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 $b \in \mathbb{R}^n$, on a comme objectif 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

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

Si on écrit les matrices A et B sous forme de lignes (notées $\boldsymbol{\ell}_i$) et colonnes (notées $\boldsymbol{c}_j$) :

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

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 :

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

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 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.

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.

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 (nonzero) pivots	A a $r < n$ pivots
A est de rang maximal : rg(A) = n.	rg(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 is symétrique définie positive [2]	A^tA n'est que semi- définie .

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

On rappelle pour une bonne lecture de ce tableau les quelques définitions suivantes :

Définition 1.1 (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).*

Définition 1.2 (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 1.3 (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 $\det\tilde{A} = -\det A$.

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

(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 1.4 (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 1.5. *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

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

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 $(e_i)_{i=1,\ldots,n}$ 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 $(\boldsymbol{e}_i)_{i=1,\ldots,n}$ à la base $(\boldsymbol{u}_i)_{i=1,\ldots,n}$; (il est bien clair que la i-ème colonne de P est constituée des composantes de $\boldsymbol{u}_i$ dans la base canonique $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$.
La matrice P est inversible car les vecteurs propres forment une base, et on peut donc aussi écrire :

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

■

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 1.6 (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*

$$\begin{aligned} E \times E &\to \mathbb{R}, \\ (x, y) &\to (x \mid y)_E, \end{aligned}$$

qui vérifie :

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

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

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^n$, le produit scalaire canonique de $x = (x_1, \dots, x_n)^t$ et $y = (y_1, \dots, y_n)^t$ est défini par $(x \mid y)_E = x \cdot y = \sum_{i=1}^n x_i y_i$. Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(x) = Ax$ est linéaire, et : $(Tx|y) = Ax \cdot y = x \cdot A^t y = x \cdot Ay = (x \mid Ty)$. 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 $(e_1, \dots, e_n)$ base canonique dans $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ dont la i-ieme colonne de P est constituée des coordonnées de f_i dans $(e_1 \dots e_n)$. On a : $Pe_i = f_i$. On a alors $P^{-1}APe_i = P^{-1}A\boldsymbol{f}_i = P^{-1}(\lambda_i f_i) = \lambda_i e_i = \mathrm{diag}(\lambda_1, \dots, \lambda_n)e_i$, 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,

$$P^tPe_i \cdot e_j = Pe_i \cdot Pe_j = (f_i|f_j) = \delta_{i,j} \quad \forall i, j \in \{1 \dots n\},$$

et donc $(P^tPe_i - e_i) \cdot e_j = 0 \quad \forall j \in \{1 \dots n\} \quad \forall i \in \{1, \dots n\}$. On en déduit $P^tPe_i = e_i$ pour tout $i = 1, \dots n$, *i.e.* $P^tP = PP^t = Id$.

DÉMONSTRATION *du lemme 1.6* 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 $e \in E$, $e \neq 0$, alors $E = \mathbb{R}e = \mathbb{R}f_1$ avec $\boldsymbol{f}_1 = \dfrac{e}{\|e\|}$. 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 $dimE < n$. 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 :

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

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

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

donc $\lambda(e + ty \mid e + ty)_E \geq (T(e + ty) \mid e + ty)$. En développant on obtient :

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

Comme $t > 0$, ceci donne :

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

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

Soit $F = \{x \in E; (x \mid e) = 0\}$, on a donc $F \neq E$, et $E = F \bigoplus \mathbb{R}e$: On peut décomposer $x \in E$ comme $x = x - (x \mid e)e + (x \mid e)e$. Si $x \in F$, on a aussi $T(x) \in F$ (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. ∎

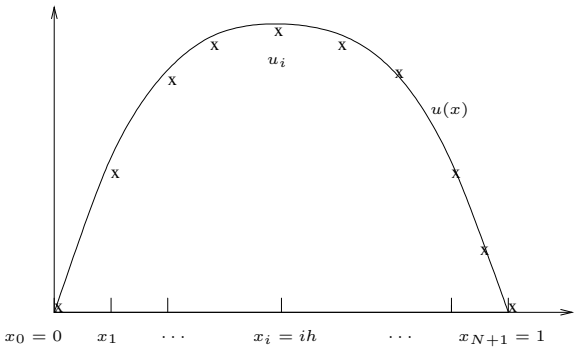

FIGURE 1.1: Solution exacte et approchée de $-u'' = f$

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 1.7 (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, 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 :

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

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 $(u_i)_{i=1,\dots,n}$ 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 :

$$\begin{cases} k_{i,i} &= \dfrac{2}{h^2}, \ \forall\, i = 1, \dots, n, \\ k_{i,j} &= -\dfrac{1}{h^2}, \ \forall\, i = 1, \dots, n, \ \ j = i \pm 1, \\ k_{i,j} &= 0, \ \forall\, i = 1, \dots, n, \ |i - j| > 1. \end{cases} \tag{1.10}$$

On remarque immédiatement que K_n est tridiagonale.
On peut montrer que K_n est symétrique définie positive (voir exercice 12 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 :

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

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 $\max_{i=1,\dots,n} e_i$ 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 52, où l'on montre que le schéma est convergent, et qu'on a l'estimation d'erreur suivante :

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

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 $i = 1,\dots,M$ et $j = 1,\dots,M$ 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 11), 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 $b = (b_1,\dots,b_n)^t$. Les composantes $b_1,\dots,b_n$ sont définies par :pour $i,j = 1,\dots,M$, on pose $k = j + (i-1)M$ et $b_k = f(x_i, y_j)$.
Les coefficients de $A = (a_{k,\ell})_{k,l=1,n}$ peuvent être calculés de la manière suivante :

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

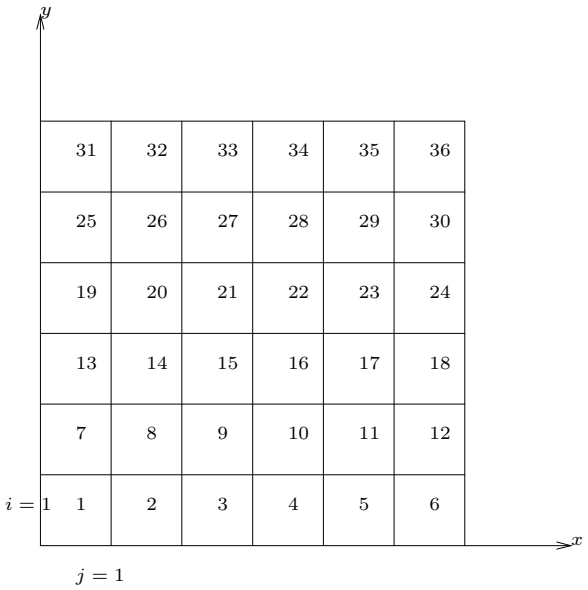

FIGURE 1.2: Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

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 1.9 qu'une matrice A est monotone si et seulement si elle est inversible et à inverse positive.

Définition 1.8 (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 1.9 (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 10. 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.