2023 - 2024

Analyse numérique

Cours

Pierre Sochala

École Spéciale des Travaux Publics, du Bâtiment et de l'Industrie

Table des matières

Chapitre 1

Résolution des systèmes matriciels par méthodes directes

Plusieurs problèmes d'analyse numérique aboutissent à un système matriciel

$$Ax = b, \tag{1.1}$$

où A est la matrice du système, b le second membre et x le vecteur inconnu dont les composantes sont à déterminer. Nous présentons les méthodes directes qui calculent la solution en un *nombre fini* d'opérations. Nous détaillons 1) le conditionnement qui traduit la sensibilité d'un système matriciel aux variations des données, 2) les méthodes de substitution directes et indirectes utilisées respectivement pour les matrices triangulaires inférieures et supérieures, 3) la méthode d'élimination de Gauss avec ses variantes obtenues avec pivotage partiel ou total, 4) la factorisation LU qui décompose une matrice en un produit de matrices inférieure et supérieure, 5) le calcul de l'inverse d'une matrice. Nous précisons que seuls les éléments non nuls sont indiqués dans les matrices et que la matrice identité est notée I.

1.1 Normes et conditionnement

1.1.1 Normes matricielles

Pour définir le conditionnement, nous introduisons la notion de norme matricielle dont la propriété spécifique est la sous-multiplicativité,

Définition. *Une norme matricielle est une application* $||\cdot|| : \mathbb{R}^{m\times n} \to \mathbb{R}$ *vérifiant pour tout* $A, B \in \mathbb{R}^{m\times n}$,

1. $||A|| \geq 0$ *et* $||A|| = 0 \Leftrightarrow A = 0$,
2. $\forall \alpha \in \mathbb{R},\ ||\alpha A|| = |\alpha|\ ||A||$ *(homogénéité),*
3. $||A + B|| \leq ||A|| + ||B||$ *(inégalité triangulaire).*

De plus, une norme matricielle est sous-multiplicative si $\forall A \in \mathbb{R}^{m\times n}$ *et* $\forall B \in \mathbb{R}^{n\times q}$,

4. $||AB|| \leq ||A||\ ||B||$

La norme p d'une matrice est définie à partir de la norme p vectorielle (on parle de norme subordonnée) de la façon suivante

$$\boxed{||A||_p \overset{\text{def}}{=} \max_{||x||_p \neq 0} \frac{||Ax||_p}{||x||_p} = \max_{||x||_p = 1} ||Ax||_p}$$

Cette relation signifie que la norme d'une matrice est le maximum de la norme du vecteur résultant de la multiplication de cette matrice par l'ensemble des vecteurs de norme unitaire. Sous cette forme, il est difficile de calculer les normes matricielles mais il existe les équivalences suivantes pour les normes 1, 2 et infinie,

$$||A||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \tag{1.2}$$

$$||A||_2 = \sqrt{\rho(A^\top A)} \tag{1.3}$$

$$||A||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|, \tag{1.4}$$

où $\rho(A)$ est le *rayon spectral* de la matrice A. Dans le cas d'une matrice symétrique,

$$\boxed{||A||_2 = \rho(A).}$$

1.1.2 Conditionnement

Nous considérons le système matriciel suivant dont la solution est le vecteur unitaire

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \quad \Rightarrow \quad x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

De faibles perturbations introduites dans la matrice (représentant des erreurs d'arrondi par exemple) modifient considérablement la solution,

$$\begin{bmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{bmatrix} \begin{bmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \\ x_3 + \delta x_3 \\ x_4 + \delta x_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \quad \Rightarrow \quad x + \delta x = \begin{bmatrix} -81 \\ 137 \\ -34 \\ 22 \end{bmatrix}.$$

De même, de faibles perturbations dans le second membre changent significativement la solution

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{bmatrix} x_1 + \delta x'_1 \\ x_2 + \delta x'_2 \\ x_3 + \delta x'_3 \\ x_4 + \delta x'_4 \end{bmatrix} = \begin{bmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{bmatrix} \quad \Rightarrow \quad x + \delta x' = \begin{bmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{bmatrix}.$$

Lorsque de petites perturbations sont introduites dans la matrice ou le second membre, nous attendons que les deux solutions soient proches de la solution du système initial. Cela n'est pas le cas ici car la matrice est *mal conditionnée*.

Définition. *Soit une norme matricielle subordonnée notée $||\cdot||$. Le conditionnement relatif à cette norme d'une matrice inversible A est le réel $K(A)$ défini par*

$$\boxed{K(A) \stackrel{\text{def}}{=} ||A||\ ||A^{-1}||} \tag{1.5}$$

Ce nombre mesure la sensibilité du système aux perturbations des données (matrice ou second membre) et plus le conditionnement d'une matrice est grand, plus la solution du système linéaire est sensible aux données.

Théorème. *Soit A une matrice inversible et b un vecteur non nul.*

1. *Soient x et $x+\delta x$ les solutions respectives des systèmes*

$$Ax = b \quad et \quad A(x+\delta x) = b+\delta b,$$

Alors

$$\frac{||\delta x||}{||x||} \leq K(A)\frac{||\delta b||}{||b||}. \tag{1.6}$$

2. *Soient x et $x+\delta x$ les solutions respectives des systèmes*

$$Ax = b \quad et \quad (A+\delta A)(x+\delta x) = b,$$

Alors

$$\frac{||\delta x||}{||x+\delta x||} \leq K(A)\frac{||\delta A||}{||A||}. \tag{1.7}$$

Démonstration. Pour le premier résultat, nous utilisons que

$$\begin{aligned} Ax = b \quad &\Rightarrow \quad ||b|| \leq ||A||\ ||x||, \\ A\delta x = \delta b \quad &\Rightarrow \quad ||\delta x|| \leq ||A^{-1}||\ ||\delta b||. \end{aligned}$$

Pour le second résultat, nous utilisons que

$$A\delta x + \delta A(x+\delta x) = 0 \quad \Rightarrow \quad ||\delta x|| \leq ||A^{-1}||\ ||\delta A||\ ||x+\delta x||.$$

□

Le conditionnement relatif à la norme 2 d'une matrice symétrique est le *conditionnement spectral*,

$$\boxed{K_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}},} \tag{1.8}$$

où $\lambda_{\max}$ est la valeur propre de plus grand module et $\lambda_{\min}$ la valeur propre de plus petit module.

Propriétés. *Soit A une matrice inversible. Le conditionnement $K(A)$ vérifie les propriétés*

1. $K(A) = K(A^{-1})$,
2. $K(A) \geq 1$,
3. $\forall \alpha \neq 0, K(\alpha A) = K(A)$.

Démonstration.

1. $K(A) = ||A||\ ||A^{-1}|| = ||A^{-1}||\ ||A|| = ||A^{-1}||\ ||(A^{-1})^{-1}|| = K(A^{-1})$,
2. $AA^{-1} = I \ \Rightarrow\ ||AA^{-1}|| = 1 \ \Rightarrow\ K(A) \geq 1$ d'après la sous-multiplicativité,
3. $\forall \alpha \neq 0, K(\alpha A) = ||\alpha A||\ ||(\alpha A)^{-1}|| = \alpha||A||\ \frac{1}{\alpha}||A^{-1}|| = K(A)$.

□

La propriété 2 indique que le conditionnement est d'autant meilleur que sa valeur est proche de 1. La propriété 3 indique que la multiplication d'une matrice par un scalaire (petit par exemple) ne change pas son conditionnement.

1.2 Systèmes triangulaires

1.2.1 Matrice triangulaire inférieure

La matrice L (lower) du système est supposée triangulaire inférieure,

$$\begin{bmatrix} l_{11} & & \\ \vdots & \ddots & \\ l_{n1} & \dots & l_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

La résolution d'un tel système est évidente : on détermine x_1 par la première équation puis x_2 par la deuxième puisque x_1 est connue puis x_3 par la troisième puisque x_1 et x_2 sont connus et ainsi de suite. Cette méthode de *substitution directe* est caractérisée par les *formules de descente* suivantes

$$\boxed{x_1 = \frac{b_1}{l_{11}} \quad \text{et} \quad x_i = \frac{1}{l_{ii}}\left(b_i - \sum_{j=1}^{i-1} l_{ij}x_j\right), \quad 2 \leq i \leq n.} \tag{1.9}$$

1.2.2 Matrice triangulaire supérieure

La matrice U (upper) du système est supposée triangulaire supérieure,

$$\begin{bmatrix} u_{11} & \dots & u_{1n} \\ & \ddots & \vdots \\ & & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

La résolution d'un tel système est évidente : on détermine x_n par la dernière équation puis x_{n-1} par l'avant-dernière puisque x_n est connue puis x_{n-2} par l'antépénultième puisque x_n et x_{n-1} sont connus et ainsi de suite. Cette méthode de *substitution indirecte* est caractérisée par les *formules de remontée* suivantes

$$\boxed{x_n = \frac{b_n}{u_{nn}} \quad \text{et} \quad x_i = \frac{1}{u_{ii}}\left(b_i - \sum_{j=i+1}^{n} u_{ij}x_j\right), \quad n-1 \leq i \leq 1.} \tag{1.10}$$

1.3 Méthode de Gauss

Le principe de la méthode de Gauss est de transformer le système initial $Ax = b$ en un système triangulaire supérieur $Ux = \widehat{b}$ résolvable par remontée. Cet algorithme d'élimination utilise que la solution d'un système linéaire reste inchangée lorsqu'on ajoute à une équation une combinaison linéaire des autres équations. Nous allons construire une suite de matrices $A^{(k)}$ et $b^{(k)}$ initialisée par

$$A^{(1)} = A, \quad b^{(1)} = b,$$

et aboutissant à l'étape finale à un système triangulaire supérieur

$$A^{(n)} = U, \quad b^{(n)} = \widehat{b}.$$

Nous présentons l'algorithme sans pivotage et ses variantes plus robustes lorsque le pivot est judicieusement choisi.

1.3.1 Algorithme sans pivotage

• ***Etape 1***. Cette étape élimine les coefficients de la première colonne des lignes 2 à n à l'aide des multiplicateurs

$$\forall\ 2 \leq i \leq n, \quad m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \tag{1.11}$$

où $a_{11}^{(1)}$ est le *pivot* de la première étape. En effet, l'inconnue x_1 s'élimine des lignes 2 à n en leur retranchant m_{i1} fois la première ligne, qu'il faut aussi soustraire au second membre. Les coefficients

$$\forall\ 2 \leq i,j \leq n, \quad a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}, \tag{1.12}$$

$$\forall\ 2 \leq i \leq n, \quad b_{i}^{(2)} = b_{i}^{(1)} - m_{i1}b_{1}^{(1)}, \tag{1.13}$$

définissent les coefficients des lignes 2 à n du nouveau système dont la matrice et le second membre sont

$$A^{(2)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & \vdots & & \vdots \\ & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix} \quad \text{et} \quad b^{(2)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}.$$

• ***Etape k***. Cette étape élimine les coefficients de la k-ième colonne des lignes $k+1$ à n à l'aide des multiplicateurs

$$\forall\ k+1 \leq i \leq n, \quad m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \tag{1.14}$$

où $a_{kk}^{(k)}$ est le *pivot* de l'étape k. En effet, l'inconnue x_k s'élimine des lignes $k+1$ à n en leur retranchant m_{ik} fois la ligne k, qu'il faut aussi soustraire au second membre. Les coefficients

$$\forall\ k+1 \leq i,j \leq n, \quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, \tag{1.15}$$

$$\forall\ k+1 \leq i \leq n, \quad b_{i}^{(k+1)} = b_{i}^{(k)} - m_{ik}b_{k}^{(k)}, \tag{1.16}$$

définissent les coefficients des lignes $k+1$ à n du nouveau système dont la matrice et le second membre sont

$$A^{(k+1)} = \begin{bmatrix} a_{11}^{(1)} & \dots & & \dots & a_{1n}^{(1)} \\ & \ddots & & & \vdots \\ & & a_{(k+1)(k+1)}^{(k+1)} & \dots & a_{(k+1)n}^{(k+1)} \\ & & \vdots & & \vdots \\ & & a_{n(k+1)}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{bmatrix} \quad \text{et} \quad b^{(k+1)} = \begin{bmatrix} b_1^{(1)} \\ \vdots \\ b_{k+1}^{(k+1)} \\ \vdots \\ b_n^{(k+1)} \end{bmatrix}.$$

• ***Etape n-1***. Cette étape élimine le coefficients de l'avant dernière colonne de la dernière ligne à l'aide du multiplicateur

$$m_{n(n-1)} = \frac{a_{n(n-1)}^{(n-1)}}{a_{(n-1)(n-1)}^{(n-1)}}, \tag{1.17}$$

où $a_{(n-1)(n-1)}^{(n-1)}$ est le *pivot* de l'étape $n-1$. En effet, l'inconnue x_{n-1} s'élimine de la dernière ligne en lui retranchant $m_{n(n-1)}$ fois l'avant dernière ligne, qu'il faut aussi soustraire au second

membre. Les coefficients

$$a_{nn}^{(n)} = a_{nn}^{(n-1)} - m_{nn-1} a_{(n-1)n}^{(n-1)} \tag{1.18}$$

$$b_{n}^{(n)} = b_{n}^{(n-1)} - m_{n(n-1)} b_{(n-1)}^{(n-1)}, \tag{1.19}$$

définissent les coefficients de la dernière ligne du nouveau système dont la matrice et le second membre sont

$$A^{(n)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix} \quad \text{et} \quad b^{(n)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{bmatrix}.$$

1.3.2 Pivotage

Il est clair qu'un coefficient multiplicatif de l'algorithme sans pivotage n'est pas défini si le pivot associé est nul. Ainsi, une amélioration nécessaire de cet algorithme est de rechercher un pivot non nul à chaque étape. Pour obtenir une meilleur stabilité numérique (propagation des erreurs d'arrondis), il faut choisir le plus grand pivot en valeur absolue. Le pivotage est

— *partiel* si le pivot est choisi dans la k-ième colonne sous la diagonale,

— *total* si le pivot est choisi dans la sous-matrice inférieure diagonale de dimension $(n-k)^2$.

A chaque étape, les algorithmes avec pivotage possèdent donc une sous-étape supplémentaire de recherche du pivot. Cela se traduit par une permutation de deux lignes pour le pivot partiel et une permutation de deux lignes et deux colonnes pour le pivot total.

1.3.3 Matrices d'éliminations et de permutations

Nous terminons cette section en précisant que l'étape d'élimination de l'étape $k+1$ s'effectue en multipliant (à gauche) la matrice de l'étape k par la matrice d'élimination $E^{(k)}$ définie comme la somme de la matrice identité et d'une matrice ayant ses coefficients sous la diagonale de la colonne k égaux aux opposés des multiplicateurs,

$$E^{(k)} = I + \begin{bmatrix} 0 & 0 & 0 \\ & \vdots & \\ \vdots & 0 & \vdots \\ & -m_{(k+1)k} & \\ & \vdots & \\ 0 & -m_{nk} & 0 \end{bmatrix},$$

de sorte que

$$A^{(k+1)} = E^{(k)} A^{(k)} \quad \text{et} \quad b^{(k+1)} = E^{(k)} b^{(k)}.$$

De même, la sous-étape de permutation s'effectue à l'aide d'une matrice de permutation du type

$$P = \begin{bmatrix} \ddots & & & & & \\ & 1 & & & & \\ & & 0 & & 1 & \\ & & & 1 & & \\ & & & & \ddots & \\ & & 1 & & & \\ & & & & & 0 & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \end{bmatrix}.$$

Une multiplication à droite permute les colonnes alors qu'une multiplication à gauche permute les lignes. Ces matrices (creuses) d'éliminations et de permutations ne sont pas assemblées lors de l'implémentation de l'algorithme du pivot de Gauss mais servent plutôt dans certaines démonstrations comme celle de la factorisation LU.

1.4 Factorisation LU

La méthode *LU* factorise la matrice A en un produit de deux matrices triangulaires

$$A = LU, \tag{1.20}$$

où L est triangulaire inférieure *à diagonale unité* et U triangulaire supérieure. La solution du système linéaire $Ax = b$ est alors obtenue par la résolution successive de deux systèmes triangulaires,

$$\begin{cases} Ly = b, \\ Ux = y \end{cases} \tag{1.21}$$

Il s'agit en fait du même algorithme que celui de l'élimination de Gauss dans le cas particulier sans pivotage. Dans ce cas, la dernière étape de la méthode de Gauss s'écrit

$$A^{(n)} = E^{(n-1)} \cdots E^{(1)} A \quad \Rightarrow \quad A = \left(E^{(n-1)} \cdots E^{(1)}\right)^{-1} A^{(n)}$$

Par identification, nous avons

$$\boxed{L = \left(E^{(n-1)} \cdots E^{(1)}\right)^{-1} \quad \text{et} \quad U = A^{(n)}.}$$

La matrice $A^{(n)}$ est triangulaire supérieure par construction et la matrice $\left(E^{(n-1)} \cdots E^{(1)}\right)^{-1}$ est triangulaire inférieure (car le produit et l'inverse de matrices triangulaires inférieures est une matrice triangulaire inférieure). Un autre moyen de déterminer la factorisation *LU* est de poser

$$L = \begin{bmatrix} 1 & & & \\ l_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \dots & l_{nn-1} & 1 \end{bmatrix} \quad \text{et} \quad U = \begin{bmatrix} u_{11} & \dots & \dots & u_{1n} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & u_{nn} \end{bmatrix}$$

et d'identifier terme à terme le produit *LU* avec A.

1.5 Inverse d'une matrice

La détermination de l'inverse A^{-1} d'une matrice A s'effectue en calculant ses colonnes c_i qui sont solutions des systèmes linéaires,

$$\forall\ 1 \leq i \leq n, \quad \boxed{Ac_i = e_i,}$$

où e_i est le i-ème vecteur de base (les composantes de e_i sont nulles sauf sa composante i égale à 1). En effet, les relations précédentes s'écrivent

$$A[c_1, \dots, c_n] = [e_1, \dots, e_n] = I,$$

ce qui signifie que la matrice $[c_1, \dots, c_n]$ est l'inverse de A. En pratique, on procède selon les deux étapes :

1. factorisation *LU* de A,
2. résolution des n systèmes $LUc_i = e_i$ par double substitution.

Chapitre 2

Approximation

L'approximation consiste à trouver une représentation d'un nuage de points par un polynôme. L'intérêt d'un tel polynôme est la synthèse de données expérimentales dont le nombre peut être élevé. Cette méthode nécessite de fixer *a priori* le degré du polynôme d'approximation.

2.1 Régression linéaire

Soient n couples de points (x_i, y_i) associés à deux grandeurs x et y. Nous souhaitons trouver une droite qui soit une « bonne représentation » de ces couples.

FIGURE 2.1 – Exemple de droite de régression pour 6 points.

Nous supposons qu'il existe une relation *linéaire* entre x et y de sorte que l'on cherche les coefficients a et b vérifiant

$$y_i \simeq ax_i + b, \quad 1 \leq i \leq n.$$

2.1.1 Principe de la méthode

Il est clair que ce système n'admet pas de solution puisqu'il est sur-déterminé (lorsque $n > 2$). Nous allons ainsi tolérer un écart, appelé *résidu*, pour chaque valeur y_i ce qui nous conduit à estimer les coefficients a et b satisfaisant

$$y_i = ax_i + b + r_i, \quad 1 \leq i \leq n \tag{2.1}$$

La quantité $\widehat{y}_i \overset{\text{def}}{=} ax_i + b$ désigne la *prédiction linéaire* de y en x_i et le système (**??**) admet plusieurs solutions car il est sous-déterminé ($n + 2$ inconnues et n équations). La méthode des moindres carrés consiste à minimiser la norme 2 du résidu définie par

$$||r||_2^2 \overset{\text{def}}{=} \sum_{i=1}^{n} r_i^2 \overset{(??)}{=} \sum_{i=1}^{n} (y_i - (ax_i + b))^2,$$

2.1.2 Expression du résidu

L'ensemble des équations décrit par la relation (**??**) s'écrit matriciellement

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} + \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \quad \Leftrightarrow \quad y = Mp + r \tag{2.2}$$

en posant M matrice de Vandermonde de dimension $n \times 2$ de terme général $m_{i,j} \stackrel{\text{def}}{=} x_i^{j-1}$,
$y \stackrel{\text{def}}{=} (y_1, \cdots, y_n)^\top$, $r \stackrel{\text{def}}{=} (r_1, \cdots, r_n)^\top$, $p \stackrel{\text{def}}{=} (b, a)^\top$.

En utilisant que $||r||_2^2 = r^\top r$, l'équation (**??**) permet de reformuler la norme 2 du résidu,

$$\begin{aligned} r^\top r &\stackrel{(??)}{=} (y - Mp)^\top (y - Mp) \\ &= (y^\top - p^\top M^\top)(y - Mp) \\ &= y^\top y - y^\top Mp - p^\top M^\top y + p^\top M^\top Mp \end{aligned}$$

Comme $y^\top Mp = p^\top M^\top y$ (puisque $y^\top Mp$ est scalaire), l'expression précédente peut se réécrire

$$\boxed{r^\top r = p^\top Qp - 2p^\top s + y^\top y}$$

avec $Q \stackrel{\text{def}}{=} M^\top M$ matrice de dimension 2×2 et $s \stackrel{\text{def}}{=} M^\top y$ vecteur de dimension 2.

2.1.3 Formulation matricielle du problème

Le problème revient ainsi à minimiser la fonction convexe $f(p) \stackrel{\text{def}}{=} p^\top Qp - 2p^\top s + y^\top y$ ce qui est équivalent à annuler son gradient

$$\min_{p \in \mathbb{R}^2} f(p) \;\Leftrightarrow\; \nabla f(p) = 0, \tag{2.3}$$

où $\nabla \stackrel{\text{def}}{=} (\partial_b, \partial_a)^\top$. La linéarité du gradient donne $\nabla f(p) = \nabla(p^\top Qp) - 2\nabla(p^\top s) + \nabla(y^\top y)$ et il faut évaluer les gradients de ces trois termes qui sont respectivement d'ordre deux, un et zéro. En notant q_{ij} le terme général de la matrice Q, nous obtenons

$$p^\top Qp = q_{11}b^2 + 2q_{12}ab + q_{22}a^2 \quad \Rightarrow \quad \nabla(p^\top Qp) = 2Qp \tag{2.4}$$

puisque Q est symétrique par construction. En notant s_1 et s_2 les deux composantes de s, nous avons

$$p^\top s = bs_1 + as_2 \quad \Rightarrow \quad \nabla(p^\top s) = s. \tag{2.5}$$

Le gradient de $y^\top y$ est nul puisqu'il ne dépend pas de a et b,

$$\nabla(y^\top y) = 0. \tag{2.6}$$

Le problème de minimisation aboutit finalement au système d'*équations normales* suivant

$$\boxed{Qp = s,} \tag{2.7}$$

où les expressions générales de Q et s sont

$$Q = \begin{bmatrix} n & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 \end{bmatrix} \quad \text{et} \quad s = \begin{bmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} y_k x_k \end{bmatrix}.$$

En revenant aux notations initiales M et y, les équations normales s'écrivent

$$M^\top M p = M^\top y \quad \Leftrightarrow \quad p = \left(M^\top M\right)^{-1} M^\top y, \tag{2.8}$$

où $M^\dagger \stackrel{\text{def}}{=} (M^\top M)^{-1} M^\top$ est la *pseudo-inverse* de M.

2.1.4 Expression analytique des coefficients

Il est possible d'obtenir une expression analytique des coefficients a et b définissant la droite de régression. La première équation normale s'écrit

$$b + a\bar{x} = \bar{y}, \tag{2.9}$$

où $\bar{\cdot}$ désigne l'opérateur de moyenne empirique $\bar{x} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k$. En injectant cette expression de b dans la première équation normale, on obtient

$$n(\bar{y} - a\bar{x}) + a\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k \quad \Rightarrow \quad \sum_{k=1}^{n} (y_k - \bar{y} - a(x_k - \bar{x})) = 0. \tag{2.10}$$

En utilisant la seconde équation normale, on obtient

$$\sum_{k=1}^{n} x_k (y_k - a x_k - b) = 0 \quad \stackrel{(??)}{\Rightarrow} \quad \sum_{k=1}^{n} x_k (y_k - \bar{y} - a(x_k - \bar{x})) = 0 \tag{2.11}$$

En multipliant (**??**) par $\bar{x}$ et en retranchant à (**??**), nous obtenons

$$\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y} - a(x_k - \bar{x})) = 0 \quad \Rightarrow \quad a = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}. \tag{2.12}$$

Les coefficients a et b s'écrivent finalement

$$\boxed{a = \frac{\sigma_{xy}}{\sigma_x^2}} \quad \text{et} \quad \boxed{b = \bar{y} - a\bar{x}} \tag{2.13}$$

où $\sigma_x^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$ désigne la variance de x et $\sigma_{xy} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$ est la covariance entre x et y.

Il est clair que la droite de régression passe par le centre du nuage de points :

$$\widehat{y} = ax + b \quad \stackrel{(??)}{\Rightarrow} \quad \widehat{y} - \bar{y} = a(x - \bar{x}), \tag{2.14}$$

de sorte que $x = \bar{x} \Rightarrow \widehat{y} = \bar{y}$. Bien évidemment, la droite de régression $\widehat{y} = ax + b$ est différente de celle $\widehat{x} = \alpha y + \beta$.

2.1.5 Interprétation géométrique

La méthode des moindres carrés peut s'interpréter comme la projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace engendré par les vecteurs colonnes de M. En effet les relations (**??**) et (**??**) donnent

$$y = M\big(M^\top M\big)^{-1} M^\top y + r, \tag{2.15}$$

de sorte que la prédiction $\widehat{y}$ s'obtient directement à partir de y par la relation suivante

$$\widehat{y} = Hy,$$

où la *hat* matrice $H \stackrel{\text{def}}{=} M\big(M^\top M\big)^{-1} M^\top$ est une *matrice de projection* (puisque $H^2 = H$). L'expression du résidu en fonction de H est

$$r = (I - H)y.$$

2.1.6 Unicité du problème de minimisation

Il convient de préciser la relation d'équivalence (**??**) entre la minimisation de f et l'annulation de son gradient. Le théorème utilisé, valable en optimisation sans contrainte, portant sur les conditions d'optimalité est le suivant

Théorème. *Soit f une fonction convexe et différentiable sur $\mathbb{R}^n$. Une condition nécessaire et suffisante pour que x^* soit un minimum global de f est que x^* soit un point critique de f, c'est-à-dire qu'il vérifie*

$$\nabla f(x^*) = 0.$$

Démonstration. La démonstration (non détaillée ici) requiert trois points. La condition nécessaire utilise qu'un minimum local d'une fonction différentiable annule son gradient (on parle de point critique), et l'équivalence local-global ainsi que la condition suffisante utilisent la convexité de f. □

Nous rappelons les définitions d'ensemble et de fonction convexes.

Définitions. *On dit qu'un ensemble $K \subset \mathbb{R}^n$ est convexe si et seulement si*

$$\forall (x_1, x_2) \in K^2, \quad \forall t \in [0,1], \quad tx_1 + (1-t)x_2 \in K.$$

Une fonction $f : K \subset \mathbb{R}^n \mapsto \mathbb{R}$ est convexe si et seulement si

$$\forall (x_1, x_2) \in K^2, \quad \forall t \in [0,1], \quad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

Dans le cas des moindres carrés, la fonction f est évidemment convexe et l'on utilise la propriété suivante pour la démonstration

Propriété. *Si $f : \mathbb{R}^n \mapsto \mathbb{R}$ est deux fois différentiable, on a l'équivalence suivante*
(i) f est convexe,
(ii) $\forall x \in \mathbb{R}^n, \mathcal{H}(f(x))$ est semi-definie positive, où $\mathcal{H}$ désigne la matrice hessienne de f.

La fonction $f(p)$ est une *forme quadratique* que l'on peut réécrire (au facteur $1/2$ près),

$$f(p) \stackrel{\text{def}}{=} \frac{1}{2}\langle Qp, p\rangle - \langle s, p\rangle + \frac{1}{2}\langle y, y\rangle \quad \Rightarrow \quad \mathcal{H}(f) = Q,$$

où $\langle\cdot\rangle$ désigne le produit scalaire sur $\mathbb{R}^n$. La matrice hessienne Q est clairement semi-définie positive,

$$\forall x \in \mathbb{R}^n, \quad \langle Qx, x\rangle = \langle M^\top Mx, x\rangle = \langle Mx, Mx\rangle \geq 0.$$

2.1.7 Évaluation de la qualité de la régression

Le *coefficient de corrélation linéaire* r_{xy} entre les variables x et y mesure la qualité de l'approximation d'un nuage par la droite de régression,

$$\boxed{r_{xy} \overset{\text{def}}{=} \frac{\sigma_{xy}}{\sigma_x \sigma_y,}} \tag{2.16}$$

où σ_x (resp. σ_y) désigne l'écart-type de x (resp. y) de moyenne $\bar{x}$ (resp. $\bar{y}$) et σ_{xy} représente la covariance entre x et y,

$$\sigma_x \overset{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2} \quad \text{et} \quad \sigma_{xy} \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) \quad \text{avec} \quad \bar{x} \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Ce coefficient mesure la dispersion du nuage de points autour de la droite de régression et varie entre -1 (points alignés sur une droite de pente négative) et 1 (points alignés sur une droite de pente positive). En général, la régression est considérée de bonne qualité lorsque $|r_{xy}|$ est supérieur à $\frac{\sqrt{3}}{2} = 0.86$. Cette valeur découle de l'interprétation géométrique du coefficient de corrélation. En effet, r_{xy} représente le cosinus de l'angle entre les vecteurs centrés $(x_1 - \bar{x}, \cdots, x_n - \bar{x})^\top$ et $(y_1 - \bar{y}, \cdots, y_n - \bar{y})^\top$. Une corrélation est acceptable si cet angle est inférieur à 30.

Il est important de représenter les données et la droite de régression. Le quartet de la figure **??** a été construit par le statisticien Anscombe en 1973. Il comprend quatre ensemble de données ayant les mêmes propriétés statistiques (et la même droite de régression) mais très différents. Le premier ensemble (haut, gauche) est distribué au hasard et les données sont peu corrélées. Les données du deuxième ensemble (haut, droite) sont fortement corrélées non linéairement. Le troisième ensemble (bas, gauche) possède une donnée aberrante faussant la régression linéaire. La quatrième ensemble (bas, droite) montre qu'une donnée aberrante suffit pour obtenir un coefficient de corrélation élevé même si les variables ne sont pas linéairement corrélées.

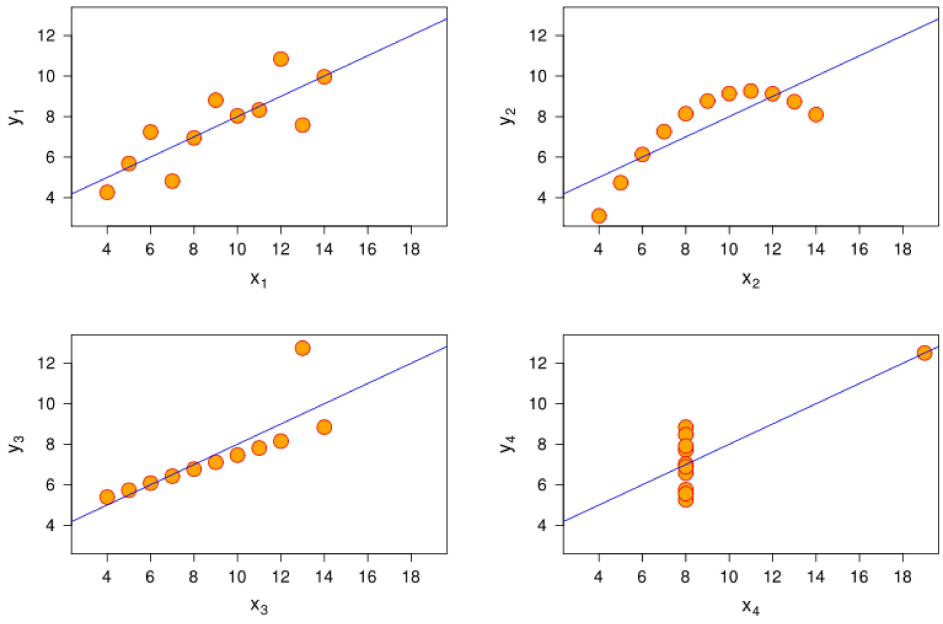


FIGURE 2.2 – Quartet d'Anscombe.

2.2 Cas général de la méthode des moindres carrés

Nous sommes maintenant dans un cadre plus général puisque nous souhaitons trouver un polynôme qui soit une « bonne représentation » d'un nuage de points.

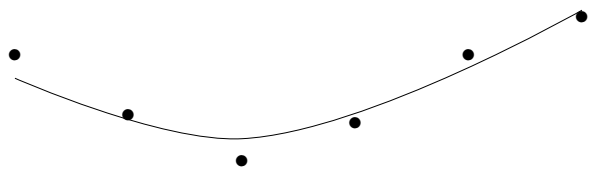


FIGURE 2.3 – Exemple d'un polynôme d'approximation pour 6 points.

Ainsi, nous supposons qu'il existe une relation *polynomiale* de degré d ($d < n$) entre x et y de sorte que l'on cherche les coefficients a_j ($0 \leq j \leq d$) vérifiant

$$y_i \simeq \sum_{j=0}^{d} a_j x_i^j, \quad 1 \leq i \leq n.$$

2.2.1 Principe de la méthode

Ce système n'admet pas de solution puisqu'il est sur-déterminé (lorsque $n > d+1$). Comme dans le cas linéaire, nous considérons une erreur pour chaque valeur y_i,

$$y_i = \sum_{j=0}^{d} a_j x_i^j + r_i, \quad 1 \leq i \leq n \tag{2.17}$$

Ce nouveau système sous-déterminé ($n+d+1$ inconnues et n équations) est résolu en minimisant la norme 2 du résidu

$$||r||_2^2 \overset{(??)}{=} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{d} a_j x_i^j \right)^2.$$

2.2.2 Expression du résidu

L'ensemble des équations décrit par la relation (**??**) s'écrit matriciellement

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1^0 & \dots & x_1^d \\ \vdots & & \vdots \\ x_n^0 & \dots & x_n^d \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_d \end{bmatrix} + \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \quad \Leftrightarrow \quad y = Mp + r \tag{2.18}$$

en posant M matrice de Vandermonde de dimension $n \times (d+1)$ de terme général $m_{i,j} \overset{\text{def}}{=} x_i^{j-1}$,
$y \overset{\text{def}}{=} (y_1, \cdots, y_n)^\top$, $r \overset{\text{def}}{=} (r_1, \cdots, r_n)^\top$, $p \overset{\text{def}}{=} (a_0, \cdots, a_d)^\top$.

Comme dans le cas linéaire, la norme 2 du résidu peut s'écrire

$$\boxed{r^\top r = p^\top Q p - 2p^\top S + y^\top y}$$

avec $Q \overset{\text{def}}{=} M^\top M$ matrice de dimension $(d+1) \times (d+1)$ et $S \overset{\text{def}}{=} M^\top y$ vecteur de dimension $(d+1)$.

2.2.3 Formulation matricielle du problème

La minimisation la fonction f, définie précédemment, revient à annuler son gradient

$$\nabla f = \nabla(p^\top Q p) - 2\nabla(p^\top s) + \nabla(y^\top y) = 0,$$

où $\nabla \stackrel{\text{def}}{=} (\partial_{a_0}, \ldots, \partial_{a_d})^\top$. Il reste maintenant à évaluer les gradients des deux premiers termes puisque le troisième est indépendant des coefficients a_j. En notant q_{ij} le terme général de la matrice Q, nous obtenons compte-tenu de la symétrie de la cette matrice,

$$p^\top Q p = \sum_{k=0}^{d} a_k \Big(\sum_{j=0}^{d} q_{k+1j+1} a_j \Big) \;\Rightarrow\; \partial_{a_i}(p^\top Q p) = 2 \sum_{j=0}^{d} a_j q_{i+1j+1} \;\Rightarrow\; \nabla(p^\top Q p) = 2Qp$$

En notant s_i le terme général de s, nous avons

$$p^\top s = \sum_{i=0}^{d} a_i s_{i+1} \quad \Rightarrow \quad \nabla(p^\top s) = s.$$

Le problème de minimisation aboutit finalement aux équations normales

$$Qp = s,$$

où les expressions générales de Q et s sont

$$Q = \begin{bmatrix} n & \cdots & \sum_{k=1}^{n} x_k^d \\ \vdots & \sum_{k=1}^{n} x_k^{i+j-2} & \vdots \\ \sum_{k=1}^{n} x_k^d & \cdots & \sum_{k=1}^{n} x_k^{2d} \end{bmatrix} \quad \text{et} \quad s = \begin{bmatrix} \sum_{k=1}^{n} y_k x_k^0 \\ \vdots \\ \sum_{k=1}^{n} y_k x_k^d \end{bmatrix}.$$

Chapitre 3

Interpolation et intégration

L'interpolation polynomiale consiste à trouver un polynôme passant par un certain nombre de points. L'intérêt d'obtenir ce polynôme est la synthétisation de données expérimentales dont le nombre peut être élevé. Une autre application intéressante est le remplacement d'une fonction analytique par un polynôme plus simple en vue d'un calcul numérique.

L'intégration numérique consiste à approcher des intégrales de fonctions dont les primitives ne sont pas connues analytiquement.

3.1 Interpolation

3.1.1 Motivation

Nous souhaitons trouver le *polynôme d'interpolation* P_n de degré n associé aux $n+1$ couples (x_i, y_i), c'est-à-dire vérifiant

$$P_n(x_i) = y_i, \quad 0 \leq i \leq n \tag{3.1}$$

Ces $n+1$ égalités s'appellent les *contraintes d'interpolation.*

FIGURE 3.1 – Exemple de polynôme d'interpolation P_5 passant par 6 points.

Nous désignons par $\mathbb{P}^n[X]$ l'ensemble des polynômes de degré n.

Théorème d'unicité. *Soient $n+1$ couples de points (x_i, y_i), il existe un unique polynôme $P_n \in \mathbb{P}^n[X]$ tel que*

$$P_n(x_i) = y_i, \quad \forall\ 0 \leq i \leq n$$

Démonstration. La forme générale de P_n est $P_n(x) = \sum_{k=0}^{n} a_k x^k$ de sorte que les contraintes d'interpolation (**??**) s'écrivent

$$\sum_{k=0}^{n} a_k x_i^k = y_i, \quad 0 \leq i \leq n$$

ce qui permet d'obtenir le système matriciel suivant

$$\begin{bmatrix} x_0^0 & \cdots & x_0^n \\ \vdots & & \vdots \\ x_n^0 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}.$$

La détermination de l'ensemble des coefficients du polynôme d'interpolation $\{a_k\}, 0 \leq k \leq n$ nécessite donc la résolution d'un système matriciel dans lequel la matrice est inversible si les x_i sont distincts puisque le déterminant de la matrice de Vandermonde M_V est

$$det(M_V) = \prod_{0 \leq j < i \leq n} (x_i - x_j).$$

□

3.1.2 Méthode de Lagrange

Polynômes de Lagrange. *Soient $n+1$ réels distincts notés $\{x_i\}_{0 \leq i \leq n}$. Les $n+1$ polynômes de Lagrange associés à ces points sont définis par*

$$\boxed{L_i(x) \stackrel{\text{def}}{=} \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}}, \quad 0 \leq i \leq n \tag{3.2}$$

Les propriétés suivantes sont vérifiées :
- le numérateur est un polynôme de degré n,
- le dénominateur est une constante,
- $L_i(x_j) = \delta_{ij}$,
- $\{L_i(x)\}_{0 \leq i \leq n}$ forment une base de $\mathbb{P}^n[X]$.

Pour démontrer que les polynômes de Lagrange forment une base de $\mathbb{P}^n[X]$, il suffit de montrer que ce système de polynômes est libre puisqu'il est formé de $n+1$ éléments dans un espace de dimension $n+1$. Pour cela, nous écrivons simplement

$$\forall 0 \leq j \leq n, \quad \sum_{i=1}^{n} \alpha_i L_i(x_j) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i \delta_{ij} = 0 \quad \Rightarrow \quad \alpha_j = 0. \tag{3.3}$$

Formule de Lagrange. *Le polynôme d'interpolation est défini à partir des polynômes de Lagrange L_i selon la relation suivante*

$$\boxed{P_n(x) = \sum_{i=0}^{n} y_i L_i(x)} \tag{3.4}$$

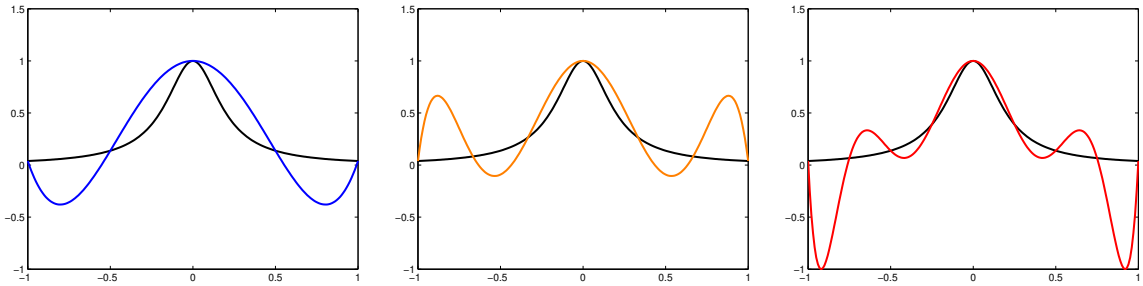
La démonstration est triviale et repose sur la relation $L_i(x_j) = \delta_{ij}$. Bien qu'intuitive, cette relation n'est pas flexible puisqu'il est nécessaire de recalculer l'ensemble des polynômes de Lagrange lorsque l'on ajoute un couple de points à interpoler. L'interpolation basée sur les polynômes de Newton s'affranchit de cette contrainte.

3.1.3 Phénomène de Runge

On peut penser de prime abord que le polynôme d'interpolation va approcher de mieux en mieux la fonction f lorsque le nombre de points augmente mais la fonction de Runge est un contre-exemple célèbre. Cette fonction définie par

$$f(x) = \frac{1}{1+25x^2}, \tag{3.5}$$

montre que l'écart maximal entre la fonction et son polynôme d'interpolation augmente indéfiniment avec n comme l'illustre la figure **??**. Les oscillations observées s'expliquent par le fait que le terme de dérivée n+unième ($f^{n+1}(c)$) dans l'erreur d'interpolation (voir le théorème) n'est pas borné en général. En d'autres termes, la suite des polynômes d'interpolation converge uniformément vers f sous l'hypothèse de régularité restrictive $f \in \mathcal{C}^\infty([a,b])$ qui est rarement vérifiée en pratique. Pour pallier cette difficulté, on peut augmenter la densité des points d'interpolation proche des bords de l'intervalle en utilisant les nœuds de Tchebychev. On peut aussi choisir l'interpolation par morceaux de façon linéaire sous chaque sous-intervalle ou avec des splines pour être dérivable.



(a) $n = 5$ (b) $n = 7$ (c) $n = 9$

FIGURE 3.2 – Interpolation de la fonction de Runge (en noir) avec noeuds équirépartis.

Théorème sur l'erreur d'interpolation. *Soit $f \in \mathcal{C}^{n+1}([a,b])$, alors $\forall x \in [a,b]$, il existe un réel $c \in]a,b[$ tel que*

$$f(x) - P_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

Démonstration. La démonstration repose sur l'application successive du théorème de Rolle sur les sous-intervalles $[x_{i-1}, x_i]$. □

3.2 Intégration numérique

3.2.1 Principe

Soit $I(f)$ l'intégrale d'une fonction f continue sur l'intervalle $[a,b]$,

$$I(f) \stackrel{\text{def}}{=} \int_a^b f(x)dx. \tag{3.6}$$

Nous souhaitons approcher cette intégrale par le nombre $\tilde{I}_n(f) \simeq I(f)$,

$$\boxed{\tilde{I}_n(f) = \sum_{i=1}^{n} w_i f(x_i).} \tag{3.7}$$

Cette relation est une *formule quadrature* dans laquelle
- les x_i sont les *points d'intégration*,
- les w_i sont les *poids d'intégration*,
- n désigne le nombre de points de la quadrature.

3.2.2 Théorème de convergence

Théorème de convergence. *Soient $[a,b]$ un intervalle fermé borné de $\mathbb{R}$ et $f \in \mathcal{C}([a,b])$. Nous considérons la formule de quadrature*

$$\tilde{I}_n(f) = \sum_{i=1}^{n} w_i f(x_i),$$

vérifiant les deux propriétés

1. $\exists M > 0$ *tel que* $\forall n,\ \sum_{i=1}^{n} |w_i| \leq M$,
2. *Pour tout polynôme* p, $\lim_{n\to\infty} \tilde{I}_n(p) = I(p)$,

Alors la formule de quadrature converge,

$$\lim_{n\to\infty} \tilde{I}_n(f) = I(f)$$

Démonstration. D'après le théorème d'approximation de Weierstrass,

$$\forall \epsilon > 0, \quad \exists p \in \mathbb{P}[X], \quad \max_{x\in[a,b]} |f(x) - p(x)| \leq \epsilon.$$

Compte-tenu de la linéarité de I et de $\tilde{I}_n$, la différence entre l'intégrale et la formule de quadrature se réécrit

$$\begin{aligned} I(f) - \tilde{I}_n(f) &= I(f) - I(p) + I(p) - \tilde{I}_n(p) + \tilde{I}_n(p) - \tilde{I}_n(f) \\ &= I(f-p) + I(p) - \tilde{I}_n(p) + \tilde{I}_n(p-f). \end{aligned}$$

ce qui permet d'obtenir l'inégalité suivante

$$|I(f) - \tilde{I}_n(f)| \leq |I(f-p)| + |I(p) - \tilde{I}_n(p)| + |\tilde{I}_n(p-f)|.$$

Les majorations des trois termes sont

1. $|I(f-p)| \leq I(|f-p|) \leq (b-a)\epsilon$ d'après le théorème de Weierstrass,
2. $\exists N(\epsilon)/\ \forall n \geq N,\ |I(p) - \tilde{I}_n(p)| \leq \epsilon$ d'après la propriété 2,
3. $|\tilde{I}_n(p-f)| \leq \sum_{i=1}^{n} |w_i||p(x_i) - f(x_i)| \leq M\epsilon$ d'après la propriété 1.

On obtient finalement $\forall n \geq N(\epsilon),\ |I(f) - \tilde{I}_n(f)| \leq (b-a+1+M)\epsilon$. □

3.2.3 Ordre d'une formule de quadrature

Une formule de quadrature est d'orde d si elle est exacte pour les polynômes de degré d,

$$\forall p \in \mathbb{P}^d[X], \quad \boxed{\tilde{I}_n(p) = I(p).} \tag{3.8}$$

Théorème. *Une formule de quadrature a un ordre d si et seulement si*

$$\forall 1 \leq q \leq d+1, \quad \sum_{i=1}^{n} w_i x_i^{q-1} = \frac{b^q - a^q}{q} \tag{3.9}$$

Démonstration. $\Rightarrow$ Une formule d'ordre d est exacte pour les polynômes de degré maximal d et en prenant $p(x) = x^{q-1}$, nous obtenons

$$\sum_{i=1}^{n} w_i x_i^{q-1} = \int_a^b x^{q-1} dx = \frac{b^q - a^q}{q}.$$

$\Leftarrow$ Soit $p(x) = \sum_{k=0}^{d} a_k x^k$ un polynôme de degré d dont l'intégration numérique est

$$\tilde{I}_n(p) = \sum_{i=1}^{n} w_i p(x_i) = \sum_{i=1}^{n} w_i \sum_{k=0}^{d} a_k x_i^k = \sum_{k=0}^{d} a_k \sum_{i=1}^{n} w_i x_i^k = \sum_{k=0}^{d} a_k \left(\frac{b^{k+1} - a^{k+1}}{k+1} \right) = I(p).$$

□

En fixant les points d'intégration, les conditions données par la relation (**??**) (pour $d+1 = n$) permettent d'obtenir les poids d'intégration en résolvant le système matriciel suivant

$$\begin{bmatrix} x_1^0 & \cdots & x_n^0 \\ \vdots & & \vdots \\ x_1^{n-1} & \cdots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{avec} \quad y_i \stackrel{\text{def}}{=} \frac{b^i - a^i}{i}. \tag{3.10}$$

3.2.4 Changement de variable

Les formules de quadrature sont données sur les intervalles $[0,1]$ et $[-1,1]$ et il est souvent nécessaire d'effectuer un changement de variable pour se ramener sur l'un de ces intervalles.

Théorème de changement de variable. *Soit f une fonction continue et φ une fonction $\mathcal{C}_1$ sur un intervalle $[\alpha, \beta]$ et dont l'image est contenue dans le domaine de définition de f. Alors*

$$\boxed{\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx = \int_{\alpha}^{\beta} f \circ \varphi(t)\varphi'(t)dt}$$

Nous en déduisons les deux relations suivantes

$$\int_a^b f(x)dx = (b-a)\int_0^1 f \circ \varphi(t)dt \quad \text{avec} \quad \varphi(t) = (b-a)t + a,$$
$$\int_a^b f(x)dx = \frac{(b-a)}{2}\int_{-1}^1 f \circ \varphi(t)dt \quad \text{avec} \quad \varphi(t) = \frac{(b-a)}{2}t + \frac{b+a}{2}.$$

3.2.5 Quadratures interpolantes

Les quadratures interpolantes exploitent le théorème de convergence précédent qui suggère d'intégrer un polynôme d'interpolation P_n de la fonction f pour en approcher son intégrale. Ces quadratures reposent sur l'approximation suivante

$$\boxed{I(f) \simeq I(P_n)}$$

Dans ces quadratures, les points d'intégration sont répartis uniformément sur l'intervalle d'intégration et les poids sont obtenus en intégrant le polynôme d'interpolation.
L'intégration du polynôme d'interpolation constant donne *la formule du rectangle*,

$$\left.\begin{array}{r} x_1 = \dfrac{a+b}{2} \\ w_1 = b-a \end{array}\right\} \quad \Rightarrow \quad I(f) \simeq (b-a) f\left(\frac{a+b}{2}\right).$$

L'intégration du polynôme d'interpolation affine donne *la formule du trapèze*,

$$\left.\begin{array}{r} x_1 = a,\ x_2 = b \\ w_1 = w_2 = \dfrac{b-a}{2} \end{array}\right\} \quad \Rightarrow \quad I(f) \simeq \frac{(b-a)}{2}(f(a)+f(b)).$$

L'intégration du polynôme d'interpolation parabolique donne *la formule de Simpson*,

$$\left.\begin{array}{r} x_1 = a,\ x_2 = \dfrac{a+b}{2},\ x_3 = b \\ w_1 = w_3 = \dfrac{b-a}{6},\ w_2 = \dfrac{2(b-a)}{3} \end{array}\right\} \quad \Rightarrow \quad I(f) \simeq \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

La figure **??** représente schématiquement les trois formules précédentes qui sont respectivement d'ordre 1, 1 et 3.

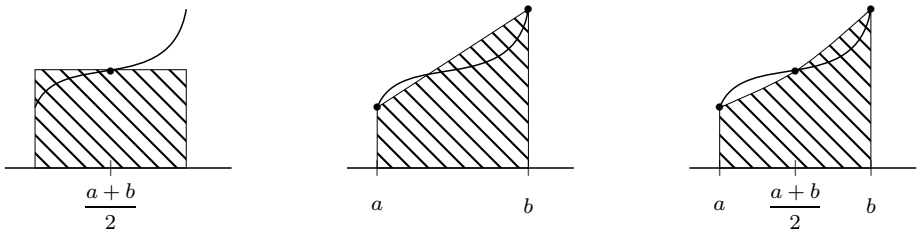


FIGURE 3.3 – Illustration des méthodes du rectangle, du trapèze et de Simpson.

Dans le cas général, nous avons la formule suivante

Formule de Newton–Cotes. *Soit un ensemble de n points distincts $\{x_i\}, 1 \leq i \leq n$ dans $[a,b]$. L'intégrale sur $[a,b]$ du polynôme d'interpolation P_{n-1} de $f \in \mathcal{C}([a,b])$ associé à ces points est*

$$I(P_{n-1}) = \sum_{i=1}^{n} w_i f(x_i) \quad \textit{avec} \quad w_i \stackrel{\text{def}}{=} \int_a^b L_i(x)dx,$$

où $L_i(x)$ désigne le i-ème polynôme de Lagrange associé aux x_i.

Démonstration. Le polynôme d'interpolation s'écrit dans la base des polynômes de Lagrange

$$P_{n-1}(x) = \sum_{i=1}^{n} f(x_i) L_i(x) \quad \Rightarrow \quad \int_a^b P_{n-1}(x)dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)dx.$$

□

Notons qu'une quadrature de type Newton-Cotes a un ordre supérieur ou égal au nombre de points n intervenant dans cette quadrature.

3.2.6 Quadratures de Gauss

Les quadratures interpolantes fixent *a priori* les points d'intégration et déterminent ensuite la valeurs des poids par l'intégration des polynômes de Lagrange ou la résolution du système (**??**). Les quadratures de Gauss imposent la répartition des points d'intégration pour avoir ***l'ordre d'intégration maximal***.

Théorème. *L'ordre d'une formule de quadrature à n points est au plus $2n-1$.*

L'optimalité des formules de Gauss implique des conditions d'orthogonalité satisfaites par les polynômes de Legendre définis par la relation de récurrence suivante,

$$\forall k > 1, \quad \boxed{(k+1)P_{k+1}(x) = (2k+1)xP_n(x) - kP_{k-1}(x)} \quad \text{avec} \quad P_0(x) = 0 \text{ et } P_1(x) = x.$$

Théorème. *Pour chaque entier positif n, il existe une formule de quadrature unique à n points d'ordre $2n-1$:*
- *les points x_i sont les racines de $P_k(2x-1)$.*
- *les poids w_i sont données par le système (**??**) .*

Les trois premières formules de quadrature de Gauss définies sur $[0,1]$ sont indiquées dans le tableau ci-dessous.

n	d	Poids	Points
1	1	1	$1/2$
2	3	$1/2, 1/2$	$1/2 - \sqrt{3}/6,\ 1/2 + \sqrt{3}/6$
3	5	$5/18, 4/9, 5/18$	$1/2 - \sqrt{15}/10, 1/2, 1/2 + \sqrt{15}/10$

TABLE 3.1 – Formules de quadrature de Gauss.

3.2.7 Formules composites

Les formules composites sont des formules d'intégration basées sur un découpage de l'intervalle d'intégration $[a,b]$ en m sous-intervalles de longueur égale h. Les points de cette subdivision uniforme représentés à la figure (**??**) sont notés S_h,

$$S_h \stackrel{\text{def}}{=} \left\{ x_{k+1} = a + kh, \quad 0 \le k \le m, \quad h \stackrel{\text{def}}{=} (b-a)/m \right\}.$$

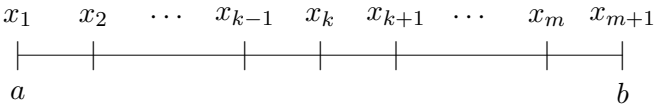


FIGURE 3.4 – Subdivision uniforme de $[a, b]$.

Le principe des formules composites repose sur la relation de Chasles,

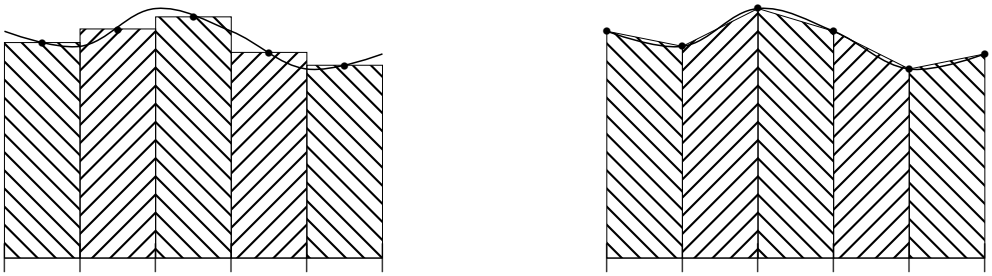
$$I(f) = \int_a^b f(x)dx = \sum_{k=1}^{m} \int_{x_k}^{x_{k+1}} f(x)dx,$$

dans laquelle on utilise l'une des formules d'intégration précédentes sur chaque sous-intervalle. La formule du rectangle sur chaque sous-intervalle donne la *formule composite du rectangle*,

$$I(f) \simeq h \sum_{k=1}^{m} f\left(x_{k+\frac{1}{2}}\right), \quad x_{k+\frac{1}{2}} \stackrel{\text{def}}{=} \frac{x_k + x_{k+1}}{2}, \quad 1 \leq k \leq m.$$

La formule du trapèze sur chaque sous-intervalle donne la *formule composite du trapèze*,

$$I(f) \simeq \frac{h}{2} \sum_{k=1}^{m} (f(x_k) + f(x_{k+1})).$$



Méthode des rectangles Méthode des trapèzes

FIGURE 3.5 – Illustration de formules composites.

Chapitre 4

Résolution des EDP par la méthode des différences finies

Les équations aux dérivées partielles (EDP) modélisent de nombreux phénonèmes dans des domaines variés de la physique (mécaniques des solides et des fluides, électromagnétisme, acoustique, thermodynamique, mécanique quantique, chimie, ...) mais aussi de la biologie et de la finance. La méthode des différences finies permet d'approcher les solutions des EDP qui ne sont jamais connues dans les cas réels d'études. Cette méthode conduit à remplacer le problème continu initial (de dimension infinie) par un problème discret (de dimension finie) en approchant les opérateurs par des quotients convenablement choisis. Nous effectuons la présentation et l'analyse sur le laplacien mais l'extension à d'autres opérateurs est possible.

4.1 Introduction

4.1.1 Quelques définitions

Une *équation aux dérivées partielles* (EDP) est une équation faisant intervenir une fonction inconnue de plusieurs variables (temps et espace) ainsi que certaines de ses dérivées. L'application construite à partir des dérivées partielles s'appelle l'*opérateur* et une EDP peut ainsi s'écrire formellement

$$\mathcal{L}(u) = f, \tag{4.1}$$

où $\mathcal{L}$ est l'opérateur, u la fonction inconnue et f le second membre.

Une EDP est
- *linéaire* si $\mathcal{L}$ est linéaire par rapport aux dérivées partielles de la fonction inconnue,
- *d'ordre* d, où d désigne l'ordre de la plus grande dérivée de $\mathcal{L}$ (on distingue l'ordre en espace et l'ordre en temps),
- *instationnaire* si $\mathcal{L}$ dépend du temps,
- *homogène* lorsque son second membre est nul.

Une EDP stationnaire est
- posée sur un *domaine d'espace* $\Omega \subset \mathbb{R}^d$ où d est la dimension de l'espace $(1 \leq d \leq 3)$,
- complétée par une *condition aux limites* g imposée sur la *frontière* $\partial\Omega = \overline{\Omega} \setminus \Omega$ de Ω,

$$\begin{cases} \mathcal{L}\left(u(x)\right) = f & \text{dans } \Omega, \\ u(x) = g & \text{sur } \partial\Omega. \end{cases}$$

Une EDP instationnaire est
— posée sur $\Omega \times]0,T]$ où $]0,T]$ est le *domaine temporel* et T désigne le temps final d'étude,
— complétée par une *condition aux limites* g imposée sur la *frontière* $\partial\Omega$ de Ω,
— complétée par une *condition initiale* u^0 imposée à l'instant initial,

$$\begin{cases} \mathcal{L}\left(u(x,t)\right) = f & \text{dans } \Omega \times [0,T], \\ u(x,t) = g & \text{sur } \partial\Omega \times [0,T], \\ u(x,0) = u^0 & \text{dans } \Omega \times \{0\}. \end{cases}$$

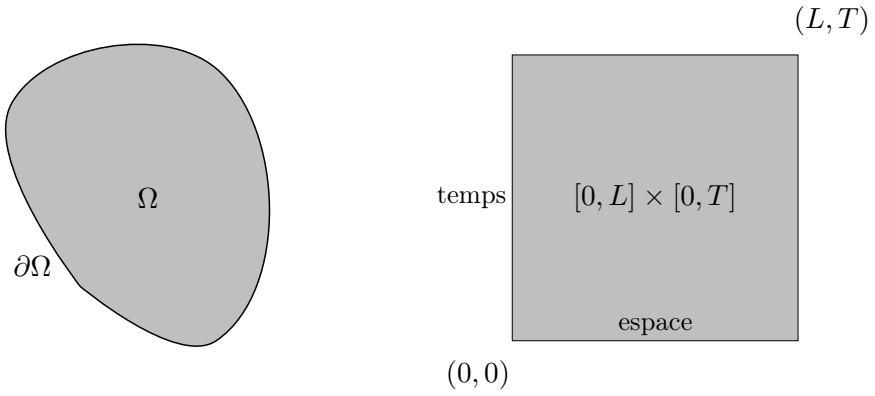


FIGURE 4.1 – Domaines associés à une EDP stationnaire 2d (gauche) et instationnaire 1d (droite).

4.1.2 Exemples d'EDP

4.1.2.1 Equation de Laplace

La conservation de la masse pour un fluide incompressible s'écrit

$$div(v) = 0.$$

Lorsque l'écoulement est irrotationnel, le champ de vitesse v dérive d'un potentiel φ,

$$rot(v) = 0 \quad \Leftrightarrow \quad \exists\, \varphi, \quad v = \nabla\varphi$$

Ce potentiel est une fonction harmonique (*i.e.* dont le Laplacien est nul) puisque

$$div(\nabla\varphi) = \boxed{\Delta\varphi = 0} \tag{4.2}$$

4.1.2.2 Equation de la chaleur

La conservation de l'énergie sans thermogénèse, et lorsque $v = 0$, s'écrit

$$\frac{\partial(\rho e)}{\partial t} = -div(q),$$

où ρ désigne la masse volumique du milieu considéré, e l'énergie et q le flux thermique. Considérons la masse volumique constante et supposons qu'il n'y a pas de transformation chimique ou physique de la matière ce qui permet de vérifier la relation $e = c_{\mathrm{p}}T$ où c_{p} est la capacité thermique massique et T la température. Par ailleurs, la loi de Fourier s'écrit $q = -\lambda\nabla T$, où λ désigne la conductivité thermique. La température vérifie l'équation de la chaleur,

$$\rho c_{\mathrm{p}}\frac{\partial T}{\partial t} = -div(-\lambda\nabla T) \quad \Leftrightarrow \quad \boxed{\frac{\partial T}{\partial t} - D\Delta T = 0} \tag{4.3}$$

où $D \stackrel{\text{def}}{=} \dfrac{\lambda}{\rho c_{\mathrm{p}}}$ est la diffusivité thermique.

4.1.2.3 Equation de Navier–Stokes

L'équation de conservation de la quantité de mouvement s'écrit

$$\rho\frac{dv}{dt} = \rho F + div(\sigma),$$

où ρ désigne la masse volumique du milieu considéré, v le champ de vitesse, F les forces extérieures et σ le tenseur des contraintes. Pour les fluides newtoniens incompressibles, σ s'exprime uniquement en fonction de la pression p et du tenseur des taux de déformations $D \stackrel{\text{def}}{=} \frac{1}{2}\big(\nabla v + (\nabla v)^{\top}\big)$,

$$\sigma = -pI + 2\mu D \quad \Rightarrow \quad div(\sigma) = -\nabla p + \mu\Delta v.$$

L'équation d'équilibre pour un fluide newtonien incompressible devient alors

$$\rho\frac{dv}{dt} = \rho F - \nabla p + \mu\Delta v.$$

En divisant cette équation par la masse volumique ρ et en explicitant la dérivée particulaire de la vitesse, nous obtenons l'équation de Navier–Stokes,

$$\boxed{\frac{\partial v}{\partial t} + v\cdot\nabla v + \frac{1}{\rho}\nabla p = F + \nu\Delta v} \tag{4.4}$$

dans laquelle $\nu \stackrel{\text{def}}{=} \mu/\rho$ s'appelle la viscosité cinématique.

4.1.2.4 Equation des ondes acoustiques

Considérons la linéarisation de la masse volumique, de la pression et de la vitesse :

$$p = p_0 + \alpha p_1, \quad \rho = \rho_0 + \beta\rho_1, \quad \text{et} \quad v = \gamma v_1,$$

avec $\alpha, \beta, \gamma \ll 1$. L'équation linéarisée de la conservation de la masse s'écrit

$$\frac{\partial \rho_1}{\partial t} + \rho_0 div(v_1) = 0.$$

L'équation linéarisée de la conservation de la quantité de mouvement en l'absence de forces extérieures est

$$\frac{\partial v_1}{\partial t} = -\frac{1}{\rho_0}\nabla p_1.$$

La linéarisation de la condition d'adiabaticité est

$$\frac{dp}{d\rho} = \frac{1}{\rho\chi} \quad \Rightarrow \quad \frac{p_1}{\rho_1} = \frac{1}{\rho_0\chi} \quad \Rightarrow \quad \rho_1 = p_1\rho_0\chi$$

où χ désigne la compressibilité adiabatique. La combinaison de la dérivée temporelle de la première équation avec la divergence de la deuxième devient, compte-tenu de l'expression de la masse volumique,

$$\boxed{\frac{\partial^2 p_1}{\partial t^2} - \frac{1}{\rho_0\chi}\Delta p_1 = 0} \tag{4.5}$$

4.2 Opérateurs aux différences finies

Nous rappelons la formule de Taylor qui permet d'obtenir des approximations des opérateurs différentiels. Des exemples sont donnés pour approcher la dérivée première, le laplacien et le bilaplacien. Les notions de stencil et d'ordre de consistance d'un opérateur discret sont précisées.

4.2.1 Formule de Taylor

Formule de Taylor–Young. *Soit f définie sur un intervalle $I = [a, b]$. Si $f \in \mathcal{C}^{n+1}(I)$ alors $\forall h \in \mathbb{R}$ tel que $x + h \in I$,*

$$f(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + o(h^n).$$

La démonstration repose sur la formule de Taylor avec reste intégral (démontrée par récurrence).

4.2.2 Dérivée

La dérivée première peut être approchée par l'opérateur décentré à droite $\mathcal{D}_h^+$, l'opérateur décentré à gauche $\mathcal{D}_h^-$ ou l'opérateur centré $\mathcal{D}_h$ définis respectivement par

$$\boxed{\mathcal{D}_h^+ f(x) \stackrel{\text{def}}{=} \frac{f(x+h) - f(x)}{h}} \tag{4.6}$$

$$\boxed{\mathcal{D}_h^- f(x) \stackrel{\text{def}}{=} \frac{f(x) - f(x-h)}{h}} \tag{4.7}$$

$$\boxed{\mathcal{D}_h f(x) \stackrel{\text{def}}{=} \frac{f(x+h) - f(x-h)}{2h}} \tag{4.8}$$

Ces expressions sont trivialement obtenues grâce à la formule de Taylor. On observe que l'opérateur centré est égal à la moyenne des opérateurs décentrés à droite et à gauche,

$$\mathcal{D}_h f(x) = \frac{1}{2}\left(\mathcal{D}_h^+ + \mathcal{D}_h^-\right) f(x). \tag{4.9}$$

Il est possible de construire des approximations plus précises en augmentant le nombre de points. La formule de Taylor donne

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6} f^{(3)}(x) - \frac{h^4}{5!} f^{(5)}(x) + o(h^5), \tag{4.10}$$

$$f'(x) = \frac{f(x+2h) - f(x-2h)}{4h} - \frac{2h^2}{3} f^{(3)}(x) - \frac{2h^4}{15} f^{(5)}(x) + o(h^5). \tag{4.11}$$

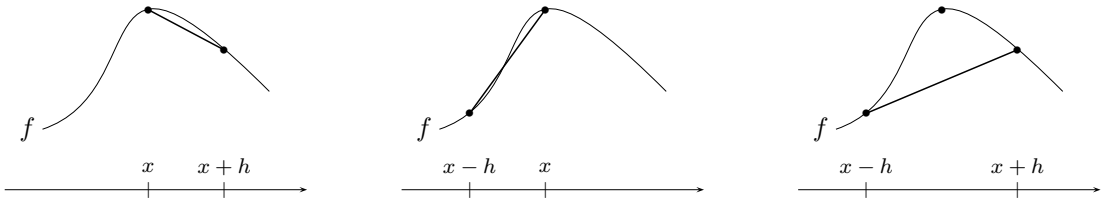


FIGURE 4.2 – Interprétations géométriques de $\mathcal{D}_h^+ f(x)$, $\mathcal{D}_h^- f(x)$ et $\mathcal{D}_h f(x)$.

Toute combinaison linéaire (consistante) α(**??**)+$(1-\alpha)$(**??**) est une approximation de la dérivée, le réel α étant choisi pour annuler le terme d'ordre 3 : $[\alpha + 4(1-\alpha)]\frac{h^2}{6} f^{(3)}(x) = 0 \;\Rightarrow\; \alpha = \frac{4}{3}$. On obtient ainsi l'opérateur centré d'ordre 4 suivant

$$\mathcal{D}_h^* f(x) \stackrel{\text{def}}{=} \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h} \tag{4.12}$$

4.2.3 Dérivées d'ordre supérieur

On peut également approcher par des différences finies les dérivées d'ordre supérieur. Concernant la dérivée seconde, l'opérateur classiquement utilisé est

$$\boxed{\mathcal{D}_h^{(2)} f(x) \stackrel{\text{def}}{=} \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}} \tag{4.13}$$

Cet opérateur centré peut être obtenu en composant les différents opérateurs approchant la dérivée première,

$$\mathcal{D}_h^{(2)} f(x) = \mathcal{D}_{h/2} \circ \mathcal{D}_{h/2} f(x) = \mathcal{D}_h^- \circ \mathcal{D}_h^+ f(x) = \mathcal{D}_h^+ \circ \mathcal{D}_h^- f(x).$$

De même, une approximation de la dérivée quatrième est

$$\boxed{\mathcal{D}_h^{(4)} f(x) \stackrel{\text{def}}{=} \frac{f(x-2h) - 4f(x-h) + 6f(x) - 4f(x+h) + f(x+2h)}{h^4}} \tag{4.14}$$

4.2.4 Représentation symbolique

Un opérateur discret est entièrement caractérisé par son *stencil*, ses *coefficients* et son terme de *scaling*. Le stencil est l'ensemble des positions des valeurs discrètes définissant l'opérateur, les coefficients pondérent les valeurs discrètes indépendamment de h et le terme de scaling est le coefficient multiplicatif lié à h. Ce terme est $1/h^d$ où d est l'ordre de l'opérateur continu ($d = 1$ pour le gradient, la divergence et le rotationnel, $d = 2$ pour le laplacien, $d = 4$ pour le bilaplacien). Les représentations symboliques des opérateurs introduits précédemment sont indiquées à la figure **??**.

4.2.5 Ordre de consistance

Définition. *L'opérateur $\mathcal{L}_h$ est une approximation consistante d'ordre p de l'opérateur $\mathcal{L}$ si pour toute fonction φ suffisamment régulière,*

$$\boxed{|\mathcal{L}(\varphi) - \mathcal{L}_h(\varphi)| \leq C h^p,} \tag{4.15}$$

avec C constante indépendante de h. $\mathcal{L}_h$ approche d'autant mieux $\mathcal{L}$ que son ordre est élevé.

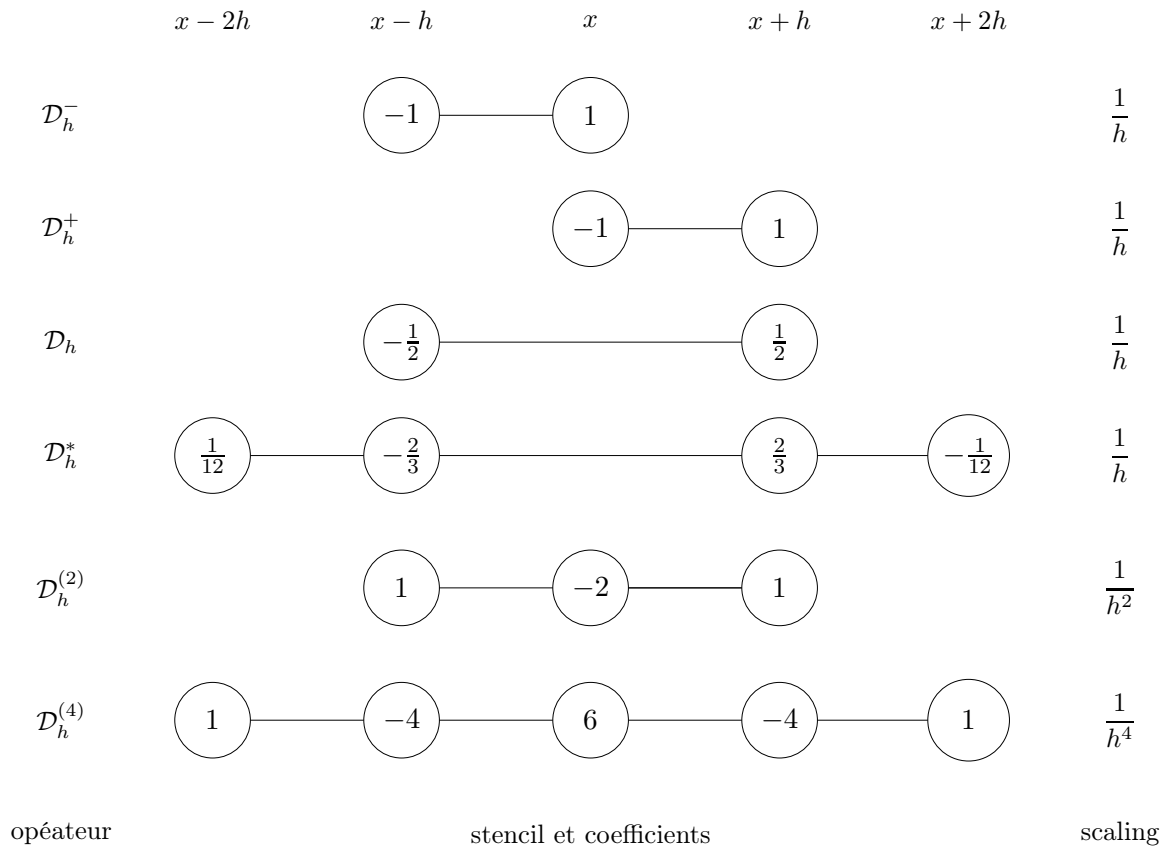


FIGURE 4.3 – Représentation symbolique des opérateurs en dimension un.

Les opérateurs $\mathcal{D}_h^-$ et $\mathcal{D}_h^+$ sont d'ordre 1 puisque les inégalités suivantes sont vérifiées,

$$\forall \varphi \in \mathcal{C}^2(\Omega), \quad |\varphi' - \mathcal{D}_h^-(\varphi)| \leq \frac{h}{2}\|\varphi''\|_\infty \tag{4.16}$$

$$\forall \varphi \in \mathcal{C}^2(\Omega), \quad |\varphi' - \mathcal{D}_h^+(\varphi)| \leq \frac{h}{2}\|\varphi''\|_\infty \tag{4.17}$$

Les opérateurs $\mathcal{D}_h, \mathcal{D}_h^{(2)}$ et $\mathcal{D}_h^{(4)}$ sont d'ordre 2 puisque les inégalités suivantes sont vérifiées,

$$\forall \varphi \in \mathcal{C}^3(\Omega), \quad |\varphi' - \mathcal{D}_h(\varphi)| \leq \frac{h^2}{6}\|\varphi^{(3)}\|_\infty \tag{4.18}$$

$$\forall \varphi \in \mathcal{C}^4(\Omega), \quad |\varphi'' - \mathcal{D}_h^{(2)}(\varphi)| \leq \frac{h^2}{12}\|\varphi^{(4)}\|_\infty \tag{4.19}$$

$$\forall \varphi \in \mathcal{C}^6(\Omega), \quad |\varphi^{(4)} - \mathcal{D}_h^{(4)}(\varphi)| \leq \frac{h^2}{6}\|\varphi^{(6)}\|_\infty \tag{4.20}$$

L'opérateur $\mathcal{D}_h^*$ est d'ordre 4 puisque l'inégalité suivante est vérifiée,

$$\forall \varphi \in \mathcal{C}^5(\Omega), \quad |\varphi' - \mathcal{D}_h^*(\varphi)| \leq \frac{h^4}{30}\|\varphi^{(5)}\|_\infty \tag{4.21}$$

La norme $\|\cdot\|_\infty$ désigne la norme infinie définie par $\|\varphi\|_\infty \stackrel{\text{def}}{=} \sup_\Omega(|\varphi|)$.

4.3 Laplacien en dimension un

Nous proposons de résoudre l'équation de diffusion sur le domaine $\Omega = [0, L]$ complétée de conditions aux limites de type Dirichlet,

$$\left(\mathcal{P}^1\right) \begin{cases} -u''(x) = f(x), & \forall x \in]0, L[\\ u(0) = \alpha, \\ u(L) = \beta. \end{cases}$$

La fonction $u(x) \in \mathcal{C}^2(\Omega)$ est l'inconnue alors que la fonction $f(x)$, les valeurs α et β ainsi que la longueur du domaine L sont données.

4.3.1 Solution continue

La solution exacte (ou solution analytique) $u(x) \in \mathcal{C}^2(\Omega)$ du problème continu $\left(\mathcal{P}^1\right)$ est

$$\boxed{u(x) = \alpha + \frac{x}{L}\left(\beta - \alpha + \int_0^L (L-s)f(s)ds\right) - \int_0^x (x-s)f(s)ds.} \tag{4.22}$$

Démonstration. Le théorème fondamental de l'analyse donne

$$u'(x) = c_2 - \int_0^x f(t)dt = c_2 - F(x), \quad \text{avec} \quad F(x) \overset{\text{def}}{=} \int_0^x f(t)dt.$$

En appliquant à nouveau le théorème, nous obtenons

$$u(x) = c_1 + c_2 x - \int_0^x F(s)ds. \tag{4.23}$$

Le terme intégral est calculé grâce à une intégration par parties,

$$\int_0^x F(s)ds = [sF(s)]_0^x - \int_0^x sF'(s)ds$$

En utilisant la définition de F, on obtient

$$\int_0^x F(s)ds = xF(x) - \int_0^x sf(s)ds = x\int_0^x f(s)ds - \int_0^x sf(s)ds = \int_0^x (x-s)f(s)ds \tag{4.24}$$

Les constantes c_1 et c_2 sont déterminées par les conditions aux limites :

$$u(0) = \alpha \quad \Rightarrow \quad c_1 = \alpha, \tag{4.25}$$

$$u(L) = \beta \quad \Rightarrow \quad \alpha + c_2 L - \int_0^L (L-s)f(s)ds = \beta,$$

$$\Rightarrow \quad c_2 = \frac{1}{L}\left(\beta - \alpha + \int_0^L (L-s)f(s)ds\right). \tag{4.26}$$

La preuve se termine en reportant (**??**), (**??**) et (**??**) dans (**??**). □

4.3.2 Solution discrète

4.3.2.1 Maillage

Nous considérons une subdivision Ω_h de Ω en $N+1$ intervalles de longueur constante h. Les points x_i de cette subdivision s'appellent des *noeuds* dont l'ensemble constitue un *maillage* de Ω,

$$\boxed{\Omega_h \overset{\text{def}}{=} \{x_i = ih, 0 \leq i \leq N+1\}.} \tag{4.27}$$

Le *pas* du maillage h est défini par

$$h \overset{\text{def}}{=} \frac{1}{N+1}. \tag{4.28}$$

Lorsque h est constant, comme sur la figure **??**, le maillage est *uniforme* mais il est parfois utile de considérer une subdivison en intervalles de longueurs différentes.

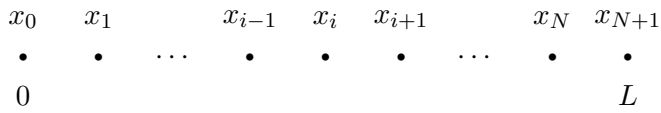


FIGURE 4.4 – Maillage uniforme Ω_h de $[0, L]$.

4.3.2.2 Schéma numérique

Le domaine de résolution de l'équation de diffusion est l'ensemble des noeuds internes du maillage

$$-u''(x_i) = f(x_i), \quad \{x_i = ih\}_{1 \leq i \leq N} \tag{4.29}$$

La dérivée seconde étant remplacée par son approximation discrète, la solution continue u est aussi remplacée par la solution discrète u_h de sorte que le problème devient

$$-\mathcal{D}_h^{(2)} u_h(x_i) = f(x_i), \quad \{x_i = ih\}_{1 \leq i \leq N} \tag{4.30}$$

En explicitant l'expression de l'opérateur discret (**??**), nous obtenons

$$\frac{-u_h(x_i - h) + 2u_h(x_i) - u_h(x_i + h)}{h^2} = f(x_i), \quad \{x_i = ih\}_{1 \leq i \leq N} \tag{4.31}$$

Avec les notations $u_i \overset{\text{def}}{=} u_h(x_i)$ et $f_i \overset{\text{def}}{=} f(x_i)$, nous obtenons le *schéma numérique* suivant

$$\boxed{\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i,} \quad 1 \leq i \leq N \tag{4.32}$$

Les valeurs u_{i-1}, u_i et u_{i+1} sont à déterminer sauf u_{i-1} lorsque $i = 1$ et u_{i+1} lorsque $i = N$. En effet, ces valeurs sont connues puisqu'il s'agit des conditions aux limites de sorte que $u_0 = \alpha$ et $u_{N+1} = \beta$. Les première et dernière relations s'écrivent donc

$$\frac{2u_1 - u_2}{h^2} = f_1 + \frac{\alpha}{h^2} \quad \text{et} \quad \frac{-u_{N-1} + 2u_N}{h^2} = f_N + \frac{\beta}{h^2}. \tag{4.33}$$

La terminologie « problème discret » est maintenant claire puisque le problème continu initial est remplacé par un problème discret $\left(\mathcal{P}_h^1\right)$ dont la solution u_h est évaluée en un nombre fini N de points (les noeuds du maillage) :

$$\left(\mathcal{P}_h^1\right) \begin{cases} -u_{i-1} + 2u_i - u_{i+1} = h^2 f_i \quad , 2 \leq i \leq N-1 \\ 2u_1 - u_2 = h^2 f_1 + \alpha, \\ -u_{N-1} + 2u_N = h^2 f_N + \beta. \end{cases}$$

4.3.2.3 Formulation matricielle

Il est évident que les N relations précédentes s'écrivent sous la forme matricielle suivante

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & -1 & 2 & -1 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_i \\ \vdots \\ f_{N-1} \\ f_N \end{bmatrix} + \frac{1}{h^2}\begin{bmatrix} \alpha \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ \beta \end{bmatrix}, \tag{4.34}$$

ce qui est équivalent à la forme compacte

$$\boxed{Ax_h = b} \tag{4.35}$$

en posant $A \overset{\text{def}}{=} \dfrac{1}{h^2} tridiag(-1, 2, -1)$, matrice tridiagonale,

$x_h \overset{\text{def}}{=} (u_1, \cdots, u_N)^\top$, vecteur constitué des inconnues nodales,

$b \overset{\text{def}}{=} \left(f_1 + \dfrac{\alpha}{h^2}, f_2, \cdots, f_{N-1}, f_N + \dfrac{\beta}{h^2}\right)^\top$.

4.3.2.4 Unicité du problème discret

Une question importante qu'il convient de se poser est de savoir si le problème discret $\left(\mathcal{P}_h^1\right)$ admet une unique solution. Pour cela, nous allons démontrer l'inversibilité de la matrice A à l'aide de la propriété suivante :

Propriété. *Toute matrice symétrique définie positive est inversible.*

Il est évident que A est symétrique. A est aussi définie positive puisque

$$\forall v \in \mathbb{R}^N,\ v \neq 0, \quad (Av, v) = \frac{1}{h^2}\Big(2v_1^2 - v_2v_1 - v_1v_2 + 2v_2^2 - v_3v_2 - \cdots \\ - v_{i-1}v_i + 2v_i^2 - v_{i+1}v_i - \cdots - v_{N-1}v_N + 2v_N^2\Big), \tag{4.36}$$

qui se réécrit

$$\forall v \in \mathbb{R}^N, \quad (Av, v) = \frac{1}{h^2}\Big(v_1^2 + v_N^2 + \sum_{i=2}^{N}(v_i - v_{i-1})^2\Big) > 0. \tag{4.37}$$

4.4 Analyse de la méthode

4.4.1 Consistance

La consistance d'un schéma numérique est la propriété de l'opérateur discret à approcher l'opérateur continu.

Définition. *L'erreur de consistance (ou résidu) r d'un schéma numérique s'obtient en remplaçant la solution approchée par la solution exacte dans le problème discret*

$$r \stackrel{\text{def}}{=} Ax - b \tag{4.38}$$

Un schéma est consistant si l'erreur de consistance tend vers 0 avec le pas du maillage h

$$\lim_{h \to 0} \|r\|_\infty = 0.$$

La consistante est d'ordre p en norme l_∞ s'il existe une constante $C \in \mathbb{R}^+$ vérifiant

$$\|r\|_\infty \leq Ch^p.$$

Propriété. *Si $u \in \mathcal{C}^4([0,1])$, le lapacien discret $\mathcal{D}_h^{(2)}$ est consistant d'ordre 2 puisque*

$$\|r\|_\infty \leq \frac{h^2}{12} \|u^{(4)}\|_\infty \tag{4.39}$$

Démonstration. La formule de Taylor–Lagrange sur les intervalles $[x_{i-1}, x_i]$ et $[x_i, x_{i+1}]$ donne

$$u''(x_i)h^2 - u(x_{i+1}) + 2u(x_i) - u(x_{i-1}) = -\frac{h^4}{24}\left(u^{(4)}(\nu_i) + u^{(4)}(\mu_i)\right),$$

avec $\nu_i \in [x_{i-1}, x_i]$ et $\mu_i \in [x_i, x_{i+1}]$. En remplaçant $u''(x_i)$ par $-f(x_i)$ car u est la solution exacte, nous obtenons

$$\frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} - f(x_i) = -\frac{h^2}{24}\left(u^{(4)}(\nu_i) + u^{(4)}(\mu_i)\right).$$

La définition de l'erreur de consistance donne

$$r_i = -\frac{h^2}{24}\left(u^{(4)}(\nu_i) + u^{(4)}(\mu_i)\right).$$

□

4.4.2 Stabilité

La stabilité d'un schéma numérique est la propriété de l'opérateur discret à fournir une solution discrète bornée (*i.e.* qui n'explose pas).

Définition. *Un schéma numérique est stable pour la norme $\|\cdot\|$ si la solution discrète est continue par rapport aux données*

$$\|x_h\| \leq C\|b\|. \tag{4.40}$$

Propriété. *Le laplacien discret $\mathcal{D}_h^{(2)}$ est stable pour la norme $\|\cdot\|_\infty$ puisque*

$$\|x_h\|_\infty \leq \frac{1}{8}\|b\|_\infty \tag{4.41}$$

Démonstration. Nous allons montrer que $\|A^{-1}\|_\infty \leq \frac{1}{8}$ puisque

$$Ax_h = b \quad \Rightarrow \quad x_h = A^{-1}b \quad \Rightarrow \quad \|x_h\|_\infty \leq \|A^{-1}\|_\infty \|b\|_\infty.$$

La clef de la démonstration repose sur l'égalité $\|A^{-1}\|_\infty = \|A^{-1}v_1\|_\infty$ avec $v_1 = \{1, \cdots, 1\}^\top$ puisque A est à diagonale dominante. Pour terminer, $\|A^{-1}v_1\|_\infty = 1/8$ car $A^{-1}v_1$ est la solution discrète lorsque $f = 1$ qui correspond exactement à la solution continue puisque les dérivées d'ordre supérieur ou égal à trois s'annulent, $u(x) = x(1-x)/2 \ \Rightarrow \ \sup|u(x)|_{0\leq x\leq 1} = 1/8$. □

4.4.3 Convergence

La convergence d'un schéma numérique est la propriété de la solution discrète à approcher la solution continue.

Définition. *L'erreur de convergence e d'un schéma numérique est définie par*

$$e \stackrel{\text{def}}{=} x - x_h \tag{4.42}$$

Un schéma est convergent si l'erreur de convergence tend vers 0 avec le pas du maillage h

$$\lim_{h\to 0} \|e\|_\infty = 0.$$

La convergence est d'ordre p en norme l_∞ s'il existe une constante $C \in \mathbb{R}^+$

$$\|e\|_\infty \leq Ch^p.$$

Propriété. *Si $u \in \mathcal{C}^4([0,1])$, le lapacien discret $\mathcal{D}_h^{(2)}$ est convergent d'ordre 2 puisque*

$$\|e\|_\infty \leq \frac{h^2}{96}\|u^{(4)}\|_\infty \tag{4.43}$$

Démonstration. L'erreur de convergence se réécrit en fonction de l'erreur de consistance,

$$e \stackrel{(??)}{=} x - x_h = A^{-1}(Ax - Ax_h) = A^{-1}(Ax - b) \stackrel{(??)}{=} A^{-1}r,$$

d'où

$$\|e\|_\infty \leq \|A^{-1}\|_\infty \|r\|_\infty \leq \frac{1}{8} \cdot \frac{h^2}{12}\|u^{(4)}\|_\infty.$$

□

Théorème de Lax. *Un schéma numérique aux différences finies stable et consistant converge. De plus, la consistance à l'ordre p et la stabilité impliquent la convergence à l'ordre p.*

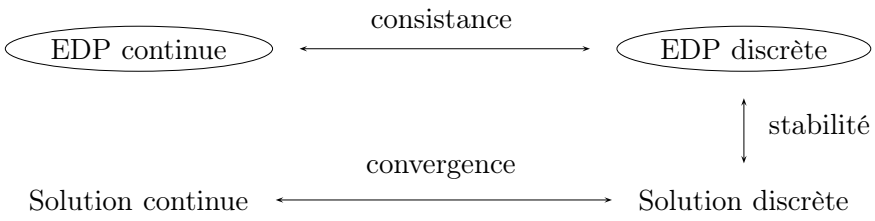


FIGURE 4.5 – Relations entre consistance, stabilité et convergence.

4.5 Laplacien en dimension deux

Nous proposons de résoudre l'équation de diffusion sur le domaine $\Omega = [0, L]^2$ complétée de conditions aux limites de type Dirichlet,

$$\left(\mathcal{P}^2\right) \begin{cases} -\Delta u(x,y) = f(x,y), & \forall (x,y) \in]0, L[^2 \\ u(x,y) = \alpha & \text{sur } \partial\Omega \end{cases}$$

La fonction $u(x, y) \in \mathcal{C}^2(\Omega)$ est l'inconnue alors que $f(x, y)$, $\alpha(x, y)$ et L sont donnés.

4.5.1 Solution continue de l'équation homogène

La solution du problème $\left(\mathcal{P}^2\right)$ sur le carré unité ($L = 1$), sans terme source ($f = 0$) et avec des conditions aux limites nulles sauf au bord supérieur où elle est unitaire ($u(0, y) = u(x, 0) = u(1, y) = 0$ et $u(x, 1) = 1$) est la série suivante

$$\boxed{u(x,y) = \sum_{n=1}^{\infty} 2\frac{1-(-1)^n}{n\pi sh(n\pi)} \sin(n\pi x) sh(n\pi y).} \tag{4.44}$$

Démonstration. Nous utilisons la méthode de séparation des variables en cherchant une solution de la forme $u(x, y) = X(x)Y(y)$ que nous introduisons dans l'équation $\Delta u = 0$,

$$X''(x)Y(y) + X(x)Y''(y) = 0 \quad \Rightarrow \quad \frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} = -k^2 \quad \text{si } u \neq 0,$$

où la forme de la constante $-k^2$ simplifie la suite des calculs. Les formes générales des fonctions X et Y sont,

$$\begin{cases} X''(x) = -k^2 X(x) \\ Y''(y) = k^2 Y(y) \end{cases} \Rightarrow \begin{cases} X(x) = C_1 \cos(kx) + C_2 \sin(kx) \\ Y(x) = C_3 e^{ky} + C_4 e^{-ky} \end{cases}$$

Ainsi, la forme générale de la solution est

$$u(x,y) = (C_1 \cos(kx) + C_2 \sin(kx)) \left(C_3 e^{ky} + C_4 e^{-ky}\right)$$

Les trois conditions aux limites homogènes donnent

$$u(0,y) = 0 \;\Rightarrow\; C_1 = 0, \qquad u(x,0) = 0 \;\Rightarrow\; C_3 = -C_4, \qquad u(1,y) = 0 \;\Rightarrow\; k = n\pi,$$

la solution générale devenant

$$u(x,y) = C \sin(n\pi x) \mathrm{sh}(n\pi y).$$

La condition sur le bord supérieur est moins directe à imposer. L'équation étant linéaire, toute combinaison linéaire est également solution

$$u(x,y) = \sum_{n=1}^{\infty} C_n \sin(n\pi x) \mathrm{sh}(n\pi y).$$

La condition sur le bord supérieur s'écrit

$$\begin{aligned} \sum_{n=1}^{\infty} C_n \sin(n\pi x) \mathrm{sh}(n\pi) = 1 \quad &\Rightarrow \quad \sum_{n=1}^{\infty} C_n \mathrm{sh}(n\pi) \int_0^1 \sin(n\pi x) \sin(m\pi x) = \int_0^1 \sin(m\pi x) \\ &\Rightarrow \quad \sum_{n=1}^{\infty} C_n \mathrm{sh}(n\pi) \frac{\delta_{mn}}{2} = \frac{1-(-1)^m}{m\pi}, \end{aligned}$$

où δ_{mn} est le symbole de Kronecker : $\delta_{mn} = 1$ si $m = n$ et $\delta_{mn} = 0$ si $m \neq n$. □

4.5.2 Solution discrète

4.5.2.1 Maillage

Nous considérons une *grille cartésienne* Ω_h de Ω constitué de $N+1$ segments de longueurs constante h suivant chaque direction. Le *maillage* de Ω est ainsi constitué de l'ensemble des noeuds de la grille,

$$\boxed{\Omega_h \stackrel{\text{def}}{=} \{(x_i, y_j) = (ih, jh), 0 \leq i, j \leq N+1\}.} \tag{4.45}$$

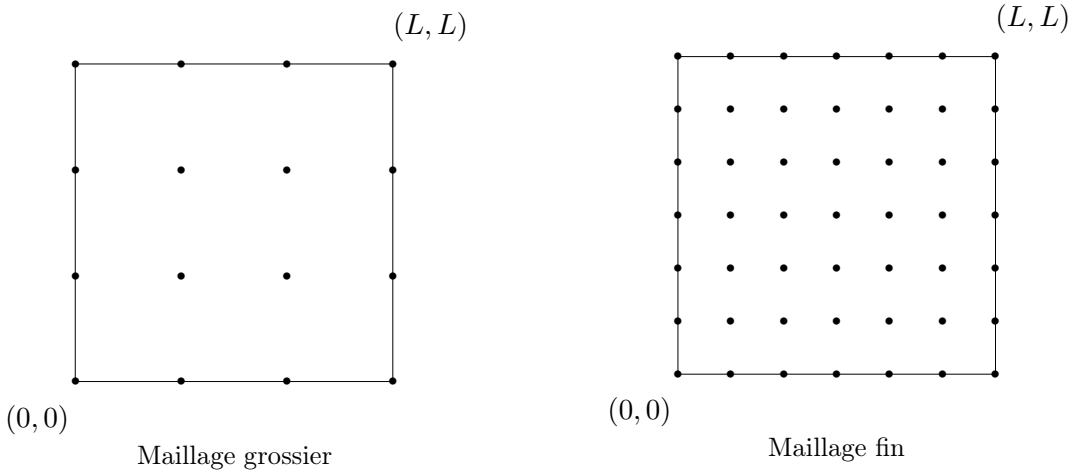


FIGURE 4.6 – Deux exemples de maillage uniforme Ω_h de $[0, L]^2$.

4.5.2.2 Opérateur discret

On montre aisément qu'un opérateur centré approchant le laplacien en dimension deux est

$$\Delta f(x,y) \simeq \boxed{\Delta_h f(x,y) \stackrel{\text{def}}{=} \frac{f(x-h,y) + f(x,y-h) - 4f(x,y) + f(x+h,y) + f(x,y+h)}{h^2}} \tag{4.46}$$

4.5.2.3 Schéma numérique

Le domaine de résolution du laplacien est l'ensemble des noeuds internes du maillage,

$$-\Delta u(x_i, y_j) = f(x_i, y_j), \quad \{(x_i, y_j)\}_{1 \leq i,j \leq N} \tag{4.47}$$

L'opérateur continu étant remplacé par son approximation discrète, la solution continue u est remplacée par la solution discrète u_h de sorte que le problème devient

$$-\Delta_h u_h(x_i, y_j) = f(x_i, y_j), \quad \{(x_i, y_j)\}_{1 \leq i,j \leq N} \tag{4.48}$$

En explicitant l'expression du laplacien discret, nous obtenons

$$\frac{-u_h(x_i-h, y_j) - u_h(x_i, y_j-h) + 4u_h(x_i, y_j) - u_h(x_i+h, y_j) - u_h(x_i, y_j+h)}{h^2} = f(x_i, y_j). \tag{4.49}$$

Avec les notations $u_{i,j} \overset{\text{def}}{=} u_h(x_i, y_j)$ et $f_{i,j} \overset{\text{def}}{=} f(x_i, y_j)$, nous obtenons le *schéma numérique*

$$\boxed{\frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f_{i,j},} \quad 1 \leq i, j \leq N \tag{4.50}$$

Les valeurs $u_{i-1,j}, u_{i,j-1}, u_{i,j}, u_{i+1,j}$ et $u_{i,j+1}$ sont à déterminer sauf les valeurs sur les bords,

- $u_{i-1,j}$ lorsque $i = 1$ (bord à gauche),
- $u_{i,j-1}$ lorsque $j = 1$ (bord en bas),
- $u_{i+1,j}$ lorsque $i = N$ (bord à droite),
- $u_{i,j+1}$ lorsque $j = N$ (bord en haut).

Le problème discret $\left(\mathcal{P}_h^2\right)$ approchant $\left(\mathcal{P}^2\right)$ s'écrit

$$\left(\mathcal{P}_h^2\right) \begin{cases} -u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1} &= h^2 f_{i,j}, & 2 \leq i, j \leq N-1 \\ -u_{1,j-1} + 4u_{1,j} - u_{2,j} - u_{1,j+1} &= h^2 f_{1,j} + \alpha, & 2 \leq j \leq N-1 \\ -u_{i-1,1} + 4u_{i,1} - u_{i+1,1} - u_{i,2} &= h^2 f_{i,1} + \alpha, & 2 \leq i \leq N-1 \\ -u_{N-1,j} - u_{N,j-1} + 4u_{N,j} - u_{N,j+1} &= h^2 f_{N,j} + \alpha, & 2 \leq j \leq N-1 \\ -u_{i-1,N} - u_{i,N-1} + 4u_{i,N} - u_{i+1,N} &= h^2 f_{i,N} + \alpha, & 2 \leq i \leq N-1 \\ 4u_{1,1} - u_{2,1} - u_{1,2} &= h^2 f_{1,1} + 2\alpha, & \\ -u_{1,N-1} + 4u_{1,N} - u_{2,N} &= h^2 f_{1,N} + 2\alpha, & \\ -u_{N-1,1} + 4u_{N,1} - u_{N,2} &= h^2 f_{N,1} + 2\alpha, & \\ -u_{N-1,N} - u_{N,N-1} + 4u_{N,N} &= h^2 f_{N,N} + 2\alpha, & \end{cases}$$

La première équation est la résolution du laplacien discret sur les noeuds internes de $\mathring{\Omega}_h$ (*i.e.* l'ouvert de Ω_h), les quatre équations suivantes correspondent aux bords (gauche, bas, droit et haut) et les quatre dernières correspondent aux coins (bas-gauche, haut-gauche, bas-droite et haut-droite).

4.5.2.4 Formulation matricielle

En dimension un, la numérotation des inconnues correspond naturellement à celle des noeuds du maillage. En revanche, plusieurs numérotations sont possibles en dimension supérieure. En dimension deux, le choix le plus courant pour passer du double indiçage (i, j) à un simple indiçage k est

$$\forall 1 \leq i, j \leq N, \quad k(i, j) = i + (j-1)N.$$

Il s'agit de la numérotation ligne par ligne illustrée à la figure (**??**). La numérotation colonne par colonne est $k(i, j) = j + (i-1)N$. Dans les deux cas, nous avons $u_{k(i,j)} = u_{i,j}, \forall 1 \leq k \leq N^2$.

Le problème discret $\left(\mathcal{P}_h^2\right)$ s'écrit sous la forme matricielle (structure bloc) suivante

$$\frac{1}{h^2} \begin{bmatrix} D & -I & & & & & & \\ -I & D & -I & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & -I & D & -I & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & -I & D & -I & \\ & & & & & -I & D & \end{bmatrix} \begin{bmatrix} u_{:,1} \\ u_{:,2} \\ \vdots \\ u_{:,j} \\ \vdots \\ u_{:,N-1} \\ u_{:,N} \end{bmatrix} = \begin{bmatrix} f_{:,1} \\ f_{:,2} \\ \vdots \\ f_{:,j} \\ \vdots \\ f_{:,N-1} \\ f_{:,N} \end{bmatrix} + \frac{1}{h^2} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \vdots \\ \alpha_2 \\ \alpha_1 \end{bmatrix},$$

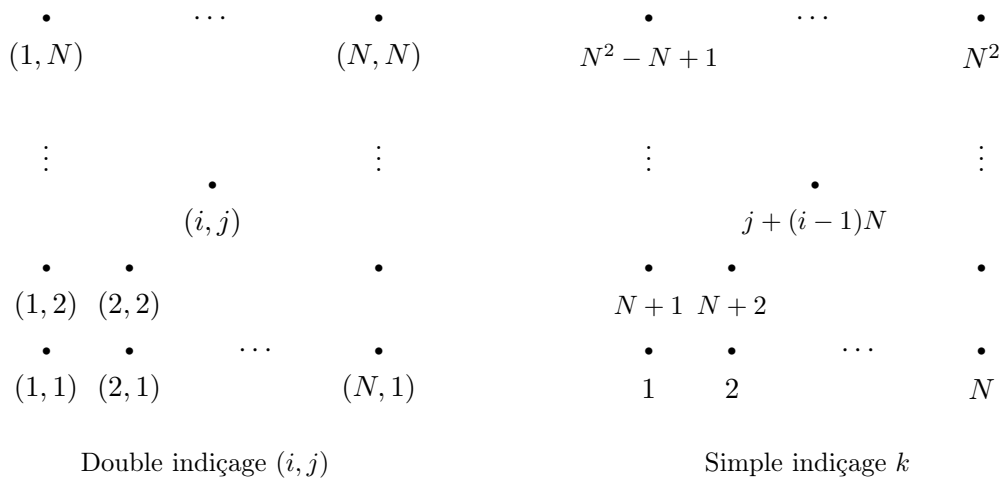


FIGURE 4.7 – Numérotation des inconnues ligne par ligne.

en posant $D \stackrel{\text{def}}{=} tridiag(-1, 4, -1)$, matrice tridiagonale,

$I \stackrel{\text{def}}{=} diag(1)$, matrice identité,

$u_{:,j} \stackrel{\text{def}}{=} (u_{1,j}, \cdots, u_{N,j})^\top,\ 1 \leq j \leq N$, vecteur des inconnues de la j-ème ligne de Ω_h,

$f_{:,j} \stackrel{\text{def}}{=} (f_{1,j}, \cdots, f_{N,j})^\top,\ 1 \leq j \leq N$,

$\alpha_1 \stackrel{\text{def}}{=} (2\alpha, \alpha, \cdots, \alpha, 2\alpha)^\top$,

$\alpha_2 \stackrel{\text{def}}{=} (\alpha, 0, \cdots, 0, \alpha)^\top$.

Chapitre 5

Résolution des EDO

De nombreux problèmes en sciences appliquées mènent à des équations différentielles ordinaires (EDO). Une EDO se définit comme une relation entre une ou plusieurs fonctions inconnues et leurs dérivées. Dans le cas d'une seule fonction y, la forme générale d'une EDO est

$$F(x, y(x), y'(x), y''(x) \cdots, y^{(n)}(x)) = 0, \tag{5.1}$$

où le degré maximal de dérivation n s'appelle l'*ordre* de l'EDO.

5.1 Introduction

5.1.1 Problème de Cauchy

Définition. *Le problème de Cauchy est la résolution d'une équation différentielle complétée d'une condition initiale y_0 appelée condition de Cauchy,*

$$\begin{cases} y'(t) = f(t, y(t)), & \forall t \in]0, T] \\ y(0) = y_0. \end{cases}$$

La fonction $y \in \mathcal{C}^1([0,T])$ est inconnue alors que f et y_0 sont données. La fonction f est continue par rapport à t ($f \in \mathcal{C}([0,T])$) et est localement lipschitzienne[a] *par rapport à y.*

a. $\forall t \in [0,T], \forall y_1, y_2 \in \mathcal{C}^1([0,T]), \exists k \geq 0, ||f(t,y_1) - f(t,y_2)|| \leq k||y_1 - y_2||$

Nous allons présenter plusieurs schémas pour résoudre ce problème. Précisons que si f ne dépend pas explicitement de t (*i.e.* $f(t,y) = f(y)$), l'équation différentielle est dite *autonome*.

5.1.2 Exemples

5.1.2.1 Mécanique vibratoire

La position y de l'amortissement d'une voiture peut êre modélisée par l'EDO d'ordre 2 suivante

$$\begin{cases} my'' + cy' + ky = 0 \quad \text{sur }]0, T], \\ y(0) = 1, \\ y'(0) = 0, \end{cases}$$

où m est la masse de la voiture, c le coefficient d'amortissement et k la force de rappel. Cette EDO s'écrit sous la forme d'un système d'ordre 1 en posant $x = (y, y')^\top$,

$$\begin{cases} x' = Ax & \text{sur }]0,T], \quad \text{avec} \quad A \stackrel{\text{def}}{=} \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} \\ x(0) = (1,0)^\top, \end{cases}$$

La matrice A s'appelle aussi matrice d'état.

5.1.2.2 Equation de la chaleur

En discrétisant le laplacien par différences finies et la condition initiale u^0 dans l'équation de la chaleur avec CL de Dirichlet homogène,

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = 0 & \text{dans } \Omega\times]0,T], \\ u(x,t) = 0 & \text{sur } \partial\Omega\times]0,T], \\ u(\cdot,0) = u^0 & \text{dans } \Omega, \end{cases}$$

nous obtenons la forme *semi-discrète* suivante

$$\begin{cases} \dfrac{dx_h}{dt} = Ax_h & \text{sur }]0,T], \\ x_h^0 = u^0(:) & \text{dans } \Omega, \end{cases}$$

où A est tridiagonale en 1D et tridiagonale par blocs en 2D, x_h est le vecteur des inconnues nodales et $u^0(:)$ désigne l'ensemble des valeurs de $u^0(x)$ aux nœuds du maillage.

5.1.3 Maillage

Nous considérons une subdivision de $[0,T]$ en N_T intervalles de longueur constante δt. Les points t^n de cette subdivision sont des multiples du pas de temps δt,

$$\forall 0 \le n \le N_T, \quad t^n = n\delta t, \quad \text{avec} \quad \delta t \stackrel{\text{def}}{=} \frac{T}{N_T}.$$

Lorsque δt est constant, comme sur la figure **??**, la discrétisation en temps est *uniforme* mais il est parfois utile de considérer une subdivison en intervalles de longueurs différentes.

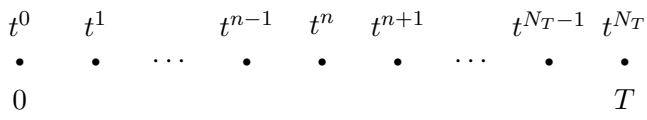


FIGURE 5.1 – Subdivision uniforme de $[0,T]$.

5.2 Quelques schémas

5.2.1 Obtention des schémas

Pour un schéma numérique, la dérivée première étant remplacée par un opérateur discret, la solution continue y est aussi remplacée par la solution discrète y_h et nous désignons par y^n la solution approchée à l'instant t^n,

$$y^n \stackrel{\text{def}}{=} y_h(t^n). \tag{5.2}$$

Nous désignons formellement par $\mathcal{D}_{\delta t}$ l'opérateur permettant d'exprimer y^{n+1} et distinguons deux types de schémas pour résoudre les EDO,

— les schémas *explicites* expriment y^{n+1} en fonction de valeurs connues de y_h et s'écrivent

$$y^{n+1} = \mathcal{D}_{\delta t}(y^n, y^{n-1}, \cdots), \tag{5.3}$$

— les schémas *implicites* expriment y^{n+1} en fonction de y^{n+1} et de valeurs connues de y_h et s'écrivent

$$y^{n+1} = \mathcal{D}_{\delta t}(y^{n+1}, y^n, y^{n-1}, \cdots). \tag{5.4}$$

Plusieurs méthodes existent pour résoudre le problème de Cauchy :

1. la dérivée première est remplacée par un opérateur approché $\mathcal{D}_{\delta t}$,

$$y'(t) \simeq \mathcal{D}_{\delta t} y(t) \quad \Rightarrow \quad \mathcal{D}_{\delta t} y(t) \simeq f(t, y(t)), \tag{5.5}$$

2. l'EDO est intégrée entre t^n et $t^n + \delta t$,

$$y'(t) = f(t, y(t)) \quad \Rightarrow \quad y(t^{n+1}) - y(t^n) = \int_{t^n}^{t^{n+1}} f(t, y(t)), \tag{5.6}$$

   et une formule de quadrature estime le second membre,

3. les méthodes précédentes sont combinées.

5.2.2 Schémas explicites

5.2.2.1 Euler explicite

La derivée temporelle est approchée par l'opérateur décentré à droite

$$y'(t) \simeq \mathcal{D}^+_{\delta t} y(t) \stackrel{\text{def}}{=} \frac{y(t+\delta t) - y(t)}{\delta t}. \tag{5.7}$$

de sorte que nous obtenons le problème discret suivant

$$\forall\ 0 \leq n \leq N_T - 1, \quad \frac{y_h(t^n + \delta t) - y_h(t^n)}{\delta t} = f(t^n, y_h(t^n)). \tag{5.8}$$

Le schéma d'Euler explicite s'écrit finalement

$$\boxed{y^{n+1} = y^n + \delta t f(t^n, y^n).} \tag{5.9}$$

5.2.2.2 Saute-mouton

La derivée temporelle est approchée par l'opérateur centré

$$y'(t) \simeq \mathcal{D}_{\delta t} y(t) \stackrel{\text{def}}{=} \frac{y(t+\delta t) - y(t-\delta t)}{2\delta t}. \tag{5.10}$$

de sorte que nous obtenons le problème discret suivant

$$\forall\ 1 \leq n \leq N_T - 1, \quad \frac{y_h(t^n + \delta t) - y_h(t^n - \delta t)}{2\delta t} = f(t^n, y_h(t^n)). \tag{5.11}$$

Le schéma saute-mouton s'écrit finalement

$$\boxed{y^{n+1} = y^{n-1} + 2\delta t f(t^n, y^n).} \tag{5.12}$$

5.2.3 Schémas implicites

5.2.3.1 Euler implicite

La derivée temporelle est approchée par l'opérateur décentré à gauche

$$y'(t) \simeq \mathcal{D}_{\delta t}^{-} y(t) \stackrel{\text{def}}{=} \frac{y(t) - y(t-\delta t)}{\delta t}. \tag{5.13}$$

de sorte que nous obtenons le problème discret suivant

$$\forall\ 0 \leq n \leq N_T - 1, \quad \frac{y_h(t^n) - y_h(t^n - \delta t)}{\delta t} = f(t^n, y_h(t^n)). \tag{5.14}$$

Le schéma d'Euler implicite s'écrit finalement (en remplaçant l'indice muet n par $n+1$ dans l'équation ci-dessus)

$$\boxed{y^{n+1} = y^n + \delta t f(t^{n+1}, y^{n+1}).} \tag{5.15}$$

5.2.3.2 Crank–Nicolson

Ce shéma est la moyenne des schémas d'Euler explicite et implicite

$$\boxed{y^{n+1} = y^n + \frac{\delta t}{2}\Big(f(t^n, y^n) + f(t^{n+1}, y^{n+1})\Big).} \tag{5.16}$$

Il peut aussi être obtenu avec la formule des trapèzes pour estimer l'intégrale dans (**??**).

5.3 Analyse des schémas

5.3.1 Convergence

Définition. *Un schéma numérique converge à l'ordre p s'il existe une constante $C \in \mathbb{R}^+$ indépendante de δt tel que*

$$\forall n > 0, \quad |y(n\delta t) - y_h(n\delta t)| \leq C\delta t^p.$$

Le tableau ci-dessous donne les ordres des principales méthodes précédentes.

Méthode	ordre
Euler explicite	1
saute-mouton	2
Euler implicite	1
Crank–Nicolson	2

TABLE 5.1 – Ordre de convergence des schémas présentés.

5.3.2 Stabilité absolue

5.3.2.1 Problème modèle

Définition. *Le problème de Cauchy linéaire suivant est appelé problème modèle*

$$\begin{cases} y'(t) = \lambda y(t), \quad t > 0, \quad \lambda \in \mathbb{C}, \\ y(0) = 1, \end{cases}$$

La solution de ce problème est $y(t) = \exp(\lambda t) = \exp((\lambda_\mathrm{r} + i\lambda_\mathrm{i})t) = \exp(\lambda_\mathrm{r} t)\big(\cos(\lambda_\mathrm{i} t) + i\sin(\lambda_\mathrm{i} t)\big)$ et son comportement varie suivant la valeur de la partie réelle de λ :

- $\lambda_\mathrm{r} > 0$: la solution croît exponentiellement et le problème est instable,
- $\lambda_\mathrm{r} = 0$: la solution oscille,
- $\lambda_\mathrm{r} < 0$: la solution decroît exponentiellement et la stabilité absolue traduit que la solution numérique a ce même comportement.

5.3.2.2 Condition de stabilité

Définition. *Un schéma est absolument stable pour le problème modèle si*

$$\lim_{t^n \to \infty} |y^n| = 0. \tag{5.17}$$

Définition. *La région de stabilité absolue d'un schéma est le sous-ensemble $\mathcal{A}$ du plan complexe défini par*

$$\mathcal{A} \overset{\text{def}}{=} \{z = \lambda\delta t \in \mathbb{C} \text{ tel que } \lim_{t^n \to \infty} |y^n| = 0\}. \tag{5.18}$$

Définition. *La fonction d'amplification $G(\lambda\delta t)$ d'un schéma est définie par*

$$y^{n+1} \overset{\text{def}}{=} G(\lambda\delta t) y^n \tag{5.19}$$

Propriété. *La région de stabilité absolue $\mathcal{A}$ est déterminée avec la fonction d'amplification*

$$\mathcal{A} = \{z \in \mathbb{C} \text{ tel que } |G(z)| < 1\}. \tag{5.20}$$

Démonstration. La condition de stabilité absolue traduit que le module de la fonction d'amplification doit être inférieur à 1,

$$\lim_{t^n \to \infty} |y^n| = 0 \quad \Leftrightarrow \quad |G(\lambda\delta t)| < 1$$

puisqu'une simple récurrence donne $y^{n+1} = \big(G(\lambda\delta t)\big)^{n+1}$. □

5.3.2.3 Exemples

Nous allons déterminer les régions de stabilité des schémas d'Euler explicite et implicite. Le schéma d'Euler explicite (Ee) appliqué au problème modèle s'écrit

$$y^{n+1} = y^n + \lambda\delta t y^n \quad \Rightarrow \quad G_\mathrm{Ee}(\lambda\delta t) = 1 + \lambda\delta t \quad \Rightarrow \quad \boxed{\mathcal{A}_\mathrm{Ee} = \{z \text{ tel que } |1+z| < 1\}}$$

ce qui signifie que la région de stabilité du schéma d'Euler explicite est l'intérieur du cercle de centre -1 et de rayon 1.

Le schéma d'Euler implicite (Ei) appliqué au problème modèle s'écrit

$$y^{n+1} = y^n + \lambda\delta t y^{n+1} \quad \Rightarrow \quad G_{\mathrm{Ei}}(\lambda\delta t) = \frac{1}{1-\lambda\delta t} \quad \Rightarrow \quad \boxed{\mathcal{A}_{\mathrm{Ei}} = \{z \text{ tel que } |1-z| > 1\}}$$

ce qui signifie que la région de stabilité du schéma d'Euler implicite est l'extérieur du cercle de centre 1 et de rayon 1.

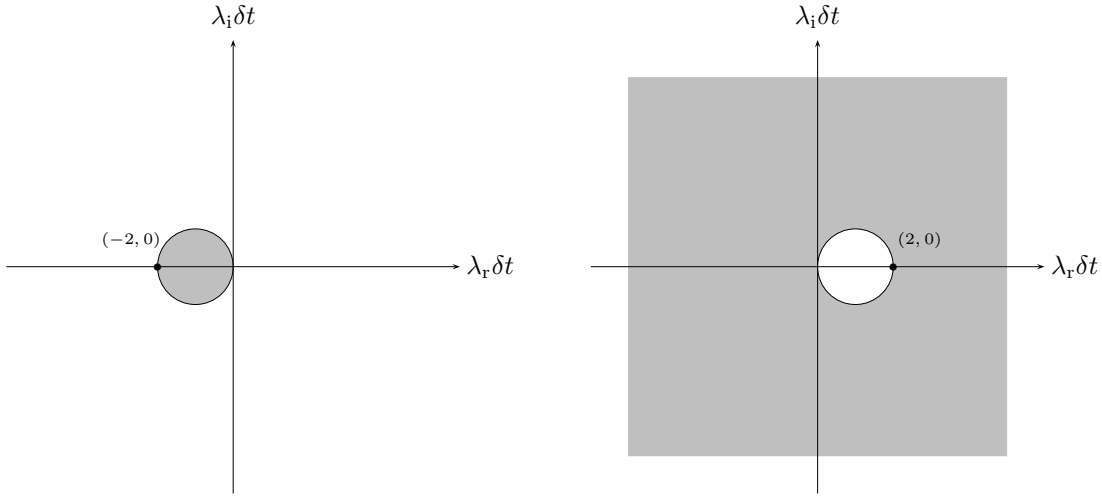


FIGURE 5.2 – Domaines de stabilité (en gris) des méthodes d'Euler explicite et implicite.

5.4 Application aux EDP : résolution de l'équation de la chaleur

Nous proposons de résoudre l'équation de la chaleur en dimension une d'espace sur le domaine $[0, L]$ avec des CL de Dirichlet homogène,

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = f & \text{dans }]0, L[\times]0, T], \\ u(0,\cdot) = u(L,\cdot) = 0, & \text{sur }]0, T], \\ u(\cdot, 0) = u^0 & \text{sur }]0, L[, \end{cases}$$

Les pas d'espace et de temps sont notés h et δt et nous adoptons la notation

$$u_h(x_i, t^n) = u_i^n.$$

La forme semi-discrète (*i.e.* la discrétisation est uniquement en espace) de l'équation de la chaleur s'écrit

$$1 \leq i \leq N, \quad \frac{du_i(n\delta t)}{dt} = \frac{u_{i-1}(n\delta t) - 2u_i(n\delta t) + u_{i+1}(n\delta t)}{h^2} + f_i. \tag{5.21}$$

Nous allons comparer les schémas d'Euler explicite et implicite pour intégrer cette EDO.

5.4.1 Euler explicite

Le schéma d'Euler explicite s'écrit pour $0 \leq n \leq N_T - 1$ et $1 \leq i \leq N$,

$$\frac{u_i^{n+1} - u_i^n}{\delta t} = \frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{h^2} + f_i. \tag{5.22}$$

ce qui se réécrit

$$\boxed{u_i^{n+1} = \frac{\delta t}{h^2} u_{i-1}^n + \left(1 - 2\frac{\delta t}{h^2}\right) u_i^n + \frac{\delta t}{h^2} u_{i+1}^n + \delta t f_i.} \tag{5.23}$$

Ce schéma est stable si la condition suivante, appelée *condition CFL*, est vérifiée

$$1 - 2\frac{\delta t}{h^2} > 0 \quad \Rightarrow \quad \boxed{\delta t < \frac{h^2}{2}.} \tag{5.24}$$

Le schéma d'Euler explicite permet d'obtenir la solution directement à partir de la solution à l'instant précédent mais nécessite une condition de stabilité sur le pas de temps.

5.4.2 Euler implicite

Le schéma d'Euler implicite s'écrit pour $0 \leq n \leq N_T - 1$ et $1 \leq i \leq N$,

$$\frac{u_i^{n+1} - u_i^n}{\delta t} = \frac{u_{i-1}^{n+1} - 2u_i^{n+1} + u_{i+1}^{n+1}}{h^2} + f_i. \tag{5.25}$$

ce qui se réécrit

$$u_i^{n+1} - \delta t \frac{u_{i-1}^{n+1} - 2u_i^{n+1} + u_{i+1}^{n+1}}{h^2} = u_i^n + \delta t f_i. \tag{5.26}$$

L'écriture matricielle de cette équation est

$$0 \leq n \leq N_T - 1, \quad \boxed{(I + \delta t A) x_h^{n+1} = b^n,} \tag{5.27}$$

en posant $I \stackrel{\text{def}}{=} diag(1)$, matrice identité,
$A \stackrel{\text{def}}{=} \frac{1}{h^2} tridiag(-1, 2, -1)$, matrice tridiagonale,
$x_h^n \stackrel{\text{def}}{=} (u_1^n, \cdots, u_N^n)^\top$, vecteur constitué des inconnues nodales à l'instant $n\delta t$,
$b^n \stackrel{\text{def}}{=} \left(u_1^n + \delta t f_1, \cdots, u_N^n + \delta t f_N\right)^\top$.

Le schéma d'Euler implicite nécessite la résolution d'un systèmes linéaire (**??**) pour trouver la solution à un instant donné mais sans condition de stabilité sur le pas de temps.

Annexe A

EDO

Nous rappelons dans cette annexe les solutions des équations différentielles ordinaires (EDO) à coefficients constants du premier et second ordre. Les équations étant linéaires, le principe de superposition s'applique : *la solution générale d'une EDO linéaire avec second membre est la somme de la solution générale de l'EDO homogène et d'une solution particulière de l'EDO avec second membre.*

A.1 EDO linéaire d'ordre un

La forme générale d'une EDO linéaire d'ordre un est

$$u'(x) + au(x) = f(x), \tag{A.1}$$

où u est la fonction inconnue, f est un terme source et a un coefficient. La solution est déterminée par une condition initiale $u(0) = u^0$.

A.1.1 Equation homogène

La solution générale de l'équation homgène est $u(x) = \alpha \exp(-ax), \alpha \in \mathbb{R}$.

A.1.2 Méthode de variation de la constante

On cherche une solution particulière de l'équation complète sous la forme $u(x) = \alpha(x) \exp(-ax)$, ce qui conduit à la résolution de

$$\alpha'(x) \exp(-ax) = f(x) \quad \Rightarrow \quad \alpha(x) = \int f(x) \exp(ax).$$

La solution particulière nécessite donc de connaitre une primitive de $f(x) \exp(ax)$.

A.1.3 Exemple

L'intensité I d'un circuit électrique constitué d'une inductance L et d'une résistance R s'écrit sous la forme générale suivante

$$LI' + RI = E_0 \sin(\omega t) \quad \Leftrightarrow \quad I' + \frac{R}{L} I = \frac{E_0}{L} \sin(\omega t),$$

où le terme source $E_0 \sin(\omega t)$ représente une force électromotrice de pulsation ω qui alimente le circuit. La solution générale s'écrit

$$I(t) = \alpha \exp(-R/Lt) + \frac{E_0}{R^2 + \omega^2 L^2}(R\sin(\omega t) - \omega L \cos(\omega t)) = \alpha \exp(-R/Lt) + \frac{E_0}{Z^2}\sin(\omega t - \varphi),$$

où $Z \stackrel{\text{def}}{=} \sqrt{R^2 + \omega^2 L^2}$ est l'impédance du circuit et $\varphi \stackrel{\text{def}}{=} \arcsin(\omega L/Z)$ son déphasage. La condition initiale $I(0) = I^0$ donne finalement

$$I(t) = \left(I_0 + \frac{E_0 \omega L}{Z^2}\right) \exp(-R/Lt) + \frac{E_0}{Z^2}\sin(\omega t - \varphi),$$

où le premier terme est le régime transitoire et le second terme le régime permanent.

A.2 EDO linéaire d'ordre deux

La forme générale d'une EDO linéaire d'ordre deux est

$$au''(x) + bu'(x) + cu(x) = f(x) \tag{A.2}$$

où u est la fonction inconnue, f est un terme source et a, b, c des coefficients. La solution est déterminée par deux conditions initiales $u(0) = u^0$ et $u'(0) = v^0$.

A.2.1 Equation homogène

L'équation caractéristique permet de déterminer le discriminant Δ,

$$ar^2 + br + c = 0 \quad \Rightarrow \quad \Delta = b^2 - 4ac.$$

Le signe du discriminant donne la forme de la solution générale

- $\Delta > 0$ $\Rightarrow$ 2 racines réelles r_1 et r_2 : $u(x) = \alpha \exp(r_1 x) + \beta \exp(r_2 x)$,
- $\Delta = 0$ $\Rightarrow$ 1 racine réelle r : $u(x) = (\alpha x + \beta)\exp(rx)$,
- $\Delta < 0$ $\Rightarrow$ 2 racines complexes $\lambda^{\pm} = u \pm iv$: $u(x) = \alpha \exp(\lambda^+ x) + \beta \exp(\lambda^- x)$ dans $\mathbb{C}$,
  $u(x) = \exp(ux)(A\cos(vx) + B\sin(vx))$ dans $\mathbb{R}$.

A.2.2 Méthode de variation de la constante

Soient u_1 et u_2 les solutions de l'EDO homogène. On cherche une solution particulière de l'équation complète telle que

$$y(x) = C_1(x)u_1(x) + C_2(x)u_2(x) \quad \text{et} \quad y'(x) = C_1(x)u_1'(x) + C_2(x)u_2'(x).$$

Ces deux relations impliquent que les fonctions $C_1(x)$ et $C_2(x)$ vérifient le système suivant

$$\begin{cases} C_1'(x)u_1(x) + C_2'(x)u_2(x) = 0, \\ C_1'(x)u_1'(x) + C_2'(x)u_2'(x) = \dfrac{f(x)}{a}. \end{cases}$$

L'intégration de $C_1'(x)$ et $C_2'(x)$ donne la solution générale dans laquelle les constantes d'intégration sont déterminées par les conditions initiales.

A.2.3 Exemple (cas homogène)

Les oscillations libres d'un système amorti sont modèlisées par

$$u'' + 2\lambda u' + \omega_0^2 u = 0, \tag{A.4}$$

où λ est le coefficient d'amortissement et ω_0 la pulsation propre. L'allongement du ressort d'un système masse-ressort, l'angle avec la verticale d'un pendule et la tension d'un circuit RLC vérifient cette équation différentielle. L'équation caractéristique de (**??**) est

$$r^2 + 2\lambda r + \omega_0^2 = 0 \quad \Rightarrow \quad \Delta = 4(\lambda^2 - \omega_0^2)$$

Les valeurs relatives de λ et ω_0 permettent de distinguer les trois régimes suivants

- *pseudo-périodique* si $\lambda < \omega_0$: $u(t) = \exp(-\lambda t)\left(\alpha\cos(\omega t) + \beta\sin(\omega t)\right),\ \omega = \sqrt{\omega_0^2 - \lambda^2}$,
- *critique* si $\lambda = \omega_0$: $u(t) = \exp(-\lambda t)(\alpha t + \beta)$,
- *apériodique* si $\lambda > \omega_0$: $u(t) = \exp(-\lambda t)\left(\alpha\exp(\omega t) + \beta\exp(-\omega t)\right),\ \omega = \sqrt{\lambda^2 - \omega_0^2}$,

les constantes α et β étant déterminées par les conditions initiales. La figure **??** représente la solution pour $u(0) = 1$ et $u'(0) = 0$ ainsi que pour des valeurs du coefficient d'amortissement λ conduisant aux différents régimes : le régime pseudo-périodique correspond à un faible amortissement tandis que le régime apériodique correspond à fort amortissement.

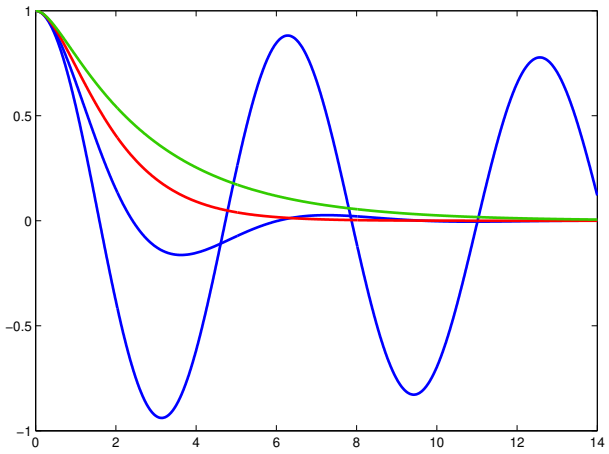


FIGURE A.1 – Influence de l'amortissement ($\lambda \in \{0.02, 0.5, 1, 1.5\}$) sur la solution de (**??**).

Annexe B

Algèbre matriciel

B.1 Norme vectorielle

Définition. *Une norme vectorielle est une application* $||\cdot|| : \mathbb{R}^n \to \mathbb{R}$ *vérifiant* $\forall u, v \in \mathbb{R}^n$,

1. $||v|| \geq 0$ *et* $||v|| = 0 \Leftrightarrow v = 0$,
2. $\forall \alpha \in \mathbb{R},\ ||\alpha v|| = |\alpha|\ ||v||$ *(homogénéité),*
3. $||u + v|| \leq ||u|| + ||v||$ *(inégalité triangulaire).*

La norme p d'un vecteur est définie par

$$\boxed{||v||_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^{n} |v_i^p| \right)^{\frac{1}{p}}}$$

Les normes les plus utilisées sont les normes 1, euclidienne (qui représente la longueur du vecteur) et infinie :

$$||v||_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} |v_i|, \quad ||v||_2 \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n} v_i^2} \quad \text{et} \quad ||v||_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} |v_i|.$$

B.2 Matrices

B.2.1 Rappels généraux

Une matrice réelle A de *dimension* $n \times p$ et de *terme général* $a_{i,j}$ est un tableau rectangulaire de coefficients réels à n lignes et p colonnes,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & a_{ij} & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix}.$$

La somme de deux matrices A et B de dimensions $n \times p$ est une matrice C de dimension $n \times p$ de terme général

$$c_{ij} = a_{ij} + b_{ij},$$

où a_{ij} et b_{ij} sont les termes généraux de A et B.

Le produit de deux matrices A et B de dimensions respectives $n \times p$ et $p \times m$ est une matrice C de dimension $n \times m$ de terme général

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}.$$

Nous considérons maintenant que la matrice A est une matrice carrée. Le déterminant de A est

$$det(A) = \sum_{s \in \sigma_n} \epsilon(s) \prod_{i=1}^{n} a_{s(i)i},$$

où σ_n est l'ensemble des permutations de $\{1, \cdots, n\}$ et $\epsilon(s)$ désigne la signature de la permutation (qui vaut 1 si la permutation est paire et -1 si la permutation est impaire).
Les *valeurs propres* λ et les *vecteurs propres* v ($\neq 0$) de A vérifient

$$Av = \lambda v.$$

Les valeurs propres sont obtenues en résolvant $det(A - \lambda I) = 0$.
Le *spectre* $Sp(A)$ de A est l'ensemble de ces valeurs propres et le *rayon spectral* $\rho(A)$ est la valeur propre de plus grand module,

$$Sp(A) = \{\lambda \in \mathbb{C}, Av = \lambda v, v \neq 0\} \quad \text{et} \quad \rho(A) = \max_{\lambda \in Sp(A)} |\lambda|.$$

La matrice A est *diagonalisable* ssi $A = PDP^{-1}$ avec D matrice diagonale constituée des valeurs propres ($d_{ii} = \lambda_i$) et P matrice de passage constituée des vecteurs propres.
La matrice A est inversible ssi $det(A) \neq 0$ et l'unique inverse A^{-1} de A vérifie

$$AA^{-1} = A^{-1}A = I,$$

où I est la matrice identité.

B.2.2 Définitions

La décomposition par *blocs* d'une matrice de dimension $n \times p$ s'écrit

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1s} \\ \vdots & A_{ij} & \vdots \\ A_{q1} & \cdots & A_{qs} \end{bmatrix},$$

où les A_{ij} sont des matrices de dimension $n_i \times n_j$ vérifiant $\sum_{i=1}^{s} n_i = n$ et $\sum_{j=1}^{q} n_j = p$.
La *transposée* $A^\top$ d'une matrice A est obtenue en permutant les lignes et les colonnes,

$$A^\top = \begin{bmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & a_{ij} & \vdots \\ a_{1p} & \cdots & a_{np} \end{bmatrix}$$

Une matrice carrée A est *symétrique* ssi $A = A^\top$,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{1n} & \cdots & a_{nn} \end{bmatrix} \quad \Leftrightarrow \quad \forall (i,j) \in \{1, \cdots, n\}^2,\ a_{ji} = a_{ij}.$$

Une matrice carrée A est *antisymétrique* ssi $A = -A^\top$,

$$A = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ -a_{12} & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & a_{n-1n} \\ -a_{1n} & \cdots & -a_{n-1n} & 0 \end{bmatrix} \quad \Leftrightarrow \quad \forall (i,j) \in \{1, \cdots, n\}^2,\ a_{ji} = -a_{ij}.$$

La diagonale d'une matrice antisymétrique est nulle.
Une matrice *pleine* contient beaucoup de coefficients non nuls tandis qu'une matrice *creuse* comporte essentiellement des coefficients nuls. Une matrice *diagonale* a des coefficients extra-diagonaux nuls,

$$a_{i,j} \neq 0 \text{ si } i = j \quad \text{et} \quad a_{i,j} = 0 \text{ si } i \neq j,$$

et une matrice *tridiagonale* a des coefficients non-nuls sur la diagonale, la sous-diagonale et la sur-diagonale,

$$a_{i,j} \neq 0 \text{ si } |i-j] \leq 1 \quad \text{et} \quad a_{i,j} = 0 \text{ si } |i-j] > 1.$$

Une matrice *triangulaire inférieure* à ses coefficients au-dessus de la diagonale qui sont nuls,

$$a_{i,j} \neq 0 \text{ si } i \geq j \quad \text{et} \quad a_{i,j} = 0 \text{ si } i < j,$$

et une matrice *triangulaire supérieure* à ses coefficients au-dessous de la diagonale qui sont nuls,

$$a_{i,j} \neq 0 \text{ si } i \leq j \quad \text{et} \quad a_{i,j} = 0 \text{ si } i > j.$$

Les positions des coefficients non nuls de ces différentes matrices sont indiquées sur la figure (**??**).
La matrice A est *définie positive* si et seulement si

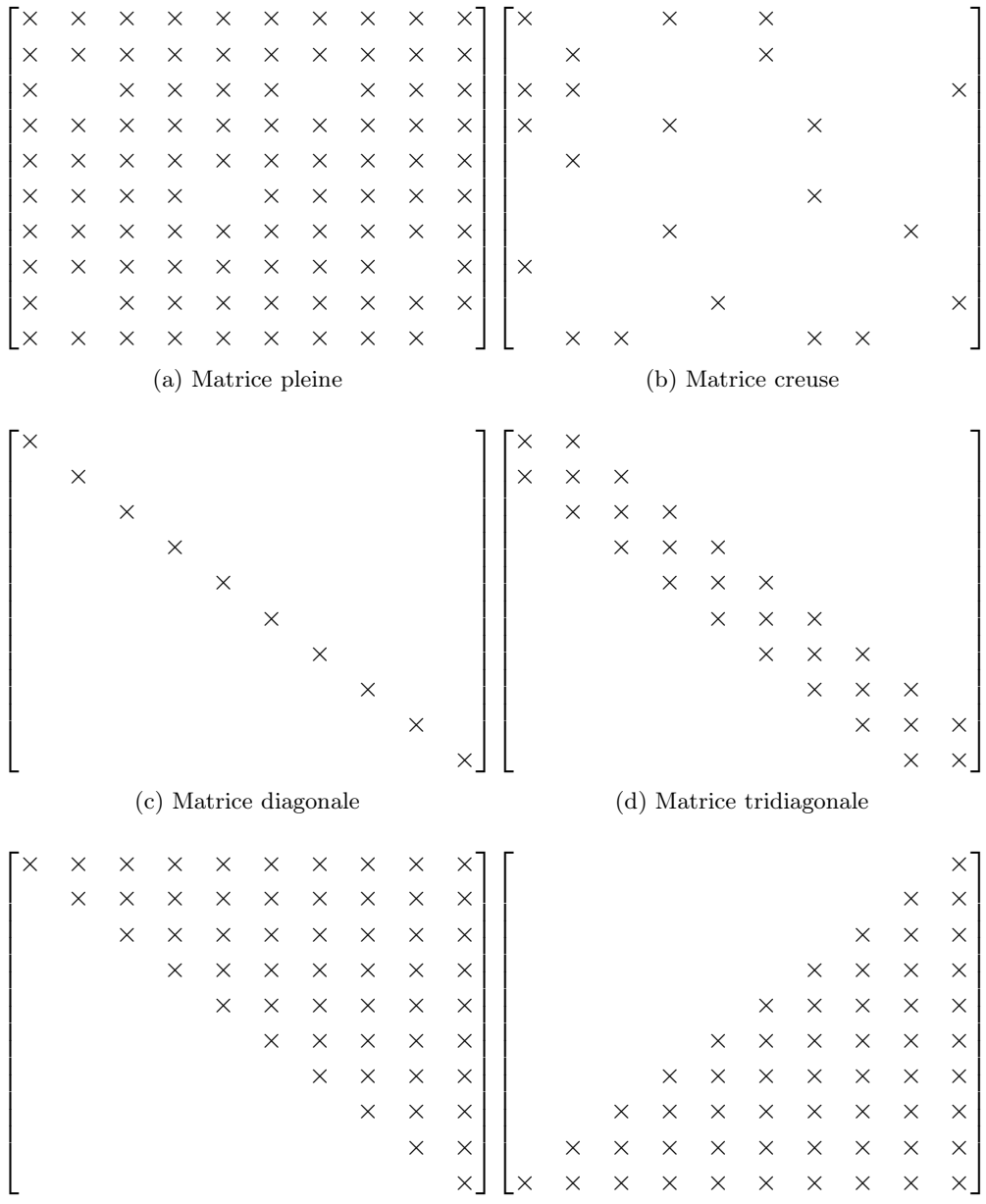
$$\forall x \in \mathbb{R}^n, \quad x \neq 0, \quad \langle Ax, x \rangle > 0.$$

La matrice A est définie positive si et seulement si toutes ses valeurs propres sont positives.

Démonstration. Supposons A définie positive. Soit (λ, v) un couple de valeur propre, vecteur propre ($v \neq 0$) associé à A,

$$Av = \lambda v \quad \Rightarrow \quad \langle Av, v \rangle = \lambda \langle v, v \rangle \quad \Rightarrow \quad \lambda = \frac{\langle Av, v \rangle}{\langle v, v \rangle} > 0.$$

Réciproquement, si $\lambda > 0$, $\langle Av, v \rangle = \lambda \langle v, v \rangle > 0$. Les vecteurs propres formant une base de $\mathbb{R}^n$, l'inégalité est vérifiée pour tout vecteur non nul de $\mathbb{R}^n$. □



(a) Matrice pleine

(b) Matrice creuse

(c) Matrice diagonale

(d) Matrice tridiagonale

(e) Matrice triangulaire supérieure

(f) Matrice triangulaire inférieure

FIGURE B.1 – Matrices particulières selon le nombre et la position des coefficients non nuls $\times$.

Annexe C

Opérateurs différentiels

Les opérateurs différentiels sont essentiels puisqu'ils sont utilisés dans la modélisation de nombreux domaines de la physique (comme la mécanique des fluides, la mécanique des solides, la thermique, l'électromagnétisme, l'acoustique, ...) *via* des équations aux dérivées partielles. Nous les donnons en coordonnées cartésiennes.

C.1 Définitions

C.1.1 Gradient

Cet opérateur est noté *grad* ou ∇. Le gradient d'un champ scalaire f est défini par le vecteur

$$\nabla f \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix}.$$

Le gradient est une extension de la dérivée classique à une fonction de plusieurs variables et indique la façon dont varie la fonction dans l'espace. Les flèches de la figure **??** représentent le gradient de la couleur la plus foncée vers la couleur la plus claire pour un champ constant suivant la direction verticale et un champ radial.

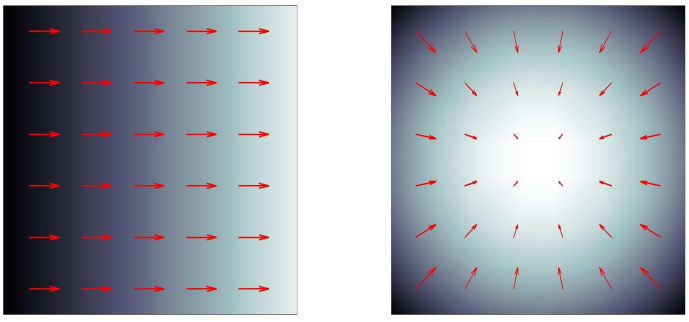


FIGURE C.1 – Exemples de gradient de champs scalaires plans.

C.1.2 Divergence

Cet opérateur est noté div ou $\nabla \cdot$. La divergence d'un champ vectoriel F est définie par le scalaire suivant

$$div(F) \stackrel{\text{def}}{=} \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z}.$$

La divergence d'un champ v mesure l'intensité de ces variations locales qui sont positives en cas d'expansion et négatives en cas de contraction comme l'illustre la figure **??**.

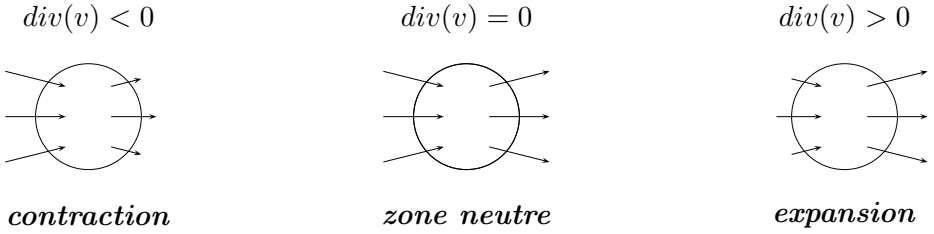


FIGURE C.2 – Types de variations locales d'un champ v selon le signe de $div(v)$.

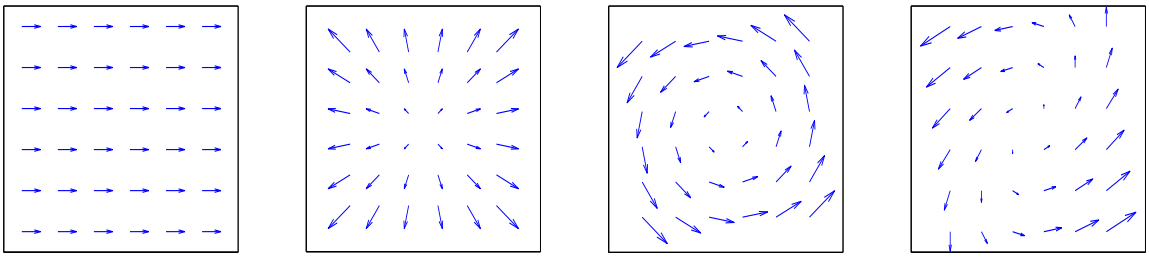
C.1.3 Rotationnel

Cet opérateur est noté rot ou $\nabla\wedge$ et le rotationnel d'un champ vectoriel F est défini par

$$rot(F) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \\ \frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \\ \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \end{pmatrix}.$$

Parmi les champs de la figure **??**, deux ont une divergence nulle et deux ont un rotationnel nul :

- v_1 vérifie $div(v_1) = 0$ et $rot(v_1) = 0$, ce champ est unidirectionnel ($v_x = \text{cste}, v_y = 0$),
- v_2 vérifie $div(v_2) \neq 0$ et $rot(v_2) = 0$, ce champ est radial ($v_x = x, v_y = y$),
- v_3 vérifie $div(v_3) = 0$ et $rot(v_3) \neq 0$, ce champ est orthoradial ($v_x = -y, v_y = x$),
- v_4 vérifie $div(v_4) \neq 0$ et $rot(v_4) \neq 0$, ($v_x = x - y, v_y = x$).



(a) v_1 (b) v_2 (c) v_3 (d) v_4

FIGURE C.3 – Exemples de champs vectoriels plans.

C.1.4 Laplacien

Cet opérateur de diffusion est noté Δ. Le laplacien d'un champ scalaire f est défini par le scalaire

$$\Delta f \stackrel{\text{def}}{=} \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

C.2 Propriétés

Nous donnons quelques propriétés des opérateurs sans indiquer les démonstrations qui sont évidentes. Soient f et g deux champs scalaires et F et G deux champs vectoriels.

Les opérateurs gradient, divergence, rotationnel et laplacien sont linéaires :

- $\nabla(f+g) = \nabla f + \nabla g$,
- $div(F+G) = div(F) + div(G)$,
- $rot(F+G) = rot(F) + rot(G)$,
- $\Delta(f+g) = \Delta f + \Delta g$.

Il est également possible d'effectuer des compositions entre les différents opérateurs :

- $div(\nabla f) = \Delta f$,
- $rot(rot(F)) = \nabla(div(F)) - \Delta F$,
- $rot(\nabla f) = 0$,
- $div(rot(F)) = 0$.