure de génie civil à 3089 chargements différents.

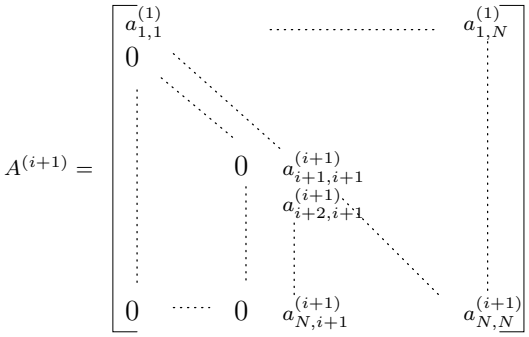

FIGURE 1.3: Allure de la matrice de Gauss à l'étape $i+1$

1. Pour $k \leq i$ et pour $j = 1, \ldots, n$, on pose $a_{k,j}^{(i+1)} = a_{k,j}^{(i)}$ et $b_k^{(i+1)} = b_k^{(i)}$.
2. Pour $k > i$, **si** $a_{i,i}^{(i)} \neq 0$, on pose :

$$a_{k,j}^{(i+1)} = a_{k,j}^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} a_{i,j}^{(i)}, \text{ pour } j = i, \ldots, n, \tag{1.28}$$

$$b_k^{(i+1)} = b_k^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} b_i^{(i)}. \tag{1.29}$$

La matrice $A^{(i+1)}$ est de la forme donnée sur la figure 1.3. Remarquons que le système $A^{(i+1)}\boldsymbol{x} = \boldsymbol{b}^{(i+1)}$ est bien équivalent au système $A^{(i)}\boldsymbol{x} = \boldsymbol{b}^{(i)}$.
Si la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour $i = 1$ à n, on obtient par le procédé de calcul ci-dessus un système linéaire $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$ équivalent au système $A\boldsymbol{x} = \boldsymbol{b}$, avec une matrice $A^{(n)}$ triangulaire supérieure facile à inverser. On verra un peu plus loin les techniques de pivot qui permettent de régler le cas où la condition $a_{i,i}^{(i)} \neq 0$ n'est pas vérifiée.

Etape de remontée Il reste à résoudre le système $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$; ceci est une étape facile. Comme $A^{(n)}$ est une matrice inversible, on a $a_{i,i}^{(i)} \neq 0$ pour tout $i = 1, \ldots, n$, et comme $A^{(n)}$ est une matrice triangulaire supérieure, on peut donc calculer les composantes de $\boldsymbol{x}$ en "remontant", c'est–à–dire de la composante x_n à la composante x_1 :

$$x_n = \frac{b_n^{(n)}}{a_{n,n}^{(n)}},$$

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

Il est important de savoir mettre sous forme algorithmique les opérations que nous venons de décrire : c'est l'étape clef avant l'écriture d'un programme informatique qui nous permettra de faire faire le boulot par l'ordinateur !

Algorithme 1.12 (Gauss sans permutation)**.**

1. *(Factorisation et descente) Pour commencer, on pose $u_{i,j} = a_{i,j}$ et $y_i = b_i$ pour pour $i, j \in \{1, \ldots, n\}$. Puis, pour i allant de 1 à $n-1$, on effectue les calculs suivants :*
   (a) *On ne change pas la i-ème ligne (qui est la ligne du pivot)*
   (b) *On modifie les lignes $i+1$ à n et le second membre $\boldsymbol{y}$ en utilisant la ligne i. Pour k allant de $i+1$ à n :*

$\ell_{k,i} = \frac{u_{k,i}}{u_{i,i}}$ *(si* $a_{i,i} = 0$*, prendre la méthode avec pivot partiel).*
Pour j *allant de* $i+1$ *à* n*,*
$u_{k,j} = u_{k,j} - \ell_{k,i} u_{i,j}$
Fin pour
$y_k = y_k - \ell_{k,i} y_i$
Fin pour

2. *(Remontée) On calcule* x *:*
$x_n = \dfrac{y_n}{u_{n,n}}$
Pour i *allant de* $n-1$ *à* 1*,*
$x_i = y_i$
Pour j *allant de* $i+1$ *à* n*,*
$x_i = x_i - u_{i,j} x_j$
Fin pour
$x_i = \dfrac{1}{u_{i,i}} x_i$
Fin pour

Coût de la méthode de Gauss (nombre d'opérations) On peut montrer (on fera le calcul de manière détaillée pour la méthode de Choleski dans la section suivante, le calcul pour Gauss est similaire) que le nombre d'opérations nécessaires n_G pour effectuer les étapes de factorisation, descente et remontée est $\frac{2}{3}n^3 + O(n^2)$; on rappelle qu'une fonction f de $\mathbb{N}$ dans $\mathbb{N}$ est $O(n^2)$ veut dire qu'il existe un réel constant C tel que $f(n) \leq Cn^2$. On a donc $\lim_{n\to+\infty} \frac{n_G}{n^3} = \frac{2}{3}$: lorsque n est grand, le nombre d'opérations se comporte comme $(2/3)n^3$.
En ce qui concerne la place mémoire, on peut très bien stocker les itérés $A^{(i)}$ dans la matrice A de départ, ce qu'on n'a pas voulu faire dans le calcul précédent, par souci de clarté.

Décomposition LU Si le système $A\boldsymbol{x} = \boldsymbol{b}$ doit être résolu pour plusieurs second membres $\boldsymbol{b}$, on a déjà dit qu'on a intérêt à ne faire l'étape de factorisation (*i.e.* le calcul de $A^{(n)}$), qu'une seule fois, alors que les étapes de descente et remontée (*i.e.* le calcul de $\boldsymbol{b}^{(n)}$ et $\boldsymbol{x}$) seront faits pour chaque vecteur $\boldsymbol{b}$. L'étape de factorisation peut se faire en décomposant la matrice A sous la forme LU. Supposons toujours pour l'instant que lors de l'algorithme de Gauss, la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour tout $i = 1, \ldots, n$. La matrice L a comme coefficients $\ell_{k,i} = \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}}$ pour $k > i$, $\ell_{i,i} = 1$ pour tout $i = 1, \ldots, n$, et $\ell_{i,j} = 0$ pour $j > i$, et la matrice U est égale à la matrice $A^{(n)}$. On peut vérifier que $A = LU$ grâce au fait que le système $A^{(n)}x = b^{(n)}$ est équivalent au système $A\boldsymbol{x} = \boldsymbol{b}$. En effet, comme $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$ et $\boldsymbol{b}^{(n)} = L^{-1}\boldsymbol{b}$, on en déduit que $LU\boldsymbol{x} = \boldsymbol{b}$, et comme A et LU sont inversibles, on en déduit que $A^{-1}\boldsymbol{b} = (LU)^{-1}\boldsymbol{b}$ pour tout $\boldsymbol{b} \in \mathbb{R}^n$. Ceci démontre que $A = LU$. La méthode LU se déduit donc de la méthode de Gauss en remarquant simplement que, ayant conservé la matrice L, on peut effectuer les calculs sur $\boldsymbol{b}$ après les calculs sur A, ce qui donne :

Algorithme 1.13 (LU simple (sans permutation))**.**

1. *(Factorisation)*
On pose $u_{i,j} = a_{i,j}$ *pour pour* $i, j \in \{1, \ldots, n\}$*.*
Pour i *allant de* 1 *à* $n-1$*, on effectue les calculs suivants :*
   (a) *On ne change pas la* i*-ème ligne*
   (b) *On modifie les lignes* $i+1$ *à* n *((mais pas le second membre) en utilisant la ligne* i*.*
   Pour k *allant de* $i+1$ *à* n *:*
   $\ell_{k,i} = \frac{u_{k,i}}{u_{i,i}}$ *(si* $u_{i,i} = 0$*, prendre la méthode avec pivot partiel).*
   Pour j *allant de* $i+1$ *à* n*,*
   $u_{k,j} = u_{k,j} - \ell_{k,i} u_{i,j}$

Fin pour

Fin pour

2. *(Descente) On calcule y (avec $Ly = b$)*
   Pour i allant de 1 à n,
   $y_i = b_i - \sum_{k=1}^{i-1} \ell_{i,k} y_k$ *(on a ainsi implicitement $\ell_{i,i} = 1$)*
   Fin pour
3. *(Remontée) On calcule x (avec $Ux = y$)*
   Pour i allant de n à 1,
   $x_i = \frac{1}{u_{i,i}}(y_i - \sum_{j=i+1}^{n} u_{i,j} x_j)$
   Fin pour

Remarque 1.14 (Optimisation mémoire)**.** *L'introduction des matrices L et U et des vecteurs $\boldsymbol{y}$ et $\boldsymbol{x}$ n'est pas nécessaire. Tout peut s'écrire avec la matrice A et le vecteur $\boldsymbol{b}$, que l'on modifie au cours de l'algorithme. A la fin de la factorisation, U est stockée dans la partie supérieure de A (y compris la diagonale) et L dans la partie strictement inférieure de A (c'est-à-dire sans la diagonale, la diagonale de L est connue car toujours formée de 1). Dans l'algorithme précédent, on remplaçe donc tous les "u" et "l" par "a". De même, on remplaçe tous les "x" et "y" par "b". A la fin des étapes de descente et de remontée, la solution du problème est alors stockée dans $\boldsymbol{b}$.*
L'introduction de L, U, $\boldsymbol{x}$ et $\boldsymbol{y}$ peut toutefois aider à comprendre la méthode.

Nous allons maintenant donner une condition nécessaire et suffisante (CNS) pour qu'une matrice A admette une décomposition LU avec U inversible et sans permutation. Commençons par un petit lemme technique qui va nous permettre de prouver cette CNS.

Lemme 1.15 (Décomposition LU de la matrice principale d'ordre k)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ et $k \in \{1, \dots, n\}$. On appelle matrice principale d'ordre k de A la matrice $A_k \in \mathcal{M}_k(\mathbb{R})$ définie par $(A_k)_{i,j} = a_{i,j}$ pour $i = 1, \dots, k$ et $j = 1, \dots, k$. On suppose qu'il existe une matrice $L_k \in \mathcal{M}_k(\mathbb{R})$ triangulaire inférieure de coefficients diagonaux tous égaux à 1 et une matrice triangulaire supérieure $U_k \in \mathcal{M}_k(\mathbb{R})$ inversible, telles que $A_k = L_k U_k$. Alors A s'écrit sous la forme "par blocs" suivante :*

$$A = \begin{bmatrix} L_k & 0_{k\times(n-k)} \\ C_k & \mathrm{Id}_{n-k} \end{bmatrix} \begin{bmatrix} U_k & B_k \\ 0_{(n-k)\times k} & D_k \end{bmatrix}, \tag{1.30}$$

où $0_{p,q}$ désigne la matrice nulle de dimension $p \times q$, $B_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$ et $C_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $D_k \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$; de plus, la matrice principale d'ordre $k+1$ s'écrit sous la forme

$$A_{k+1} = \begin{bmatrix} L_k & 0_{1\times k} \\ \boldsymbol{c}_k & 1 \end{bmatrix} \begin{bmatrix} U_k & \boldsymbol{b}_k \\ 0_{k\times 1} & d_k \end{bmatrix} \tag{1.31}$$

où $\boldsymbol{b} \in \mathcal{M}_{k,1}(\mathbb{R})$ est la première colonne de la matrice B_k, $\boldsymbol{c}_k \in \mathcal{M}_{1,k}$ est la première ligne de la matrice C_k, et d_k est le coefficient de la ligne 1 et colonne 1 de D_k.

DÉMONSTRATION – On écrit la décomposition par blocs de A :

$$A = \begin{bmatrix} A_k & E_k \\ F_k & G_k \end{bmatrix},$$

avec $A_k \in \mathcal{M}_k(\mathbb{R})$, $E_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$, $F_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $G_k \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$. Par hypothèse, on a $A_k = L_k U_k$. De plus L_k et U_k sont inversibles, et il existe donc une unique matrice $B_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$ (resp. $C_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$) telle que $L_k B_k = E_k$ (resp $C_k U_k = F_k$). En posant $D_k = G_k - C_k B_k$, on obtient (1.30). L'égalité (1.31) en découle immédiatement. ∎

Proposition 1.16 (CNS pour LU sans permutation)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$. Les deux propriétés suivantes sont équivalentes.*

(P1) Il existe un unique couple (L, U), avec L matrice triangulaire inférieure de coefficients égaux à 1 et U une matrice inversible triangulaire supérieure, tel que $A = LU$.

(P2) Les mineurs principaux[5] *de A sont tous non nuls.*

DÉMONSTRATION – Si $A = LU$ avec L triangulaire inférieure de coefficients égaux à 1 et U inversible triangulaire supérieure, alors $A_k = L_k U_k$ où les matrices L_k et U_k les matrices principales d'ordre k de L et U, qui sont encore respectivement triangulaire inférieure de coefficients égaux à 1 et inversible triangulaire supérieure. On a donc

$$\det(A_k) = \det(L_k)\det(U_k) \neq 0 \text{ pour tout } k = 1, \dots, n,$$

et donc (P1) $\Rightarrow$ (P2).

Montrons maintenant la réciproque. On suppose que les mineurs sont non nuls, et on va montrer que $A = LU$. On va en fait montrer que pour tout $k = 1, \dots n$, on a $A_k = L_k U_k$ où L_k triangulaire inférieure de coefficients égaux à 1 et U_k inversible triangulaire supérieure. Le premier mineur est non nul, donc $a_{11} = 1 \times a_{11}$, et la récurrence est bien initialisée. On la suppose vraie à l'étape k. Par le lemme 1.15, on a donc A_{k+1} qui est de la forme (1.31), et donc une $A_{k+1} = L_{k+1}U_{k+1}$ Comme $\det(A_{k+1}) \neq 0$, la matrice U_{k+1} est inversible, et l'hypothèse de récurrence est vérifiée à l'ordre $k + 1$. On a donc bien (P2) $\Rightarrow$ (P1) (l'unicité de L et U est laissée en exercice). ∎

Que faire en cas de pivot nul : la technique de permutation ou de "pivot partiel" La caractérisation que nous venons de donner pour qu'une matrice admette une décomposition LU sans permutation est intéressante mathématiquement, mais de peu d'intérêt en pratique. On ne va en effet jamais calculer n déterminants pour savoir si on doit ou non permuter. En pratique, on effectue la décomposition LU sans savoir si on a le droit ou non de le faire, avec ou sans permutation. Au cours de l'élimination, si $a_{i,i}^{(i)} = 0$, on va permuter la ligne i avec une des lignes suivantes telle que $a_{k,i}^{(i)} \neq 0$. Notons que si le "pivot" $a_{i,i}^{(i)}$ est très petit, son utilisation peut entraîner des erreurs d'arrondi importantes dans les calculs et on va là encore permuter. En fait, même dans le cas où la CNS donnée par la proposition 1.16 est verifiée, la plupart des fonctions de libraries scientifiques vont permuter.
Plaçons-nous à l'itération i de la méthode de Gauss. Comme la matrice $A^{(i)}$ est forcément non singulière, on a :

$$\det(A^{(i)}) = a_{1,1}^{(i)} a_{2,2}^{(i)} \cdots a_{i-1,i-1}^{(i)} \det \begin{bmatrix} a_{i,i}^{(i)} & \dots & a_{i,n}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{n,i}^{(i)} & \dots & a_{n,n}^{(i)} \end{bmatrix} \neq 0.$$

On a donc en particulier

$$\det \begin{bmatrix} a_{i,i}^{(i)} & \dots & a_{i,n}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{n,i}^{(i)} & \dots & a_{n,n}^{(i)} \end{bmatrix} \neq 0.$$

On déduit qu'il existe $i_0 \in \{i, \dots, n\}$ tel que $a_{i_0,i}^{(i)} \neq 0$. On choisit alors $i_0 \in \{i, \dots, n\}$ tel que $|a_{i_0,i}^{(i)}| = \max\{|a_{k,i}^{(i)}|, k = i, \dots, n\}$. Le choix de ce max est motivé par le fait qu'on aura ainsi moins d'erreur d'arrondi. On échange alors les lignes i et i_0 (dans la matrice A et le second membre $\boldsymbol{b}$) et on continue la procédure de Gauss décrite plus haut.
L'intérêt de cette stratégie de pivot est qu'on aboutit toujours à la résolution du système (dès que A est inversible).

Remarque 1.17 (Pivot total)**.** *La méthode que nous venons de d'écrire est souvent nommée technique de pivot "partiel". On peut vouloir rendre la norme du pivot encore plus grande en considérant tous les coefficients restants et pas uniquement ceux de la colonne i. A l'etape i, on choisit maintenant i_0 et $j_0 \in \{i, \dots, n\}$ tels que $|a_{i_0,j_0}^{(i)}| = \max\{|a_{k,j}^{(i)}|, k = i, \dots, n, j = i, \dots, n\}$, et on échange alors les lignes i et i_0 (dans la matrice A et le second*

5. On rappelle que le mineur principal d'ordre k est le déterminant de la matrice prinicipale d'ordre k.

membre $\boldsymbol{b}$), les colonnes i et j_0 de A et les inconnues x_i et x_{j_0}. La stratégie du pivot total permet une moins grande sensibilité aux erreurs d'arrondi. L'inconvénient majeur est qu'on change la structure de A : si, par exemple la matrice avait tous ses termes non nuls sur quelques diagonales seulement, ceci n'est plus vrai pour la matrice $A^{(n)}$.

Ecrivons maintenant l'algorithme de la méthode LU avec pivot partiel ; pour ce faire, on va simplement remarquer que l'ordre dans lequel les équations sont prises n'a aucune importance pour l'algorithme. Au départ de l'algorithme, on initialise la bijection t de $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ par l'identité, c.à.d. $t(i)=i$; cette bijection t va être modifiée au cours de l'algorithme pour tenir compte du choix du pivot.

Algorithme 1.18 (LU avec pivot partiel)**.**

1. *(Initialisation de t) Pour i allant de 1 à n, $t(i)=i$. Fin pour*
2. *(Factorisation)*
   Pour i allant de 1 à n, on effectue les calculs suivants :
   (a) Choix du pivot (et de $t(i)$) : on cherche $i^*\in\{i,\dots,n\}$ t.q. $|a_{t(i^*),i}|=\max\{|a_{t(k),i}|, k\in\{i,\dots,n\}\}$ (noter que ce max est forcément non nul car la matrice est inversible).
   On modifie alors t en inversant les valeurs de $t(i)$ et $t(i^*)$.
   $p=t(i^*)$; $t(i^*)=t(i)$; $t(i)=p$.
   On ne change pas la ligne $t(i)$:
   $u_{t(i),j}=a_{t(i),j}$ *pour* $j=i,\dots,n$,
   (b) On modifie les lignes $t(k)$, $k>i$ (et le second membre), en utilisant la ligne $t(i)$.
   Pour $k=i+1,\dots,$ (noter qu'on a uniquement besoin de connaître l'ensemble , et pas l'ordre) :
   $\ell_{t(k),i}=\dfrac{a_{t(k),i}}{a_{t(i),i}}$
   Pour j allant de $i+1$ à n,
   $a_{t(k),j}=a_{t(k),j}-\ell_{t(k),i}u_{t(i),j}$
   Fin pour
   Fin pour
3. *(Descente) On calcule y*
   Pour i allant de 1 à n,
   $y_{t(i)}=b_{t(i)}-\sum_{j=1}^{i-1}\ell_{t(j),k}y_k$
   Fin pour
4. *(Remontée) On calcule x*
   Pour i allant de n à 1,
   $x_{(t(i)}=\frac{1}{u_{t(i),i}}(y_i-\sum_{j=i+1}^{n}u_{t(i),j}x_j)$
   Fin pour

NB : On a changé l'ordre dans lequel les équations sont considérées (le tableau t donne cet ordre, et donc la matrice P). On a donc aussi changé l'ordre dans lequel interviennent les composantes du second membre : le système $A\boldsymbol{x}=\boldsymbol{b}$ est devenu $PA\boldsymbol{x}=P\boldsymbol{b}$. Par contre, on n'a pas touché à l'ordre dans lequel interviennent les composantes de $\boldsymbol{x}$ et $\boldsymbol{y}$.

Il reste maintenant à signaler la propriété magnifique de cet algorithme...Il est inutile de connaitre *a priori* la bijection pour cet algorithme. A l'étape i de l'item 1 (et d'ailleurs aussi à l'étape i de l'item 2), il suffit de connaître $t(j)$ pour j allant de 1 à i, les opérations de $1(b)$ se faisant alors sur toutes les autres lignes (dans un ordre quelconque). Il suffit donc de partir d'une bijection arbitraire de $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ (par exemple l'identité) et de la modifier à chaque étape. Pour que l'algorithme aboutisse, il suffit que $a_{t(i),i}\neq 0$ (ce qui toujours possible car A est inversible).

Remarque 1.19 (Ordre des équations et des inconnues)**.** *L'algorithme se ramène donc à résoudre $LU\boldsymbol{x} = \boldsymbol{b}$, en résolvant d'abord $L\boldsymbol{y} = b$ puis $Ux = \boldsymbol{y}$. Notons que lors de la résolution du système $L\boldsymbol{y} = \boldsymbol{b}$, les équations sont dans l'ordre $t(1), \ldots, t(k)$ (les composantes de $\boldsymbol{b}$ sont donc aussi prises dans cet ordre), mais le vecteur $\boldsymbol{y}$ est bien le vecteur de composantes $(y_1, \ldots, y_n)$, dans l'ordre initial. Puis, on résout $U\boldsymbol{x} = \boldsymbol{y}$, et les équations sont encore dans l'ordre $t(1), \ldots, t(k)$ mais les vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ ont comme composantes respectives $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$.*

Le théorème d'existence L'algorithme LU avec pivot partiel nous permet de démontrer le théorème d'existence de la décomposition LU pour une matrice inversible.

Théorème 1.20 (Décomposition LU d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, il existe une matrice de permutation P telle que, pour cette matrice de permutation, il existe un et un seul couple de matrices (L, U) où L est triangulaire inférieure de termes diagonaux égaux à 1 et U est triangulaire supérieure, vérifiant*

$$PA = LU.$$

DÉMONSTRATION –
1.**L'existence** de la matrice P et des matrices L U peut s'effectuer en s'inspirant de l'algorithme "LU avec pivot partiel" 1.18). Posons $A^{(0)} = A$.
À chaque étape i de l'algorithme 1.18 peut s'écrire comme $A^{(i)} = E^{(i)}P^{(i)}A^{(i-1)}$, où $P^{(i)}$ est la matrice de permutation qui permet le choix du pivot partiel, et $E^{(i)}$ est une matrice d'élimination qui effectue les combinaisons linéaires de lignes permettant de mettre à zéro tous les coefficients de la colonne i situés en dessous de la ligne i. Pour simplifier, raisonnons sur une matrice 4×4 (le raisonnement est le même pour une matrice $n \times n$. On a donc en appliquant l'algorithme de Gauss :

$$E^{(3)}P^{(3)}E^{(2)}P^{(2)}E^{(1)}P^{(1)}A = U$$

Les matrices $P^{(i+1)}$ et $E^{(i)}$ ne commutent en général pas. Prenons par exemple E_2, qui est de la forme

$$E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \end{bmatrix}$$

Si $P^{(3)}$ est la matrice qui échange les lignes 3 et 4, alors

$$P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ et } P^{(3)}E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 0 & 1 \\ 0 & a & 1 & 0 \end{bmatrix}, \text{ alors que } E^{(2)}P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 0 & 1 \\ 0 & b & 1 & 0 \end{bmatrix}$$

Mais par contre, comme la multiplication *à gauche* par $P^{(i+1)}$ permute *les lignes* $i+1$ et $i+k$, pour un certain $k \geq 1$, et que la multiplication *à droite* permute *les colonnes* $i+1$ et $i+k$, la matrice $\widetilde{E^{(i)}} = P^{(i+1)}E^{(i)}P^{(i+1)}$ est encore une matrice triangulaire inférieure avec la même structure que $E^{(i)}$: on a juste échangé les coefficients extradiagonaux des lignes $i+1$ et $i+k$. On a donc

$$P^{(i+1)}E^{(i)} = \widetilde{E^{(i)}}P^{(i+1)}. \tag{1.32}$$

Dans l'exemple précédent, on effectue le calcul :

$$P^{(3)}E^{(2)}P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 1 & 0 \\ 0 & a & 0 & 1 \end{bmatrix} = \widetilde{E^{(2)}},$$

qui est une matrice triangulaire inférieure de coefficients tous égaux à 1, et comme $P^{(3)}P^{(3)} = \mathrm{Id}$, on a donc :

$$P^{(3)}E^{(2)} = \widetilde{E^{(2)}}P^{(3)}.$$

Pour revenir à notre exemple $n = 4$, on peut donc écrire :

$$E^{(3)}\widetilde{E^{(2)}}P^{(3)}\widetilde{E^{(1)}}P^{(2)}P^{(1)}A = U$$

Mais par le même raisonnement que précédemment, on a $P^{(3)}\widetilde{E^{(1)}} = \widetilde{\widetilde{E^{(1)}}}P^{(3)}$ où $\widetilde{\widetilde{E^{(1)}}}$ est encore une matrice triangulaire inférieure avec des 1 sur la diagonale. On en déduit que

$$E^{(3)}\widetilde{E^{(2)}}\widetilde{\widetilde{E^{(1)}}}P^{(3)}P^{(2)}P^{(1)}A = U, \text{ soit encore } PA = LU$$

où $P = P^{(3)}P^{(2)}P^{(1)}$ bien une matrice de permutation, et $L = (E^{(3)}\widetilde{E^{(2)}}\widetilde{\widetilde{E^{(1)}}})^{-1}$ est une matrice triangulaire inférieure avec des 1 sur la diagonale.

Le raisonnement que nous venons de faire pour $n = 3$ se généralise facilement à n quelconque. Dans ce cas, l'échelonnement de la matrice s'écrit sous la forme

$$U = E^{(n-1)}P^{(n-1)}\dots E^{(2)}P^{(2)}E^{(1)}P^{(1)}A,$$

et se transforme grâce à (1.32) en

$$U = F^{(n-1)}\dots F^{(2)}F^{(1)}P^{(n-1)}\dots P^{(2)}P^{(1)}A,$$

où les matrices $F^{(i)}$ sont des matrices triangulaires inférieures de coefficients diagonaux tous égaux à 1. Plus précisément, $F^{(n-1)} = E^{(n-1)}$, $F^{(n-2)} = \widetilde{E^{(n-2)}}$, $F^{(n-3)} = \widetilde{\widetilde{E^{(n-3)}}}$, etc... On montre ainsi par récurrence l'existence de la décomposition LU (voir aussi l'exercice 23 page 46).

2. Pour montrer **l'unicité** du couple (L, U) à P donnée, supposons qu'il existe une matrice P et des matrices L_1, L_2, triangulaires inférieures et U_1, U_2, triangulaires supérieures, telles que

$$PA = L_1U_1 = L_2U_2$$

Dans ce cas, on a donc $L_2^{-1}L_1 = U_2U_1^{-1}$. Or la matrice $L_2^{-1}L_1$ est une matrice triangulaire inférieure dont les coefficients diagonaux sont tout égaux à 1, et la matrice $U_2U_1^{-1}$ est une matrice triangulaire supérieure. On en déduit que $L_2^{-1}L_1 = U_2U_1^{-1} = \mathrm{Id}$, et donc que $L_1 = L_2$ et $U_1 = U_2$. ∎

Remarque 1.21 (Décomposition LU pour les matrices non inversibles)**.** *En fait n'importe quelle matrice carrée admet une décomposition de la forme $PA = LU$. Mais si la matrice A n'est pas inversible, son échelonnement va nous donner des lignes de zéros pour les dernières lignes . La décomposition LU n'est dans ce cas pas unique. Cette remarque fait l'objet de l'exercice 32.*

1.3.3 Méthode de Choleski

On va maintenant étudier la méthode de Choleski, qui est une méthode directe adaptée au cas où A est symétrique définie positive. On rappelle qu'une matrice $A \in \mathcal{M}_n(\mathbb{R})$ de coefficients $(a_{i,j})_{i=1,n,j=1,n}$ est symétrique si $A = A^t$, où A^t désigne la transposée de A, définie par les coefficients $(a_{j,i})_{i=1,n,j=1,n}$, et que A est définie positive si $Ax \cdot x > 0$ pour tout $x \in \mathbb{R}^n$ tel que $x \neq 0$. Dans toute la suite, $x \cdot y$ désigne le produit scalaire des deux vecteurs x et y de $\mathbb{R}^n$. On rappelle (exercice) que si A est symétrique définie positive elle est en particulier inversible.

Description de la méthode

Commençons par un exemple. On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$, qui est symétrique. Calculons sa décomposition LU. Par échelonnement, on obtient

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

La structure LU ne conserve pas la symétrie de la matrice A. Pour des raisons de coût mémoire, il est important de pouvoir la conserver. Une façon de faire est de décomposer U en sa partie diagonale fois une matrice triangulaire. On obtient

$$U = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 \end{bmatrix}$$

On a donc $U = DL^t$, et comme tous les coefficients de D sont positifs, on peut écrire $D = \sqrt{D}\sqrt{D}$, où $\sqrt{D}$ est la matrice diagonale dont les éléments diagonaux sont les racines carrées des éléments diagonaux de A. On a donc $A = L\sqrt{D}\sqrt{D}L^t = \widetilde{L}\widetilde{L}^t$, avec $\widetilde{L} = L\sqrt{D}$. Notons que la matrice $\widetilde{L}$ est toujours triangulaire inférieure, mais ses coefficients diagonaux ne sont plus astreints à être égaux à 1. C'est la décomposition de Choleski de la matrice A.

De fait, la méthode de Choleski consiste donc à trouver une décomposition d'une matrice A symétrique définie positive de la forme $A = LL^t$, où L est triangulaire inférieure de coefficients diagonaux strictement positifs. On résout alors le système $A\boldsymbol{x} = \boldsymbol{b}$ en résolvant d'abord $L\boldsymbol{y} = \boldsymbol{b}$ puis le système $L^t\boldsymbol{x} = \boldsymbol{y}$. Une fois la matrice A "factorisée", c'est–à–dire la décomposition LL^t obtenue (voir paragraphe suivant), on effectue les étapes de "descente" et "remontée" :

1. Etape 1 : "descente" Le système $L\boldsymbol{y} = \boldsymbol{b}$ s'écrit :

$$L\boldsymbol{y} = \begin{bmatrix} \ell_{1,1} & 0 & \\ \vdots & \ddots & \vdots \\ \ell_{n,1} & \dots & \ell_{n,n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

   Ce système s'écrit composante par composante en partant de $i = 1$.

$$\begin{aligned}
\ell_{1,1}y_1 = b_1, \text{ donc} \qquad & y_1 = \frac{b_1}{\ell_{1,1}} \\
\ell_{2,1}y_1 + \ell_{2,2}y_2 = b_2, \text{ donc} \qquad & y_2 = \frac{1}{\ell_{2,2}}(b_2 - \ell_{2,1}y_1) \\
\vdots \qquad & \vdots \\
\sum_{j=1,i} \ell_{i,j}y_j = b_i, \text{ donc} \qquad & y_i = \frac{1}{\ell_{i,i}}(b_i - \sum_{j=1,i-1} \ell_{i,j}y_j) \\
\vdots \qquad & \vdots \\
\sum_{j=1,n} \ell_{n,j}y_j = b_n, \text{ donc} \qquad & y_n = \frac{1}{\ell_{n,n}}(b_n - \sum_{j=1,n-1} \ell_{n,j}y_j).
\end{aligned}$$

   On calcule ainsi $y_1, y_2, \dots, y_n$.

2. Etape 2 : "remontée" On calcule maintenant x solution de $L^t x = y$.

$$L^t x = \begin{bmatrix} \ell_{1,1} & \ell_{2,1} & \dots & \ell_{n,1} \\ 0 & \ddots & & \\ \vdots & & & \vdots \\ 0 & \dots & & \ell_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \\ \vdots \\ \\ y_n \end{bmatrix}.$$

   On a donc :

$$\begin{aligned}
& \ell_{n,n}\, x_n = y_n \text{ donc } x_n = \frac{y_n}{\ell_{n,n}} \\
& \ell_{n-1,n-1}x_{n-1} + \ell_{n,n-1}x_n = y_{n-1} \text{ donc } x_{n-1} = \tfrac{y_{n-1} - \ell_{n,n-1}x_n}{\ell_{n-1,n-1}} \\
& \vdots \\
& \sum_{j=1,n} \ell_{j,1}x_j = y_1 \text{ donc } x_1 = \frac{y_1 - \sum_{j=2,n} \ell_{j,1}x_j}{\ell_{1,1}}.
\end{aligned}$$

   On calcule ainsi $x_n, x_{n-1}, \dots, x_1$.

Existence et unicité de la décomposition

Soit A une matrice symétrique définie positive. On sait déjà par le théorème 1.20 page 36, qu'il existe une matrice de permutation et L triangulaire inférieure et U triangulaire supérieure telles que $PA = LU$. L'avantage dans le cas où la matrice est symétrique définie positive, est que la décomposition est toujours possible sans permutation. On prouve l'existence et unicité en construisant la décomposition, c.à.d. en construisant la matrice L.
Pour comprendre le principe de la preuve, commençons d'abord par le cas $n = 2$. Dans ce cas on peut écrire $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$. On sait que $a > 0$ car A est s.d.p. . L'échelonnement de A donne donc

$$A = LU = \begin{bmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{bmatrix} \begin{bmatrix} a & b \\ 0 & c - \frac{b^2}{a} \end{bmatrix}$$

En extrayant la diagonale de U, on obtient :

$$A = LU = \begin{bmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & c - \frac{b^2}{a} \end{bmatrix} \begin{bmatrix} a & \frac{b}{a} \\ 0 & 1 \end{bmatrix}.$$

Et donc

$$A = \widetilde{L}\widetilde{L}^t \text{ avec } \widetilde{L} = \begin{bmatrix} \sqrt{a} & 0 \\ b\sqrt{\frac{ac-b^2}{a}} & \end{bmatrix}.$$

Théorème 1.22 (Décomposition de Choleski)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ ($n \geq 1$) une matrice symétrique définie positive. Alors il existe une unique matrice $L \in \mathcal{M}_n(\mathbb{R})$, $L = (\ell_{i,j})_{i,j=1}^n$, telle que :*

1. *L est triangulaire inférieure (c'est–à–dire $\ell_{i,j} = 0$ si $j > i$),*
2. *$\ell_{i,i} > 0$, pour tout $i \in \{1, \ldots, n\}$,*
3. *$A = LL^t$.*

DÉMONSTRATION –

I- Existence de L : démonstration par récurrence sur n

1. Dans le cas $n = 1$, on a $A = (a_{1,1})$. Comme A est symétrique définie positive, on a $a_{1,1} > 0$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = \sqrt{a_{1,1}}$, et on a bien $A = LL^t$.
2. On suppose que la décomposition de Choleski s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive, pour $1 \leq p \leq n$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive. Soit donc $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.33}$$

où $B \in \mathcal{M}_n(\mathbb{R})$ est symétrique, $a \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$. Montrons que B est définie positive, c.à.d. que $By \cdot y > 0$, pour tout $y \in \mathbb{R}^n$ tel que $y \neq 0$. Soit donc $y \in \mathbb{R}^n \setminus \{0\}$, et $x = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n+1}$. Comme A est symétrique définie positive, on a :

$$0 < Ax \cdot x = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} By \\ \hline a^t y \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = By \cdot y$$

donc B est définie positive. Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_n(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^n$ telle que :

(a) $m_{i,j} = 0$ si $j > i$

(b) $m_{i,i} > 0$

(c) $B = MM^t$.

On va chercher L sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \tag{1.34}$$

avec $b \in \mathbb{R}^n$, $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$. Pour déterminer $\boldsymbol{b}$ et λ, calculons LL^t où L est de la forme (1.34) et identifions avec A :

$$LL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & \lambda \end{array}\right] = \left[\begin{array}{c|c} MM^t & Mb \\ \hline b^tM^t & b^tb+\lambda^2 \end{array}\right]$$

On cherche $b \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$Mb = a \text{ et } b^tb + \lambda^2 = \alpha.$$

Comme M est inversible (en effet, le déterminant de M s'écrit $\det(M) = \prod_{i=1}^n m_{i,i} > 0$), la première égalité ci-dessus donne : $b = M^{-1}a$ et en remplaçant dans la deuxième égalité, on obtient : $(M^{-1}a)^t(M^{-1}a) + \lambda^2 = \alpha$, donc $a^t(M^t)^{-1}M^{-1}a + \lambda^2 = \alpha$ soit encore $a^t(MM^t)^{-1}a + \lambda^2 = \alpha$, c'est–à–dire :

$$a^tB^{-1}a + \lambda^2 = \alpha \tag{1.35}$$

Pour que (1.35) soit vérifiée, il faut que

$$\alpha - a^tB^{-1}a > 0 \tag{1.36}$$

Montrons que la condition (1.36) est effectivement vérifiée : Soit $z = \begin{pmatrix} B^{-1}a \\ -1 \end{pmatrix} \in \mathbb{R}^{n+1}$. On a $z \neq 0$ et donc $Az \cdot z > 0$ car A est symétrique définie positive. Calculons Az :

$$Az = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \left[\begin{array}{c} B^{-1}a \\ \hline -1 \end{array}\right] = \left[\begin{array}{c} 0 \\ \\ a^tB^{-1}a - \alpha \end{array}\right].$$

On a donc $Az \cdot z = \alpha - a^tB^{-1}a > 0$ ce qui montre que (1.36) est vérifiée. On peut ainsi choisir $\lambda = \sqrt{\alpha - a^tB^{-1}a}$ (> 0) de telle sorte que (1.35) est vérifiée. Posons :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline (M^{-1}a)^t & \lambda \end{array}\right].$$

La matrice L est bien triangulaire inférieure et vérifie $\ell_{i,i} > 0$ et $A = LL^t$.

On a terminé ainsi la partie "existence".

II- Unicité et calcul de L. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive ; on vient de montrer qu'il existe $L \in \mathcal{M}_n(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} > 0$ et $A = LL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{n} \ell_{i,k}\ell_{j,k}, \quad \forall\, (i,j) \in \{1 \dots n\}^2. \tag{1.37}$$

1. Calculons la 1-ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = \ell_{1,1}\ell_{1,1} \text{ donc } \ell_{1,1} = \sqrt{a_{1,1}} \quad (a_{1,1} > 0 \text{ car } \ell_{1,1} \text{ existe }), \\ &a_{2,1} = \ell_{2,1}\ell_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{\ell_{1,1}}, \\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2, \dots, n\}. \end{aligned}$$

2. On suppose avoir calculé les q premières colonnes de L. On calcule la colonne $(q+1)$ en prenant $j = q+1$ dans (1.37)

   Pour $i = q+1$, $a_{q+1,q+1} = \sum_{k=1}^{q+1} \ell_{q+1,k}\ell_{q+1,k}$ donc

   $$\ell_{q+1,q+1} = (a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}^2)^{1/2} > 0. \tag{1.38}$$

   Notons que $a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}^2 > 0$ car L existe : il est indispensable d'avoir d'abord montré l'existence de L pour pouvoir exhiber le coefficient $\ell_{q+1,q+1}$.
   On procède de la même manière pour $i = q+2, \ldots, n$; on a :

   $$a_{i,q+1} = \sum_{k=1}^{q+1} \ell_{i,k}\ell_{q+1,k} = \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k} + \ell_{i,q+1}\ell_{q+1,q+1}$$

   et donc

   $$\ell_{i,q+1} = \left(a_{i,q+1} - \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k}\right) \frac{1}{\ell_{q+1,q+1}}. \tag{1.39}$$

   On calcule ainsi toutes les colonnes de L. On a donc montré que L est unique par un moyen constructif de calcul de L.

■

Remarque 1.23 (Choleski et LU)**.** *Considérons une matrice A symétrique définie positive. Alors une matrice P de permutation dans le théorème 1.22 possible n'est autre que l'identité. Il suffit pour s'en convaincre de remarquer qu'une fois qu'on s'est donné la bijection $t = \mathrm{Id}$ dans l'algorithme 1.18, celle-ci n'est jamais modifiée et donc on a $P = \mathrm{Id}$. Les théorèmes d'existence et d'unicité 1.20 et 1.22 nous permettent alors de remarquer que $A = LU = \tilde{L}\tilde{L}^t$ avec $\tilde{L} = L\sqrt{D}$, où D est la matrice diagonale extraite de U, et $\sqrt{D}$ désigne la matrice dont les coefficients sont les racines carrées des coefficients de D (qui sont tous positifs). Voir à ce sujet l'exercice 33 page 48.*

La décomposition LU permet de caractériser les matrices symétriques définies positives.

Proposition 1.24 (Caractérisation des matrices symétriques définies positives par la décomposition LU)**.** *Soit A une matrice symétrique admettant une décomposition LU sans permutation, c'est-à-dire qu'on suppose qu'il existe L triangulaire inférieure de coefficients diagonaux tous égaux à 1, et U triangulaire supérieure telle que $A = LU$. Alors A est symérique définie positive si et seulement si tous les pivots (c'est-à-dire les coefficients diagonaux de la matrice U) sont strictement positifs.*

DÉMONSTRATION – Soit A une matrice symétrique admettant une décomposition LU sans permutation. Si A est symétrique définie positive, le théorème 1.22 de décomposition de Choleski donne immédiatement le résultat.
Montrons maintenant la réciproque : supposons que $A = LU$ avec tous les pivots strictement positifs. On a donc $A = LU$, et U est inversible car c'est une matrice triangulaire supérieure dont tous les coefficients diagonaux sont strictement positifs. Donc A est aussi inversible, et la décomposition LU est donc unique, par le théorème 1.20 de décomposition LU d'une matrice inversible. On a donc $A = LU = LD\tilde{L}^t$ où D est la matrice diagonale dont la diagonale est celle de U, et $\tilde{L}$ est la matrice triangulaire inférieure de coefficients diagonaux tous égaux à 1 définie par $\tilde{L}^t = D^{-1}U$. On a donc aussi par symétrie de A

$$A^t = \tilde{L}DL^t = A = LU$$

et par unicité de la décomposition LU, on en déduit que $\tilde{L} = L$ et $DL^t = U$, ce qui entraîne que $A = LDL^t = CC^t$ avec $C = L\sqrt{D}$. On a donc pour tout $\boldsymbol{x} \in \mathbb{R}^n$, $A\boldsymbol{x} \cdot \boldsymbol{x} = CC^t\boldsymbol{x} \cdot \boldsymbol{x} = \|C\boldsymbol{x}\|^2$ et donc que A est symétrique définie positive. ■

Attention : la proposition précédente est fausse si la décomposition est avec permutation, méditer pour s'en convaincre l'exemple $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, voir aussi exercice 35.

Remarque 1.25 (Pivot partiel et Choleski)**.** *Considérons une matrice A symétrique définie positive. On a vu dans le théorème qu'on n'a pas besoin de permutation pour obtenir la décomposition LL^t d'une matrice symétrique définie positive. Par contre, on utilise malgré tout la technique de pivot partiel pour minimiser les erreurs d'arrondi. On peut illustrer cette raison par l'exemple suivant :*

$$A = \begin{bmatrix} -10^{-n} & 1 \\ 1 & 1 \end{bmatrix}$$

À titre d'illustration, pour $n = 12$ en FORTRAN (double précision), on obtient la bonne solution, c.à.d. $(-1, 1)$, avec le programme `gausslupivot` *donné plus haut, alors que le programme sans pivot* `gausslu` *donne comme solution $(0, 1)$.*

Calcul du coût de la méthode de Choleski

Calcul du coût de calcul de la matrice L. Dans le procédé de calcul de L exposé ci-dessus, le nombre d'opérations pour calculer la première colonne est n. Calculons, pour $p = 0, \ldots, n-1$, le nombre d'opérations pour calculer la $(p+1)$-ième colonne : pour la colonne $(p+1)$, le nombre d'opérations par ligne est $2p+1$, car le calcul de $\ell_{p+1,p+1}$ par la formule (1.38) nécessite p multiplications, p soustractions et une extraction de racine, soit $2p+1$ opérations ; le calcul de $\ell_{i,p+1}$ par la formule (1.39) nécessite p multiplications, p soustractions et une division, soit encore $2p+1$ opérations. Comme les calculs se font des lignes $p+1$ à n (car $\ell_{i,p+1} = 0$ pour $i \leq p$), le nombre d'opérations pour calculer la $(p+1)$-ième colonne est donc $(2p+1)(n-p)$. On en déduit que le nombre d'opérations N_L nécessaires au calcul de L est :

$$\begin{aligned} N_L &= \sum_{p=0}^{n-1} (2p+1)(n-p) = 2n \sum_{p=0}^{n-1} p - 2\sum_{p=0}^{n-1} p^2 + n\sum_{p=0}^{n-1} 1 - \sum_{p=0}^{n-1} p \\ &= (2n-1)\frac{n(n-1)}{2} + n^2 - 2\sum_{p=0}^{n-1} p^2. \end{aligned}$$

(On rappelle que $2\sum_{p=0}^{n-1} p = n(n-1)$.) Il reste à calculer $C_n = \sum_{p=0}^{n} p^2$, en remarquant par exemple que

$$\begin{aligned} \sum_{p=0}^{n} (1+p)^3 &= \sum_{p=0}^{n} 1 + p^3 + 3p^2 + 3p = \sum_{p=0}^{n} 1 + \sum_{p=0}^{n} p^3 + 3\sum_{p=0}^{n} p^2 + 3\sum_{p=0}^{n} p \\ &= \sum_{p=1}^{n+1} p^3 = \sum_{p=0}^{n} p^3 + (n+1)^3. \end{aligned}$$

On a donc $3C_n + 3\frac{n(n+1)}{2} + n + 1 = (n+1)^3$, d'où on déduit que

$$C_n = \frac{n(n+1)(2n+1)}{6}.$$

On a donc :

$$\begin{aligned} N_L &= (2n-1)\frac{n(n-1)}{2} - 2C_{n-1} + n^2 \\ &= n\left(\frac{2n^2+3n+1}{6}\right) = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} = \frac{n^3}{3} + 0(n^2). \end{aligned}$$

Coût de la résolution d'un système linéaire par la méthode LL^t. Nous pouvons maintenant calculer le coût (en termes de nombre d'opérations élémentaires) nécessaire à la résolution de (1.1) par la méthode de Choleski pour $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive. On a besoin de N_L opérations pour le calcul de L, auquel il faut rajouter le nombre d'opérations nécessaires pour les étapes de descente et remontée. Le calcul de y solution de $Ly = b$ s'effectue en résolvant le système :

$$\begin{bmatrix} \ell_{1,1} & & 0 \\ \vdots & \ddots & \vdots \\ \ell_{n,1} & \dots & \ell_{n,1} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

Pour la ligne 1, le calcul $y_1 = \dfrac{b_1}{\ell_{1,1}}$ s'effectue en une opération.

Pour les lignes $p = 2$ à n, le calcul $y_p = \left(b_p - \sum_{i=1}^{p-1} \ell_{i,p} y_i\right) / \ell_{p,p}$ s'effectue en $(p-1)$ (multiplications) $+(p-2)$ (additions) $+1$ soustraction $+1$ (division) $= 2p - 1$ opérations. Le calcul de y (descente) s'effectue donc en $N_1 = \sum_{p=1}^{n}(2p-1) = n(n+1) - n = n^2$. On peut calculer de manière similaire le nombre d'opérations nécessaires pour l'étape de remontée $N_2 = n^2$. Le nombre total d'opérations pour calculer x solution de (1.1) par la méthode de Choleski est $N_C = N_L + N_1 + N_2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} + 2n^2 = \frac{n^3}{3} + \frac{5n^2}{2} + \frac{n}{6}$. L'étape la plus coûteuse est donc la factorisation de A.

Remarque 1.26 (Décomposition LDL^t)**.** *Dans les programmes informatiques, on préfère implanter la variante suivante de la décomposition de Choleski : $A = \tilde{L}D\tilde{L}^t$ où D est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}^2$, $\tilde{L}_{i,i} = L\tilde{D}^{-1}$, où $\tilde{D}$ est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}$. Cette décomposition a l'avantage de ne pas faire intervenir le calcul de racines carrées, qui est une opération plus compliquée que les opérations "élémentaires" ($\times$, $+$, $-$).*

1.3.4 Quelques propriétés

Comparaison Gauss/Choleski

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible, la résolution de (1.1) par la méthode de Gauss demande $2n^3/3 + 0(n^2)$ opérations (exercice). Dans le cas d'une matrice symétrique définie positive, la méthode de Choleski est donc environ deux fois moins chère.

Et la méthode de Cramer ?

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible. On rappelle que la méthode de Cramer pour la résolution de (1.1) consiste à calculer les composantes de x par les formules :

$$x_i = \frac{\det(A_i)}{\det(A)},\ i = 1, \dots, n,$$

où A_i est la matrice carrée d'ordre n obtenue à partir de A en remplaçant la i-ème colonne de A par le vecteur $\boldsymbol{b}$, et $\det(A)$ désigne le déterminant de A.
Le calcul du déterminant d'une matrice carrée d'ordre n en utilisant les formules "usuelles" (c'est-à-dire en développant par rapport à une ligne ou une colonne) nécessite au moins $n!$ opérations (voir cours L1-L2, ou livres d'algèbre linéaire proposés en avant-propos). Par exemple, pour $n = 10$, la méthode de Gauss nécessite environ 700 opérations, la méthode de Choleski environ 350 et la méthode de Cramer (avec les formules usuelles de calcul du déterminant) plus de 4 000 000.... Cette dernière méthode est donc à proscrire.

Conservation du profil de A

Dans de nombreuses applications, par exemple lors de la résolution de systèmes linéaires issus de la discrétisation [6] de problèmes réels, la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est "creuse", au sens où un grand nombre de ses coefficients sont nuls. Il est intéressant dans ce cas pour des raisons d'économie de mémoire de connaître le "profil" de la matrice, donné dans le cas où la matrice est symétrique, par les indices $j_i = \min\{j \in \{1, \dots, n\} \text{ tels que } a_{i,j} \neq 0\}$. Le profil de la matrice est donc déterminé par les diagonales contenant des coefficients non nuls qui sont les plus éloignées de la diagonale principale. Dans le cas d'une matrice creuse, il est avantageux de faire un stockage "profil" de A, en stockant, pour chaque ligne i la valeur de j_i et des coefficients $a_{i,k}$, pour $k = i - j_i, \dots, i$, ce qui peut permettre un large gain de place mémoire.
Une propriété intéressante de la méthode de Choleski est de conserver le profil. On peut montrer (en reprenant les calculs effectués dans la deuxième partie de la démonstration du théorème 1.22) que $\ell_{i,j} = 0$ si $j < j_i$. Donc si on a adopté un stockage "profil" de A, on peut utiliser le même stockage pour L.

Matrices non symétriques

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible. On ne suppose plus ici que A est symétrique. On cherche à calculer $x \in \mathbb{R}^n$ solution de (1.1) par la méthode de Choleski. Ceci est possible en remarquant que : $Ax = b \Leftrightarrow A^tAx = A^tb$ car $\det(A) = \det(A^t) \neq 0$. Il ne reste alors plus qu'à vérifier que A^tA est symétrique définie positive. Remarquons d'abord que pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$, la matrice AA^t est symétrique. Pour cela on utilise le fait que si $B \in \mathcal{M}_n(\mathbb{R})$, alors B est symétrique si et seulement si $Bx \cdot y = x \cdot By$ et $Bx \cdot y = x \cdot B^ty$ pour tout $(x, y) \in (\mathbb{R}^n)^2$. En prenant $B = A^tA$, on en déduit que A^tA est symétrique. De plus, comme A est inversible, $A^tAx \cdot x = Ax \cdot Ax = |Ax|^2 > 0$ si $x \neq 0$. La matrice A^tA est donc bien symétrique définie positive.
La méthode de Choleski dans le cas d'une matrice non symétrique consiste donc à calculer A^tA et A^tb, puis à résoudre le système linéaire$A^tA \cdot x = A^tb$ par la méthode de Choleski "symétrique".
Cette manière de faire est plutôt moins efficace que la décomposition LU puisque le coût de la décomposition LU est de $2n^3/3$ alors que la méthode de Choleski dans le cas d'une matrice non symétrique nécessite au moins $4n^3/3$ opérations (voir exercice 27).

Systèmes linéaires non carrés

On considère ici des matrices qui ne sont plus carrées. On désigne par $\mathcal{M}_{M,n}(\mathbb{R})$ l'ensemble des matrices réelles à M lignes et n colonnes. Pour $A \in \mathcal{M}_{M,n}(\mathbb{R})$, $M > n$ et $b \in \mathbb{R}^M$, on cherche $x \in \mathbb{R}^n$ tel que

$$Ax = b. \tag{1.40}$$

Ce système contient plus d'équations que d'inconnues et n'admet donc en général pas de solution. On cherche $x \in \mathbb{R}^n$ qui vérifie le sytème (1.40) "au mieux". On introduit pour cela une fonction f définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par :

$$f(x) = |Ax - b|^2,$$

où $|x| = \sqrt{x \cdot x}$ désigne la norme euclidienne sur $\mathbb{R}^n$. La fonction f ainsi définie est évidemment positive, et s'il existe x qui annule f, alors x est solution du système (1.40). Comme on l'a dit, un tel x n'existe pas forcément, et on cherche alors un vecteur x qui vérifie (1.40) "au mieux", au sens où $f(x)$ soit le plus proche de 0. On cherche donc $x \in \mathbb{R}^n$ satisfaisant (1.40) en minimisant f, c.à.d. en cherchant $x \in \mathbb{R}^n$ solution du problème d'optimisation :

$$f(x) \leq f(y) \quad \forall y \in \mathbb{R}^n \tag{1.41}$$

On peut réécrire f sous la forme : $f(x) = A^tAx \cdot x - 2b \cdot Ax + b \cdot b$. On montrera au chapitre III que s'il existe une solution au problème (1.41), elle est donnée par la résolution du système linéaire suivant :

$$AA^tx = A^tb \in \mathbb{R}^n, \tag{1.42}$$

6. On appelle discrétisation le fait de se ramener d'un problème où l'inconnue est une fonction en un problème ayant un nombre fini d'inconnues scalaires.

qu'on appelle équations normales du problème de minimisation. La résolution approchée du problème (1.40) par cette procédure est appelée méthode des moindres carrés . La matrice AA^t étant symétrique, on peut alors employer la méthode de Choleski pour la résolution du système (1.42).

1.3.5 Exercices (méthodes directes)

Exercice 19 (Vrai ou faux ?)**.** *Corrigé en page 50*

Les propositions suivantes sont-elles vraies ou fausses ?

1. La matrice $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ admet une décomposition de Choleski.
2. La matrice $B = \begin{bmatrix} 1 & -2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ est symétrique définie positive.
3. La matrice B ci-dessus admet une décomposition LU.
4. La matrice $\begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}$ s'écrit C^tC.
5. La matrice $A = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ admet une décomposition de Choleski $A = C^tC$ avec $C = \begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$.
6. Soit $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$
   (a) La matrice AA^t admet une décomposition de Choleski.
   (b) La matrice A^tA admet une décomposition de Choleski.

Exercice 20 (Elimination de Gauss)**.** On cherche la solution du système linéaire $Ax = b$ avec

$$A = \begin{bmatrix} 1 & 0 & 6 & 2 \\ 8 & 0 & -2 & -2 \\ 2 & 9 & 1 & 3 \\ 2 & 1 & -3 & 10 \end{bmatrix} \text{ et } b = \begin{bmatrix} 6 \\ -2 \\ -8 \\ -4 \end{bmatrix}.$$

1. Pourquoi la méthode de Gauss sans permutation ne fonctionne pas pour résoudre ce système linéaire ?
2. Donner une permutation de lignes de A permettant d'utiliser ensuite la méthode de Gauss.
3. Donner la solution de ce système linéaire. (NB : La solution prend ses valeurs dans $\mathbb{Z}\ldots$)

Exercice 21 (LU)**.** *Corrigé en page 50*

1. Donner la décomposition LU de la matrice $A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix}$.
2. Montrer que la matrice $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ vérifie $PA = LU$ avec P une matrice de permutation, L triangulaire inférieure et U triangulaire supérieure à déterminer.
3. Calculer la décomposition LU de la matrice $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

Exercice 22 (Décomposition LU et mineurs principaux)**.**
Soit $n \geq 1$. On considère la matrice A de $\mathcal{M}_n(\mathbb{R})$ dont les coefficients sont :

$$a_{ij} = \begin{cases} -1 \text{ si } i > j, \\ 1 \text{ si } i = j, \\ 1 \text{ si } j = n, \\ 0 \text{ sinon.} \end{cases}$$

1. Montrer que $\det A = 2^{n-1}$. [On pourra par exemple raisonner par récurrence et remarquer que $\det A = \det B$ où B est obtenue en ajoutant, pour tout $i \in \{2, \ldots, n\}$, la première ligne de A à la i-ieme ligne de A, ce qui correspond à la première étape de l'algorithme de décomposition LU.]
2. Montrer que A admet une décomposition LU sans permutation et calculer les coefficients diagonaux de la matrice U.

Exercice 23 (Existence de la décomposition LU à une permutation près)**.** *Suggestions en page 49, corrigé en page 51*

L'objet de cet exercice est de démontrer par récurrence le résultat suivant (voir aussi théorème 1.20) :

Lemme 1.27 (Décomposition LU d'une matrice inversible par technique du pivot partiel)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible ; il existe une matrice de permutation $P \in \mathcal{M}_n(\mathbb{R})$ au sens de la définition 1.1, une matrice $L \in \mathcal{M}_n(\mathbb{R})$ triangulaire inférieure inversible et une matrice triangulaire supérieure $U \in \mathcal{M}_n(\mathbb{R})$ de coefficients diagonaux tous égaux à 1, telles que l'on ait la relation $PA = LU$ (décomposition LU de la matrice PA).*

Pour cela, nous allons démontrer par récurrence la propriété suivante : pour tout $k \in \{1, \ldots, n\}$, il existe une matrice de permutation $P^{(k)} \in \mathcal{M}_n(\mathbb{R})$, une matrice $L_k \in \mathcal{M}_k(\mathbb{R})$ triangulaire inférieure inversible et une matrice triangulaire supérieure $U_k \in \mathcal{M}_k(\mathbb{R})$ de coefficients diagonaux tous égaux à 1, telles que la matrice $A^{(k)} = P^{(k)}A$ vérifie $A_k^{(k)} = L_k U_k$, en notant $A_k^{(k)} \in \mathcal{M}_k(\mathbb{R})$ la matrice définie par $(A_k^{(k)})_{i,j} = a_{i,j}^{(k)}$ pour $i = 1, \ldots, k$ et $j = 1, \ldots, k$.

1. Montrer que l'hypothèse de récurrence est vrai au rang $k = 1$.

On suppose maintenant que la propriété de récurrence est vérifiée au rang $k \in \{1, \ldots, n-1\}$, et on va prouver qu'elle est encore vraie au rang $k+1$.

2. Montrer que la matrice $A^{(k)} = P^{(k)}A$ peut s'écrire sous la forme par blocs suivante :

$$A^{(k)} = \begin{bmatrix} L_k & 0_{k\times(n-k)} \\ C & D \end{bmatrix} \begin{bmatrix} U_k & V \\ 0_{(n-k)\times k} & \mathrm{Id}_{n-k} \end{bmatrix}, \tag{1.43}$$

où $0_{p,q}$ désigne la matrice nulle de dimension $p \times q$, $V \in \mathcal{M}_{k,n-k}(\mathbb{R})$ et $C \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $D \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$.

On appelle $\boldsymbol{c}_1(D)$, $\boldsymbol{c}_1(V)$, $\boldsymbol{c}_1(E)$ et $\boldsymbol{c}_1(G)$ les premières colonnes respectives des matrices D, V, E et G.

3. Montrer que $\boldsymbol{c}_1(D) \neq 0_{(n-k)\times 1}$.

Soit $i^* \in \{k+1, \ldots, n\}$ t.q. $|d_{i^*,1}| = \max\{|d_{i,1}|, 1 \in \{k+1, \ldots, n\}\}$. On pose $P^{(k+1)} = P^{(i^* \leftrightarrow k+1)}P^{(k)}$, $A^{(k+1)} = P^{(i^* \leftrightarrow k+1)}A^{(k)} = P^{(k+1)}A$, et

$$L_{k+1} = \begin{bmatrix} L_k & 0_{k\times 1} \\ \boldsymbol{\ell}_{i^*}(C) & d_{i^*,1} \end{bmatrix},\ U_{k+1} = \begin{bmatrix} U_k & \boldsymbol{c}_1(V) \\ 0_{1\times k} & 1 \end{bmatrix},\ A_{k+1}^{(k+1)} = \begin{bmatrix} A_k^{(k)} & \boldsymbol{c}_1(E) \\ \boldsymbol{\ell}_{i^*}(F) & g_{i^*,1} \end{bmatrix}, \tag{1.44}$$

où $\boldsymbol{\ell}_{i^*}(C)$ (resp. $\boldsymbol{\ell}_{i^*}(F)$) désigne la i^*-ème ligne de la matrice C (resp. F).

4. Montrer que les matrices $P^{(k+1)}$, L_{k+1} et U_{k+1} vérifient l'hypothèse de récurrence par construction, et conclure la démonstration du lemme 1.27.

Exercice 24 (Conservation du profil)**.** On considère des matrices A et $B \in \mathcal{M}_4(\mathbb{R})$ de la forme suivante, où x en position (i,j) de la matrice signifie que le coefficient $a_{i,j}$ est non nul et 0 en position (i,j) de la matrice signifie que $a_{i,j} = 0$)

$$A = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 \\ 0 & \mathrm{x} & \mathrm{x} & 0 \\ 0 & 0 & \mathrm{x} & \mathrm{x} \end{bmatrix} \text{ et } B = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 \\ \mathrm{x} & \mathrm{x} & 0 & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{bmatrix}.$$

Pour chacune de ces matrices, quels sont les coefficients nuls (notés 0 dans les matrices) qui resteront nécessairement nuls dans les matrices L et U de la factorisation LU sans permutation (si elle existe) ?

Exercice 25 (Une méthode directe particulière)**.** Soit $n \geq 1$, $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathbb{R}^n$ (on rappelle que $\mathbb{R}^n$ est identifié à $\mathcal{M}_{n,1}(\mathbb{R})$), $C \in \mathcal{M}_{1,n}$ et $D \in \mathbb{R}$. On note $\bar{A}$ la matrice appartenant à $\mathcal{M}_{n+1}(\mathbb{R})$ définie (par blocs) par :

$$\bar{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

On suppose que la matrice A est inversible.
On note x_B le vecteur de $\mathbb{R}^n$ tel que $Ax_B = B$.

1. Montrer que $\bar{A}$ est inversible si et seulement si $D - Cx_B \neq 0$.
2. On suppose maintenant que $\bar{A}$ est inversible. Soit $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On note x_b le vecteur de $\mathbb{R}^n$ tel que $Ax_b = b$. Montrer que la solution de $\bar{A}x = \begin{bmatrix} b \\ c \end{bmatrix}$ est donnée par $x = \begin{bmatrix} y \\ z \end{bmatrix}$ avec $z = \dfrac{c - Cx_b}{D - Cx_B}$ et $y = x_b - zx_B$.

Exercice 26 (Matrices définies positives et décomposition LU)**.** On rappelle que les mineurs principaux d'une matrice $A \in \mathcal{M}_n(\mathbb{R})$, sont les déterminants Δ_p des matrices $A_p = A(1:p, 1:p)$ extraites de la matrice A.

1. Montrer qu'une matrice symétrique définie positive a tous ses mineurs pricipaux strictement positifs.
2. En déduire que toute matrice symétrique définie positive admet une décomposition LU.

Exercice 27 (Sur la méthode LL^t)**.** *Corrigé détaillé en page 51.*
Soit A une matrice carrée d'ordre n, symétrique définie positive et pleine. On cherche à résoudre le système $A^2x = b$.
On propose deux méthodes de résolution de ce système :

1. Calculer A^2, effectuer la décomposition LL^t de A^2, résoudre le système $LL^tx = b$.
2. Calculer la décomposition LL^t de A, résoudre les systèmes $LL^ty = b$ et $LL^tx = y$.

Calculer le nombre d'opérations élémentaires nécessaires pour chacune des deux méthodes et comparer.

Exercice 28 (Décomposition LU d'une matrice à paramètres)**.** *Corrigé en page 52.*
Soient a, b, c et d des nombres réels. On considère la matrice suivante :

$$A = \begin{bmatrix} a & a & a & a \\ a & b & b & b \\ a & b & c & c \\ a & b & c & d \end{bmatrix}.$$

Appliquer l'algorithme d'élimination de Gauss à A pour obtenir sa décomposition LU (si elle existe).
Donner les conditions sur a, b, c et d pour que la matrice A soit inversible.

Exercice 29 (Echelonnement et factorisation LU et LDU)**.** *Corrigé en page 53.*
Echelonner les matrices suivantes (c.à.d. appliquer l'algorithme de Gauss), et lorsqu'elle existe, donner leur décomposition LU et LDU

$$A = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix}; \qquad B = \begin{bmatrix} 1. & 2. & 1. & 2. \\ -1. & -1. & 0. & -2. \\ 1. & 2. & 2. & 3. \\ -1. & -1. & 1. & 0. \end{bmatrix}.$$

Exercice 30 (Décomposition de Choleski d'une matrice particulière)**.**
Soit $n \in \mathbb{N} \setminus \{0\}$. On considère la matrice A_n carrée d'ordre n dont les coefficients sont donnés par $(A_n)_{i,j}$: $\min(i, j)$, et qui s'écrit donc :

$$A_n = \begin{bmatrix} 1 & 1 & \cdots & \cdots & 1 \\ 1 & 2 & \cdots & \cdots & 2 \\ \vdots & \vdots & & & \\ \vdots & \vdots & & n-1 & n-1 \\ 1 & 2 & & n-1 & n \end{bmatrix}$$

1. Écrire et échelonner les matrices A_2 et A_3. Montrer que A_2 et A_3 sont des matrices symétriques définies positives et donner leur décomposition de Choleski.
2. En déduire la décomposition de Choleski de la matrice A_n.

Exercice 31 (LU et Choleski sur un exemple)**.** Soit $M = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 8 & 10 \\ 1 & 10 & 18 \end{bmatrix}$.

1. Calculer les mineurs principaux de M. En déduire que M admet des décompositions LU et de Choleski.
2. Donner la décomposition LU de M.
3. Donner la décomposition de Choleski de M.

Exercice 32 (Matrices non inversibles et décomposition LU)**.**

1. Matrices 2×2
   (a) Soit $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ On suppose que $a_{11} \neq 0$.
      i. Echelonner la matrice A et en déduire qu'il existe une matrice $\widetilde{L}$ triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et une matrice $\widetilde{U}$ triangulaire supérieure telles que $A = \widetilde{L}\widetilde{U}$.
      ii. Montrer que $\widetilde{L}$ et $\widetilde{U}$ sont uniques.
      iii. Donner une condition nécessaire et suffisante sur les coefficients de A pour que la matrice $\widetilde{U}$ soit inversible.
   (b) On pose maintenant $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Trouver deux matrices $\widetilde{L}_1$ et $\widetilde{L}_2$ distinctes, toutes deux triangulaires inférieures et dont les coefficients diagonaux sont égaux à 1, et des matrices $\widetilde{U}_1$ et $\widetilde{U}_2$ triangulaires supérieures avec $A = \widetilde{L}_1\widetilde{U}_1 = \widetilde{L}_2\widetilde{U}_2$.
2. Matrices 3×3
   (a) Echelonner la matrice $A = \begin{bmatrix} 1. & 2. & 3. \\ 5. & 7. & 9. \\ 12. & 15. & 18. \end{bmatrix}$. et en déduire que la matrice A peut se décomposer en $A = \widetilde{L}\widetilde{U}$ où $\widetilde{L}$ est une matrice triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et $\widetilde{U}$ est une matrice triangulaire supérieure.
   (b) Soit $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$. Montrer que si $a_{11} \neq 0$ et que la matrice $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ est inversible, alors il existe un unique couple de matrices $(\widetilde{L}, \widetilde{U})$ tel que $A = \widetilde{L}\widetilde{U}$, où $\widetilde{L}$ est triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et une matrice $\widetilde{U}$ triangulaire supérieure.
3. Matrices $n \times n$.
   (a) Généraliser le résultat de la question précédente à une matrice de dimension n : donner le résultat espéré sous forme de théorème et le démontrer.
   (b) Soit maintenant A une matrice de dimensions $n \times n$. . Montrer qu'il existe une matrice de permutation P et des matrices $\widetilde{L}$ triangulaire inférieure et de coefficients diagonaux égaux à 1, et $\widetilde{U}$ triangulaire supérieure, telles que $PA = LU$. (On pourra commencer par le cas où est de rang égal à $n - 1$.)

Exercice 33 (Décomposition LL^t "pratique")**.** *Corrigé en page 54.*
1. Soit A une matrice symétrique définie positive. Montrer que la décomposition de Choleski $\tilde{L}\tilde{L}^t$ de la matrice A est obtenue à partir de sa décomposition LU en posant $\tilde{L} = L\sqrt{D}$ où D est la matrice diagonale extraite de U. (Voir remarque 1.23.)

En déduire la décomposition LL^t de la matrice particulière $A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$.

2. Que deviennent les coefficients nuls dans la décomposition LL^t ci-dessus ? Quelle est la propriété vue en cours qui est ainsi vérifiée ?

Exercice 34 (Factorisation de Choleski sur un exemple)**.** Calculer la factorisation de Choleski de la matrice suivante :

$$A = \begin{bmatrix} 4 & 4 & 2 & 0 \\ 4 & 5 & 0 & 0 \\ 2 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

Exercice 35 (Décomposition LDL^t et LL^t)**.** *Corrigé en page 56*

1. Soit $A = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$. Calculer la décomposition LDL^t de A. Existe-t-il une décomposition LL^t de A ?
2. Montrer que toute matrice de $\mathcal{M}_n(\mathbb{R})$ symétrique définie positive admet une décomposition LDL^t.
3. Ecrire l'algorithme de décomposition LDL^t. La matrice $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ admet-elle une décomposition LDL^t ?

Exercice 36 (Décomposition LL^t d'une matrice tridiagonale symétrique)**.** Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive et tridiagonale (i.e. $a_{i,j} = 0$ si $i - j > 1$).

1. Montrer que A admet une décomposition LL^t, où L est de la forme

$$L = \begin{bmatrix} \alpha_1 & 0 & \dots & & 0 \\ \beta_2 & \alpha_2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \dots & 0 \\ \vdots & \ddots & \ddots & \dots & \vdots \\ 0 & \cdots & 0 & \beta_n & \alpha_n \end{bmatrix}.$$

2. Donner un algorithme de calcul des coefficients α_i et β_i, en fonction des coefficients $a_{i,j}$, et calculer le nombre d'opérations élementaires nécessaires dans ce cas.
3. En déduire la décomposition LL^t de la matrice :

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

4. L'inverse d'une matrice inversible tridiagonale est elle tridiagonale ?

Exercice 37 (Choleski pour matrice bande)**.** Suggestions en page 50, corrigé en page 58
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive.

1. On suppose ici que A est tridiagonale. Estimer le nombre d'opérations de la factorisation LL^t dans ce cas.
2. Même question si A est une matrice bande (c'est-à-dire p diagonales non nulles).
3. En déduire une estimation du nombre d'opérations nécessaires pour la discrétisation de l'équation $-u'' = f$ vue page 11. Même question pour la discrétisation de l'équation $-\Delta u = f$ présentée page 13.

1.3.6 Suggestions

Exercice 23 page 46 (Existence de la décomposition LU à une permutation près)

2. Ecrire $A^{(k)} = P^{(k)}A$ sous une forme par blocs.

3. Procéder par contradiction.

Exercice 37 page 49

2. Soit q le nombre de sur- ou sous-diagonales ($p = 2q + 1$). Compter le nombre c_q d'opérations nécessaires pour le calcul des colonnes 1 à q et $n - q + 1$ à n, puis le nombre d_n d'opérations nécessaires pour le calcul des colonnes $n = q + 1$ $n - q$. En déduire l'estimation sur le nombre d'opérations nécessaires pour le calcul de toutes les colonnes, $Z_p(n)$, par :

$$2c_q \leq Z_p(n)2c_q + \sum_{n=q+1}^{n-q} c_n.$$

1.3.7 Corrigés

Exercice 19 page 45 (Vrai ou faux ?)

1. La matrice A est symétrique, sa trace est égale à 3 et son déterminant à 1, donc elle est s.d.p. et donc elle admet une décomposition de Choleski.
   Autre argument, ses deux mineurs principaux sont strictement positifs.
   Autre argument, A admet une décomposition LU avec 2 pivots strictement positifs
2. La matrice B n'est pas symétrique.
3. L'élimination de Gauss donne $A = LU$ avec

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

   La matrice B ci-dessus admet une décomposition LU.
4. Non car elle n'est pas symétrique.
5. La matrice $A = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ admet une décomposition de Choleski $A = C^tC$ avec $C = \begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$. Non la décomposition de Choleski fait apparaître des termes positifs sur la diagonale. Elle s'écrit

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

6. (a) FAUX. La matrice est d'ordre 3, mais de rang au plus 2, donc elle n'est pas inversible.
   (b) VRAI. $A^tA = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ qui est symétrique définie positive (trace et déterminants strictement positifs, par exemple).

Exercice 21 page 45 (Décomposition LU)

1. L'échelonnement donne

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0. & 0. \\ 0 & 0 & 1 & 0. \\ 1 & 1 & 1 & 1. \end{bmatrix} \text{ et } U = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

2. La matrice A est une matrice de permutation (des lignes 2 et 3). Donc on a $P = A$ et $PA = \mathrm{Id} = LU$ avec $L = U = \mathrm{Id}$.

3. Calculer la décomposition LU de la matrice $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$ L'échelonnement donne

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{3} & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} 2 & 1 & 0 \\ 0 & \frac{3}{2} & 1 \\ 0 & 0 & \frac{4}{3} \end{bmatrix}$$

Exercice 23 page 46 (Existence de la décomposition LU à une permutation près)

1. Vérifions la propriété de récurrence au rang $k = 1$. Soit $i^* \in \{1, \dots, n\}$ t.q. $|a_{i^*,1}| = \max\{|a_{i,1}|, 1 \in \{1, \dots, n\}\}$ (noter que ce max est forcément non nul car la matrice est inversible). Soit $P^{(1)} = P^{(1 \leftrightarrow i^*)}$ (voir Définition 1.1). On a alors $A_1^{(1)} = \begin{bmatrix} a_{i^*,1} \end{bmatrix}$, $L_1 = A_1^{(1)}$ et $U_1 = \begin{bmatrix} 1 \end{bmatrix}$.

2. Il suffit d'écrire la décomposition par blocs de $A^{(k)}$:

$$A^{(k)} = \begin{bmatrix} A_k^{(k)} & E \\ F & G \end{bmatrix},$$

avec $A_k^{(k)} \in \mathcal{M}_k(\mathbb{R})$, $E \in \mathcal{M}_{k,n-k}(\mathbb{R})$, $F \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $G \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$. Par hypothèse de récurrence, on a $A_k^{(k)} = L_k U_k$. De plus L_k et U_k sont inversibles, et il existe donc une unique matrice $V \in \mathcal{M}_{k,n-k}(\mathbb{R})$ (resp. $C \in \mathcal{M}_{n-k,k}(\mathbb{R})$) telle que $L_k V = E$ (resp $C U_k = F$). En posant $D = G - CV$, on obtient l'égalité (1.43).

3. En effet, si $\boldsymbol{c}_1(D) = 0_{(n-k)\times 1}$, alors $\boldsymbol{c}_1(G) = C\boldsymbol{c}_1(V) = FU^{-1}\boldsymbol{c}_1(V)$ et en même temps $\boldsymbol{c}_1(E) = L\boldsymbol{c}_1(V) = A_k^{(k)} U^{-1}\boldsymbol{c}_1(V)$. On obtient alors que la colonne $k+1$ de la matrice $A^{(k)}$, composée des deux vecteurs $\boldsymbol{c}_1(E)$ et $\boldsymbol{c}_1(G)$, est obtenue par la combinaison linéaire avec les coefficients $U^{-1}\boldsymbol{c}_1(V)$ des k premières colonnes de la matrice $A^{(k)}$, constituées des matrices $A_k^{(k)}$ et F. C'est impossible, puisque la matrice $A^{(k)}$ est le produit des deux matrices inversibles $P^{(k)}$ et A.

4. On a bien

1. $L_k v_{\cdot,1} = \boldsymbol{c}_1(E)$,
2. $\boldsymbol{\ell}_{i^*}(C) U_k = \boldsymbol{\ell}_{i^*}(F)$,
3. $\boldsymbol{\ell}_{i^*}(C)\boldsymbol{c}_1(V) + d_{i^*,1} = g_{i^*,1}$.

La conclusion du lemme est alors obtenue pour $k = n$.

Exercice 27 page 47 (Sur la méthode LL^t)

Calculons le nombre d'opérations élémentaires nécessaires pour chacune des méthodes :

1. Le calcul de chaque coefficient nécessite n multiplications et $n-1$ additions, et la matrice comporte n^2 coefficients. Comme la matrice est symétrique, seuls $n(n+1)/2$ coefficients doivent être calculés. Le calcul de A^2 nécessite nécessite donc $\frac{(2n-1)n(n+1)}{2}$ opérations élémentaires.

   Le nombre d'opérations élémentaires pour effectuer la décomposition LL^t de A^2 nécessite $\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$ (cours).

   La résolution du système $A^2 x = b$ nécessite $2n^2$ opérations (n^2 pour la descente, n^2 pour la remontée, voir cours).

   Le nombre total d'opérations pour le calcul de la solution du système $A^2 x = b$ par la première méthode est donc $\frac{(2n-1)n(n+1)}{2} + \frac{n^3}{3} + \frac{3n^2}{2} + \frac{n}{6} = \frac{4n^3}{3} + O(n^2)$ opérations.

2. La décomposition LL^t de A nécessite $\frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{6}$, et la résolution des systèmes $LL^ty=b$ et $LL^tx=y$ nécessite $4n^2$ opérations. Le nombre total d'opérations pour le calcul de la solution du système $A^2x=b$ par la deuxième méthode est donc $\frac{n^3}{3}+\frac{9n^2}{2}+\frac{n}{6}=\frac{n^3}{3}+O(n^2)$ opérations.

Pour les valeurs de n assez grandes, il est donc avantageux de choisir la deuxième méthode.

Exercice 28 page 47 (Décomposition LU d'une matrice à paramètres)

Appliquons l'algorithme de Gauss ; la première étape de l'élimination consiste à retrancher la première ligne à toutes les autres, c.à.d. à multiplier A à gauche par E_1, avec

$$E_1=\begin{bmatrix}1&0&0&0\\-1&1&0&0\\-1&0&1&0\\-1&0&0&1\end{bmatrix}.$$

On obtient :

$$E_1A=\begin{bmatrix}a&a&a&a\\0&b-a&b-a&b-a\\0&b-a&c-a&c-a\\0&b-a&c-a&d-a\end{bmatrix}.$$

La deuxième étape consiste à multiplier A à gauche par E_2, avec

$$E_2=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&-1&1&0\\0&-1&0&1\end{bmatrix}.$$

On obtient :

$$E_2E_1A=\begin{bmatrix}a&a&a&a\\0&b-a&b-a&b-a\\0&0&c-b&c-b\\0&0&c-b&d-b\end{bmatrix}.$$

Enfin, la troisième étape consiste à multiplier A à gauche par E_3, avec

$$E_3=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&-1&1\end{bmatrix}.$$

On obtient :

$$E_3E_2E_1A=\begin{bmatrix}a&a&a&a\\0&b-a&b-a&b-a\\0&0&c-b&c-b\\0&0&0&d-c\end{bmatrix}.$$

On $A=LU$ avec $L=(E_3E_2E_1)^{-1}=(E_1)^{-1}(E_2)^{-1}(E_3)^{-1}$; les matrices $(E_1)^{-1}$, $(E_2)^{-1}$ et $(E_3)^{-1}$ sont faciles à calculer : la multiplication à gauche par $(E_1)^{-1}$ consiste à ajouter la première ligne à toutes les suivantes ; on calcule de la même façon $(E_2)^{-1}$ et $(E_3)^{-1}$. On obtient (sans calculs !) :

$$(E_1)^{-1}=\begin{bmatrix}1&0&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\end{bmatrix},\ (E_2)^{-1}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&1&1&0\\0&1&0&1\end{bmatrix},\ (E_3)^{-1}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&1&1\end{bmatrix},$$

$$\text{et donc } L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & 0 & d-c \end{bmatrix}.$$

La matrice L est inversible car produit de matrices élémentaires, et la matrice A est donc inversible si et seulement si la matrice U l'est. Or U est une matrice triangulaire qui est inversible si et seulement si ses éléments diagonaux sont non nuls, c.à.d. $a \neq 0$, $b \neq c$ et $c \neq d$.

Exercice 29 page 47 (Echelonnement et factorisation LU et LDU)

Pour la première matrice, on donne le détail de l'élimination de Gauss sur cette matrice, et on montre ainsi qu'on peut stocker les multiplicateurs qu'on utilise au fur et à mesure dans la matrice L pour chaque étape k.
Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_3 \leftarrow \lambda_3 + + \lambda_1}{\overset{\lambda_2 \leftarrow \lambda_2 - 2\lambda_1}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} = A^{(2)}$$

où $\lambda_i \leftarrow \lambda_i - \alpha\lambda_j$ veut dire qu'on a soustrait α fois la ligne j à la ligne i. On a donc, sous forme matricielle,

$$A^{(2)} = E^{(1)}A^{(1)} \text{ avec } E^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Notons que $A = A^{(1)} = (E^{(1)})^{-1}A^{(2)}$ avec $(E^{(1)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ et donc $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & \mathrm{x} & 1 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 \end{bmatrix}$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_4 \leftarrow \lambda_4 - 3\lambda_2}{\overset{\lambda_3 \leftarrow \lambda_3 - \lambda_2}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = A^{(3)} = E^{(2)}A^{(2)} \text{ avec } E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}.$$

Notons que $A^{(2)} = (E^{(2)})^{-1}A^{(3)}$ avec $(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$ et donc $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$.

Et la vie est belle... car $A^{(3)}$ est déjà triangulaire supérieure, avec tous les coefficients diagonaux non nuls (ce qui prouve A est inversible). On n'a donc pas besoin d'étape 4 :

$$U = A^{(3)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

On a également $U = A^{(3)} = E^{(2)}E^{(1)}A$, soit encore $A = (E^{(1)})^{-1}(E^{(2)})^{-1}U = LU$ avec

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$$

On peut vérifier par le calcul qu'on a bien $A = LU$. Une fois que le mécanisme d'élimination est bien compris, il est inutile de calculer les matrices $E^{(k)}$: on peut directement stocker les multiplicateurs de l'élimination de Gauss dans la matrice L.

Pour la seconde matrice, l'élimination donne :

$$L = \begin{bmatrix} 1. & 0. & 0. & 0. \\ -1. & 1. & 0. & 0. \\ 1. & 0. & 1. & 0. \\ -1. & 1. & 1. & 1. \end{bmatrix}, U = \begin{bmatrix} 1. & 2. & 1. & 2. \\ 0. & 1. & 1. & 0. \\ 0. & 0. & 1. & 1. \\ 0. & 0. & 0. & 1. \end{bmatrix}$$

Exercice 33 page 48 (Décomposition LL^t "pratique")

1. Ecrivons l'élimination de Gauss sur cette matrice, en stockant les multiplicateurs qu'on utilise au fur et à mesure dans la matrice $E^{(k)}$ pour chaque étape k.
Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_3 \leftarrow \lambda_3 + +\lambda_1}{\overset{\lambda_2 \leftarrow \lambda_2 - 2\lambda_1}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} = A^{(2)}$$

où $\lambda_i \leftarrow \lambda_i - \alpha\lambda_j$ veut dire qu'on a soustrait α fois la ligne j à la ligne i. On a donc, sous forme matricielle,

$$A^{(2)} = E^{(1)}A^{(1)} \text{ avec } E^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Notons que $A = A^{(1)} = (E^{(1)})^{-1}A^{(2)}$ avec $(E^{(1)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_4 \leftarrow \lambda_4 - 3\lambda_2}{\overset{\lambda_3 \leftarrow \lambda_3 - \lambda_2}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = A^{(3)} = E^{(2)}A^{(2)} \text{ avec } E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}.$$

Notons que $A^{(2)} = (E^{(2)})^{-1}A^{(3)}$ avec $(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$.

Et la vie est belle... car $A^{(3)}$ est déjà triangulaire supérieure, avec tous les coefficients diagonaux non nuls (ce qui prouve A est inversible). On n'a donc pas besoin d'étape 4 :

$$U = A^{(3)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

On a également $U = A^{(3)} = E^{(2)}E^{(1)}A$, soit encore $A = (E^{(1)})^{-1}(E^{(2)})^{-1}U = LU$ avec

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$$

2. Si A est une matrice symétrique définie positive, on sait par le théorème 1.20 et la remarque 1.23 qu'il existe une unique décomposition LU : $A = LU$. Le théorème 1.22 nous donne l'existence (et l'unicité) de la décomposition $A = \tilde{L}\tilde{L}^t$. Soit $\tilde{D}$ la matrice diagonale extraite de $\tilde{L}$, qui est strictement positive par construction de $\tilde{L}$; on pose $\bar{L} = \tilde{L}\tilde{D}^{-1}$. On a donc $A = \bar{L}\tilde{D}\tilde{D}\bar{L}^t = \bar{L}\bar{U}$, avec $\bar{U} = \tilde{D}^2\bar{L}^t$. La matrice $\bar{D} = \tilde{D}^2$ est donc la diagonale de la matrice $\bar{U}$. Par unicité de la décomposition LU, on a $\bar{L} = L$, $\bar{U} = U$ et $\bar{D} = D$, et donc $\tilde{L} = L\sqrt{D}$.

Montrons maintenant que $A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$ est s.d.p (symétrique définite positive). Elle est évidemment symétrique. Soit $x = (a, b, c, d) \in \mathbb{R}^4$. Calculons $Ax \cdot x$:

$$Ax \cdot x = \begin{bmatrix} 2a - b \\ -a + 2b - c \\ -b + 2c - d \\ -c + 2d \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Donc $Ax \cdot x = 2a^2 - ab - ab + 2b^2 - bc - bc + 2c^2 - cd - cd + 2d^2 = a^2 + (a-b)^2 + (b-c)^2 + (c-d)^2 + d^2 \geq 0$. De plus $Ax \cdot x = 0$ ssi $a = b = c = d = 0$. Donc A est sdp.

On peut soit appliquer ici l'algorithme de construction de la matrice donné dans la partie unicité de la preuve du théorème 1.22 d'existence et d'unicité de la décomposition de Choleski, soit procéder comme en 1, calculer la décomposition LU habituelle, puis calculer la décomposition de $A = LU$, écrire $A = \tilde{L}\tilde{L}^t$ avec $\tilde{L} = L\sqrt{D}$, où $\sqrt{D}$et D la matrice diagonale extraite de U, comme décrit plus haut. Nous allons procéder selon le deuxième choix, qui est un peu plus rapide à écrire. (on utilise ici la notation $\tilde{L}$ parce que les matrices L dans les décompositions LU et LL^t ne sont pas les mêmes...)

Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_2 \leftarrow \lambda_2 + \frac{1}{2}\lambda_1}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = A^{(2)}$$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_3 \leftarrow \lambda_3 + \frac{2}{3}\lambda_2}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = A^{(3)}$$

Etape $k = 3$

$$A^{(3)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_4 \leftarrow \lambda_4 + \frac{3}{4}\lambda_3}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & 0 & \frac{5}{4} \end{bmatrix} = A^{(4)}$$

On vérifie alors qu'on a bien $U = A^{(4)} = DL^t$ où L est la matrice inverse du produit des matrices élémentaires utilisées pour transformer A en une matrice élémentaire (même raisonnement qu'en 1), c.à.d.

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & -\frac{2}{3} & 1 & 0 \\ 0 & 0 & -\frac{3}{4} & 1 \end{bmatrix}$$

On en déduit la décomposition $A = \tilde{L}\tilde{L}^t$ avec

$$\tilde{L} = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} & 0 & 0 \\ 0 & -\frac{\sqrt{6}}{3} & \frac{2\sqrt{3}}{3} & 0 \\ 0 & 0 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{5}}{2} \end{bmatrix}$$

3. Que deviennent les coefficients nuls dans la décomposition LL^t ci-dessus ? Quelle est la propriété vue en cours qui est ainsi vérifiée ?

Ils restent nuls : le profil est préservé, comme expliqué dans le cours page 17.

Exercice 35 page 49 (Décompositions LL^t et LDL^t)

1. On pose $L = \begin{bmatrix} 1 & 0 \\ \gamma & 1 \end{bmatrix}$ et $D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$. Par identification, on obtient $\alpha = 2$, $\beta = -\frac{1}{2}$ et $\gamma = \frac{1}{2}$.

Si maintenant on essaye d'écrire $A = LL^t$ avec $L = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix}$, on obtient $c^2 = -\frac{1}{2}$ ce qui est impossible dans $\mathbb{R}$.

En fait, on peut remarquer qu'il est normal que A n'admette pas de décomposition LL^t, car elle n'est pas définie positive. En effet, soit $\boldsymbol{x} = (x_1, x_2)^t \in \mathbb{R}^2$, alors $A\boldsymbol{x} \cdot \boldsymbol{x} = 2x_1(x_1 + x_2)$, et en prenant $\boldsymbol{x} = (1, -2)^t$, on a $A\boldsymbol{x} \cdot \boldsymbol{x} < 0$.

2. 2. Reprenons en l'adaptant la démonstration du théorème 1.3. On raisonne donc par récurrence sur la dimension.

1. Dans le cas $n = 1$, on a $A = (a_{1,1})$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = 1$, $D = (a_{1,1})$, $d_{1,1} \neq 0$, et on a bien $A = LDL^t$.
2. On suppose que, pour $1 \leq p \leq n$, la décomposition $A = LDL^t$ s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive ou négative, avec $d_{i,i} \neq 0$ pour $1 \leq i \leq n$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ou négative. Soit donc $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ou négative ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.45}$$

où $B \in \mathcal{M}_n(\mathbb{R})$ est symétrique définie positive ou négative (calculer $Ax \cdot x$ avec $x = (y, 0)^t$, avec $y \in \mathbb{R}^n$ pour le vérifier), $a \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$.

Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_n(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^n$ et une matrice diagonale $\tilde{D} = \mathrm{diag}(d_{1,1}, d_{2,2}, \ldots, d_{n,n})$ dont les coefficients sont tous non nuls, telles que :

(a) $m_{i,j} = 0$ si $j > i$

(b) $m_{i,i} = 1$

(c) $B = M\tilde{D}M^t$.

On va chercher L et D sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right], \; D = \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right], \tag{1.46}$$

avec $b \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$ tels que $LDL^t = A$. Pour déterminer b et λ, calculons LDL^t avec L et D de la forme (1.46) et identifions avec A :

$$LDL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right] \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b+\lambda \end{array}\right]$$

On cherche $b \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ tels que $LDL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$M\tilde{D}b = a \text{ et } b^t\tilde{D}b + \lambda = \alpha.$$

La matrice M est inversible (en effet, le déterminant de M s'écrit $det(M) = \prod_{i=1}^n 1 = 1$). Par hypothèse de récurrence, la matrice $\tilde{D}$ est aussi inversible. La première égalité ci-dessus donne : $b = \tilde{D}^{-1}M^{-1}a$. On calcule alors $\lambda = \alpha - b^tM^{-1}a$. Remarquons qu'on a forcément $\lambda \neq 0$, car si $\lambda = 0$,

$$A = LDL^t = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right]$$

qui n'est pas inversible. En effet, si on cherche $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ solution de

$$\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right] \left[\begin{array}{c} x \\ \hline y \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right],$$

on se rend compte facilement que tous les couples de la forme $(-M^{-t}by, y)^t$, $y \in \mathbb{R}$, sont solutions. Le noyau de la matrice n'est donc pas réduit à $\{0\}$ et la matrice n'est donc pas inversible. On a ainsi montré que $d_{n+1,n+1} \neq 0$ ce qui termine la récurrence.

3. On reprend l'algorithme de décomposition LL^t :
Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive ou négative ; on vient de montrer qu'il existe une matrice $L \in \mathcal{M}_n(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} = 1$, et une matrice $D \in \mathcal{M}_n(\mathbb{R})$ diagonale inversible, telles que et $A = LDL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{n} \ell_{i,k}d_{k,k}\ell_{j,k}, \quad \forall\,(i,j) \in \{1,\dots,n\}^2. \tag{1.47}$$

1. Calculons la 1ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = d_{1,1} \text{ donc } d_{1,1} = a_{1,1}, \\ &a_{2,1} = \ell_{2,1}d_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{d_{1,1}}, \\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{d_{1,1}} \ \forall i \in \{2,\dots,n\}. \end{aligned}$$

2. On suppose avoir calculé les n premières colonnes de L. On calcule la colonne $(k+1)$ en prenant $j = n+1$ dans (1.37).

   Pour $i = n+1$, $a_{n+1,n+1} = \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k} + d_{n+1,n+1}$ donc

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}. \tag{1.48}$$

On procède de la même manière pour $i = n+2, \ldots, n$; on a :

$$a_{i,n+1} = \sum_{k=1}^{n+1} \ell_{i,k} d_{k,k} \ell_{n+1,k} = \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} + \ell_{i,n+1} d_{n+1,n+1} \ell_{n+1,n+1},$$

et donc, comme on a montré dans la question 2 que les coefficients $d_{k,k}$ sont tous non nuls, on peut écrire :

$$\ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} \right) \frac{1}{d_{n+1,n+1}}. \tag{1.49}$$

3. Procédons par identification, en posant comme à la première question $L = \begin{bmatrix} 1 & 0 \\ \gamma & 1 \end{bmatrix}$ et $D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$. Pour que $A = LDL^t$, il faut $\alpha = 0$, $\beta = 0$ et $\alpha\gamma = 1$ ce qui est impossible. Cet exemple montre qu'une matrice symétrique (non définie positive) n'admet pas forcément une décomposition LDL^t, voir à ce propos la proposition 1.24.

Exercice 37 page 49 (Décomposition LL^t d'une matrice bande)

On utilise le résultat de conservation du profil de la matrice énoncé dans le cours, voir aussi exercice 24. Comme A est symétrique, le nombre p de diagonales de la matrice A est forcément impair si A ; notons $q = \frac{p-1}{2}$ le nombre de sous- et sur-diagonales non nulles de la matrice A, alors la matrice L aura également q sous-diagonales non nulles.

1. Cas d'une matrice tridiagonale. Si on reprend l'algorithme de construction de la matrice L vu en cours, on remarque que pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, on a le nombre d'opérations suivant :

— Calcul de $\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k} \ell_{n+1,k})^{1/2} > 0$:
une multiplication, une soustraction, une extraction de racine, soit 3 opérations élémentaires.

— Calcul de $\ell_{n+2,n+1} = \left(a_{n+2,n+1} - \sum_{k=1}^{n} \ell_{n+2,k} \ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}$:
une division seulement car $\ell_{n+2,k} = 0$.

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, est de 4.
Or le nombre d'opérations pour la première et dernière colonnes est inférieur à 4 (2 opérations pour la première colonne, une seule pour la dernière). Le nombre $Z_1(n)$ d'opérations élémentaires pour la décomposition LL^t de A peut donc être estimé par : $4(n-2) \leq Z_1(n) \leq 4n$, ce qui donne que $Z_1(n)$ est de l'ordre de $4n$ (le calcul exact du nombre d'opérations, inutile ici car on demande une estimation, est $4n-3$.)

2. Cas d'une matrice à p diagonales.
On cherche une estimation du nombre d'opérations $Z_p(n)$ pour une matrice à p diagonales non nulles (ou q sous-diagonales non nulles) en fonction de n.
On remarque que le nombre d'opérations nécessaires au calcul de

$$\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k} \ell_{n+1,k})^{1/2} > 0, \tag{1.50}$$

$$\text{et } \ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} \ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}, \tag{1.51}$$

est toujours inférieur à $2q+1$, car la somme $\sum_{k=1}^{n}$ fait intervenir au plus q termes non nuls.
De plus, pour chaque colonne $n+1$, il y a au plus $q+1$ coefficients $\ell_{i,n+1}$ non nuls, donc au plus $q+1$ coefficients à calculer. Donc le nombre d'opérations pour chaque colonne peut être majoré par $(2q+1)(q+1)$.
On peut donc majorer le nombre d'opérations z_q pour les q premières colonnes et les q dernières par $2q(2q+1)(q+1)$, qui est indépendant de n (on rappelle qu'on cherche une estimation en fonction de n, et donc le nombre z_q est $O(1)$ par rapport à n.)
Calculons maintenant le nombre d'opérations x_n nécessaires une colonne $n=q+1$ à $n-q-1$. Dans (1.50) et (1.51), les termes non nuls de la somme sont pour $k=i-q,\dots,n$, et donc on a $(n-i+q+1)$ multiplications et additions, une division ou extraction de racine. On a donc

$$\begin{aligned} x_n &= \sum_{i=n+1}^{n+q+1} \big(2(n-i+q+1)+1\big) \\ &= \sum_{j=1}^{q+1} \big(2(-j+q+1)+1\big) \\ &= (q+1)(2q+3) - 2\sum_{j=1}^{q+1} j \\ &= (q+1)^2. \end{aligned}$$

Le nombre z_i d'opérations nécessaires pour les colonnes $n=q+1$ à $n-q-1$ est donc

$$z_i = (q+1)^2(n-2q).$$

Un encadrement du nombre d'opérations nécessaires pour la décomposition LL^t d'une matrice à p diagonales est donc donnée par :

$$(q+1)^2(n-2q) \le Z_p(n) \le (q+1)^2(n-2q) + 2q(2q+1)(q+1), \tag{1.52}$$

et que, à q constant, $Z_p(n) = O((q+1)^2 n))$. Remarquons qu'on retrouve bien l'estimation obtenue pour $q=1$.

3. Dans le cas de la discrétisation de l'équation $-u'' = f$ (voir page 11), on a $q=1$ et la méthode de Choleski nécessite de l'ordre de $4n$ opérations élémentaires, alors que dans le cas de la discrétisation de l'équation $-\Delta u = f$ (voir page 13), on a $q=\sqrt{n}$ et la méthode de Choleski nécessite de l'ordre de n^2 opérations élémentaires (dans les deux cas n est le nombre d'inconnues).

On peut noter que l'encadrement (1.52) est intéressant dès que q est d'ordre inférieur à n^α, $\alpha < 1$.

1.4 Normes et conditionnement d'une matrice

Dans ce paragraphe, nous allons définir la notion de conditionnement d'une matrice, qui peut servir à établir une majoration des erreurs d'arrondi dues aux erreurs sur les données. Malheureusement, nous verrons également que cette majoration n'est pas forcément très utile dans des cas pratiques, et nous nous efforcerons d'y remédier. La notion de conditionnement est également utilisée dans l'étude des méthodes itératives que nous verrons plus loin. Pour l'étude du conditionnement comme pour l'étude des erreurs, nous avons tout d'abord besoin de la notion de norme et de rayon spectral, que nous rappelons maintenant.

1.4.1 Normes, rayon spectral

Définition 1.28 (Norme matricielle, norme induite)**.** *On note $\mathcal{M}_n(\mathbb{R})$ l'espace vectoriel (sur $\mathbb{R}$) des matrices carrées d'ordre n.*

1. *On appelle norme matricielle sur $\mathcal{M}_n(\mathbb{R})$ une norme $\|\cdot\|$ sur $\mathcal{M}_n(\mathbb{R})$ t.q.*

$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_n(\mathbb{R}) \tag{1.53}$$

2. *On considère $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$. On appelle norme matricielle induite (ou norme induite) sur $\mathcal{M}_n(\mathbb{R})$ par la norme $\|\cdot\|$, encore notée $\|\cdot\|$, la norme sur $\mathcal{M}_n(\mathbb{R})$ définie par :*

$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}, \ \forall A \in \mathcal{M}_n(\mathbb{R}) \tag{1.54}$$

Proposition 1.29 (Propriétés des normes induites)**.** *Soit $\mathcal{M}_n(\mathbb{R})$ muni d'une norme induite $\|\cdot\|$. Alors pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$, on a :*

1. $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$,
2. $\|A\| = \max\{\|A\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{R}^n\}$,
3. $\|A\| = \max\left\{\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \ ; \ \boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}\right\}$.
4. *$\|\cdot\|$ est une norme matricielle.*

DÉMONSTRATION –

1. Soit $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$, posons $\boldsymbol{y} = \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, alors $\|\boldsymbol{y}\| = 1$ donc $\|A\boldsymbol{y}\| \leq \|A\|$. On en déduit que $\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|$ et donc que $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$. Si maintenant $\boldsymbol{x} = 0$, alors $A\boldsymbol{x} = 0$, et donc $\|\boldsymbol{x}\| = 0$ et $\|A\boldsymbol{x}\| = 0$; l'inégalité $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$ est encore vérifiée.
2. L'application φ définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par : $\varphi(\boldsymbol{x}) = \|A\boldsymbol{x}\|$ est continue sur la sphère unité $S_1 = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$ qui est un compact de $\mathbb{R}^n$. Donc φ est bornée et atteint ses bornes : il existe $\boldsymbol{x}_0 \in S_1$ tel que $\|A\| = \|A\boldsymbol{x}_0\|$.
3. Cette égalité résulte du fait que
$$\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|A\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\| \text{ et } \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in S_1 \text{ et } \boldsymbol{x} \neq 0.$$
4. Soient A et $B \in \mathcal{M}_n(\mathbb{R})$, on a $\|AB\| = \max\{\|AB\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{R}^n\}$. Or
$$\|AB\boldsymbol{x}\| \leq \|A\|\|B\boldsymbol{x}\| \leq \|A\|\|B\|\|\boldsymbol{x}\| \leq \|A\|\|B\|.$$
On en déduit que $\|\cdot\|$ est une norme matricielle.

■

Définition 1.30 (Rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice. On appelle rayon spectral de A la quantité $\rho(A) = \max\{|\lambda|; \lambda \in \mathbb{C}, \lambda$ valeur propre de $A\}$.*

La proposition suivante caractérise les principales normes matricielles induites.

Proposition 1.31 (Caractérisation de normes induites)**.** *Soit $A = (a_{i,j})_{i,j\in\{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$.*

1. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Alors*

$$\|A\|_\infty = \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{i,j}|. \tag{1.55}$$

2. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Alors*

$$\|A\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^{n} |a_{i,j}| \tag{1.56}$$

3. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$.*

$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.57}$$

En particulier, si A est symétrique, $\|A\|_2 = \rho(A)$.

DÉMONSTRATION – La démonstration des points 1 et 2 fait l'objet de l'exercice 39 page 70. On démontre ici uniquement le point 3.
Par définition de la norme 2, on a :

$$\|A\|_2^2 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A\boldsymbol{x}\cdot A\boldsymbol{x} = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A^tA\boldsymbol{x}\cdot\boldsymbol{x}.$$

Comme A^tA est une matrice symétrique positive (car $A^tA\boldsymbol{x}\cdot\boldsymbol{x} = A\boldsymbol{x}\cdot A\boldsymbol{x} \geq 0$), il existe une base orthonormée $(\boldsymbol{f}_i)_{i=1,\dots,n}$ et des valeurs propres $(\mu_i)_{i=1,\dots,n}$, avec $0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ tels que $A\boldsymbol{f}_i = \mu_i\boldsymbol{f}_i$ pour tout $i \in \{1,\dots,n\}$. Soit $\boldsymbol{x} = \sum_{i=1,\dots,n} \alpha_i\boldsymbol{f}_i \in \mathbb{R}^n$. On a donc :

$$A^tA\boldsymbol{x}\cdot\boldsymbol{x} = \Big(\sum_{i=1,\dots,n} \mu_i\alpha_i\boldsymbol{f}_i\Big)\cdot\Big(\sum_{i=1,\dots,n} \alpha_i\boldsymbol{f}_i\Big) = \sum_{i=1,\dots,n} \alpha_i^2\mu_i \leq \mu_n\|\boldsymbol{x}\|_2^2.$$

On en déduit que $\|A\|_2^2 \leq \rho(A^tA)$.
Pour montrer qu'on a égalité, il suffit de considérer le vecteur $\boldsymbol{x} = \boldsymbol{f}_n$; on a en effet $\|\boldsymbol{f}_n\|_2 = 1$, et $\|A\boldsymbol{f}_n\|_2^2 = A^tA\boldsymbol{f}_n\cdot\boldsymbol{f}_n = \mu_n = \rho(A^tA)$.

■

Nous allons maintenant comparer le rayon spectral d'une matrice avec des normes. Rappelons d'abord le théorème de triangularisation (ou trigonalisation) des matrices complexes. On rappelle d'abord qu'une matrice unitaire $Q \in \mathcal{M}_n(\mathbb{C})$ est une matrice inversible telle que $Q^* = Q^{-1}$; ceci est équivalent à dire que les colonnes de Q forment une base orthonormale de $\mathbb{C}^n$. Une matrice carrée orthogonale est une matrice unitaire à coefficients réels ; on a dans ce cas $Q^* = Q^t$, et les colonnes de Q forment une base orthonormale de $\mathbb{R}^n$.

Théorème 1.32 (Décomposition de Schur, triangularisation d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ ou $\mathcal{M}_n(\mathbb{C})$ une matrice carrée quelconque, réelle ou complexe ; alors il existe une matrice complexe Q unitaire (c.à.d. une matrice telle que $Q^* = Q^{-1}$ et une matrice complexe triangulaire supérieure T telles que $A = QTQ^{-1}$.*
Ce résultat s'énonce de manière équivalente de la manière suivante : Soit ψ une application linéaire de E dans E, où E est un espace vectoriel de dimension finie n sur $\mathbb{C}$, muni d'un produit scalaire. Alors il existe une base orthonormée $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i=1,\dots,n,j=1,\dots,n,j\geq i}$ telles que $\psi(\boldsymbol{f}_i) = t_{i,i}\boldsymbol{f}_i + \sum_{k<i} t_{k,i}\boldsymbol{f}_k$. De plus $t_{i,i}$ est valeur propre de ψ et de A pour tout $i \in \{1,\dots,n\}$.
Les deux énoncés sont équivalents au sens où la matrice A de l'application linéaire ψ s'écrit $A = QTQ^{-1}$, où T est la matrice triangulaire supérieure de coefficients $(t_{i,j})_{i,j=1,\dots,n,j\geq i}$ et Q la matrice unitaire dont la colonne j est le vecteur $\boldsymbol{f}_j$).

DÉMONSTRATION – On démontre cette propriété par récurrence sur n. Elle est évidemment vraie pour $n = 1$. Soit $n \geq 1$, on suppose la propriété vraie pour n et on la démontre pour $n+1$. Soit donc E un espace vectoriel sur $\mathbb{C}$ de dimension $n+1$, muni d'un produit scalaire. Soit ψ une application linéaire de E dans E. On sait qu'il existe $\lambda \in \mathbb{C}$ (qui résulte du caractère algébriquement clos de $\mathbb{C}$) et $\boldsymbol{f}_1 \in E$ tels que $\psi(\boldsymbol{f}_1) = \lambda\boldsymbol{f}_1$ et $\|\boldsymbol{f}_1\| = 1$; on pose $t_{1,1} = \lambda$ et on note F le sous espace vectoriel de E supplémentaire orthogonal de $\mathbb{C}\boldsymbol{f}_1$. Soit $\boldsymbol{u} \in F$, il existe un unique couple $(\mu,\boldsymbol{v}) \in \mathbb{C}\times F$ tel que $\psi(\boldsymbol{u}) = \mu\boldsymbol{f}_1 + \boldsymbol{v}$. On note $\tilde{\psi}$ l'application qui à $\boldsymbol{u}$ associe $\boldsymbol{v}$. On peut appliquer l'hypothèse de

récurrence à $\tilde{\psi}$ (car $\tilde{\psi}$ est une application linéaire de F dans F, F est de dimension n et le produit scalaire sur E induit un produit scalaire sur F). Il existe donc une base orthonormée $\boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n+1}$ de F et $(t_{i,j})_{j\geq i\geq 2}$ tels que

$$\tilde{\psi}(\boldsymbol{f}_i) = \sum_{2\leq j\leq i} t_{j,i}\boldsymbol{f}_j,\ i = 2, \ldots, n+1.$$

On en déduit que

$$\psi(\boldsymbol{f}_i) = \sum_{1\leq j\leq i\leq n} t_{j,i}\boldsymbol{f}_j,\ i = 1, \ldots, n+1.$$

Le fait que l'ensemble des $t_{i,i}$ est l'ensemble des valeurs propores de A, comptées avec leur multiplicité. vient de l'égalité $\det(A - \lambda I) = \det(T - \lambda I)$ pour tout $\lambda \in \mathbb{C}$.

■

Dans la proposition suivante, nous montrons qu'on peut toujours trouver une norme (qui dépend de la matrice) pour approcher son rayon spectral d'aussi près que l'on veut par valeurs supérieures.

Théorème 1.33 (Approximation du rayon spectral par une norme induite)**.**
1. Soit $\|\cdot\|$ une norme induite. Alors

$$\rho(A) \leq \|A\|,\ \textit{pour tout } A \in \mathcal{M}_n(\mathbb{R}).$$

2. Soient maintenant $A \in \mathcal{M}_n(\mathbb{R})$ et $\varepsilon > 0$, alors il existe une norme sur $\mathbb{R}^n$ (qui dépend de A et ε) telle que la norme induite sur $\mathcal{M}_n(\mathbb{R})$, notée $\|\cdot\|_{A,\varepsilon}$, vérifie $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$.

DÉMONSTRATION – 1 Soit $\lambda \in \mathbb{C}$ valeur propre de A telle que $|\lambda| = \rho(A)$.
On suppose tout d'abord que $\lambda \in \mathbb{R}$. Il existe alors un vecteur non nul de $\mathbb{R}^n$, noté $\boldsymbol{x}$, tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Comme $\|\cdot\|$ est une norme induite, on a

$$\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\| = \|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{x}\|.$$

On en déduit que $|\lambda| \leq \|A\|$ et donc $\rho(A) \leq \|A\|$.

Si $\lambda \in \mathbb{C} \setminus \mathbb{R}$, la démonstration est un peu plus compliquée car la norme considérée est une norme dans $\mathbb{R}^n$ (et non dans $\mathbb{C}^n$). On montre tout d'abord que $\rho(A) < 1$ si $\|A\| < 1$.
En effet, Il existe $x \in \mathbb{C}^n$, $x \neq 0$, tel que $Ax = \lambda x$. En posant $x = y + iz$, avec $y, z \in \mathbb{R}^n$, on a donc pour tout $k \in \mathbb{N}$, $\lambda^k x = A^k x = A^k y + iA^k z$. Comme $\|A^k y\| \leq \|A\|^k\|y\|$ et $\|A^k z\| \leq \|A\|^k\|z\|$, on a, si $\|A\| < 1$, $A^k y \to 0$ et $A^k z \to 0$ (dans $\mathbb{R}^n$) quand $k \to +\infty$. On en déduit que $\lambda^k x \to 0$ dans $\mathbb{C}^n$. En choisissant une norme sur $\mathbb{C}^n$, notée $\|\cdot\|_a$, on a donc $|\lambda|^k\|x\|_a \to 0$ quand $k \to +\infty$, ce qui montre que $|\lambda| < 1$ et donc $\rho(A) < 1$.

Pour traiter le cas général (A quelconque dans $\mathcal{M}_n(\mathbb{R})$), il suffit de remarquer que la démonstration précédente donne, pour tout $\eta > 0$, $\rho(A/(\|A\| + \eta)) < 1$ (car $\|A/(\|A\| + \eta)\| < 1$). On a donc $\rho(A) < \|A\| + \eta$ pour tout $\eta > 0$, ce qui donne bien $\rho(A) \leq \|A\|$.
2. Soit $A \in \mathcal{M}_n(\mathbb{R})$, alors par le théorème de triangularisation de Schur (théorème 1.32 ppécédent), il existe une base $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i,j=1,\ldots,n, j\geq i}$ telles que $A\boldsymbol{f}_i = \sum_{j\leq i} t_{j,i}\boldsymbol{f}_j$. Soit $\eta \in]0, 1[$, qu'on choisira plus précisément plus tard. Pour $i = 1, \ldots, n$, on définit $\boldsymbol{e}_i = \eta^{i-1}\boldsymbol{f}_i$. La famille $(\boldsymbol{e}_i)_{i=1,\ldots,n}$ forme une base de $\mathbb{C}^n$. On définit alors une norme sur $\mathbb{R}^n$ par $\|\boldsymbol{x}\| = (\sum_{i=1}^n \alpha_i\overline{\alpha_i})^{1/2}$, où les α_i sont les composantes de $\boldsymbol{x}$ dans la base $(\boldsymbol{e}_i)_{i=1,\ldots,n}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon > 0$; montrons que pour η bien choisi, on a $\|A\| \leq \rho(A) + \varepsilon$. Remarquons d'abord que

$$A\boldsymbol{e}_i = A(\eta^{i-1}\boldsymbol{f}_i) = \eta^{i-1}A\boldsymbol{f}_i = \eta^{i-1}\sum_{j\leq i} t_{j,i}\boldsymbol{f}_j = \eta^{i-1}\sum_{j\leq i} t_{j,i}\eta^{1-j}\boldsymbol{e}_j = \sum_{1\leq j\leq i} \eta^{i-j}t_{j,i}\boldsymbol{e}_j,$$

Soit maintenant $\boldsymbol{x} = \sum_{i=1,\ldots,n} \alpha_i\boldsymbol{e}_i$. On a

$$A\boldsymbol{x} = \sum_{i=1}^n \alpha_i A\boldsymbol{e}_i = \sum_{i=1}^n \sum_{1\leq j\leq i} \eta^{i-j}t_{j,i}\alpha_i\boldsymbol{e}_j = \sum_{j=1}^n \Big(\sum_{i=j}^n \eta^{i-j}t_{j,i}\alpha_i\Big)\boldsymbol{e}_j.$$

On en déduit que

$$\begin{aligned}\|A\boldsymbol{x}\|^2 &= \sum_{j=1}^{n}\big(\sum_{i=j}^{n}\eta^{i-j}t_{j,i}\alpha_i\big)\big(\sum_{i=j}^{n}\eta^{i-j}\overline{t}_{j,i}\overline{\alpha}_i\big),\\ &= \sum_{j=1}^{n}t_{j,j}\overline{t_{j,j}}\alpha_j\overline{\alpha_j} + \sum_{j=1}^{n}\sum_{\substack{k,\ell\geq j\\(k,\ell)\neq(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}}\alpha_k\overline{\alpha_\ell}\\ &\leq \rho(A)^2\|\boldsymbol{x}\|^2 + \max_{k=1,\dots,n}|\alpha_k|^2\sum_{j=1}^{n}\sum_{\substack{k,\ell\geq j\\(k,\ell)\neq(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}}.\end{aligned}$$

Comme $\eta\in[0,1]$ et $k+\ell-2j\geq 1$ dans la dernière sommation, on a

$$\sum_{j=1}^{n}\sum_{\substack{k,\ell\geq j\\(k,\ell)\neq(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}}\leq \eta C_T n^3,$$

où $C_T=\max_{j,k,\ell=1,\dots,n}|t_{j,k}||t_{j,l}|$ ne dépend que de la matrice T, qui elle même ne dépend que de A. Comme

$$\max_{k=1,\dots,n}|\alpha_k|^2\leq\sum_{k=1,\dots,n}|\alpha_k|^2=\|\boldsymbol{x}\|^2,$$

on a donc, pour tout x dans $\mathbb{C}^n$, $x\neq 0$,

$$\frac{\|A\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2}\leq\rho(A)^2+\eta C_T n^3.$$

On en déduit que $\|A\|^2\leq\rho(A)^2+\eta C_T n^3$ et donc

$$\|A\|\leq\rho(A)\big(1+\frac{\eta C_T n^3}{\rho(A)^2}\big)^{\frac{1}{2}}\leq\rho(A)(1+\frac{\eta C_T n^3}{\rho(A)^2}).$$

D'où le résultat, en prenant $\|\cdot\|_{A,\varepsilon}=\|\cdot\|$ et η tel que $\eta=\min\left(1,\dfrac{\rho(A)\varepsilon}{C_T n^3}\right)$.

■

Corollaire 1.34 (Convergence et rayon spectral)**.** *Soit $A\in\mathcal{M}_n(\mathbb{R})$. Alors :*

$$\rho(A)<1 \textit{ si et seulement si } A^k\to 0 \textit{ quand } k\to\infty.$$

DÉMONSTRATION – Si $\rho(A)<1$, grâce au résultat d'approximation du rayon spectral de la proposition précédente, il existe $\varepsilon>0$ tel que $\rho(A)<1-2\varepsilon$ et une norme induite $\|.\|_{A,\varepsilon}$ tels que $\|A\|_{A,\varepsilon}=\mu\leq\rho(A)+\varepsilon=1-\varepsilon<1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon}\leq\mu^k\to 0$ lorsque $k\to\infty$. Comme l'espace $\mathcal{M}_n(\mathbb{R})$ est de dimension finie, toutes les normes sont équivalentes, et on a donc $\|A^k\|\to 0$ lorsque $k\to\infty$.
Montrons maintenant la réciproque : supposons que $A^k\to 0$ lorsque $k\to\infty$, et montrons que $\rho(A)<1$. Soient λ une valeur propre de A et $\boldsymbol{x}$ un vecteur propre associé. Alors $A^k\boldsymbol{x}=\lambda^k\boldsymbol{x}$, et si $A^k\to 0$, alors $A^k\boldsymbol{x}\to 0$, et donc $\lambda^k\boldsymbol{x}\to 0$, ce qui n'est possible que si $|\lambda|<1$.

■

Remarque 1.35 (Convergence des suites)**.** *Une conséquence immédiate du corollaire précédent est que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ définie par $\boldsymbol{x}^{(k+1)}=A\boldsymbol{x}^{(k)}$ converge vers $\mathbf{0}$ (le vecteur nul) pour tout $\boldsymbol{x}^{(0)}$ donné si et seulement si $\rho(A)<1$.*

Proposition 1.36 (Convergence et rayon spectral)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A\in\mathcal{M}_n(\mathbb{R})$. Alors*

$$\rho(A)=\lim_{k\to\infty}\|A^k\|^{\frac{1}{k}}. \tag{1.58}$$

DÉMONSTRATION – La démonstration se fait par des arguments d'homogénéité, en trois étapes. Rappelons tout d'abord que

$$\limsup_{k\to+\infty} u_k = \lim_{k\to+\infty} \sup_{n\geq k} u_n,$$
$$\liminf_{k\to+\infty} u_k = \lim_{k\to+\infty} \inf_{n\geq k} u_n,$$

et que si $\limsup_{k\to+\infty} u_k \leq \liminf_{k\to+\infty} u_k$, alors la suite $(u_k)_{k\in\mathbb{N}}$ converge vers $\lim_{k\to+\infty} u_k = \liminf_{k\to+\infty} u_k = \limsup_{k\to+\infty} u_k$.

Etape 1. On montre que

$$\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1. \tag{1.59}$$

En effet, si i $\rho(A) < 1$, d'après le corollaire 1.34 on a : $\|A^k\| \to 0$ donc il existe $K \in \mathbb{N}$ tel que pour $k \geq K$, $\|A^k\| < 1$. On en déduit que pour $k \geq K$, $\|A^k\|^{1/k} < 1$, et donc en passant à la limite sup sur k, on obtient bien que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq 1.$$

Etape 2. On montre maintenant que

$$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1. \tag{1.60}$$

Pour démontrer cette assertion, rappelons que pour toute suite $(u_k)_{k\in\mathbb{N}}$ d'éléments de $\mathbb{R}$ ou $\mathbb{R}^n$, la limite inférieure $\liminf_{k\to+\infty} u_k$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, donc qu'il existe une suite extraite $(u_{k_n})_{n\in\mathbb{N}}$ telle que $u_{k_n} \to \liminf_{k\to+\infty} u_k$ lorsque $n \to +\infty$. Or $\liminf_{k\to+\infty} \|A^k\|^{1/k} < 1$; donc il existe une sous-suite $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ telle que $\|A^{k_n}\|^{1/k_n} \to \ell < 1$ lorsque $n \to +\infty$. Soit $\eta \in]l, 1[$ il existe donc n_0 tel que pour $n \geq n_0$, $\|A^{k_n}\|^{1/k_n} \leq \eta$. On en déduit que pour $n \geq n_0$, $\|A^{k_n}\| \leq \eta^{k_n}$, et donc que $A^{k_n} \to 0$ lorsque $n \to +\infty$. Soient λ une valeur propre de A et x un vecteur propre associé, on a : $A^{k_n}x = \lambda^{k_n}x$; on en déduit que $|\lambda| < 1$, et donc que $\rho(A) < 1$.

Etape 3. On montre que $\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.
Soit $\alpha \in \mathbb{R}_+$ tel que $\rho(A) < \alpha$. Alors $\rho(\frac{1}{\alpha}A) < 1$, et donc grâce à (1.59),

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \alpha,\ \forall \alpha > \rho(A).$$

En faisant tendre α vers $\rho(A)$, on obtient donc :

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A). \tag{1.61}$$

Soit maintenant $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. On a alors $\liminf_{k\to+\infty} \|(\frac{1}{\beta}A)^k\|^{\frac{1}{k}} < 1$ et donc en vertu de (1.60), $\rho(\frac{1}{\beta}A) < 1$, donc $\rho(A) < \beta$ pour tout $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. En faisant tendre β vers $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}$, on obtient donc

$$\rho(A) \leq \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.62}$$

De (1.61) et (1.62), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.63}$$

■

Un corollaire important de la proposition 1.36 est le suivant.

Corollaire 1.37 (Comparaison rayon spectral et norme)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme **matricielle**, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

$$\rho(A) \leq \|A\|.$$

Par conséquent, si $M \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, pour montrer que la suite $\boldsymbol{x}^{(k)}$ définie par $\boldsymbol{x}^{(k)} = M^k\boldsymbol{x}^{(0)}$ converge vers $\mathbf{0}$ dans $\mathbb{R}^n$, il suffit de trouver une norme matricielle $\|\cdot\|$ telle que $\|M\| < 1$.

DÉMONSTRATION – Si $\|.\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc par la caractérisation (1.58) du rayon spectral donnée dans la proposition précédente, on obtient que $\rho(A) \leq \|A\|$. ■

Ce dernier résultat est évidemment bien utile pour montrer la convergence de la suite A^k, ou de suites de la forme $A^k\boldsymbol{x}^{(0)}$ avec $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. Une fois qu'on a trouvé une norme matricielle pour laquelle A est de norme strictement inférieure à 1, on a gagné. Attention cependant au piège suivant : pour toute matrice A, on peut toujours trouver une norme pour laquelle $\|A\| < 1$, alors que la série de terme général A^k peut ne pas être convergente.
Prenons un exemple dans $\mathbb{R}$, $\|x\| = \frac{1}{4}|x|$. Pour $x = 2$ on a $\|x\| = \frac{1}{2} < 1$. Et pourtant la série de terme général x^k n'est pas convergente ; le problème ici est que la norme choisie n'est pas une norme matricielle (on n'a pas $\|xy\| \leq \|x\|\|y\|$).
De même, on peut trouver une matrice et une norme telles que $\|A\| \geq 1$, alors que la série de terme général A^k converge...

Nous donnons maintenant un théorème qui nous sera utile dans l'étude du conditionnement, ainsi que plus tard dans l'étude des méthodes itératives.

Théorème 1.38 (Matrices de la forme $Id + A$)**.**

1. *Soit une norme matricielle induite,* Id *la matrice identité de* $\mathcal{M}_n(\mathbb{R})$ *et* $A \in \mathcal{M}_n(\mathbb{R})$ *telle que* $\|A\| < 1$. *Alors la matrice* $Id + A$ *est inversible et*
$$\|(Id + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$
2. *Si une matrice de la forme* $Id + A \in \mathcal{M}_n(\mathbb{R})$ *est singulière, alors* $\|A\| \geq 1$ *pour toute norme matricielle* $\|\cdot\|$.

DÉMONSTRATION –

1. La démonstration du point 1 fait l'objet de l'exercice 44 page 71.
2. Si la matrice $Id + A \in \mathcal{M}_n(\mathbb{R})$ est singulière, alors $\lambda = -1$ est valeur propre, et donc $\rho(A) \geq 1$. En utilisant le corollaire 1.37, on obtient que $\|A\| \geq \rho(A) \geq 1$.

■

1.4.2 Le problème des erreurs d'arrondis

Soient $A \in \mathcal{M}_n(\mathbb{R})$ inversible et $\boldsymbol{b} \in \mathbb{R}^n$; supposons que les données A et $\boldsymbol{b}$ ne soient connues qu'à une erreur près. Ceci est souvent le cas dans les applications pratiques. Considérons par exemple le problème de la conduction thermique dans une tige métallique de longueur 1, modélisée par l'intervalle $[0, 1]$. Supposons que la température u de la tige soit imposée aux extrémités, $u(0) = u_0$ et $u(1) = u_1$. On suppose que la température dans la tige satisfait à l'équation de conduction de la chaleur, qui s'écrit $(k(x)u'(x))' = 0$, où k est la conductivité thermique. Cette équation différentielle du second ordre peut se discrétiser par exemple par différences finies (on verra une description de la méthode page 11), et donne lieu à un système linéaire de matrice A. Si la conductivité k n'est connue qu'avec une certaine précision, alors la matrice A sera également connue à une erreur près, notée δ_A. On aimerait que l'erreur commise sur les données du modèle (ici la conductivité thermique k) n'ait pas une conséquence trop grave sur le calcul de la solution du modèle (ici la température u). Si par exemple 1% d'erreur sur k entraîne 100% d'erreur sur u, le modèle ne sera pas d'une utilité redoutable...
L'objectif est donc d'estimer les erreurs commises sur $\boldsymbol{x}$ solution de (1.1) à partir des erreurs commises sur $\boldsymbol{b}$ et A. Notons $\boldsymbol{\delta_b} \in \mathbb{R}^n$ l'erreur commise sur $\boldsymbol{b}$ et $\delta_A \in \mathcal{M}_n(\mathbb{R})$ l'erreur commise sur A. On cherche alors à évaluer $\boldsymbol{\delta_x}$ où $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution (si elle existe) du système :
$$\begin{cases} \boldsymbol{x} + \boldsymbol{\delta_x} \in \mathbb{R}^n \\ (A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}. \end{cases} \tag{1.64}$$
On va montrer que si δ_A "n'est pas trop grand", alors la matrice $A + \delta_A$ est inversible, et qu'on peut estimer $\boldsymbol{\delta_x}$ en fonction de δ_A et $\boldsymbol{\delta_b}$.

1.4.3 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.39 (Conditionnement). *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On appelle conditionnement de A par rapport à la norme $\|\cdot\|$ le nombre réel positif $\mathrm{cond}(A)$ défini par :*

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

Proposition 1.40 (Propriétés générales du conditionnement). *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, alors $\mathrm{cond}(A) \geq 1$.*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\alpha \in \mathbb{R}^*$, alors $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.*
3. *Soient A et $B \in \mathcal{M}_n(\mathbb{R})$ des matrices inversibles, alors $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.*

DÉMONSTRATION – 1. Comme $\|\cdot\|$ est une norme induite, c'est donc une norme matricielle. On a donc pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$,

$$\|\mathrm{Id}\| \leq \|A\| \, \|A^{-1}\|$$

ce qui prouve que $\mathrm{cond}(A) \geq 1$.

2. Par définition,

$$\begin{aligned} \mathrm{cond}(\alpha A) &= \|\alpha A\| \, \|(\alpha A)^{-1}\| \\ &= |\alpha| \, \|A\| \frac{1}{|\alpha|} \|A^{-1}\| = \mathrm{cond}(A) \end{aligned}$$

3. Soient A et B des matrices inversibles, alors AB est une matrice inversible et comme $\|\cdot\|$ est une norme matricielle,

$$\begin{aligned} \mathrm{cond}(AB) &= \|AB\| \, \|(AB)^{-1}\| \\ &= \|AB\| \, \|B^{-1}A^{-1}\| \\ &\leq \|A\| \, \|B\| \, \|B^{-1}\| \, \|A^{-1}\|. \end{aligned}$$

Donc $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$. ∎

Proposition 1.41 (Caractérisation du conditionnement pour la norme 2). *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note σ_n [resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Alors*

$$\mathrm{cond}_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. *Si de plus A une matrice symétrique définie positive, alors*

$$\mathrm{cond}_2(A) = \frac{\lambda_n}{\lambda_1},$$

où λ_n [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.

DÉMONSTRATION – On rappelle que si A a comme valeurs propres $\lambda_1, \ldots, \lambda_n$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$ et A^t a comme valeurs propres $\lambda_1, \ldots, \lambda_n$.
1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$. Or par le point 3. de la proposition 1.31 que $\|A\|_2 = (\rho(A^tA))^{1/2} = \sqrt{\sigma_n}$. On a donc

$$\|A^{-1}\|_2 = (\rho((A^{-1})^tA^{-1}))^{1/2} = \left(\rho(AA^t)^{-1}\right)^{1/2} \text{ ; or } \rho((AA^t)^{-1}) = \frac{1}{\tilde{\sigma}_1},$$

où $\tilde{\sigma}_1$ est la plus petite valeur propre de la matrice AA^t. Mais les valeurs propres de AA^t sont les valeurs propres de A^tA : en effet, si λ est valeur propre de AA^t associée au vecteur propre x alors λ est valeur propre de AA^t associée au vecteur propre A^tx. On a donc

$$cond_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. Si A est s.d.p., alors $A^tA = A^2$ et $\sigma_i = \lambda_i^2$ où λ_i est valeur propre de la matrice A. On a dans ce cas $cond_2(A) = \dfrac{\lambda_n}{\lambda_1}$. ∎

Les propriétés suivantes sont moins fondamentales, mais cependant intéressantes !

Proposition 1.42 (Propriétés du conditionnement pour la norme 2)**.** *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Alors $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.*
3. *Si A et B sont deux matrices symétriques définies positives, alors*

$$\mathrm{cond}_2(A+B) \leq \max(\mathrm{cond}_2(A), \mathrm{cond}_2(B)).$$

La démonstration de la proposition 1.42 fait l'objet de l'exercice 47 page 71.
On va maintenant majorer l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur $\delta_{\boldsymbol{b}}$ sur le second membre $\boldsymbol{b}$.

Proposition 1.43 (Majoration de l'erreur relative pour une erreur sur le second membre)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\boldsymbol{\delta_b} \in \mathbb{R}^n$. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de*

$$A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}, \tag{1.65}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \tag{1.66}$$

DÉMONSTRATION – En retranchant (1.1) à (1.65), on obtient :

$$A\boldsymbol{\delta_x} = \boldsymbol{\delta_b}$$

et donc

$$\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\boldsymbol{\delta_b}\|. \tag{1.67}$$

Cette première estimation n'est pas satisfaisante car elle porte sur l'erreur globale ; or la notion intéressante est celle d'erreur relative. On obtient l'estimation sur l'erreur relative en remarquant que $\boldsymbol{b} = A\boldsymbol{x}$, ce qui entraîne que $\|\boldsymbol{b}\| \leq \|A\|\|\boldsymbol{x}\|$. On en déduit que

$$\frac{1}{\|\boldsymbol{x}\|} \leq \frac{\|A\|}{\|\boldsymbol{b}\|}.$$

En multipliant membre à membre cette dernière inégalité et (1.67), on obtient le résultat souhaité. ∎

Remarquons que l'estimation (1.66) est optimale. En effet, on va démontrer qu'on peut avoir égalité dans (1.66). Pour cela, il faut choisir convenablement $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$. On sait déjà que si $\boldsymbol{x}$ est solution de (1.1) et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de (1.64), alors

$$\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}, \text{ et donc } \|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\|.$$

Soit $\boldsymbol{x} \in \mathbb{R}^n$ tel que $\|\boldsymbol{x}\| = 1$ et $\|A\boldsymbol{x}\| = \|A\|$. Notons qu'un tel $\boldsymbol{x}$ existe parce que

$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\} = \max\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\}$$

(voir proposition 1.29 page 60). On a donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|A\boldsymbol{x}\|}.$$

Posons $\boldsymbol{b} = A\boldsymbol{x}$; on a donc $\|\boldsymbol{b}\| = \|A\|$, et donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|b\|}.$$

De même, grâce à la proposition 1.29, il existe $\boldsymbol{y} \in \mathbb{R}^n$ tel que $\|\boldsymbol{y}\| = 1$, et $\|A^{-1}\boldsymbol{y}\| = \|A^{-1}\|$. On choisit alors $\boldsymbol{\delta_b}$ tel que $\delta_b = \boldsymbol{y}$. Comme $A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}$, on a $\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}$ et donc :

$$\|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\| = \|A^{-1}\boldsymbol{y}\| = \|A^{-1}\| = \|\boldsymbol{\delta_b}\| \ \|A^{-1}\|.$$

On en déduit que

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{\delta_x}\| = \|\boldsymbol{\delta_b}\| \ \|A^{-1}\| \ \ \frac{\|A\|}{\|\boldsymbol{b}\|} \text{ car } \|\boldsymbol{b}\| = \|A\| \text{ et } \|\boldsymbol{x}\| = 1.$$

Par ce choix de $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$ on a bien égalité dans (1.66) qui est donc optimale.
Majorons maintenant l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur δ_A sur la matrice A.

Proposition 1.44 (Majoration de l'erreur relative pour une erreur sur la matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\delta_A \in \mathcal{M}_n(\mathbb{R})$; on suppose que $A + \delta_A$ est une matrice inversible. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \delta_{\boldsymbol{x}}$ est solution de*

$$(A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} \tag{1.68}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x} + \boldsymbol{\delta_x}\|} \leq \text{cond}(A)\frac{\|\delta_A\|}{\|A\|} \tag{1.69}$$

DÉMONSTRATION – En retranchant (1.1) à (1.68), on obtient :

$$A\boldsymbol{\delta_x} = -\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x})$$

et donc

$$\boldsymbol{\delta_x} = -A^{-1}\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x}).$$

On en déduit que $\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\delta_A\|\|\boldsymbol{x} + \boldsymbol{\delta_x}\|$, d'où on déduit le résultat souhaité. ∎

On peut en fait majorer l'erreur relative dans le cas où l'on commet à la fois une erreur sur A et une erreur sur $\boldsymbol{b}$. On donne le théorème à cet effet ; la démonstration est toutefois nettement plus compliquée.

Théorème 1.45 (Majoration de l'erreur relative pour une erreur sur matrice et second membre)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq \boldsymbol{0}$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soient $\delta_A \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{\delta_b} \in \mathbb{R}^n$. On suppose que $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$. Alors la matrice $(A+\delta_A)$ est inversible et si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x}+\boldsymbol{\delta_x}$ est solution de* (1.64)*, alors*

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\mathrm{cond}(A)}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|}+\frac{\|\delta_A\|}{\|A\|}\right). \tag{1.70}$$

DÉMONSTRATION – On peut écrire $A+\delta_A = A(\mathrm{Id}+B)$ avec $B = A^{-1}\delta_A$. Or le rayon spectral de B, $\rho(B)$, vérifie $\rho(B) \leq \|B\| \leq \|\delta_A\|\,\|A^{-1}\| < 1$, et donc (voir le théorème 1.38 page 65 et l'exercice 44 page 71) $(\mathrm{Id}+B)$ est inversible et $(\mathrm{Id}+B)^{-1} = \displaystyle\sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(\mathrm{Id}+B)^{-1}\| \leq \displaystyle\sum_{n=0}^{\infty}\|B\|^n = \frac{1}{1-\|B\|} \leq \frac{1}{1-\|A^{-1}\|\ \|\delta_A\|}$. On en déduit que $A+\delta_A$ est inversible, car $A+\delta_A = A(Id+B)$ et comme A est inversible, $(A+\delta_A)^{-1} = (\mathrm{Id}+B)^{-1}A^{-1}$.

Comme A et $A+\delta_A$ sont inversibles, il existe un unique $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = b$ et il existe un unique $\boldsymbol{\delta_x} \in \mathbb{R}^n$ tel que $(A+\delta_A)(\boldsymbol{x}+\boldsymbol{\delta_x}) = \boldsymbol{b}+\boldsymbol{\delta_b}$. Comme $A\boldsymbol{x} = \boldsymbol{b}$, on a $(A+\delta_A)\boldsymbol{\delta_x} + \delta_A\boldsymbol{x} = \boldsymbol{\delta_b}$ et donc $\boldsymbol{\delta_x} = (A+\delta_A)^{-1}\boldsymbol{\delta_b} - \delta_A\boldsymbol{x}$). Or $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$, on en déduit :

$$\begin{aligned}\|(A+\delta_A)^{-1}\| &\leq \|(\mathrm{Id}+B)^{-1}\|\ \ \|A^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\ \|\delta_A\|}.\end{aligned}$$

On peut donc écrire la majoration suivante :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|x\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\boldsymbol{\delta_b}\|}{\|A\|\ \|\boldsymbol{x}\|}+\frac{\|\delta_A\|}{\|A\|}\right).$$

En utilisant le fait que $\boldsymbol{b} = A\boldsymbol{x}$ et que par suite $\|\boldsymbol{b}\| \leq \|A\|\ \|\boldsymbol{x}\|$, on obtient :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta_A\|}{\|A\|}\right),$$

ce qui termine la démonstration. ∎

1.4.4 Discrétisation d'équations différentielles, conditionnement "efficace"

On suppose encore ici que $\delta_A = 0$. On suppose que la matrice A du système linéaire à résoudre provient de la discrétisation par différences finies du problème de la chaleur unidimensionnel (1.5a). On peut alors montrer (voir exercice 55 page 74) que le conditionnement de A est d'ordre n^2, où n est le nombre de points de discrétisation. Pour $n = 10$, on a donc $\mathrm{cond}(A) \simeq 100$ et l'estimation (1.66) donne :

$$\frac{\|\delta_x\|}{\|x\|} \leq 100\ \frac{\|\delta_b\|}{\|b\|}.$$

Une erreur de 1% sur $\boldsymbol{b}$ peut donc entraîner une erreur de 100% sur x. Autant dire que dans ce cas, il est inutile de rechercher la solution de l'équation discrétisée... Heureusement, on peut montrer que l'estimation (1.66) n'est pas significative pour l'étude de la propagation des erreurs lors de la résolution des systèmes linéraires provenant de la discrétisation d'une équation différentielle ou d'une équation aux dérivées partielles[7]. Pour illustrer notre propos, reprenons l'étude du système linéaire obtenu à partir de la discrétisation de l'équation de la chaleur (1.5a) qu'on écrit : $A\boldsymbol{u} = \boldsymbol{b}$ avec $\boldsymbol{b} = (b_1, \ldots, b_n)$ et A la matrice carrée d'ordre n de coefficients $(a_{i,j})_{i,j=1,n}$ définis par (1.10). On rappelle que A est symétrique définie positive (voir exercice 14 page 20), et que

$$\max_{i=1\ldots n}\{|u_i - u(x_i)|\} \leq \frac{h^2}{96}\|u^{(4)}\|_\infty.$$

7. On appelle équation aux dérivées partielles une équation qui fait intervenir les dérivées partielles de la fonction inconnue, par exemple $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$, où u est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$

En effet, si on note $\overline{u}$ le vecteur de $\mathbb{R}^n$ de composantes $u(x_i)$, $i = 1, \ldots, n$, et R le vecteur de $\mathbb{R}^n$ de composantes R_i, $i = 1, \ldots, n$, on a par définition de R (formule (1.7)) $A(u - \overline{u}) = R$, et donc $\|u - \overline{u}\|_\infty \leq \|A^{-1}\|_\infty \|R\|_\infty$. Or on peut montrer (voir exercice 55 page 74) que $\mathrm{cond}(A) \simeq n^2$. Donc si on augmente le nombre de points, le conditionnement de A augmente aussi. Par exemple si $n = 10^4$, alors $\|\delta_x\|/\|x\| = 10^8 \|\delta_b\|/\|b\|$. Or sur un ordinateur en simple précision, on a $\|\delta_b\|/\|b\| \geq 10^{-7}$, donc l'estimation (1.66) donne une estimation de l'erreur relative $\|\delta_x\|/\|x\|$ de 1000%, ce qui laisse à désirer pour un calcul qu'on espère précis.
En fait, l'estimation (1.66) ne sert à rien pour ce genre de problème, il faut faire une analyse un peu plus poussée, comme c'est fait dans l'exercice 57 page 75. On se rend compte alors que pour f donnée il existe $C \in \mathbb{R}_+$ ne dépendant que de f (mais pas de n) tel que

$$\frac{\|\boldsymbol{\delta_u}\|}{\|\boldsymbol{u}\|} \leq C \frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \text{ avec } \boldsymbol{b} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{1.71}$$

L'estimation (1.71) est évidemment bien meilleure que l'estimation (1.66) puisqu'elle montre que l'erreur relative commise sur u est du même ordre que celle commise sur $\boldsymbol{b}$. En particulier, elle n'augmente pas avec le nombre de points de discrétisation. En conclusion, l'estimation (1.66) est peut-être optimale dans le cas d'une matrice quelconque, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.66)) mais elle n'est pas toujours significative pour l'étude des systèmes linéaires issus de la discrétisation des équations aux dérivées partielles.

1.4.5 Exercices (normes et conditionnement)

Exercice 38 (Normes de l'Identité)**.**

Soit Id la matrice "Identité" de $\mathcal{M}_n(\mathbb{R})$. Montrer que pour toute norme induite on a $\|\mathrm{Id}\| = 1$ et que pour toute norme matricielle on a $\|\mathrm{Id}\| \geq 1$.

Exercice 39 (Normes induites particulières)**.** *Suggestions en page 76, corrigé détaillé en page 77.*

Soit $A = (a_{i,j})_{i,j \in \{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$.

1. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Montrer que
$$\|A\|_\infty = \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} |a_{i,j}|.$$
2. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Montrer que
$$\|A\|_1 = \max_{j \in \{1,\ldots,n\}} \sum_{i=1}^{n} |a_{i,j}|.$$

Exercice 40 (Norme non induite)**.**

Pour $A = (a_{i,j})_{i,j \in \{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$, on pose $\|A\|_s = (\sum_{i,j=1}^n a_{i,j}^2)^{\frac{1}{2}}$.

1. Montrer que $\|\cdot\|_s$ est une norme matricielle mais n'est pas une norme induite (pour $n > 1$).
2. Montrer que $\|A\|_s^2 = tr(A^t A)$. En déduire que $\|A\|_2 \leq \|A\|_s \leq \sqrt{n}\|A\|_2$ et que $\|Ax\|_2 \leq \|A\|_s \|x\|_2$, pour tout $A \in \mathcal{M}_n(\mathbb{R})$ et tout $x \in \mathbb{R}^n$.
3. Chercher un exemple de norme non matricielle.

Exercice 41 (Valeurs propres d'un produit de matrices)**.**

Soient p et n des entiers naturels non nuls tels que $n \leq p$, et soient $A \in \mathcal{M}_{n,p}(\mathbb{R})$ et $B \in \mathcal{M}_{p,n}(\mathbb{R})$. (On rappelle que $\mathcal{M}_{n,p}(\mathbb{R})$ désigne l'ensemble des matrices à n lignes et p colonnes.)

1. Montrer que λ est valeur propre non nulle de AB si et seulement si λ est valeur propre non nulle de BA.

2. Montrer que si 0 est valeur propre de AB alors 0 est valeur propre de BA. *(Il est conseillé de distinguer les cas $Bx \neq 0$ et $Bx = 0$, où x est un vecteur propre associé à la valeur propre nulle de AB. Pour le deuxième cas, on pourra distinguer selon que $ImA = \mathbb{R}^n$ ou non.)*
3. Montrer en donnant un exemple que 0 peut être une valeur propre de BA sans être valeur propre de AB. *(Prendre par exemple $n = 1$, $p = 2$.)*
4. On suppose maintenant que $n = p$, déduire des questions 1 et 2 que l'ensemble des valeurs propres de AB est égal à l'ensemble des valeurs propres de la matrice BA.

Exercice 42 (Matrice diagonalisable et rayon spectral)**.** *Corrigé en page 78.*

Soit $A \in \mathcal{M}_n(\mathbb{R})$. Montrer que si A est diagonalisable, il existe une norme induite sur $\mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) = \|A\|$. Montrer par un contre exemple que ceci peut être faux si A n'est pas diagonalisable.

Exercice 43 (Le rayon spectral est-il une norme ou une semi-norme ?)**.** On définit les matrices carrées d'ordre 2 suivantes :

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},\ B = \begin{bmatrix} -1 & 0 \\ -1 & -1 \end{bmatrix},\ C = A + B.$$

Calculer le rayon spectral de chacune des matrices A, B et C et en déduire que le rayon spectral ne peut être ni une norme, ni même une semi-norme sur l'espace vectoriel des matrices.

Exercice 44 (Série de Neumann)**.** *Suggestions en page 76, corrigé détaillé en page 78.*

Soient $A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $\rho(A) < 1$, les matrices $Id - A$ et $Id + A$ sont inversibles.

2. Montrer que la série de terme général A^k converge (vers $(Id - A)^{-1}$) si et seulement si $\rho(A) < 1$.

3. Montrer que si $\rho(A) < 1$, et si $\|\cdot\|$ une norme matricielle telle que $\|A\| < 1$, alors $\|(Id - A)^{-1}\| \leq \frac{1}{1-\|A\|}$ et $\|(Id + A)^{-1}\| \leq \frac{1}{1-\|A\|}$.

Exercice 45 (Normes induites)**.**

Soit $\|.\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ par une norme quelconque sur $\mathbb{R}^n$, et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) < 1$ (on rappelle qu'on note $\rho(A)$ le rayon spectral de la matrice A). Pour $x \in \mathbb{R}^n$, on définit $\|x\|_*$ par :

$$\|x\|_* = \sum_{j=0}^{\infty} \|A^j x\|.$$

1. Montrer que l'application définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par $x \mapsto \|x\|_*$ est une norme.
2. Soit $x \in \mathbb{R}^n$ tel que $\|x\|_* = 1$. Calculer $\|Ax\|_*$ en fonction de $\|x\|$, et en déduire que $\|A\|_* < 1$.
3. On ne suppose plus que $\rho(A) < 1$. Soit $\varepsilon > 0$ donné. Construire à partir de la norme $\|.\|$ une norme induite $\|.\|_{**}$ telle que $\|A\|_{**} \leq \rho(A) + \varepsilon$.

Exercice 46 (Calcul de conditionnement)**.** *Corrigé détaillé en page 79.*

Calculer le conditionnement pour la norme 2 de la matrice $\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$.

Exercice 47 (Propriétés générales du conditionnement)**.** *Corrigé détaillé en page 79.*

On suppose que $\mathbb{R}^n$ est muni de la norme euclidienne usuelle $\|\cdot\| = \|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite (notée aussi $\|\cdot\|_2$. On note alors $\mathrm{cond}_2(A)$ le conditionnement d'une matrice A inversible.

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Montrer que $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Montrer que $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
3. Soit $A, B \in \mathcal{M}_n(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que

$$\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)\}.$$

Exercice 48 (Conditionnement de la matrice transposée)**.** On suppose que $A \in \mathcal{M}_n(\mathbb{R})$ est inversible.

1. Montrer que si $B \in \mathcal{M}_n(\mathbb{R})$, on a pour tout $\lambda \in \mathbb{C}$, $\det(AB - \lambda Id) = \det(BA - \lambda Id)$.
2. En déduire que les rayons spectraux des deux matrices AB et BA sont identiques.
3. Montrer que $\|A^t\|_2 = \|A\|_2$.
4. En déduire que $\mathrm{cond}_2(A) = \mathrm{cond}_2(A^t)$.
5. A-t-on $\|A^t\|_1 = \|A\|_1$?
6. Montrer que dans le cas $n = 2$, on a toujours $\mathrm{cond}_1(A) = \mathrm{cond}_1(A^t), \forall A \in M_2(\mathbb{R})$.
7. Calculer $\mathrm{cond}_1(A)$ pour $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ et conclure.

Exercice 49 (Conditionnement et normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$)**.**

1. On considère la matrice $B = (B_{ij})$ de $\mathcal{M}_n(\mathbb{R})$ définie par $B_{ii} = 1$, $B_{ij} = -1\ i < j$, $B_{ij} = 0$ sinon.
   (a) Calculer B^{-1}.
   (b) En déduire $\mathrm{cond}_1(B)$ et $\mathrm{cond}_\infty(B)$.
2. Soit A une matrice carrée de taille $n \times n$. L'objectif de cette question est de montrer que

$$\frac{1}{n^2}\mathrm{cond}_\infty(A) \leq \mathrm{cond}_1(A) \leq n^2 \mathrm{cond}_\infty(A).$$

   (a) Montrer que pour tout $x \in \mathbb{R}^n$,

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty.$$

   (b) En déduire que pour toute matrice carrée de taille $n \times n$

$$\frac{1}{n}\|A\|_\infty \leq \|A\|_1 \leq n\|A\|_\infty.$$

   (c) Conclure.

Exercice 50 (Un système par blocs)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre N inversible, $b, c, f \in \mathbb{R}^n$. Soient α et $\gamma \in \mathbb{R}$. On cherche à résoudre le système suivant (avec $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$) :

$$\begin{aligned} &Ax + b\lambda = f, \\ &c \cdot x + \alpha\lambda = \gamma. \end{aligned} \tag{1.72}$$

1. Ecrire le système (1.72) sous la forme : $My = g$, où M est une matrice carrée d'ordre $n+1$, $y \in \mathbb{R}^{n+1}$, $g \in \mathbb{R}^{n+1}$. Donner l'expression de M, y et g.
2. Donner une relation entre A, b, c et α, qui soit une condition nécessaire et suffisante pour que le système (1.72) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.72) :
   (a) Soient z solution de $Az = b$, et h solution de $Ah = f$.

(b) $x = h - \dfrac{\gamma - c\cdot h}{\alpha - c\cdot z} z, \qquad \lambda = \dfrac{\gamma - c\cdot h}{\alpha - c\cdot z}.$

Montrer que $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.72).

4. On suppose dans cette question que A est une matrice bande, dont la largeur de bande est p.
   (a) Calculer le coût de la méthode de résolution proposée ci-dessus en utilisant la méthode LU pour la résolution des systèmes linéaires.
   (b) Calculer le coût de la résolution du système $My = g$ par la méthode LU (en profitant ici encore de la structure creuse de la matrice A).
   (c) Comparer et conclure.
   Dans les deux cas, le terme d'ordre supérieur est $2nq^2$, et les coûts sont donc comparables.

Exercice 51 (Majoration du conditionnement)**.**

Soit $\|\,.\,\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\det(A) \neq 0$.

1. Montrer que si $\|A - B\| < \frac{1}{\|A^{-1}\|}$, alors B est inversible.
2. Montrer que $\text{cond}\,(A) \geq \sup_{\substack{B\in\mathcal{M}_n(\mathbb{R})\\ \det B=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 52 (Minoration du conditionnement)**.** *Corrigé détaillé en page 80.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible, $\text{cond}(A) = \|A\|\|A^{-1}\|$ le conditionnement de A, et soit $\delta_A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $A + \delta_A$ est singulière, alors
$$\text{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.73}$$
2. On suppose dans cette question que la norme $\|\cdot\|$ est la norme induite par la norme euclidienne sur $\mathbb{R}^n$. Montrer que la minoration (1.73) est optimale, c'est-à-dire qu'il existe $\delta_A \in \mathcal{M}_n(\mathbb{R})$ telle que $A + \delta_A$ soit singulière et telle que l'égalité soit vérifiée dans (1.73).
   [*On pourra chercher δ_A de la forme*
   $$\delta_A = -\frac{y\,x^t}{x^t\,x},$$
   avec $y \in \mathbb{R}^n$ convenablement choisi et $x = A^{-1}y$.]
3. On suppose ici que la norme $\|\cdot\|$ est la norme induite par la norme infinie sur $\mathbb{R}^n$. Soit $\alpha \in]0,1[$. Utiliser l'inégalité (1.73) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\text{cond}(A)$ pour la matrice
$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \alpha & -\alpha \\ 1 & \alpha & \alpha \end{pmatrix}.$$

Exercice 53 (Conditionnement du carré)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice telle que $\det A \neq 0$.

1. Quelle relation existe-t-il en général entre $\text{cond}\,(A^2)$ et $(\text{cond}\,A)^2$?
2. On suppose que A symétrique. Montrer que $\text{cond}_2\,(A^2) = (\text{cond}_2\,A)^2$.
3. On suppose que $\text{cond}_2\,(A^2) = (\text{cond}_2\,A)^2$. Peut-on conclure que A est symétrique ? (justifier la réponse.)

Exercice 54 (Calcul de l'inverse d'une matrice et conditionnement)**.** *Corrigé détaillé en page 80.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible. On cherche ici des moyens d'évaluer la précision de calcul de l'inverse de A.

1. On suppose qu'on a calculé B, approximation (en raison par exemple d'erreurs d'arrondi) de la matrice A^{-1}. On pose :
$$\left\{ \begin{array}{rclcrcl} e_1 &=& \dfrac{\|B-A^{-1}\|}{\|A^{-1}\|}, & & e_2 &=& \dfrac{\|B^{-1}-A\|}{\|A\|} \\ \\ e_3 &=& \|AB-\mathrm{Id}\|, & & e_4 &=& \|BA-\mathrm{Id}\| \end{array} \right. \tag{1.74}$$
   (a) Expliquer en quoi les quantités e_1, e_2, e_3 et e_4 mesurent la qualité de l'approximation de A^{-1}.
   (b) On suppose ici que $B = A^{-1} + E$, où $\|E\| \leq \varepsilon\|A^{-1}\|$, et que
$$\varepsilon \mathrm{cond}(A) < 1.$$
   Montrer que dans ce cas,
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon\mathrm{cond}(A)}{1-\varepsilon\mathrm{cond}(A)},\ e_3 \leq \varepsilon\mathrm{cond}(A) \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$
   (c) On suppose maintenant que $AB - Id = E'$ avec $\|E'\| \leq \varepsilon < 1$. Montrer que dans ce cas :
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon}{1-\varepsilon},\ e_3 \leq \varepsilon \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$
2. On suppose maintenant que la matrice A n'est connue qu'à une certaine matrice d'erreurs près, qu'on note δ_A.
   (a) Montrer que la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.
   (b) Montrer que si la matrice $A + \delta_A$ est inversible, alors
$$\frac{\|(A+\delta_A)^{-1} - A^{-1}\|}{\|(A+\delta_A)^{-1}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|}.$$

Exercice 55 (Conditionnement du Laplacien discret 1D)**.** *Suggestions en page 76, corrigé détaillé en page 82.*
Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 13, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.
Calculer les valeurs propres et les vecteurs propres de A. [*On pourra commencer par chercher* $\lambda \in \mathbb{R}$ *et* $\varphi \in C^2(\mathbb{R},\mathbb{R})$ *(*φ *non identiquement nulle) t.q.* $-\varphi''(x) = \lambda\varphi(x)$ *pour tout* $x \in]0,1[$ *et* $\varphi(0) = \varphi(1) = 0$].
Calculer $\mathrm{cond}_2(A)$ et montrer que $h^2\mathrm{cond}_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 56 (Conditionnement, réaction diffusion 1d.)**.**

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème (1.25) étudié à l'exercice 16, qu'on rappelle :
$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.75}$$
Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.25) aux points $\left(\frac{j}{n+1}\right)_{j=1,\ldots,n}$. On rappelle que la discrétisation par différences finies de ce problème consiste à chercher U comme solution du système linéaire $AU = \left(f(\frac{j}{N+1})\right)_{j=1,\ldots,n}$ où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est définie par $A = (N+1)^2B + \mathrm{Id}$, Id désigne la matrice identité et
$$B = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

1. (Valeurs propres de la matrice B.)
   On rappelle que le problème aux valeurs propres
   $$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.76}$$
   admet la famille $(\lambda_k, u_k)_{k\in\mathbb{N}^*}$, $\lambda_k = (k\pi)^2$ et $u_k(x) = \sin(k\pi x)$ comme solution. Montrer que les vecteurs $U_k = \left(u_k(\frac{j}{n+1})\right)_{j=1,\dots,n}$ sont des vecteurs propres de la matrice B. En déduire toutes les valeurs propres de la matrice B.
2. En déduire les valeurs propres de la matrice A.
3. En déduire le conditionnement pour la norme euclidienne de la matrice A.

Exercice 57 (Conditionnement "efficace")**.** *Suggestions en page 77.*

Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 13, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.

Pour $u \in \mathbb{R}^n$, on note $u_1, \dots, u_n$ les composantes de u. Pour $u \in \mathbb{R}^n$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1, \dots, n\}$. Pour $u, v \in \mathbb{R}^n$, on note $u \cdot v = \sum_{i=1}^n u_i v_i$.

On munit $\mathbb{R}^n$ de la norme suivante : pour $u \in \mathbb{R}^n$, $\|u\| = \max\{|u_i|,\ i \in \{1, \dots, n\}\}$. On munit alors $\mathcal{M}_n(\mathbb{R})$ de la norme induite, également notée $\|\cdot\|$, c'est-à-dire $\|B\| = \max\{\|Bu\|,\ u \in \mathbb{R}^n \text{ t.q. } \|u\| = 1\}$, pour tout $B \in \mathcal{M}_n(\mathbb{R})$.

Partie I Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de A^{-1}) Soient $b \in \mathbb{R}^n$ et $u \in \mathbb{R}^n$ t.q. $Au = b$. Remarquer que $Au = b$ peut s'écrire :
   $$\left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i,\ \forall i \in \{1, \dots, n\}, \\ u_0 = u_{n+1} = 0. \end{array} \right. \tag{1.77}$$
   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considérer $p \in \{0, \dots, n+1\}$ t.q. $u_p = \min\{u_j,\ j \in \{0, \dots, n+1\}$.]

   En déduire que A est inversible.
2. (Préliminaire) On considère la fonction $\varphi \in C([0,1], \mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\boldsymbol{\phi} = (\phi_1, \dots \phi_n) \in \mathbb{R}^n$ par $\phi_i = \varphi(ih)$ pour tout $i \in \{1, \dots, n\}$. Montrer que $(A\boldsymbol{\phi})_i = 1$ pour tout $i \in \{1, \dots, n\}$.
3. (Calcul de $\|A^{-1}\|$) Soient $\boldsymbol{b} \in \mathbb{R}^n$ et $\boldsymbol{u} \in \mathbb{R}^n$ t.q. $A\boldsymbol{u} = \boldsymbol{b}$. Montrer que $\|\boldsymbol{u}\| \leq (1/8)\|\boldsymbol{b}\|$ [Calculer $A(\boldsymbol{u} \pm \|\boldsymbol{b}\|\boldsymbol{\phi})$ avec $\boldsymbol{\phi}$ défini à la question 2 et utiliser la question 1]. En déduire que $\|A^{-1}\| \leq 1/8$ puis montrer que $\|A^{-1}\| = 1/8$.
4. (Calcul de $\|A\|$) Montrer que $\|A\| = \frac{4}{h^2}$.
5. (Conditionnement pour la norme $\|\cdot\|$). Calculer $\|A^{-1}\|\|A\|$. Soient $\boldsymbol{b}$, $\delta_b \in \mathbb{R}^n$ et soient u, $\delta_u \in \mathbb{R}^n$ t.q. $Au = b$ et $A(u + \delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \dfrac{\|\delta_b\|}{\|b\|}$.

   Montrer qu'un choix convenable de $\boldsymbol{b}$ et δ_b donne l'égalité dans l'inégalité précédente.

Partie II Borne réaliste sur l'erreur relative : Conditionnement "efficace"

On se donne maintenant $f \in C([0,1], \mathbb{R})$ et on suppose (pour simplifier...) que $f(x) > 0$ pour tout $x \in]0,1[$. On prend alors, dans cette partie, $b_i = f(ih)$ pour tout $i \in \{1, \dots, n\}$. On considère aussi le vecteur ϕ défini à la question 2 de la partie I.

1. Montrer que
$$h\sum_{i=1}^{n} b_i\phi_i \to \int_0^1 f(x)\varphi(x)dx \text{ quand } n \to \infty$$
et que
$$\sum_{i=1}^{n} b_i\phi_i > 0 \text{ pour tout } n \in \mathbb{N}^*.$$
En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h\sum_{i=1}^{n} b_i\phi_i \geq \alpha$ pour tout $n \in \mathbb{N}^*$.
2. Soit $u \in \mathbb{R}^n$ t.q. $Au = b$. Montrer que $n\|u\| \geq \sum_{i=1}^{n} u_i = u \cdot A\phi \geq \frac{\alpha}{h}$ (avec α donné à la question 1).
Soit $\delta_b \in \mathbb{R}^n$ et $\delta_u \in \mathbb{R}^n$ t.q. $A(u + \delta_u) = b + \delta_b$. Montrer que $\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}\frac{\|\delta_b\|}{\|b\|}$.
3. Comparer $\|A^{-1}\|\|A\|$ (question I.5) et $\frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$ (question II.2) quand n est "grand" (ou quand $n \to \infty$).

1.4.6 Suggestions pour les exercices

Exercice 39 page 70 (Normes induites particulières)

1. Pour montrer l'égalité, prendre x tel que $x_j = \text{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\ldots,n} |a_{i_0,j}| \geq \sum_{j=1,\ldots,n} |a_{i,j}|$, $\forall i = 1, \ldots, n$, et $\text{sign}(s)$ désigne le signe de s.

2. Pour montrer l'égalité, prendre x tel que $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\ldots,n} |a_{i,j_0}| = \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$.

Exercice 44 page 71 (Série de Neumann)

1. Montrer que si $\rho(A) < 1$, alors 0 n'est pas valeur propre de $Id + A$ et $\text{Id} - A$.

2. Utiliser le corollaire 1.34.

Exercice 47 page 71 (Propriétés générales du conditionnement)

3. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_n$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_n$ les valeurs propres de A et B (qui sont s.d.p.). Montrer d'abord que :
$$cond_2(A + B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$
Montrer ensuite que
$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a, b, c, d) \in (\mathbb{R}_+^*)^4.$$
et conclure

Exercice 55 page 74(Conditionnement du Laplacien discret 1D)

2. Chercher les vecteurs propres $\Phi \in \mathbb{R}^n$ de A sous la forme $\Phi_j = \varphi(x_j)$, $j = 1, \ldots, n$ où φ est introduite dans les indications de l'énoncé. Montrer que les valeurs propres associées à ces vecteurs propres sont de la forme :
$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}).$$

Exercice 57 page 75 (Conditionnement efficace)

Partie 1

1. Pour montrer que A est inversible, utiliser le théorème du rang.

2. Utiliser le fait que φ est un polynôme de degré 2.

3. Pour montrer que $\|A^{-1}\| = \dfrac{1}{8}$, remarquer que le maximum de φ est atteint en $x = .5$, qui correspond à un point de discrétisation car n est impair.

Partie 2 Conditionnement efficace

1. Utiliser la convergence uniforme des fonctions constantes par morceaux φ_h et f_h définies par

$$\begin{array}{rcl} \varphi_h(x) & = & \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f_h(x) & = & \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

2. Utiliser le fait que $A\phi = (1 \ldots 1)^t$.

1.4.7 Corrigés

Exercice 39 page 70 (Normes induites particulières)

1. Par définition, $\|A\|_\infty = \sup_{\substack{\boldsymbol{x} \in \mathbb{R}^n \\ \|\boldsymbol{x}\|_\infty = 1}} \|A\boldsymbol{x}\|_\infty$, et

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1,\ldots,n} a_{i,j} x_j| \leq \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}||x_j|.$$

Or $\|\boldsymbol{x}\|_\infty = 1$ donc $|x_j| \leq 1$ et

$$\|A\boldsymbol{x}\|_\infty \leq \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|.$$

Montrons maintenant que la valeur $\alpha = \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|$ est atteinte, c'est-à-dire qu'il existe $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_\infty = 1$, tel que $\|A\boldsymbol{x}\|_\infty = \alpha$. Pour $s \in \mathbb{R}$, on note $\text{sign}(s)$ le signe de s, c'est–à–dire

$$\text{sign}(s) = \begin{cases} s/|s| \text{ si } s \neq 0, \\ 0 \text{ si } s = 0. \end{cases}$$

Choisissons $\boldsymbol{x} \in \mathbb{R}^n$ défini par $x_j = \text{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\ldots,n} |a_{i_0,j}| \geq \sum_{j=1,\ldots,n} |a_{i,j}|$, $\forall i = 1, \ldots, n$. On a bien $\|\boldsymbol{x}\|_\infty = 1$, et

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1}^{n} a_{i,j} \text{sign}(a_{i_0,j})|.$$

Or, par choix de $\boldsymbol{x}$, on a

$$\sum_{j=1,\ldots,n} |a_{i_0,j}| = \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|.$$

On en déduit que pour ce choix de $\boldsymbol{x}$, on a bien $\|A\boldsymbol{x}\| = \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|$.

2. Par définition, $\|A\|_1 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \|\boldsymbol{x}\|_1=1}} \|A\boldsymbol{x}\|_1$, et

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{i,j}x_j| \leq \sum_{j=1}^{n} |x_j| \left(\sum_{i=1}^{n} |a_{i,j}|\right) \leq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{i,j}| \sum_{j=1,\ldots,n} |x_j|.$$

Et comme $\sum_{j=1}^{n} |x_j| = 1$, on a bien que $\|A\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$.
Montrons maintenant qu'il existe $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_1 = 1$, tel que $\|A\boldsymbol{x}\|_1 = \sum_{i=1,\ldots,n} |a_{i,j}|$. Il suffit de considérer pour cela le vecteur $\boldsymbol{x} \in \mathbb{R}^n$ défini par $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\ldots,n} |a_{i,j_0}| = \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$. On vérifie alors facilement qu'on a bien $\|A\boldsymbol{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$.

Exercice 42 page 71 (Rayon spectral)

Il suffit de prendre comme norme la norme définie par : $\|x\|^2 = \sum_{i=1}^{n} \alpha_i^2$ où les $(\alpha_i)_{i=1,n}$ sont les composantes de x dans la base des vecteurs propres associés à A. Pour montrer que ceci est faux dans le cas où A n'est pas diagonalisable, il suffit de prendre $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, on a alors $\rho(A) = 0$, et comme A est non nulle, $\|A\| \neq 0$.

Exercice 44 page 71 (Série de Neumann)

1. Si $\rho(A) < 1$, les valeurs propres de A sont toutes différentes de 1 et -1. Donc 0 n'est pas valeur propre des matrices $Id - A$ et $Id + A$, qui sont donc inversibles.

2. Supposons que $\rho(A) < 1$. Remarquons que

$$(\sum_{k=0}^{n} A^k)(Id - A) = \mathrm{Id} - A^{n+1}. \tag{1.78}$$

Comme $\rho(A) < 1$, d'après le corollaire 1.34, on a $A^k \to 0$ lorsque $k \to 0$. De plus, $Id - A$ est inversible. En passant à la limite dans (1.78) et on a donc

$$(Id - A)^{-1} = \sum_{k=0}^{+\infty} A^k. \tag{1.79}$$

Réciproquement, si $\rho(A) \geq 1$, la série ne peut pas converger en raison du corollaire 1.34.
3. On a démontré plus haut que si $\rho(A) < 1$, la série de terme général A^k est absolument convergente et qu'elle vérifie (1.79). On en déduit que si $\|A\| < 1$,

$$\|(Id - A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| = \frac{1}{1 - \|A\|}.$$

On a de même

$$(Id + A)^{-1} = \sum_{k=0}^{+\infty} (-1)^k A^k,$$

d'où on déduit de manière similaire que

$$\|(Id + A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| = \frac{1}{1 - \|A\|}.$$

Exercice 46 page 71 (Calcul de conditionnement)

On a $A^tA = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$. Les valeurs propres de cette matrice sont $3 \pm \sqrt{5}$ et donc $\mathrm{cond}_2(A) = \sqrt{\frac{3+\sqrt{5}}{3-\sqrt{5}}} \neq 2$.

Exercice 47 page 71 (Propriétés générales du conditionnement)

1. Si $cond_2(A) = 1$, alors $\sqrt{\frac{\sigma_n}{\sigma_1}} = 1$ et donc toutes les valeurs propres de A^tA sont égales. Comme A^tA est symétrique définie positive (car A est inversible), il existe une base orthonormée $(\boldsymbol{f}_1 \dots \boldsymbol{f}_n)$ telle que $A^tA\boldsymbol{f}_i = \sigma \boldsymbol{f}_i, \ \forall i$ et $\sigma > 0$ (car A^tA est s.d.p.). On a donc $A^tA = \sigma \mathrm{Id}\ A^t = \alpha^2 A^{-1}$ avec $\alpha = \sqrt{\sigma}$. En posant $Q = \frac{1}{\alpha}A$, on a donc $Q^t = \frac{1}{\alpha}A^t = \alpha A^{-1} = Q^{-1}$.
Réciproquement, si $A = \alpha Q$, alors $A^tA = \alpha^2 Id$, $\frac{\sigma_n}{\sigma_1} = 1$, et donc $cond_2(A) = 1$.

2. $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. On a donc $cond_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}$ où $\sigma_1 \leq \ldots \leq \sigma_n$ sont les valeurs propres de A^tA. Or $A^tA = (QR)^t(QR) = R^tQ^{-1}QR = R^tR$. Donc $cond_2(A) = cond_2(R)$.

3. Soient $0 < \lambda_1 \leq \lambda_2 \dots \leq \lambda_n$ et $0 < \mu_1 \leq \mu_2 \dots \leq \mu_n$ les valeurs propres de A et B (qui sont s.d.p.). Alors $cond_2(A+B) = \frac{\nu_n}{\nu_1}$, où $0 < \nu_1 \leq \ldots \leq \nu_n$ sont les valeurs propres de $A+B$.

a) On va d'abord montrer que

$$cond_2(A+B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$

On sait que si A est s.d.p., alors

$$cond_2(A) = \frac{\lambda_n}{\lambda_1}.$$

Or, si A est s.d.p., alors $\sup_{\|x\|_2=1} Ax \cdot x = \lambda_n$; il suffit pour s'en rendre compte de décomposer x sur la base $(\boldsymbol{f}_i)_{i=1\dots n}$. Soit $x = \sum_{i=1}^n \alpha_i \boldsymbol{f}_i$, alors :

$$Ax \cdot x = \sum_{i=1}^n \alpha_i^2 \lambda_i \leq \lambda_n \sum_{i=1}^n \alpha_i^2 = \lambda_n.$$

Et $A\boldsymbol{f}_n \cdot \boldsymbol{f}_n = \lambda_n$.
De même, $Ax \cdot x \geq \lambda_1 \sum_{i=1}^n \alpha_i^2 = \lambda_1$ et $Ax \cdot x = \lambda_1$ si $x = f_1$. Donc $\inf_{\|x\|=1} Ax \cdot x = \lambda_1$.
On en déduit que si A est s.d.p.,

$$\mathrm{cond}_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|=1} Ax \cdot x}.$$

Donc $cond_2(A+B) = \dfrac{\sup_{\|x\|=1}(A+B)x \cdot x}{\inf_{\|x\|=1}(A+B)x \cdot x}$. Or

$$\sup_{\|x\|=1}(Ax \cdot x + Bx \cdot x) \leq \sup_{\|x\|=1} Ax \cdot x + \sup_{\|x\|=1} Bx \cdot x = \lambda_n + \mu_n,$$
$$\inf_{\|x\|=1}(Ax \cdot x + Bx \cdot x) \geq \inf_{\|x\|=1} Ax \cdot x + \inf_{\|x\|=1} Bx \cdot x = \lambda_1 + \mu_1,$$

et donc

$$cond_2(A+B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$

b) On va montrer que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

Supposons que $\frac{a+b}{c+d} \geq \frac{a}{c}$ alors $(a+b)c \geq (c+d)a$ c'est–à–dire $bc \geq da$ donc $bc+bd \geq da+db$ soit $b(c+d) \geq d(a+b)$; donc $\frac{a+b}{c+d} \leq \frac{b}{d}$. On en déduit que $cond_2(A+B) \leq \max(cond_2(A), cond_2(B))$.

Exercice 52 page 73 (Minoration du conditionnement)

1. Comme A est inversible, $A+\delta_A = A(Id + A^{-1}\delta_A)$, et donc si $A+\delta_A$ est singulière, alors $Id + A^{-1}\delta_A$ est singulière. Or on a vu en cours que toute matrice de la forme $Id+B$ est inversible si $\rho(B) < 1$. On en déduit que $\rho(A^{-1}\delta_A) \geq 1$, et comme

$$\rho(A^{-1}\delta_A) \leq \|A^{-1}\delta_A\| \leq \|A^{-1}\|\|\delta_A\|,$$

on obtient

$$\|A^{-1}\|\|\delta_A\| \geq 1, \text{ soit encore } \mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}.$$

2. Soit $y \in \mathbb{R}^n$ tel que $\|y\| = 1$ et $\|A^{-1}y\| = \|A^{-1}\|$. Soit $x = A^{-1}y$, et $\delta_A = \frac{-y\ x^t}{x^t\ x}$,on a donc

$$(A+\delta_A)x = Ax - \frac{-y\ x^t}{x^t\ x}x = y - \frac{-y\ x^t x}{x^t\ x} = 0.$$

La matrice $A+\delta_A$ est donc singulière. De plus,

$$\|\delta_A\| = \frac{1}{\|x\|^2}\|y\ y^t\ A^{-t}\|.$$

Or par définition de x et y, on a $\|x\|^2 = \|A^{-1}\|^2$. D'autre part, comme il s'agit ici de la norme L^2, on a $\|A^{-t}\| = \|A^{-1}\|$. On en déduit que

$$\|\delta_A\| = \frac{1}{\|A^{-1}\|^2}\|y\|^2\|A^{-1}\| = \frac{1}{\|A^{-1}\|}.$$

On a donc dans ce cas égalité dans (1.73).

3. Remarquons tout d'abord que la matrice A est inversible. En effet, $\det A = 2\alpha^2 > 0$.
Soit $\delta_A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\alpha & \alpha \\ 0 & -\alpha & -\alpha \end{pmatrix}$. Comme $\det(A+\delta_A) = 0$, la matrice $A+\delta_A$ est singulière, et donc

$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.80}$$

Or $\|\delta_A\| = 2\alpha$ et $\|A\| = \max(3, 1+2\alpha) = 3$, car $\alpha \in]0,1[$. Donc $\mathrm{cond}(A) \geq \frac{3}{2\alpha}$.

Exercice 54 page 73 (Calcul de l'inverse d'une matrice et conditionnement)

1. (a) L'inverse de la matrice A vérifie les quatre équations suivantes :

$$\begin{cases} X - A^{-1} = 0, & X^{-1} - A = 0, \\ AX - \mathrm{Id} = 0, & XA - \mathrm{Id} = 0. \end{cases}$$

Les quantités e_1, e_2, e_3 et e_4 sont les erreurs relatives commises sur ces quatre équations lorsqu'on remplace X par B ; en ce sens, elles mesurent la qualité de l'approximation de A^{-1}.

(b) On remarque d'abord que comme la norme est matricielle, on a $\|MP\| \leq \|M\|\|P\|$ pour toutes matrices M et P de $\mathcal{M}_n(\mathbb{R})$. On va se servir de cette propriété plusieurs fois par la suite.

(α) Comme $B = A^{-1} + E$, on a

$$e_1 = \frac{\|E\|}{\|A^{-1}\|} \leq \varepsilon \frac{\|A^{-1}\|}{\|A^{-1}\|} = \varepsilon.$$

(β) Par définition,

$$e_2 = \frac{\|B^{-1} - A\|}{\|A\|} = \frac{\|(A^{-1} + E)^{-1} - A\|}{\|A\|}.$$

Or

$$\begin{aligned}(A^{-1} + E)^{-1} - A &= (A^{-1}(Id + AE))^{-1} - A \\ &= (Id + AE)^{-1}A - A \\ &= (Id + AE)^{-1}(Id - (Id + AE))A \\ &= -(Id + AE)^{-1}AEA.\end{aligned}$$

On a donc

$$e_2 \leq \|(Id + AE)^{-1}\|\|A\|\|E\|.$$

Or par hypothèse, $\|AE\| \leq \|A\|\|E\| \leq \mathrm{cond}(A)\varepsilon < 1$; on en déduit, en utilisant le théorème 1.11, que :

$$\|(Id + AE))^{-1}\| \leq \frac{1}{1 - \|AE\|}, \text{ et donc } e_2 \leq \frac{\varepsilon \mathrm{cond}(A)}{1 - \varepsilon \mathrm{cond}(A)}.$$

(γ) Par définition, $e_3 = \|AB - \mathrm{Id}\| = \|A(A^{-1} + E) - \mathrm{Id}\| = \|AE\| \leq \|A\|\|E\| \leq \|A\|\varepsilon\|A^{-1}\| = \varepsilon \mathrm{cond}(A)$.

(δ) Enfin, $e_4 = \|BA - \mathrm{Id}\| = \|(A^{-1} + E)A - \mathrm{Id}\| \leq \|EA\| \leq \|E\|\|A\| \leq \varepsilon \mathrm{cond}(A)$.

(c) (α) Comme $B = A^{-1}(Id + E')$, on a

$$e_1 = \frac{\|A^{-1}(Id + E') - A^{-1}\|}{\|A^{-1}\|} \leq \|\mathrm{Id} + E' - \mathrm{Id}\| \leq \varepsilon.$$

(β) Par définition,

$$\begin{aligned}e_2 &= \tfrac{\|(Id+E')^{-1}A - A\|}{\|A\|} \\ &= \tfrac{\|(Id+E')^{-1}(A - (Id+E')A)\|}{\|A\|} \\ &\leq \|(Id + E')^{-1}\|\|\mathrm{Id} - (Id + E')\| \leq \tfrac{\varepsilon}{1-\varepsilon}\end{aligned}$$

car $\varepsilon < 1$ (théorème 1.1).

(γ) Par définition, $e_3 = \|AB - \mathrm{Id}\| = \|AA^{-1}(Id + E') - \mathrm{Id}\| = \|E'\| \leq \varepsilon$.

(δ) Enfin, $e_4 = \|BA - \mathrm{Id}\| = \|A^{-1}(Id + E')A - \mathrm{Id}\| = \|A^{-1}(A + E'A - A)\| \leq \|A^{-1}\|\|AE'\| \leq \varepsilon \mathrm{cond}(A)$.

2. (a) On peut écrire $A + \delta_A = A(Id + A^{-1}\delta_A)$. On a vu en cours (théorème 1.11) que si $\|A^{-1}\delta_A\| < 1$, alors la matrice $Id + A^{-1}\delta_A$ est inversible. Or $\|A^{-1}\delta_A\| \leq \|A^{-1}\|\|\delta_A\|$, et donc la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.

(b) On peut écrire $\|(A + \delta_A)^{-1} - A^{-1}\| = \|(A + \delta_A)^{-1}(\mathrm{Id} - (A + \delta_A)A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\|\mathrm{Id} - \mathrm{Id} - \delta_A A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\delta_A\|\|A^{-1}\|$. On en déduit le résultat.

Exercice 55 page 74 (Conditionnement du Laplacien discret 1D)

Pour chercher les valeurs propres et vecteurs propres de A, on s'inspire des valeurs propres et vecteurs propres du problème continu, c'est–à–dire des valeurs λ et fonctions φ telles que

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & x \in]0,1[\\ \varphi(0) = \varphi(1) = 0 \end{cases} \tag{1.81}$$

(Notons que ce "truc" ne marche pas dans n'importe quel cas.)
L'ensemble des solutions de l'équation différentielle $-\varphi'' = \lambda\varphi$ est un espace vectoriel d'ordre 2. donc φ est de la forme $\varphi(x) = \alpha\cos\sqrt{\lambda}x + \beta\sin\sqrt{\lambda}x$ ($\lambda \geq 0$) et α et β dont déterminés par les conditions aux limites $\varphi(0) = \alpha = 0$ et $\varphi(1) = \alpha\cos\sqrt{\lambda} + \beta\sin\sqrt{\lambda} = 0$; on veut $\beta \neq 0$ car on cherche $\varphi \neq 0$ et donc on obtient $\lambda = k^2\pi^2$. Les couples (λ, φ) vérifiant (1.81) sont donc de la forme $(k^2\pi^2, \sin k\pi x)$.
2. Pour $k = 1$ à n, posons $\Phi_i^{(k)} = \sin k\pi x_i$, où $x_i = ih$, pour $i = 1$ à n, et calculons $A\Phi^{(k)}$:

$$(A\Phi^{(k)})_i = -\sin k\pi(i-1)h + 2\sin k\pi(ih) - \sin k\pi(i+1)h.$$

En utilisant le fait que $\sin(a+b) = \sin a\cos b + \cos a\sin b$ pour développer $\sin k\pi(1-i)h$ et $\sin k\pi(i+1)h$, on obtient (après calculs) :

$$(A\Phi^{(k)})_i = \lambda_k\Phi_i^{(k)}, \quad i = 1, \ldots, n,$$

avec

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}) \tag{1.82}$$

On a donc trouvé n valeurs propres $\lambda_1, \ldots, \lambda_n$ associées aux vecteurs propres $\Phi^{(1)}, \ldots, \Phi^{(n)}$ de $\mathbb{R}^n$ définis par $\Phi_i^{(k)} = \sin\dfrac{k\pi i}{n+1}$, $i = 1 \ldots n$.
Remarque : Lorsque $n \to +\infty$ (ou $h \to 0$), on a

$$\lambda_k^{(h)} = \frac{2}{h^2}\left(1 - 1 + \frac{k^2\pi^2h^2}{2} + O(h^4)\right) = k^2\pi^2 + O(h^2)$$

Donc

$$\lambda_k^{(h)} \to k^2\pi^2 = \lambda_k \text{ lorsque } h \to 0.$$

Calculons maintenant $\mathrm{cond}_2(A)$. Comme A est s.d.p., on a

$$cond_2(A) = \frac{\lambda_n}{\lambda_1} = \frac{1 - \cos\frac{n\pi}{n+1}}{1 - \cos\frac{\pi}{n+1}}$$

On a : $h^2\lambda_n = 2(1 - \cos\frac{n\pi}{n+1}) \to 4$ et $\lambda_1 \to \pi^2$ lorsque $h \to 0$. Donc

$$h^2 cond_2(A) \to \frac{4}{\pi^2} \text{ lorsque } h \to 0.$$

Exercice 57 page 75 (Conditionnement "efficace")

Partie I
1. Soit $u = (u_1, \ldots, u_n)^t$. On a

$$Au = b \Leftrightarrow \begin{cases} \dfrac{1}{h^2}(u_i - u_{i-1}) + \dfrac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \forall i = 1, \ldots, n, \\ u_0 = u_{n+1} = 0. \end{cases}$$

Supposons $b_i \geq 0, \forall i = 1, \ldots, n$, et soit

$$p = \min\{k \in \{0, \ldots, n+1\}; u_k = \min\{u_i, \ i = 0, \ldots, n+1\}\}.$$

Remarquons que p ne peut pas être égal à $n+1$ car $u_0 = u_{n+1} = 0$. Si $p = 0$, alors $u_i \geq 0\ \forall i = 0, n+1$ et donc $u \geq 0$.
Si $p \in \{1, \ldots, n\}$, alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0;$$

mais par définition de p, on a $u_p - u_{p-1} < 0$ et $u_p - u_{p+1} \leq 0$, et on aboutit donc à une contradiction.

Montrons maintenant que A est inversible. On vient de montrer que si $Au \geq 0$ alors $u \geq 0$. On en déduit par linéarité que si $Au \leq 0$ alors $u \leq 0$, et donc que si $Au = 0$ alors $u = 0$. Ceci démontre que l'application linéaire représentée par la matrice A est injective donc bijective (car on est en dimension finie).

2. Soit $\varphi \in C([0,1], \mathbb{R})$ tel que $\varphi(x) = \frac{1}{2}x(1-x)$ et $\phi_i = \varphi(x_i),\ i = 1, n$, où $x_i = ih$.
On remarque que $(A\phi)_i$ est le développement de Taylor à l'ordre 2 de $\varphi(x_i)$. En effet, φ est un polynôme de degré 2, sa dérivée troisième est nulle ; de plus on a $\varphi'(x) = \frac{1}{2} - x$ et $\varphi''(x) = 1$. On a donc :

$$\phi_{i+1} = \phi_i + h\varphi'(x_i) - \frac{h^2}{2}$$
$$\phi_{i-1} = \phi_i - h\varphi'(x_i) - \frac{h^2}{2}$$

On en déduit que $\frac{1}{h^2}(2\phi_i - \phi_{i+1} - \phi_{i+1}) = 1$, et donc que $(A\phi)_i = 1$.

3. Soient $b \in \mathbb{R}^n$ et $u \in \mathbb{R}^n$ tels que $Au = b$. On a :

$$(A(u \pm \|b\|\varphi))_i = (Au)_i \pm \|b\|(A\phi)_i = b_i \pm \|b\|.$$

Prenons d'abord $\tilde{b}_i = b_i + \|b\| \geq 0$, alors par la question (1),

$$u_i + \|b\|\phi_i \geq 0 \quad \forall i = 1 \ldots n.$$

Si maintenant on prend $\bar{b}_i = b_i - \|b\| \leq 0$, alors

$$u_i - \|b\|\phi_i \leq 0 \quad \forall i = 1, \ldots, n.$$

On a donc $-\|b\|\phi_i \leq u_i \leq \|b\|\phi_i$.
On en déduit que $\|u\| \leq \|b\|\ \|\phi\|$; or $\|\phi\| = \frac{1}{8}$. D'où $\|u\| \leq \frac{1}{8}\|b\|$.
On peut alors écrire que pour tout $b \in \mathbb{R}^n$,

$$\|A^{-1}b\| \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|}{\|b\|} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que $\|A^{-1}\| = \frac{1}{8}$ en prenant le vecteur $\boldsymbol{b}$ défini par $b(x_i) = 1, \forall i = 1, \ldots, n$. On a en effet $A^{-1}b = \phi$, et comme n est impair, $\exists i \in \{1, \ldots, n\}$ tel que $x_i = \frac{1}{2}$;or $\|\varphi\| = \varphi(\frac{1}{2}) = \frac{1}{8}$.

4. Par définition, on a $\|A\| = \sup_{\|x\|=1} \|Ax\|$, et donc $\|A\| = \max_{i=1,n} \sum_{j=1,n} |a_{i,j}|$, d'où le résultat.

5. Grâce aux questions 3 et 4, on a, par définition du conditionnement pour la norme $\|\cdot\|$, $\mathrm{cond}(A) = \|A\|\|A^{-1}\| = \frac{1}{2h^2}$.
Comme $A\delta_u = \delta_b$, on a :

$$\|\delta_u\| \leq \|A^{-1}\|\|\delta_b\|\frac{\|b\|}{\|b\|} \leq \|A^{-1}\|\|\delta_b\|\frac{\|A\|\|u\|}{\|b\|},$$

d'où le résultat.
Pour obtenir l'égalité, il suffit de prendre $b = Au$ où u est tel que $\|u\| = 1$ et $\|Au\| = \|A\|$, et δ_b tel que $\|\delta_b\| = 1$ et $\|A^{-1}\delta_b\| = \|A^{-1}\|$. On obtient alors

$$\frac{\|\delta_b\|}{\|b\|} = \frac{1}{\|A\|} \text{ et } \frac{\|\delta u\|}{\|u\|} = \|A^{-1}\|.$$

D'où l'égalité.

Partie 2 Conditionnement "efficace"

1. Soient φ_h et f_h les fonctions constantes par morceaux définies par

$$\begin{array}{ll} \varphi_h(x) & = \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f_h(x) & = \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

Comme $f \in C([0,1], \mathbb{R})$ et $\varphi \in C^2([0,1], \mathbb{R})$, la fonction f_h (resp. φ_h) converge uniformément vers f (resp. φ) lorsque $h \to 0$. En effet,

$$\begin{aligned} \|f - f_h\|_\infty &= \sup_{x \in [0,1]} |f(x) - f_h(x)| \\ &= \max_{i=0,\ldots,n} \sup_{x \in [x_i, x_{i+1}]} |f(x) - f_h(x)| \\ &= \max_{i=0,\ldots,n} \sup_{x \in [x_i, x_{i+1}]} |f(x) - f(x_i)| \end{aligned}$$

Comme f est continue, elle est uniformément continue sur $[0,1]$ et donc pour tout $\varepsilon > 0$, il existe $h_\varepsilon > 0$ tel que si $|s - t| \leq h_\varepsilon$, alors $|f(s) - f(t)|$. On en conclut que si l'on prend $h \leq h_\varepsilon$, on a $\|f - f_h\| \leq \varepsilon$. Le raisonnement est le même pour φ_h, et donc $f_h\varphi_h$ converge uniformément vers $f\varphi$. On peut donc passer à la limite sous l'intégrale et écrire que :

$$h \sum_{i=1}^{n} b_i \varphi_i = \int_0^1 f_h(x)\varphi_h(x)dx \to \int_0^1 f(x)\varphi(x)dx \text{ lorsque } h \to 0.$$

Comme $b_i > 0$ et $\phi_i > 0\ \forall i = 1, \ldots, n$, on a évidemment

$$S_n = \sum_{i=1}^{n} b_i \varphi_i > 0 \text{ et } S_n \to \int_0^1 f(x)\varphi(x)dx = \beta > 0 \text{ lorsque } h \to 0.$$

Donc il existe $n_0 \in \mathbb{N}$ tel que si $n \geq n_0$, $S_n \geq \dfrac{\beta}{2}$, et donc $S_n \geq \alpha = \min(S_0, S_1 \ldots S_{n_0}, \dfrac{\beta}{2}) > 0$.

2. On a $n\|u\| = n \sup_{i=1,n} |u_i| \geq \sum_{i=1}^n u_i$. D'autre part, $A\varphi = (1 \ldots 1)^t$ donc $u \cdot A\varphi = \displaystyle\sum_{i=1}^{n} u_i$; or $u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$ car A est symétrique. Donc $u \cdot A\varphi = \displaystyle\sum_{i=1}^{n} b_i\varphi_i \geq \frac{\alpha}{h}$ d'après la question 1. Comme $\delta_u = A^{-1}\delta_b$, on a donc $\|\delta_u\| \leq \|A^{-1}\|\ \|\delta_b\|$; et comme $n\|u\| \geq \dfrac{\alpha}{h}$, on obtient : $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{1}{8}\dfrac{hn}{\alpha}\|\delta_b\|\dfrac{\|f\|}{\|b\|}$. Or $hn \leq 1$ et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|}{8\alpha}\frac{\|\delta_b\|}{\|b\|}.$$

3. Le conditionnement $\mathrm{cond}(A)$ calculé dans la partie 1 est d'ordre $1/h^2$, et donc tend vers l'infini lorsque le pas de discrétisation tend vers 0, alors qu'on vient de montrer dans la partie 2 que la variation relative $\frac{\|\delta_u\|}{\|u\|}$ est inférieure à une constante multipliée par la variation relative de $\frac{\|\delta_b\|}{\|b\|}$. Cette dernière information est nettement plus utile et réjouissante pour la résolution effective du système linéaire.

Université d'Aix-Marseille, 2018-2019
Licence de Mathématiques, analyse numérique
Travaux Pratiques 1, en python

Exercice 1 (Résolution numérique de $-u'' = f$)
Pour $f \in C([0,1], \mathbb{R})$ donné, on cherche à calculer de manière approchée, par un schéma aux Différences Finies, la solution, notée u, du problème suivant :

$$\begin{aligned} -u''(x) &= f(x) \text{ pour } x \in]0,1[, \\ u(0) &= u(1) = 0. \end{aligned}$$

On note h le pas du maillage, $h = 1/(n+1)$, $n \in \mathbb{N}^\star$. le problème discrétisé est donc de la forme $A_h u_h = f_h$, où A_h est une matrice carré de taille n et u_h, f_h sont des vecteurs de taille n (voir le cours pour plus de précisions). L'erreur de discrétisation est donnée par la norme infinie du vecteur $(u_a - u_e)$ où u_e est le vecteur formé par la solution exacte prise aux points du maillage.
On choisit pour second membre la fonction f définie par $f(x) = \pi^2 \sin(\pi x)$.

1. Pour $n \in \mathbb{N}^\star$,
   (a) Ecrire un programme construisant la matrice A_h sous forme d'une matrice creuse.
   [On pourra utiliser la structure **scipy.sparse.lil_matrix**.]
   (b) Contruire le vecteur f_h.
   (c) Calculer le vecteur u_h en utilisant une résolution directe pour matrice creuse.
   [On pourra mettre la matrice A_h et vecteur f_h sous la forme **csr** et utiliser le solveur **scipy.sparse.linalg.spsolve**.]
2. Vérifier que la méthode est bien convergente d'ordre 2.
   [En prenant, par exemple, $n+1 = 100$ et $n+1 = 200$, l'erreur de discrétisation est essentiellement divisée par 4.]

Exercice 2 (Décomposition LU et Cholesky)
On pose $A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$ et $B = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}$.

1. Calculer la décomposition de Cholesky de la matrice A et constater la conservation du profil. [On pourra utiliser **linalg.cholesky**.]
2. Calculer les mineurs principaux de la matrice B et en déduire qu'on peut utiliser, pour cette matrice, la décomposition LU. [On pourra utiliser **linalg.det**.]
3. Calculer la décomposition LU de la matrice B et constater la conservation du profil. [On pourra utiliser **scipy.linalg.lu**.]

Exercice 3 (Le ballon de Foot) L'objectif de cet exercice est de déterminer le nombre de faces d'un ballon de foot. Un ballon de foot est formée de faces de forme pentagonales ou hexagonales. On notera x le nombre de pentagones et y le nombre d'hexagones qui le constituent. On notera f le nombre total de faces, a le nombre d'arêtes et s le nombre de sommets du ballon. Ces nombres sont des entiers positifs.
Pour déterminer x et y, on écrit les relations suivantes :

- chaque sommet appartient à exactement trois faces, $3s = 5x + 6y$,
- chaque arête est partagée par deux faces, $2a = 5x + 6y$,
- le nombre total de faces est égal à la somme des nombres de pentagones et hexagones, $f = x + y$,
- (relation d'Euler) le nombre total d'arêtes est égal à la somme du nombre de faces et du nombre de sommets moins 2, $a = f + s - 2$.

1. On note X le vecteur de $\mathbb{R}^5$ dont les composantes sont x, y, f, a, s. Montrer que X est solution d'un système linéaire de 4 équations (à 5 inconnues) de la forme $AX = b$.
2. Trouver (avec python) les solutions entières du système linéaire de la question précédente. On pourra (comprendre et) programmer l'algorithme suivant consistant a échelonner la matrice A en notant n le nombre de lignes de A (ici $n = 4$) et p le nombre de colonnes (ici $p = 5$). Le second membre b est donc un vecteur de $\mathbb{R}^n$. On note l_{i-1} la i-ieme ligne de A (la première ligne est la ligne l_0).

   $i = 0$

   Pour j de 0 à $p - 2$:

   choisir, si c'est possible, k entre i et $n - 1$ tel que $a_{k,j} \neq 0$.

   échanger l_i et l_k

   échanger b_i et b_k

   Pour m de $i + 1$ à $n - 1$:

   Remplacer l_m par $l_m - (a_{m,j}/a_{i,j})l_i$

   Remplacer b_m par $b_m - (a_{m,j}/a_{i,j})b_i$

   $i = i + 1$
3. Sachant que le ballon de foot correspond à $y = 20$, donner x, f, a et s.

Université d'Aix-Marseille, année 2018-2019
Licence de Mathématiques, analyse numérique
Travaux Pratiques 2, en python

Exercice 1 (Etude d'un système particulier)
On s'intéresse au système $Ax = b$ avec

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

1. Calculer les valeurs propres de A.

   [On pourra utiliser la fonction **numpy.linalg.eigvals**].

   En déduire que A est une matrice s.d.p..

2. Calculer le conditionnement de A en utilisant les valeurs propres de A. Comparer avec le résultat donné par la fonction **numpy.linalg.cond**.

3. Calculer le solution du système linéaire $Ax = b$ en utilisant la fonction **numpy.linalg.solve**.

4. Calculer le solution du système linéaire $Ax = b$ en utilisant le programme d'échelonnement du tp1 (qui, ici, consiste à utiliser la méthode de Gauss avec pivot partiel). Comparer avec la solution de la question précédente.

5. On perturbe maintenant légèrement le système en remplaçant b par $b + \delta_b$, avec $\delta_b = \begin{bmatrix} 0.1 \\ -0.1 \\ 0.1 \\ -0.1 \end{bmatrix}$. Calculer la nouvelle solution du système, notée $x + \delta_x$. Vérifier que $\frac{|\delta_x|}{|x|} \leq cond(A)\frac{|\delta_b|}{|b|}$.

On rappelle que la norme euclidienne s'obtient avec **numpy.linalg.norm**

Exercice 2 (L'inégalité sur le conditionnement est optimale)
Dans cet exercice, on construit un exemple pour lequel $\frac{|\delta_x|}{|x|} \leq cond(A)\frac{|\delta_b|}{|b|}$ avec $A \in \mathcal{M}_n(\mathbb{R})$ inversible. On pourra prendre, par exemple, $n = 10$.

1. Construire une matrice diagonale $D = diag(d_1, \ldots, d_n n)$ en choisissant d_1, $\ldots$, d_n de manière aléatoire entre 1 et 10.

   [utiliser **numpy.random.random** qui donne n nombres aléatoires compris entre 0 et 1.]

   Ordonner les nombres pour que $0 < d_1 < \ldots < d_n$ (utiliser **numpy.sort**). Enfin, utiliser **numpy.diag** pour construire D.

2. Construire une matrice Q (de $\mathcal{M}_n(\mathbb{R})$) orthogonale de la manière suivante : Construire une première matrice P avec des coefficients alétaoires (en général P est inversible, mais si P n'est pas inversible, on recommence le choix des coefficients). On obtient alors une matrice orthogonale Q en effectuant la décomposition QR de cette matrice (utiliser **numpy.linalg.qr**).

On choisit maintenant $A = QDQ^t$, avec D et Q données par les questions précédentes. La matrice A ainsi construite est symétrique définie positive. On remarque que $c_1(Q)$ est le vecteur propre associé à la plus petite valeur propre de A alors que $c_n(Q)$ est le vecteur propre associé à la plus grande valeur propre de A.

3. On pose $b = c_n(Q)$ et $\delta_b = c_1(Q)$ Calculer x et $x + \delta_x$ (tels que $Ax = b$ et $A(\delta_x) = \delta_b$) et vérifier que

$$\frac{|\delta_x|}{|x|} = cond(A)\frac{|\delta_b|}{|b|}.$$

Pour cet exemple le conditionnement mesure exactement le rapport entre les erreurs relatives. Mais, on remarque que cette égalité est obtenue pour des choix très particuliers de b et δ_b. L'exercice suivant montre que la sensibilité de la solution aux erreurs sur le second membre peut être bien meilleure que celle prédite par le conditionnement lorsque le problème provient d'une discrétisation d'une équation différentielle (ou, plus généralement, d'une équation aux dérivées partielles).

Exercice 3 (Conditionnement "efficace")
On s'intéresse, dans cet exercice, à la matrice de l'exercice 1 du tp1. Pour $n \in \mathbb{N}^\star$, on note A_n cette matrice (qui était notée A_h au tp1, avec $h = 1/(n+1)$). On considère le même problème que dans le tp1, c'est à dire celui correspondant à $f(x) = \pi^2 \sin(\pi x)$ pour $x \in [0, 1]$.
Pour $n \in \mathbb{N}^\star$ le problème discrétisé s'écrit donc $A_n x_n = b_n$.

1. Pour n variant entre 100 et 1000 :
   (a) Calculer b_n et x_n,
   (b) Choisir de manière aléatoire un vecteur de $\mathbb{R}^n$, noté δ_{b_n}, prenant ses valeurs entre 0 et 0.1. Calculer $x_n + \delta_{x_n}$ solution de $A_n(x_n + \delta_{x_n}) = b_n + \delta_{b_n}$ et calculer le nombre $cond_f(A_n)$ vérifiant $\frac{|\delta_{x_n}|}{|x_n|} = cond_f(A_n)\frac{|\delta_{b_n}|}{|b_n|}$.

2. Dessiner les graphes (Pour n variant entre 100 et 1000) des applications $n \mapsto cond(A_n)$ et $n \mapsto cond_f(A_n)$. Remarquer que contrairement au conditionnement de A_n, le rapport entre les erreurs relatives sur b et x (notée $cond_f(A_n)$) ne croit pas comme n^2 (mais reste borné).

   [Utiliser `matplotlib.pyplot.plot` et `matplotlib.pyplot.show`.]