COURS D'ANALYSE NUMERIQUE

MAM 3, Polytech Lyon, 2022-2023

Ionel Sorin CIUPERCA

L'objectif de ce cours est de fournir aux étudiants une présentation et une analyse des méthodes numériques les plus couramment utilisées pour la résolution des problèmes de mathématiques appliquées.

Table des matières

Chapitre 1

Introduction et Motivation

L'Analyse Numérique (AN) est une discipline qui se situe à l'interface entre les mathématiques et l'informatique et qui permet de résoudre de manière approximative et à l'aide des algorithmes, des problèmes de mathématiques. Comme exemples de tels problèmes on peut donner : la résolution des systèmes algébriques linéaires ou non linéaires, le calcul des intégrales ou la résolution des systèmes d'équations différentielles.

L'utilisation de l'AN est beaucoup facilitée par les ordinateurs, dont l'accroisement de la rapidité de calcul permet son application dans de nombreux domaines comme les sciences physique, l'ingénierie, la biologie et la médecine, l'économie, etc ..

Dans la suite on donne deux exemples de problèmes concrets que l'AN permet de résoudre.

1. **Dans le domaine de l'ingénierie :** pour concevoir une forme d'aile d'avion la plus optimale que possible, on souhaite calculer, pour une forme d'aile donnée, la portance générée par celle-ci
   (**portance** = la force verticale qui compense la gravitation et qui permet à l'avion de se maintenir en altitude).
   Pour calculer cette force il faut connaitre , pour une vitesse donnée de l'avion, la pression exercée par l'air environnant sur la surface de l'avion et pour cela on utilise les équations de la mécanique des fluides, qui sont des équations différentielles aux dérivées partielles, satisfaites par la vitesse et la pression de l'air autour de l'avion. La résolution de ces équations serait impossible à faire "à la main", sans la contribution des algorithmes et de l'ordinateur (sauf dans des cas très particuliers et pas toujours réalistes). L'idée est de "discrétiser" les èquations différentielles à résoudre, c'est à dire, les approcher par un système algébrique linéaire ou non linéaire qui pourra être résolu à l'aide des algorithmes spécifiques.
2. **Dans le domaine de la médecine :** on modélise souvent une maladie en modélisant l'évolution d'une ou plusieurs populations de cellules ou bactéries à l'aide d'un système d'équations différentielles ordinaires (EDO), contenant plusieurs paramètres qui sont plus ou moins connus. Nous avons alors besoin des algorithmes qui permettent de résoudre le système d'EDO, pour un jeu donné des paramètres.

Très souvent cependant, certains paramètres du système EDO ne sont pas connus et on est amené à les déduire en utilisant à la fois des résultats expérimentaux et les résultats des simulations numériques (c'est ce qu'on appelle un **problème d'identification des paramètres**). On peut aussi utiliser des arguments d'ordre probabiliste ou statistique pour évaluer la crédibilité des évaluations qu'on fait pour ces paramètres.

L'approche pédagogique de ce cours repose aussi bien sur les aspects algorithmiques et calculatoires que sur la compréhension profonde des méthodes abordées. Par exemple on donnera toujours l'erreur d'approximations que nous faisons en utilisant une méthode donnée, ceci dans le but de prévoir le degré de précision qu'on peut attendre de cette méthode.

Chapitre 2

Quelque rappels d'algèbre linéaire et calcul matriciel

2.1 Notations générales

Partout dans ce cours m, n, p sont des nombres dans $\mathbb{N}^*$.

1. Nous considérons l'ensemble des **nombres complexes** :

$$\mathbb{C} = \{x + iy, \quad x, y \in \mathbb{R}\}$$

avec i le "nombre imaginaire" satisfaisant $i^2 = -1$.
Sur $\mathbb{C}$ nous avons les opérations usuelles $+$ et $\cdot$:

$$(a + ib) + (c + id) = (a + c) + i(b + d) \quad \forall\, a, b, c, d \in \mathbb{R}$$

$$(a + ib) \cdot (c + id) = (ac - bd) + i(ad + bc) \quad \forall\, a, b, c, d \in \mathbb{R}.$$

(comme pour le produit des nombre réelles, on préfère en général ne pas utiliser le symbole "$\cdot$" pour désigner le produit des nombres complexes).
Pour un nombre complexe $z = x + iy$ avec $x, y \in \mathbb{R}$ nous définissons :
— **le conjugué complexe** de z par $\bar{z} = x - iy$.
— le **module** de z par $|z| = \sqrt{x^2 + y^2}$.
Rappelons les propriétés suivantes :

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}, \quad \forall\, z_1, z_2 \in \mathbb{C}$$

$$\overline{z_1\, z_2} = \overline{z_1}\, \overline{z_2}, \quad \forall\, z_1, z_2 \in \mathbb{C}$$

$$z\, \bar{z} = |z|^2, \quad \forall\, z \in \mathbb{C}$$

$$z = \bar{z} \;\Leftrightarrow\; z \in \mathbb{R}.$$

2. Nous noterons par $\mathbb{K}$ soit l'ensemble $\mathbb{R}$ soit l'ensemble $\mathbb{C}$.

3. Nous notons par $\mathbb{K}^n$ l'espace euclidien défini par $\mathbb{K}^n = \mathbb{K} \times \mathbb{K} \times \cdots \times \mathbb{K}$ (n fois).
4. En général un vecteur $x \in \mathbb{K}^n$ sera noté $x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$ (vecteur colonne).
5. Pour tous $i, j \in \mathbb{Z}$ on notera par δ_{ij} les **symboles de Kronecker** définis par
$$\delta_{ij} = \begin{cases} 1 & \textbf{si} \quad i = j \\ 0 & \textbf{si} \quad i \neq j \end{cases}$$
6. On note par $e_1, e_2, \ldots e_n$ les vecteurs de la **base canonique** de $\mathbb{K}^n$, c'est à dire
$$(e_i)_j = \delta_{ij}, \quad \forall\, i, j = 1, 2, \cdots n.$$
7. Pour tous $a, b \in \mathbb{Z}$ avec $a < b$ on pose
$$[[a, b]] = \{k \in \mathbb{Z}, \quad a \leq k \leq b\} = [a, b] \cap \mathbb{Z}.$$
8. Soit $M \subset \mathbb{R}$ un ensemble non vide.
a) On dit que $a \in \mathbb{R}$ est un **majorant** de M si
$$x \leq a, \quad \forall\, x \in M.$$
b) On définit $\sup(M)$ comme étant : $+\infty$ si M n'a pas de majorant (donc si M n'est pas bornée supérieurement) ou le plus petit des majorants si M a au moins un majorant.
Si $\sup(M) < +\infty$ et si on pose $b = \sup(M)$ alors b satisfait les 2 propriétés suivantes :
$$x \leq b, \quad \forall\, x \in M \tag{2.1}$$
et
$$\forall\, \epsilon > 0,\ \exists x_\epsilon \in M,\ x_\epsilon > b - \epsilon. \tag{2.2}$$
Le cas le plus simple est quand le sup est "atteint" (c'est à dire quad $b = \sup(M) \in M$) ; alors la condition (2.2) est toujours satisfaite (avec $x_\epsilon = b$). Dans ce cas il suffit de vérifier seulement (2.1) (avec, bien sûr, le fait que $b \in M$). Dans ce cas on peut utiliser le notation $\max(M)$ à la place de $\sup(M)$.
Les notions de "minorant", $\inf(M)$ et $\min(M)$ sont analogues.

Le but de la première partie du cours est la résolution numérique des systèmes algébriques linéaires du type : trouver $x_1, x_2, \cdots x_n \in \mathbb{K}$ tels que
$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n = b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n = b_2 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n = b_n \end{cases}$$

avec $n \in \mathbb{N}^*$, $A_{ij}, b_i \in \mathbb{K}$, $\forall\, i,j = 1, \cdots n$ données.
Il est plus commode d'écrire ce système sous une forme plus "compacte". Pour cela on introduit la matrice carrée $A \in \mathcal{M}_n(\mathbb{K})$ donnée par

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdot & \cdot & A_{1n} \\ A_{21} & A_{22} & \cdot & \cdot & A_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{n1} & A_{n2} & \cdot & \cdot & A_{nn} \end{pmatrix}$$

et le vecteur $b \in \mathbb{K}^n$ donné par $b = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$. On mettra ensuite les inconnues sous la forme d'un vecteur $x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \in \mathbb{K}^n$. Alors le système de départ s'écrit sous la forme matricielle :
trouver le vecteur inconnu $x \in \mathbb{K}^n$ tel que

$$Ax = b.$$

2.2 Rappels sur les matrices

Dans ce cours on va noter par $\mathcal{M}_{mn}(\mathbb{K})$ l'ensemble des matrices à m lignes et n colonnes, à éléments dans $\mathbb{K}$; on notera l'ensemble des matrices carrées de taille n par $\mathcal{M}_n(\mathbb{K})$ au lieu de $\mathcal{M}_{nn}(\mathbb{K})$. On va identifier $\mathcal{M}_{m1}(\mathbb{K})$ à $\mathbb{K}^m$, donc on vera un élément de $\mathbb{K}^m$ soit comme une matrice avec une seule colonne, soit comme un vecteur colonne.
Pour une matrice $A \in \mathcal{M}_{mn}(\mathbb{K})$ on va noter par A_{ij} l'élément de $\mathbb{K}$ qui se trouve sur la ligne i et la colonne j de la matrice.

Rappel opérations élémentaires sur les matrices :

1. Adition des matrices : si $A, B \in \mathcal{M}_{mn}(\mathbb{K})$ on définit $A+B$ comme étant la matrice dans $\mathcal{M}_{mn}(\mathbb{K})$ telle que

$$(A+B)_{ij} = A_{ij} + B_{ij}, \quad \forall\, i \in [[1,m]],\ j \in [[1,n]].$$

2. Multiplication entre un scalaire et une matrice : si $\alpha \in \mathbb{K}$ et $A \in \mathcal{M}_{mn}(\mathbb{K})$ on définit αA comme étant la matrice dans $\mathcal{M}_{mn}(\mathbb{K})$ telle que

$$(\alpha A)_{ij} = \alpha A_{ij}, \quad \forall\, i \in [[1,m]],\ j \in [[1,n]].$$

3. Multiplication des matrices : si $A \in \mathcal{M}_{mn}(\mathbb{K})$ et $B \in \mathcal{M}_{np}(\mathbb{K})$ on définit AB comme étant la matrice dans $\mathcal{M}_{mp}(\mathbb{K})$ telle que

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}, \quad \forall\, i \in [[1,m]],\ j \in [[1,p]].$$

Remarques :

1. $(AB)_{ij}$ s'obtient en multipliant la ligne i de A (qui est une matrice ligne dans $\mathcal{M}_{1n}(\mathbb{K})$) par la colonne j de B (qui est une matrice colonne dans $\mathcal{M}_{n1}(\mathbb{K})$).
2. Si $A \in \mathcal{M}_{mn}(\mathbb{K})$ et $x \in \mathbb{K}^n$, en identifiant $\mathbb{K}^n$ à $\mathcal{M}_{n1}(\mathbb{K})$ on pose $Ax \in \mathbb{K}^m$ telle que

$$(Ax)_i = \sum_{k=1}^{n} A_{ik}x_k, \quad \forall\, i \in [[1,m]].$$

3. Si $A \in \mathcal{M}_{mn}(\mathbb{K})$ et e_j est le j - ème élément de la base canonique de $\mathbb{K}^n$ alors

$$A\,e_j = A_{.j} = \text{la } j \text{ - ème colonne de } A.$$

   De même si e_i est le i - ème élément de la base canonique en $\mathbb{K}^m$ alors

$$e_i^T A = A_{i.} = \text{la } i \text{ - ème ligne de } A.$$

Définition 2.1. *Pour tout matrice $A \in \mathcal{M}_{mn}(\mathbb{K})$ on définit :*

1. *la matrice* **transposée** *de A notée $A^T \in \mathcal{M}_{nm}(\mathbb{K})$ définie par*
   $(A^T)_{ij} = A_{ji}, \quad \forall\, i \in [[1,n]],\ j \in [[1,m]]$
   (les lignes de A^T sont les colonnes de A et vice versa)
2. *la matrice* **adjointe** *de A notée $A^* \in \mathcal{M}_{nm}(\mathbb{K})$ définie par*
   $(A^*)_{ij} = \overline{A_{ji}}, \quad \forall\, i \in [[1,n]],\ j \in [[1,m]]$
   où $\overline{A_{ji}}$ représente le conjugué complexe de A_{ji}

Remarque : Si $\mathbb{K} = \mathbb{R}$ alors $A^T = A^*$.
Rappelons les résultats suivants :

Proposition 2.1. *Pour tous matrices $A, B \in \mathcal{M}_{mn}(\mathbb{K})$ on a*

1. $(A^T)^T = A$
2. $(A^*)^* = A$
3. $(A+B)^T = A^T + B^T$
4. $(A+B)^* = A^* + B^*$
5. $(AB)^T = B^T A^T$
6. $(AB)^* = B^* A^*$.

Nous avons

Définition 2.2. *1. Pour tous $x, y \in \mathbb{K}^n$ on définit le* **produit scalaire** *de x par y noté $< x, y >$ ou $\langle x, y\rangle$ ou $x \cdot y$ ou $< x \,|\, y >$ par la relation*

$$< x, y > = \langle x, y\rangle = y^* x = \sum_{i=1}^{n} x_i \overline{y_i}.$$

(Remarque : si $\mathbb{K} = \mathbb{R}$ on retrouve le produit scalaire habituel en $\mathbb{R}$: $< x, y > = \sum_{i=1}^{n} x_i y_i$).

2. On dit que deux vecteurs $x, y \in \mathbb{K}^n$ sont **orthogonaux** *noté $x \perp y$ si $< x, y > = 0$.*

3. Pour tout vecteur $x \in \mathbb{K}^n$ on définit la **norme euclidienne** *de x notée $\|x\|$ comme étant la nombre positif $\|x\| = \sqrt{< x, x >} = \sqrt{|x_1|^2 + |x_2|^2 + \cdots |x_n|^2}$.*
(remarque : si $\mathbb{K} = \mathbb{R}$ alors $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$).

On utilisera souvent le résultat suivant :

Proposition 2.2.

$$\langle Ax\,,\, y\rangle \;=\; \langle x\,,\, A^* y\rangle, \quad \forall\, A \in \mathcal{M}_{mn}(\mathbb{K}),\; x \in \mathbb{K}^n,\; y \in \mathbb{K}^m.$$

Démonstration. La preuve est très simple :

$$\langle Ax\,,\, y\rangle \;= y^* A x = (A^* y)^* x = \;\langle x\,,\, A^* y\rangle.$$

□

Rappelons maintent les notions suivantes d'algèbre linéaire :

Définition 2.3. *Soit $A \in \mathcal{M}_{mn}(\mathbb{K})$. On appelle*

1. **noyau** *de A noté $Ker(A)$ l'ensemble*

$$Ker(A) = \{x \in \mathbb{K}^n, \quad Ax = 0\} \;\subset \mathbb{K}^n$$

(rappel : $Ker(A)$ est un sous-espace vectoriel de $\mathbb{K}^n$)

2. **image** *de A noté $Im(A)$ l'ensemble*

$$Im(A) = \{Ax,\; x \in \mathbb{K}^n\} \;\subset \mathbb{K}^m$$

(rappel : $Im(A)$ est un sous-espace vectoriel de $\mathbb{K}^m$).
Remarque : *nous avons*

$$Im(A) = \left\{ \sum_{j=1}^{n} x_j A_{.j}, \quad x_1, x_2, \cdots x_n \in \mathbb{K} \right\} \quad \text{où} \quad A_{.j} \;\text{ est la }\; j \text{ - ème colonne de } A$$

ce qui nous dit que $Im(A)$ est l'ensemble des combinaisons linéaires des colonnes de A (donc $Im(A)$ est le sous-espace vectoriel de K^m engendré par les colonnes de A).

3. **rang** *de A noté $rang(A)$ la dimension de $Im(A)$, donc $rang(A) = dim(Im(A))$ (c'est le nombre maximal des colonnes de A indépendantes en $\mathbb{K}^m$).*

Rappelons le résultat suivant (qu'on appelle aussi **théorème du rang**) :

Proposition 2.3. *Pour toute matrice $A \in \mathcal{M}_{mn}(\mathbb{K})$ on a*

$$dim(Ker(A)) + rang(A) = n.$$

2.3 Le cas particulier des matrices carrées

On va considérer dans cette partie des matrices carrées (avec le même nombre de lignes et de colonnes).
On suppose connue la notion de **determinant** qu'on peut voir comme une application de $\mathcal{M}_n(\mathbb{K})$ dans $\mathbb{K}$; pour toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ on va noter par $det(A) \in \mathbb{K}$ son determinant.
Rappelons les formules suivantes :

$$det(A^T) = det(A) \quad \text{et} \quad det(A^*) = \overline{det(A)}, \quad \forall A \in \mathcal{M}_n(\mathbb{K})$$

$$det(AB) = det(BA) = det(A)\, det(B), \quad \forall\, A, B \in \mathcal{M}_n(\mathbb{K}).$$

Rappelons ensuite que nous notons par $I_n \in \mathcal{M}_n(\mathbb{K})$ la **matrice identité** définie par $(I_n)_{ij} = \delta_{ij}$ pour tous $i, j \in [[1, n]]$. Cette matrice a la propriété remarquable suivante :

$$AI_n = I_n A = A, \quad \forall\, A \in \mathcal{M}_n(\mathbb{K}).$$

Rappelons les définitions suivantes :

Définition 2.4. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que A est*

1. **triangulaire inférieure** *si $A_{ij} = 0$ pour tous i, j tels que $1 \leq i < j \leq n$*
2. **triangulaire supérieure** *si $A_{ij} = 0$ pour tous i, j tels que $1 \leq j < i \leq n$*
3. **diagonale** *si $A_{ij} = 0$ pour tous $i, j \in [[1, n]]$ avec $i \neq j$*
   (c'est à dire, elle est à la fois triangulaire inférieure et triangulaire supérieure).

Notation : On notera souvent par $diag(\alpha_1, \alpha_2, \cdots \alpha_n) \in \mathcal{M}_n(\mathbb{K})$ une matrice diagonale dont les éléments diagonaux sont dans l'ordre : $\alpha_1, \alpha_2, \cdots \alpha_n \in \mathbb{K}$.
Rappel : Pour toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ qui est soit triangulaire inférieure soit triangulaire supérieure (soit diagonale), nous avons

$$det(A) = A_{11} A_{22} \cdots A_{nn} = \prod_{i=1}^{n} A_{ii}.$$

Définition 2.5. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que A est* **inversible** *ou* **régulière** *ou* **non singulière** *s'il existe $B \in \mathcal{M}_n(\mathbb{K})$ telle que $AB = BA = I_n$.*
Alors B est unique et s'appelle **l'inverse** *de A notée A^{-1}.*
Une matrice qui n'est pas inversible s'appelle matrice **non inversible** *ou* **singulière**.

Définition 2.6. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que*

1. *A est* **hermitienne** *ou* **autoadjointe** *si $A^* = A$.*
2. *A est* **symétrique** *si $A^T = A$.*
   (si $\mathbb{K} = \mathbb{R}$ alors "hermitienne" ou "symétrique" c'est la même chose).

3. *A est* **unitaire** *si $A^*A = AA^* = I_n$ (donc $A^{-1} = A^*$). Dans le cas $\mathbb{K} = \mathbb{R}$ on dit* **orthogonale** *au lieu de "unitaire"; donc $A \in \mathcal{M}_n(\mathbb{R})$ est dite orthogonale si $A^TA = AA^T = I_n$ (donc $A^{-1} = A^T$).*
4. *A est* **normale** *si $A^*A = AA^*$.*

Remarque 2.1. *Il est facile de voir que A est unitaire si et seulement si les colonnes de A sont des vecteurs orthogonaux de norme 1 en $\mathbb{K}^n$. Cela veut dire que si on note par $A_1, A_2, \cdots A_n$ les colonnes de A alors $< A_j \ , \ A_i > = \ \delta_{ij}, \ \forall \ i, j \in [[1, n]]$*
(car $< A_j \ , \ A_i > = A_i^* A_j = (A^*A)_{ij} = (I_n)_{ij} = \delta_{ij}$)
ce qui veut dire aussi que les colonnes de A forment une **base orthonormée** *en $\mathbb{K}^n$.*

Rappelons les résultats suivants :

Proposition 2.4. *Soient $A, B \in \mathcal{M}_n(\mathbb{K})$ des matrices inversibles. Alors*

1. *Pour tout $\alpha \in \mathbb{K} \setminus \{0\}$ la matrice αA est inversible et $(\alpha A)^{-1} = \frac{1}{\alpha} A^{-1}$*
2. *AB est inversible et $(AB)^{-1} = B^{-1}A^{-1}$*
3. *A^* est inversible et $(A^*)^{-1} = (A^{-1})^*$*
   (car $A^*(A^{-1})^* = (A^{-1}A)^* = I_n^* = I_n$ et aussi $(A^{-1})^*A^* = (AA^{-1})^* = I_n^* = I_n$).
   Pour $A \in \mathcal{M}_n(\mathbb{R})$ inversible cela donne : A^T est inversible et $(A^T)^{-1} = (A^{-1})^T$.
4. *A^{-1} est inversible et $(A^{-1})^{-1} = A$.*

Proposition 2.5. *(Théorème des matrices inversibles).*
Soit $A \in \mathcal{M}_n(\mathbb{K})$. Alors les propositions suivantes sont équivalentes :

1. *A est inversible*
2. $\det(A) \neq 0$
3. *$rang(A) = n$ (donc les n colonnes de A forment une base en $\mathbb{K}^n$)*
4. *Le système linéaire homogène : trouver $x \in \mathbb{K}^n$ tel que $Ax = 0$ a pour l'unique solution $x = 0$ (c'est à dire : $Ker(A) = \{0\}$).*
5. *Pour tout $b \in \mathbb{K}^n$, le système linéaire homogène : trouver $x \in \mathbb{K}^n$ tel que $Ax = b$ a une solution unique (cette solution est donnée par $x = A^{-1}b$).*

Remarque 2.2. *Si deux matrices $A, B \in \mathcal{M}_n(\mathbb{K})$ sont tels que $AB = I_n$ alors A et B sont inversibles et on a $A = B^{-1}$ et $B = A^{-1}$*
(car $AB = I_n \Rightarrow det(A)\, det(B) = det(I_n) = 1$ donc $det(A) \neq 0$ et $det(B) \neq 0$, donc A et B sont inversibles; en multipliant l'égalité $AB = I_n$ à gauche par A^{-1} on trouve $B = A^{-1}$ et en la multipliant à droite par B^{-1} on trouve $A = B^{-1}$).

Proposition 2.6. *(Méthode de Cramer)*
Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice inversible et $b \in \mathbb{K}^n$. Alors la solution du système algébrique linéaire $Ax = b$ est donnée par

$$x_i = \frac{\det(\tilde{A}_i)}{\det(A)}, \quad \forall \ i \in [[1, n]]$$

où $\tilde{A}_i \in \mathcal{M}_n(\mathbb{K})$ est la matrice obtenue de A en remplaçant la i - ème colonne de A par la colonne b.

Rappelons dans la suite les notions de valeur propre et vecteur propre d'une matrice carrée.

Définition 2.7. *Soit $A \in \mathcal{M}_n(\mathbb{K})$.*

1. *On dit que $\lambda \in \mathbb{C}$ est une* **valeur propre** *de A si $\exists x \in \mathbb{C}^n$ avec $x \neq 0$ tel que $Ax = \lambda x$; on dira alors que x est un* **vecteur propre** *de A associé à la valeur propre λ.*
   Ceci est équivalent au fait que λ est une racine du **polynome caractéristique** *P_A, donc $P_A(\lambda) = 0$, où par définition*

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

   Rappelons que P_A est un polynome de degrée n et qu'il existe toujours au moins une racine de P_A donc au moins une valeur propre de A (conséquence du **Théorème de d'Alembert-Gauss***).*
2. *On appelle* **spectre** *de A noté $Sp(A)$ l'ensemble des valeurs propres de A, donc $Sp(A) = \{\lambda \in \mathbb{C},\ \lambda$ valeur propre de $A\}$ (remarquons que $Sp(A)$ est toujours non vide).*
3. *On appelle* **rayon spectral** *de A noté $\rho(A)$ le nombre réel positif définit par $\rho(A) = \max\{|\lambda|,\ \lambda \in Sp(A)\}$.*

Remarque 2.3. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. Alors A est inversible si et seulement si $0 \notin Sp(A)$ (car A inversible $\Leftrightarrow \det(A) \neq 0 \Leftrightarrow \det(A - 0I_n) \neq 0 \Leftrightarrow P_A(0) \neq 0 \Leftrightarrow 0 \notin Sp(A)$).*

Nous utiliserons souvent le résultat suivant :

Proposition 2.7. *Soit $A \in \mathcal{M}_n(\mathbb{K})$.*
a) *Soient $\alpha, \beta \in \mathbb{K}$ avec $\beta \neq 0$. Alors $\lambda \in \mathbb{C}$ est valeur propre de A si et seulement si $\alpha + \beta\lambda$ est valeur propre de $\alpha I_n + \beta A$.*
b) *Supposons que A est une matrice inversible. Alors $\lambda \in \mathbb{C}$ est valeur propre de A si et seulement si $\lambda \neq 0$ et $\frac{1}{\lambda}$ est valeur propre de A^{-1}.*

Démonstration. à faire en TD. □

Rappel : On peut toujours mettre le polynome caractéristique P_A sous la forme :

$$P_A(\lambda) = (-1)^n(\lambda - \mu_1)^{m_1}(\lambda - \mu_2)^{m_2} \cdots (\lambda - \mu_k)^{m_k}$$

avec $k \in [[1, n]]$ le nombre des valeurs propres distinctes de A, avec $\mu_1, \mu_2, \cdots \mu_k \in \mathbb{C}$ les valeurs propres distintes de A (en fait $Sp(A) = \{\mu_1, \mu_2, \cdots, \mu_k\}$) et $m_1, m_2, \cdots m_k \in [[1, n]]$ tels que $m_1 + m_2 + \cdots + m_k = n$ les multiplicités algébriques respectives des valeurs propres. Nous avons le résultat suivant qui se démontre très facilement :

Proposition 2.8. *Supposons que $A \in \mathcal{M}_n(\mathbb{K})$ est une matrice triangulaire inférieure ou triangulaire supérieure. Alors les valeurs propres de A (répétées selon leur multiplicitées) sont les éléments diagonaux de A.*

Démonstration. Le résultat est une conséquence immédiate de l'égalité

$$det(A - \lambda I_n) = \prod_{i=1}^{n}(A_{ii} - \lambda), \quad \forall\, \lambda \in \mathbb{C}.$$

□

2.4 Opérations par blocs sur les matrices

Une matrice $A \in \mathcal{M}_{m,n}(\mathbb{K})$ peut toujours s'écrire sous la forme

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdot & \cdot & A_{1p} \\ A_{21} & A_{22} & \cdot & \cdot & A_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{k1} & A_{k2} & \cdot & \cdot & A_{kp} \end{pmatrix}$$

où pour tout $i \in [[1,k]]$ et $j \in [[1,p]]$ A_{ij} est un bloc (ou sous-matrice) avec m_i lignes et n_j colonnes. Il est clair qu'on a

$$\sum_{i=1}^{k} m_i = m \quad \text{et} \quad \sum_{j=1}^{p} n_j = n.$$

Exemple : *en classe*

Opérations :

Addition par blocs : Si deux matrices A et B s'écrivent par blocs comme $A = \begin{pmatrix} A_{11} & \cdots & A_{1p} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ A_{k1} & \cdots & A_{kp} \end{pmatrix}$ et respectivement $B = \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ B_{k1} & \cdots & B_{kp} \end{pmatrix}$ et si A_{ij} et B_{ij} ont les mêmes dimensions alors la matrice somme $C = A + B$ peut s'écrire par blocs sous la forme $C = \begin{pmatrix} C_{11} & \cdots & C_{1p} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ C_{k1} & \cdots & C_{kp} \end{pmatrix}$ avec $C_{ij} = A_{ij} + B_{ij}, \ \forall\, i \in [[1,k]], \ j \in [[1,p]]$.

Multiplication par blocs : Si deux matrices A et B s'écrivent par blocs comme $A =$

$\begin{pmatrix} A_{11} & \cdots & A_{1p} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ A_{k1} & \cdots & A_{kp} \end{pmatrix}$ et respectivement $B = \begin{pmatrix} B_{11} & \cdots & B_{1q} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ B_{p1} & \cdots & B_{pq} \end{pmatrix}$ alors la matrice produit $C = AB$ peut s'écrire par blocs sous la forme $C = \begin{pmatrix} C_{11} & \cdots & C_{1q} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ C_{k1} & \cdots & C_{kq} \end{pmatrix}$ avec

$C_{ij} = \sum_{l=1}^{p} A_{il}B_{lj}, \quad \forall\, i \in [[1,k]],\ j \in [[1,q]],$ à condition que les multiplications matricielles $A_{il}B_{lj}$ puissent se faire (nombre des colonnes de A_{il} égal au nombre des lignes de B_{lj}).

Exemples : (souvent utilisés)

Si $A \in \mathcal{M}_{mn}(\mathbb{K})$ et $B \in \mathcal{M}_{np}(\mathbb{K})$ et si on écrit B sous la forme $B = (B_1, B_2, \cdots, B_p)$, où $B_1, \cdots B_p$ sont les colonnes de B (chaque colonne étant un élément de $\mathcal{M}_{n1}(\mathbb{K})$) alors la matrice produit AB s'écrit par colonnes comme $AB = (AB_1, AB_2, \cdots AB_p)$ (chaque colonne AB_i de AB est un élément de $\mathcal{M}_{m1}(\mathbb{K})$).

De la même manière si on écrit A sous la forme $A = \begin{pmatrix} A_1 \\ A_2 \\ \cdot \\ \cdot \\ A_m \end{pmatrix}$ où $A_1, \cdots A_m$ sont les lignes de A (chaque ligne étant un élément de $\mathcal{M}_{1n}(\mathbb{K})$) alors la matrice produit AB s'écrit par lignes comme $AB = \begin{pmatrix} A_1B \\ A_2B \\ \cdot \\ \cdot \\ A_mB \end{pmatrix}$

2.5 Réduction des matrices carrées

Rappelons les définitions suivantes :

Définition 2.8. *Soit $A, B \in \mathcal{M}_n(\mathbb{K})$. On dit que A et B sont* **semblables** *en $\mathbb{K}$ s'il existe $P \in \mathcal{M}_n(\mathbb{K})$ matrice inversible telle que $B = PAP^{-1}$.*

Remarque : Si A et B sont semblables en $\mathbb{K}$ alors aussi B et A sont semblables en $\mathbb{K}$ (car $B = PAP^{-1} \Leftrightarrow A = P^{-1}BP = P^{-1}B(P^{-1})^{-1}$).

Proposition 2.9. *Deux matrices semblables ont les mêmes polynomes caractéristiques, donc deux matrices semblables ont les mêmes valeurs propres avec les mêmes multiplicités respectives.*

Démonstration. Soient $A, B \in \mathcal{M}_n(\mathbb{K})$ tels qu'il existe $P \in \mathcal{M}_n(\mathbb{K})$ matrice inversible avec

$B = PAP^{-1}$. Alors pour tout $\lambda \in \mathbb{C}$ on a

$$P_B(\lambda) = det(B - \lambda I_n) = det(PAP^{-1} - \lambda P I_n P^{-1}) = det[P(A - \lambda I_n)P^{-1}] = \quad (2.3)$$
$$= det(P) det(A - \lambda I_n) det(P^{-1}). \quad (2.4)$$

Comme $det(P)\, det(P^{-1}) = 1$ alors $P_B(\lambda) = P_A(\lambda)$, $\forall\, \lambda \in \mathbb{C}$ ce qui donne le résultat. □

On a alors le résultat évident suivant :

Corollaire 2.1. *Si une matrice $A \in \mathcal{M}_n(\mathbb{K})$ est semblable à une matrice $B \in \mathcal{M}_n(\mathbb{K})$ qui est triangulaire supérieure ou triangulaire inférieure, alors les valeurs propres de A (répétées selon leurs multiplicitées) sont les éléments diagonaux de B.*

Définition 2.9. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que A est* **diagonalisable** *en $\mathbb{K}$ s'il existe $B \in \mathcal{M}_n(\mathbb{K})$ matrice diagonale telle que A est B sont semblables en $\mathbb{K}$ (donc $\exists P \in \mathcal{M}_n(\mathbb{K})$ matrice inversible telle que $A = PBP^{-1}$).*

Nous avons

Proposition 2.10. *$A \in \mathcal{M}_n(\mathbb{K})$ est diagonalisable en $\mathbb{K}$ si et seulement si il existe une base en $\mathbb{K}^n$ des vecteurs propres de A, avec les valeurs propres respectives en $\mathbb{K}$. En plus nous avons $A = PBP^{-1}$ avec $P, B \in \mathcal{M}_n(\mathbb{K})$ avec B matrice diagonale, P matrice inversible et pour tout $j \in [[1, n]]$ l'élément diagonal B_{jj} de B est une valeur propre de A avec vecteur propre associé la $j-$ ème colonne de P.*

Démonstration. A est diagonalisable en $\mathbb{K} \Leftrightarrow$ il existe $B \in \mathcal{M}_n(\mathbb{K})$ matrice diagonale et $P \in \mathcal{M}_n(\mathbb{K})$ matrice inversible telles que $A = PBP^{-1}$, c'est à dire (en multipliant à droite par la matrice inversible P) :

$$AP = PB. \quad (2.5)$$

On écrit P par colonnes : $P = (P_1, P_2, \cdots P_n)$. D'autre part, comme B est une matrice diagonale, on peut l'écrire par colonnes sous la forme $B = (B_{11}e_1, B_{22}e_2 \cdots B_{nn}e_n)$. Alors (2.5) s'écrit

$$(AP_1, AP_2, \cdots AP_n) = (B_{11}Pe_1, B_{22}Pe_2, \cdots B_{nn}Pe_n).$$

Remarquons que $Pe_j = P_j, \quad \forall\, j \in [[1, n]]$, donc la relation précédente est équivalente à

$$AP_j = B_{jj}\, P_j, \quad \forall\, j \in [[1, n]].$$

Nous avons aussi $P_j \neq 0, \quad \forall\, j \in [[1, n]]$ (car $det(P) \neq 0$). Alors le fait que A est diagonalisable en $\mathbb{K}$ est équivalent avec le fait que pour tout j de 1 à n la colonne j de P est un vecteur propre de la matrice A correpondant à la valeur propre B_{jj}. Comme P est inversible alors les colonnes de P forment une base en $\mathbb{K}^n$. □

Rappelons le résultat suivant, qui est une conséquence de la proposition précédente :

Proposition 2.11. *Si $A \in \mathcal{M}_n(\mathbb{K})$ admet n valeurs propres distinctes en $\mathbb{K}$ alors A est diagonalisable en $\mathbb{K}$.*

Rappelons les deux résultats suivants utiles pour des matrices qui peuvent être non diagonalisables :

Proposition 2.12. *(résultat admis) Toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ est semblable en $\mathbb{C}$ à sa forme de Jordan, c'est à dire $\exists P \in \mathcal{M}_n(\mathbb{C})$ inversible telle que $A = PJP^{-1}$ avec $J \in \mathcal{M}_n(\mathbb{C})$ qui s'écrit par blocs sous la forme*

$$J = \begin{pmatrix} J_{11} & 0 & 0 & \cdots & 0 \\ 0 & J_{22} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdot & \cdots & J_{kk} \end{pmatrix}$$

où $J_{ii} \in \mathcal{M}_{m_i}(\mathbb{C})$ avec $m_1 + m_2 + \cdots m_k = n$. Ici chaque matrice J_{ii} est un bloc de Jordan, c'est à dire
- soit $m_i = 1$ (c'est à dire, J_{ii} peut être vu comme un nombre complexe)
- soit $m_i > 1$ et J_{ii} est de la forme $J_{ii} = \begin{pmatrix} \lambda_i & 1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & \lambda_i & 1 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_i \end{pmatrix}$
(donc J est une matrice triangulaire supérieure).

Proposition 2.13. Lemme de Schur *(résultat admis) Pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ il existe $T, U \in \mathcal{M}_n(\mathbb{C})$ avec T triangulaire supérieure et U unitaire, telles que $A = UTU^*$ (donc les matrices A et T sont semblables).*

Nous avons la conséquence suivante :

Corollaire 2.2. *i) Si $A \in \mathcal{M}_n(\mathbb{C})$ est une matrice hermitienne alors elle est diagonalisable ; en plus elle est semblable en C à une matrice diagonale réelle et il existe une base orthonormée en $\mathbb{C}^n$ des vecteurs propres de A. Plus précisément A s'écrit sous la forme*

$$A = UDU^*$$

avec $U \in \mathcal{M}_n(\mathbb{C})$ matrice unitaire et $D \in \mathcal{M}_n(\mathbb{R})$ diagonale.

ii) En plus si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique alors il existe une base orthonormée des vecteurs propres de A en $\mathbb{R}^n$. Plus précisément A s'écrit sous la forme

$$A = UDU^T$$

avec $U \in \mathcal{M}_n(\mathbb{R})$ matrice orthogonale et $D \in \mathcal{M}_n(\mathbb{R})$ diagonale.

Démonstration. i) Du Lemme de Schur on déduit l'existence de $T, U \in \mathcal{M}_n(\mathbb{C})$ avec T triangulaire supérieure et U unitaire, telles que $A = UTU^*$. On a alors $A^* = (U^*)^* T^* U^* = UT^*U^*$. D'autre part nous avons $A = A^*$, donc $UTU^* = UT^*U^*$ et par simplification on déduit $T = T^*$. Comme T est triangulaire supérieure alors T est diagonale réelle ce qui nous donne le résultat.

ii) Partie admise. □

On a aussi le résultat suivant :

Proposition 2.14. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ et $k \in \mathbb{N}^*$. Si les valeurs propres de A (répétées selon leurs multiplicitées) sont $\lambda_1, \lambda_2, \cdots \lambda_n$ alors les valeurs propres de A^k (répétées selon leurs multiplicitées) sont $\lambda_1^k, \lambda_2^k, \cdots \lambda_n^k$.*

Démonstration. Du Lemme de Schur on déduit l'existence de $T, U \in \mathcal{M}_n(\mathbb{C})$ avec T triangulaire supérieure et U unitaire, telles que $A = UTU^*$. Alors $\lambda_1, \lambda_2, \cdots \lambda_n$ sont les éléments diagonaux de T. D'autre part, on montre facilement par récurrence sur k (*Exercice*) que

$$A^k = UT^kU^*.$$

D'autre part on montre facilement que T^2 est aussi une matrice triangulaire supérieure et qu'on a l'égalité $(T^2)_{ii} = (T_{ii})^2$ pour tout i. Plus généralement, on montre par récurrence sur k que T^k est une matrice triangulaire supérieure avec

$$(T^k)_{ii} = (T_{ii})^k, \quad \forall\, i \in [[1, n]].$$

Alors les valeurs propres de A^k sont les éléments diagonaux de T^k donc les $\lambda_1^k, \lambda_2^k, \cdots \lambda_n^k$. □

Rappel définition : pour une matrice $A \in \mathcal{M}_n(\mathbb{K})$ l'application

$$x \in \mathbb{K}^n \ \mapsto < Ax\ ,\ x > = \ Ax \cdot x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \overline{x_i} x_j \ \in \mathbb{K}$$

s'appelle **forme quadratique** associée à la matrice A.
On finit ce chapitre en considérant deux cas particuliers très importants des matrices. Nous commençons par préciser la convention suivante qui sera utilisée dans tout ce cours :
Si on écrit une inégalité faisant intervenir un nombre complexe alors on sous-entend qu'on affirme aussi que ce nombre est réel.

Définition 2.10. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que A est*

1. **positive** *(ou* **semi-définie positive***) si* $Ax \cdot x \geq 0, \quad \forall\, x \in \mathbb{K}^n$
2. **définie positive** *si* $Ax \cdot x > 0, \quad \forall\, x \in \mathbb{K}^n \setminus \{0\}$.

Remarques :

1. Conformément à la convention ci-dessus l'affirmation $Ax \cdot x \geq 0, \quad \forall\, x \in \mathbb{K}^n$, doit être comprise comme $Ax \cdot x \in \mathbb{R}, \quad \forall\, x \in \mathbb{K}^n$ et $Ax \cdot x \geq 0, \quad \forall\, x \in \mathbb{K}^n$.
   De même pour $Ax \cdot x > 0, \quad \forall\, x \in \mathbb{K}^n \setminus \{0\}$.
2. Une méthode usuelle pour montrer qu'une matrice est définie positive est de montrer d'abord qu'elle est positive et ensuite montrer l'implication :

$$Ax \cdot x = 0 \ \Rightarrow \ x = 0.$$

Définition 2.11. *Soit $A \in \mathcal{M}_n(\mathbb{K})$. Pour tout $p \in [[1,n]]$ la matrice $A_p \in \mathcal{M}_p(\mathbb{K})$ qui s'obtient de A en retenant seulement les premières p lignes et colonnes de A est appellée* **sous-matrice principale** *de A*

$$(donc \quad (A_p)_{kl} = A_{kl}, \quad \forall\, k,l \in [[1,p]]).$$

Proposition 2.15. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice définie positive (respectivement positive). Alors on a*

1. *Tous les éléments A_{jj} de la diagonale de A sont > 0 (respectivement ≥ 0).*
2. *Toute sous-matrice principale A_p de A est définie positive (respectivement positive).*

Démonstration. 1. Prendre $x = e_j$ dans la Definition 2.10. On a

$$< Ae_j\,,\, e_j > = A_{jj} > 0 \quad (respectivement\ \geq 0).$$

2. Considérons le cas $p < n$. Pour tout vecteur $x \in \mathbb{R}^p$ on considère le vecteur $\tilde{x} \in \mathbb{R}^n$ tel que pour tout $k \in [[1,n]]$ on a

$$(\tilde{x})_k = \begin{cases} x_k & \text{si} \quad 1 \leq k \leq p \\ 0 & \text{si} \quad k > p \end{cases}$$

(c'est à dire on a l'écriture par blocs $\tilde{x} = \begin{pmatrix} x \\ 0 \end{pmatrix}$). Nous utilisons l'écriture par blocs de A :

$$A = \begin{pmatrix} A_p & B \\ C & D \end{pmatrix}$$

avec $B \in \mathcal{M}_{p,n-p}(\mathbb{K}), C_{n-p,p}(\mathbb{K})$ et $D \in \mathcal{M}_{n-p,n-p}(\mathbb{K})$. Nous avons

$$A\tilde{x} = \begin{pmatrix} A_p & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} A_p x \\ Cx \end{pmatrix}$$

ce qui nous donne

$$\langle A\tilde{x}\,,\, \tilde{x} \rangle = \langle \begin{pmatrix} A_p x \\ Cx \end{pmatrix}\,,\, \begin{pmatrix} x \\ 0 \end{pmatrix} \rangle = \langle A_p x\,,\, x \rangle$$

On suppose $x \neq 0$, donc $\tilde{x} \neq 0$ ce qui implique $\langle A\tilde{x}\,,\, \tilde{x} \rangle > 0$ grâce à l'hypothèse sur A. L'égalité précédente nous donne alors $\langle A_p x\,,\, x \rangle > 0$ ce qui finit la preuve.

□

Nous avons aussi

Proposition 2.16. *Toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ qui est définie positive est inversible.*

Démonstration. Il suffit de montrer $Ker(A) = \{0\}$. Supposons par absurd qu'il existe $x \in \mathbb{K}^n$ avec $x \neq 0$ tel $Ax = 0$; ceci donne $< Ax\,,\, x > = 0$ ce qui contredit le fait que A est définie positive. □

Nous avons encore

Lemme 2.1. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne et considérons $\lambda_1, \lambda_2, \cdots \lambda_n$ les valeurs propres de A (en comptant les multiplicitées) rangées en ordre croissant :*

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

(rappelons que les valeurs propres de A sont réelles). Alors on a

$$Ax \cdot x \geq \lambda_1 \|x\|^2, \quad \forall\, x \in \mathbb{K}^n.$$

Démonstration. On a $A = PDP^*$ avec D matrice diagonale réelle avec $\lambda_1, \lambda_2, \ldots \lambda_n$ sur la diagonale et $P \in \mathcal{M}_n(\mathbb{K})$ matrice unitaire. Alors pour tout $x \in \mathbb{K}^n$ on a

$$< Ax, x > = x^*Ax = x^*PDP^*x = (P^*x)^*DP^*x = < DP^*x\,,\, P^*x > = < Dy\,,\, y >$$

où on a noté $y = P^*x$. Nous avons aussi

$$\|x\|^2 = x^*x = x^*PP^*x = (P^*x)^*P^*x = < P^*x\,,\, P^*x > = \|P^*x\|^2 = \|y\|^2$$

et

$$< Ax\,,\, x > = < Dy\,,\, y > = \sum_{i,j=1}^{n} D_{ij} y_j \overline{y_i} = \sum_{i=1}^{n} \lambda_i |y_i|^2 \geq \lambda_1 \sum_{i=1}^{n} |y_i|^2 = \lambda_1 \|y\|^2 = \lambda_1 \|x\|^2$$

ce qui donne le résultat. □

Remarque : Une manière équivalente d'énoncer le résultat du Lemme 2.1 est :

$$Ax \cdot x \geq \lambda_{min} \|x\|^2, \quad \forall\, x \in \mathbb{K}^n.$$

où λ_{min} est la plus petite valeur propre de A.

On peut alors montrer les deux résultats suivants :

Proposition 2.17. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne. Alors les trois propositions suivantes sont équivalentes :*

1. *A est définie positive*
2. *Les valeurs propres de A sont strictement positives*
3. *Il existe $\alpha > 0$ telle que*

$$Ax \cdot x \geq \alpha \|x\|^2, \quad \forall\, x \in \mathbb{K}^n.$$

Démonstration. 2). $\Rightarrow$ 3).
C'est une conséquence immédiate du Lemme 2.1 : prendre $\alpha = \lambda_1 > 0$.
3.) $\Rightarrow$ 1).
Si $x \in \mathbb{K}^n,\ x \neq 0$ alors $Ax \cdot x \geq \alpha \|x\|^2 > 0$.
1.) $\Rightarrow$ 2.)
On raisonne par absurd : si $\lambda \leq 0$ est une valeur propre de A alors il existe $y \in \mathbb{K}^n,\ y \neq 0$ tel que $Ay = \lambda y$. En faisant le produit scalaire avec y on trouve

$$Ay \cdot y = \lambda y \cdot y = \lambda \|y\|^2 \leq 0.$$

ce qui contredit le fait que A est définie positive. □

Proposition 2.18. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne. Alors les deux propositions suivantes sont équivalentes :*

1. *A est positive*
2. *Les valeurs propres de A sont positives.*

Démonstration. 2.) $\Rightarrow$ 1).
Du Lemme 2.1 si $x \in \mathbb{K}^n$ alors $Ax \cdot x \geq \lambda_1 \|x\|^2 \geq 0$ (car $\lambda_1 \geq 0$).
1.) $\Rightarrow$ 2.)
On raisonne par absurd : si $\lambda < 0$ est une valeur propre de A alors il existe $y \in \mathbb{K}^n,\ y \neq 0$ tel que $Ay = \lambda y$. En faisant le produit scalaire avec y on trouve

$$Ay \cdot y = \lambda y \cdot y = \lambda \|y\|^2 < 0.$$

ce qui contredit le fait que A est positive. □

Rappelons le critère suivant, facile à utiliser :

Proposition 2.19. ***(critère de Sylvester)***
Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne. Alors A est définie positive si et seulement si

$$det(A_p) > 0, \quad \forall p \in [[1, n]]$$

où $A_1, A_2, \cdots A_n$ sont les sous-matrices principales de A
(rappel : $A_p = (A_{ij})_{i,j=1,\cdots p} \in \mathcal{M}_p(\mathbb{K})$).

Chapitre 3

Approximation numérique des EDP linéaires par la méthode des différences finies

Dans la suite on se donne $\Omega \subset \mathbb{R}^n$ (très souvent nous avons $n = 3$) un ensemble ouvert et une fonction $f : \Omega \to \mathbb{R}$ continue. On cherche une fonction $u : \Omega \to \mathbb{R}$ dans $C^2(\Omega) \cap C(\overline{\Omega})$ telle que

$$-\Delta u(x) = f(x), \quad \forall\, x = (x_1, x_2, x_3) \in \Omega \tag{3.1}$$

avec condition limite sur la frontière de Ω :

$$u(x) = 0, \quad \forall\, x \in \partial\Omega \tag{3.2}$$

où on a utilisé la notation

$$\Delta u(x) = \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2} \qquad \textbf{opérateur de Laplace}.$$

L'équation (3.1) s'appelle **équation de Laplace**. La condition limite (3.2) s'appelle **condition de Dirichlet homogène**. On pourrait considérer une condition limite plus générale à la place de (3.2) :

$$u(x) = g(x), \quad \forall\, x \in \partial\Omega$$

avec $g : \partial\Omega \to \mathbb{R}$ une fonction donnée (condition de Dirichlet non-homogène).
Cette équation de Laplace est utilisé comme modèle mathématique dans plusieurs domaines scientifiques : physique, biologie, etc .. On va détailler dans la suite l'application de ce model dans le domaine de la thermodynamique : **l'équation de la chaleur**.
Ici $u(x)$ représente la température à l'équilibre au point $x \in \Omega$ et $f(x)$ est la source de chaleur interne au point x. On supposera donc que u et f ne dépendent pas du temps.
On considère la fonction $\Phi : \overline{\Omega} \to \mathbb{R}^n$ (champ des vecteurs) qui à tout $x \in \overline{\Omega}$ associe le vecteur $\Phi(x)$ qui représente le **flux de chaleur** en x. La grandeur physique "flux de chaleur" indique en quelle direction et à quelle intensité se propage la chaleur ; on l'appelle

aussi *flux thermique* ou encore *flux thermique par unité de surface.*
On considère ω un sous-ensemble de Ω ; pour un point $x \in \partial\omega$ nous notons par ν ou $\nu(x)$ le vecteur normal en x à la frontière $\partial\omega$ orienté vers l'extérieur de ω, $\nu \in \mathbb{R}^n$.
On a l'égalité suivante :

$$\int_{\partial\omega} \Phi \cdot \nu \, d\sigma = \int_{\omega} f(x) \, dx$$

Cette égalité est la traduction d'une loi physique qui dit que à l'équilibre le flux de chaleur sortant de ω est égal à la chaleur produite par la source de chaleur sur ω. D'autre part, la formule de Stokes nous donne

$$\int_{\partial\omega} \Phi \cdot \nu \, d\sigma = \int_{\omega} \nabla \cdot \Phi \, dx$$

(ou $\nabla \cdot \phi = \sum_{i=1}^{n} \frac{\partial \Phi_i}{\partial x_i}$ est la divergence de Φ).
On déduit alors

$$\int_{\omega} \nabla \cdot \Phi \, d\sigma = \int_{\omega} f(x) \, dx.$$

Comme cette égalité est valable pour tout sous-ensemble ω de Ω, on déduit

$$\nabla \cdot \Phi = f. \tag{3.3}$$

(c'est la loi de conservation de la quantité de chaleur).
D'autre part on a la loi de Fourier qui dit que le vecteur flux est proportionnel au vecteur gradient de la température avec sens opposé, ce qui se traduit mathématiquement par la relation

$$\Phi = -\alpha \nabla u \tag{3.4}$$

avec $\alpha > 0$ une constante physique appelée coefficient de diffusion (*remarque : α peut aussi être une fonction de la position x*). Pour simplifier on va prendre ici $\alpha = 1$, ce qui avec (3.3) et (3.4) nous donne $-\Delta u = f$, donc (3.1).
En supposant que sur la frontière $\partial\Omega$ la température est maintenue à une valeur constante (qui sera considérée comme valeur de référence, donc la valeur 0) on obtient la condition limite (3.2).
Remarque : D'autres conditions limite à la place de (3.2) peuvent être considérées, comme par exemple

1. $\frac{\partial u}{\partial \nu} = 0$ sur $\partial\Omega$ (condition de **Neuman homogène**) ; cette condition est utilisée quand on peut supposer que le flux de chaleur sur la frontière $\partial\Omega$ est orthogonal au vecteur ν, donc quand il n'y a pas de perte ou gain de chaleur par la frontière (*rappel :* $\frac{\partial u}{\partial \nu} = \nabla u \cdot \nu = -\Phi \cdot \nu = \Phi \cdot (-\nu)$ c'est le flux thermique **rentrant**).
2. $\frac{\partial u}{\partial \nu} = \beta(u_{ext} - u)$ sur $\partial\Omega$ avec $u_{ext} \in \mathbb{R}$ une constante qui représente la température à l'extérieur de Ω supposée connue et constante. Cette condition s'appelle condition limite de **Robin** qui dit que le flux rentrant est proportionnel à la différence entre la température extérieure et la température dans Ω ; ici $\beta > 0$ est une constante physique donnée.

3. En général on doit considérer aussi la dépendance en temps de la température et éventuellement de la source de chaleur f. Alors l'inconnue u d'épend aussi de la variable temp notée en général par t (donc $u = u(x,t)$). Dans ce cas u satisfera un **problème d'évolution** qu'on va décrire dans la suite ; le problème (3.1) - (3.2) va s'appeller **problème stationnaire**. Nous obtenons l'équation principale du problème d'évolution en ajoutant dans l'équation (3.1) le term $\frac{\partial u}{\partial t}$, donc cette équation principale s'écrit

$$\frac{\partial u}{\partial t} - \Delta u(x) = f(x,t), \quad \forall\, x = (x_1, x_2, x_3) \in \Omega,\ t \geq 0.$$

Comme dans le cas stationnaire il faut aussi une condition sur la frontière, comme par exemple une condition de Dirichlet homogène, qui va s'écrire sous la forme

$$u(x,t) = 0, \quad \forall\, x \in \partial\Omega, \quad t \geq 0.$$

A la différence du cas stationnaire il faut aussi une **condition initiale**, c'est à dire, on suppose que u est connue au moment initial supposé $t = 0$, donc

$$u(x,0) = u_0(x), \quad \forall\, x \in \Omega$$

avec $u_0 : \Omega \to \mathbb{R}$ une fonction donnée.
Mais dans la suite nous nous limitons au cas stationnaire qui est plus simple à traiter.

Dans la suite nous donnons une méthode de résolution numérique de (3.1) et (3.2), c'est à dire, une manière d'obtenir une solution approximative de (3.1) et (3.2). Cette méthode s'appelle **méthode des différences finies**.

Cas 1. On suppose ici $n = 1$ (c'est le cas unidimensionnel) ; on supposera aussi $\Omega =]a,b[$ avec $a, b \in \mathbb{R},\ a < b$. Physiquement cette hypothèse simplificatrice peut être faite dans le cas où on peut supposer que f et u ne dépendent pas de x_2 et x_3, donc elles dépendent uniquement de x_1.
Alors on se donne $f : [a,b] \to \mathbb{R}$ avec $f \in C(]a,b[)$ et on cherche $u : [a,b] \to \mathbb{R}$ avec $u \in C^2(]a,b[) \cap C([a,b])$ telle que

$$-u''(x) = f(x), \quad \forall\, x \in]a,b[\tag{3.5}$$

et

$$u(a) = u(b) = 0 \tag{3.6}$$

(pour simplifier l'écriture on note $x_1 = x$).
Remarque : *Les équations* (3.5) *et* (3.6) *se résoudent très facilement par l'intégration de* (3.5) *:*

$$-u'(x) = \int_a^x f(t)\,dt + C_1$$

et ensuite

$$-u(x) = \int_a^x \left(\int_a^y f(t)\, dt \right) dy + C_1\, x + C_2$$

et on calcule les constantes $C_1, C_2 \in \mathbb{R}$ *telles que les deux conditions de* (3.6) *soient satisfaites.*
Mais le but ici est de faire de l'approximation numérique.

L'idée est de faire une **discrétisation** de l'intervalle $[a, b]$. On considère $N \in \mathbb{N}^*$ avec N assez grand, on pose $h = \frac{b-a}{N+1}$ et ensuite

$$x_i = a + ih, \quad i \in [[0, N+1]]$$

c'est à dire

$$x_0 = a, \quad x_1 = a + h, \quad x_2 = a + 2h, \quad \cdots \quad x_{N+1} = a + (N+1)h = b.$$

On va noter par U_i une approximation numérique de $u(x_i)$ (donc u_i sera inconnu) et on pose $f_i = f(x_i)$ (f_i sera connue) et ceci pour tout $i \in [[0, N+1]]$.
Nous allons utiliser la définition de la dérivée de la fonction u dans un point x (quand elle existe) :

$$u'(x) = \lim_{t \to 0} \frac{u(x+t) - u(x)}{t}.$$

Dans notre cas on va prendre N très "grand" ce qui par définition de h va donner h très "petit".
Comme la solution u qu'on cherche est une fonction de classe C^2 alors $u'(x_i)$ peut être approché par $\frac{u(x_{i+1})-u(x_i)}{h}$ ou par $\frac{u(x_{i-1})-u(x_i)}{-h} = \frac{u(x_i)-u(x_{i-1})}{h}$ ce qu'on peut écrire par

$$u'(x_i) \approx \frac{u(x_{i+1}) - u(x_i)}{h}, \quad \text{ceci pour tout} \quad i \in [[0, N]] \tag{3.7}$$

ou respectivement

$$u'(x_i) \approx \frac{u(x_i) - u(x_{i-1})}{h}, \quad \text{ceci pour tout} \quad i \in [[1, N+1]]. \tag{3.8}$$

D'autre part, pour tout i de 1 à N, nous pouvons approcher $u''(x_i)$ par $\frac{u'(x_i)-u'(x_{i-1})}{h}$ (utiliser (3.8) avec u' à la place de u). En utilisant l'approximation (3.7) d'abord avec i et ensuite avec $i-1$ à la place de i, on déduit que $u''(x_i)$ peut être approché par

$$\frac{1}{h} \left[\frac{u(x_{i+1}) - u(x_i)}{h} - \frac{u(x_i) - u(x_{i-1})}{h} \right]$$

On trouve donc :

$$u''(x_i) \approx \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}, \quad \forall \quad i \in [[1, N]]. \tag{3.9}$$

Remarque : ce n'est pas la seule possibilité pour approcher $u''(x_i)$: on peut aussi utiliser

$$u''(x_i) \approx \frac{u'(x_{i+1}) - u'(x_i)}{h} \approx \frac{1}{h}\left[\frac{u(x_{i+2}) - u(x_{i+1})}{h} - \frac{u(x_{i+1}) - u(x_i)}{h}\right]$$

mais (3.9) est la plus utilisée pour des raisons de "symétrie".
On se propose maintenant de voir quelle erreur on fait quand on utilise l'approximation (3.9). Pour cela on suppose que $u \in C^4(]a, b[)$ et on utilise le développement de Taylor ; nous avons

$$u(x_{i+1}) = u(x_i + h) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + O(h^4)$$

Rappel : *Pour tout $k \in \mathbb{R}$ l'expression $O(h^k)$ signifie une expression qui a la propriété :*

$$\left|\frac{O(h^k)}{h^k}\right| \leq C, \quad \textit{si } h \textit{ est "proche" de } 0$$

avec $C \geq 0$ une constante indépendante de h.
De même, $o(h^k)$ signifie une expression qui a la propriété :

$$\frac{o(h^k)}{h^k} \to 0, \quad si \;\; h \to 0.$$

Nous avons aussi

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

En faisant la somme des deux expressions on déduit :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + O(h^4)$$

ce qui donne

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

On conclut alors que nous faisons une erreur d'orde h^2 (on dira : une erreur d'ordre 2) en faisant l'approximation (3.9) ; bien sûr, à condition que la solution u soit de classe C^4 pour que les développements de Taylor précédentes soient valables.
Rappelons aussi que pour tout $j \in [[0, N]]$ on approche $u(x_j)$ par $U_j \in \mathbb{R}$ qui sera une inconnue du problème. En utilisant ceci en (3.9) on arrive à l'approximation

$$u''(x_i) \approx \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2}, \quad \forall \quad i \in [[1, N]]. \tag{3.10}$$

Remarque : l'approximation (3.10) est une approximation directement implémentable, alors que l'approximation (3.9) est une approximation "théorique".
Alors une approximation de (3.5) en $x = x_i$ sera

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = f_i, \quad i = 1, 2, \cdots N$$

On a ici N équations mais avec $N+2$ inconnues qui sont $U_0, U_1, \cdots U_{N+1}$.
On utilisera maintenant les conditions aux limites (3.6). Comme la solution u du problème est connue aux points x_0 et x_{N+1} (elle est égale à 0), il est naturel d'imposer que l'approximation coincide avec la solution exacte dans ces points ; on pose alors

$$U_0 = 0 \quad \text{et} \quad U_{N+1} = 0.$$

On aura alors à résoudre le système algébrique linéaire :

$$\begin{cases} -U_{i-1} + 2U_i - U_{i+1} = h^2 f_i, & i = 1, 2, \cdots N \\ U_0 = U_{N+1} = 0 \end{cases}$$

Ce système s'écrit, en posant $b_i = h^2 f(x_i),\ i = 1, \cdots N$:

$$\begin{cases} 2U_1 - U_2 = b_1 \\ -U_1 + 2U_2 - U_3 = b_2 \\ \cdots \\ -U_{i-1} + 2U_i - U_{i+1} = b_i \\ \cdots \\ -U_{N-1} + 2U_N = b_N \end{cases} \tag{3.11}$$

On va écrire ce sytème sous forme matricielle : on pose $U = \begin{pmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ U_n \end{pmatrix}$ le vecteur inconnue,

$b = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$ un vecteur donne et on introduit la matrice $A \in \mathcal{M}_n(\mathbb{R})$ définie par

$$A_{ij} = \begin{cases} 2 & \text{si} \quad i = j \\ -1 & \text{si} \quad |i-j| = 1 \\ 0 & \text{si} \quad |i-j| \geq 2 \end{cases}$$

Cette matrice s'appelle **matrice du laplacien** en dimension 1 ou matrice du laplacien 1D.
Remarque : *La matrice A est symétrique.*
Le système (3.11) s'écrira alors sous la forme matricielle

$$\text{Trouver } U \in \mathbb{R}^N \text{ tel que} \quad AU = b. \tag{3.12}$$

Remarque : *En fait les éléments du vecteur U dépendent de N qui est le nombre des points intérieurs de discrétisation ; on pourrait noter par $U^{(N)}$ ce vecteur et par $U_i^{(N)}$ sa $i-$ ème composante ; mais pour simplifier les notations on ne mettra pas pas en évidence la dépendence en N de U.*
On admet le résultat suivat, qui justifie le fait que la solution du problème approché est "proche" de la solution exacte, dans les points de discrétisation, quand N est grand :

Proposition 3.1. *On a*

$$\max_{i=1,\cdots N} |U_i^{(N)} - u(x_i)| \to 0, \quad \textit{pour } N \to +\infty.$$

Cas 2. $n = 2$ (le cas bidimensionnel).
Physiquement cette hypothèse simplificatrice peut être faite dans le cas où on peut supposer que f et u ne dépendent pas de x_3, donc elles dépendent uniquement de x_1 et x_2.
On suppose pour simplifier qu'on est dans le cas particulier

$$\Omega =]a_1, b_1[\times]a_2, b_2[$$

avec $a_1, b_1, a_2, b_2 \in \mathbb{R},\ a_1 < b_1,\ a_2 < b_2$ (Ω est un rectangle ouvert en $\mathbb{R}^2$).
On considère encore $N \in \mathbb{N}$ très grand et on pose

$$h_1 = \frac{b_1 - a_1}{N+1} \quad \text{et} \quad h_2 = \frac{b_2 - a_2}{N+1}.$$

Nous introduisons pour tous $i, j \in [[0, N+1]]$ les points dans $\mathbb{R}^2$:

$$y_{i,j} = (a_1 + ih_1,\ a_2 + jh_2)$$

et remarquons que tous ces points sont dans $\bar{\Omega}$.
On approchera alors l'EDP (3.1) en chaque point $y_{i,j}$ et on va noter $U_{i,j} \in \mathbb{R}$ une approximation de la solution u de (3.1) au point $y_{i,j}$.
En s'inspirant de ce qu'on a fait en dimension 1, on peut écrire pour tous $i, j \in [[1, N]]$:

$$\frac{\partial^2 u}{\partial x_1^2}(y_{i,j}) \approx \frac{u(y_{i-1,j}) - 2u(y_{i,j}) + u(y_{i+1,j})}{h_1^2} \approx \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h_1^2}$$

et aussi

$$\frac{\partial^2 u}{\partial x_2^2}(y_{i,j}) \approx \frac{u(y_{i,j-1}) - 2u(y_{i,j}) + u(y_{i,j+1})}{h_2^2} \approx \frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{h_2^2}.$$

Supposons pour simplifier qu'on a $b_1 - a_1 = b_2 - a_2$, ce qui donne $h_1 = h_2$ qu'on va noter par h. En faisant la somme, on déduit

$$\Delta u(y_{i,j}) \approx \frac{U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{i,j}}{h^2}.$$

On va alors approcher le problème par le système algébrique linéaire :
trouver $U_{i,j},\ i, j \in [[0, N+1]]$ tels que

$$\begin{cases} -U_{i-1,j} - U_{i,j-1} + 4U_{i,j} - U_{i,j+1} - U_{i+1,j} = b_{i,j}, \quad \forall\, i, j \in [[1, N]] \\ U_{0,j} = U_{N+1,j} = 0, \quad \forall\, j \in [[0, N+1]] \\ U_{i,0} = U_{i,N+1} = 0, \quad \forall\, j \in [[1, N]] \end{cases} \tag{3.13}$$

où nous notons $b_{i,j} = h^2 f(y_{i,j})$. Les premières N^2 égalités de (3.13) (ligne 1 de (3.13)) viennent de (3.1) et les dernères $4N+4$ égalités (lignes 2 et 3) viennent de (3.2). Remarquons qu'on a finalement un système avec N^2 équations et N^2 inconnues : $U_{i,j},\ i,j \in [[1,N]]$.
Pour écrire ce système de manière matricielle on utilise l'écriture par blocs. Le vecteur inconnu $U \in \mathbb{R}^{N^2}$ peut s'écrire sous la forme $U = \begin{pmatrix} U^{(1)} \\ U^{(2)} \\ \cdot \\ \cdot \\ U^{(N)} \end{pmatrix}$, avec pour tout $i \in [[1,N]]$ le vecteur $U^{(i)} \in \mathbb{R}^N$ donné par $U^{(i)} = \begin{pmatrix} U_{i,1} \\ U_{i,2} \\ \cdot \\ \cdot \\ U_{i,N} \end{pmatrix}$.
Nous introduisons aussi le vecteur donné $b \in \mathbb{R}^{N^2}$ qui peut s'écrire sous la forme $b = \begin{pmatrix} b^{(1)} \\ b^{(2)} \\ \cdot \\ \cdot \\ b^{(N)} \end{pmatrix}$, avec pour tout $i \in [[1,N]]$ le vecteur $b^{(i)} \in \mathbb{R}^N$ donné par $b^{(i)} = \begin{pmatrix} b_{i,1} \\ b_{i,2} \\ \cdot \\ \cdot \\ b_{i,N} \end{pmatrix}$.
En prenant $i = 1$ dans la première ligne de (3.13) on a

$$-U_{1,j-1} + 4U_{1,j} - U_{1,j+1} - U_{2,j} = b_{1,j}, \quad \forall\ j \in [[1,N]] \tag{3.14}$$

avec $U_{1,0} = U_{1,N+1} = 0$.
Pour un i général de 2 à $N-1$ on aura

$$-U_{i-1,j} - U_{i,j-1} + 4U_{i,j} - U_{i,j+1} - U_{i+1,j} = b_{i,j}, \quad \forall\ j \in [[1,N]] \tag{3.15}$$

avec $U_{i,0} = U_{i,N+1} = 0$.
Finalement pour $i = N$ on a

$$-U_{N-1,j} - U_{N,j-1} + 4U_{N,j} - U_{N,j+1} = b_{i,j}, \quad \forall\ j \in [[1,N]] \tag{3.16}$$

avec $U_{N,0} = U_{N,N+1} = 0$.
D'autre part on note $\tilde{A} \in \mathcal{M}_n(\mathbb{R})$ la matrice définie par

$$\tilde{A}_{ij} = \begin{cases} 4 & \text{si} \quad i = j \\ -1 & \text{si} \quad |i-j| = 1 \\ 0 & \text{si} \quad |i-j| \geq 2 \end{cases}$$

On observe alors qu'on peut écrire (3.14) sous la forme matricielle

$$\tilde{A}U^{(1)} - U^{(2)} = b^{(1)}$$

ensuite (3.15) sous la forme

$$-U^{(i-1)} + \tilde{A}U^{(i)} - U^{(i+1)} = b^{(i)}, \quad \forall\, i \in [[2, N-1]]$$

et finalement (3.16) sous la forme

$$-U^{(N-1)} + \tilde{A}U^{(N)} = b^{(N)}.$$

Remarque : pour tout $i \in [[1, N]]$ on va écrire $I_N U^{(i)}$ à la place de $U^{(i)}$.
Alors le système algébrique linéaire (3.13) peut s'écrire sous la forme matricielle

$$\text{trouver} \quad U \in \mathbb{R}^{N^2} \quad \text{tel que} \quad AU = b$$

avec $A \in \mathcal{M}_{N^2}(\mathbb{R})$ une matrice qui s'écrit par blocs sous la forme

$$A = (A_{ij})_{i,j \in [[1,N]]}$$

où $A_{ij} \in \mathcal{M}_N(\mathbb{R})$ est une matrice bloc définie par

$$A_{ij} = \begin{cases} \tilde{A} & \text{si} \quad i = j \\ -I_N & \text{si} \quad |i-j| = 1 \\ 0 & \text{si} \quad |i-j| \geq 2 \end{cases}$$

Cette matrice A s'appelle **matrice du laplacien 2D**.
Remarque : *Il est possible aussi de discrétiser le problème* (3.1) - (3.2) *posé sur un domaine bidimensionnel* Ω *borné général. Dans ce cas l'idée serait de considérer un carré* $D \subset \mathbb{R}^2$ *tel que* $\Omega \subset D$. *On discrétise alors le carrée* D *comme on vient de le faire ci-dessus dans le cas où* Ω *était un carré. Il faut alors écrire des équations linéaire comme dans la ligne 1 de* (3.13) *pour les points* $y_{i,j}$ *de* Ω *qui sont "loin" de* $\partial\Omega$ *et prendre* $U_{i,j} = 0$ *pour les points* $y_{i,j} \in \Omega$ *qui sont "proches" de* $\partial\Omega$. *On n'associe pas des inconnues aux points de* D *qui ne sont pas dans* Ω.

Chapitre 4

Méthodes de résolution numérique des systèmes algébriques linéaires

4.1 Normes des matrices

4.1.1 Rappel définition et équivalence des normes

Rappelons d'abord la définition générale d'une norme.

Définition 4.1. *Soit E un espace vectoriel sur $\mathbb{K}$. On dit qu'une application $\|\cdot\| : E \to \mathbb{R}$ est une* **norme** *sur E si elle satisfait les 3 conditions suivantes :*

1. $$\|x\| \geq 0, \quad \forall\, x \in E$$
   avec en plus
   $$\|x\| = 0 \;\Leftrightarrow\; x = 0$$
2. $$\|\lambda x\| = |\lambda|\,\|x\|, \quad \forall\, \lambda \in \mathbb{K},\ \forall\, x \in E$$
   (homogénéité)
3. $$\|x + y\| \leq \|x\| + \|y\|, \quad x, y \in E.$$
   (inégalité triangulaire).

On a

Définition 4.2. *On dit que deux normes $\|\cdot\|_a$ et $\|\cdot\|_b$ sur un espace vectoriel E sont* **équivalentes** *s'il existe deux contantes C_1, C_2 avec $0 < C_1 \leq C_2$, tels que*

$$C_1\|x\|_a \leq \|x\|_b \leq C_2\|x\|_a, \quad \forall\, x \in E.$$

Dans la suite on donne deux exemples d'espaces vectoriels sur lesquels on définit des normes.

4.1.2 Normes dans l'espace euclidien

On considère ici $E = \mathbb{K}^n$ (l'espace euclidien) ; nous rappelons les normes qu'on peut définir sur cet espace vectoriel (on admet sans preuve que ce sont des normes) :

1. La norme euclidienne, notée $\|\cdot\|_2$: pour tout $x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \in \mathbb{K}^n$ on pose

$$\|x\|_2 = \sqrt{< x, x >} = \sqrt{|x_1|^2 + |x_2|^2 + \cdots |x_n|^2} = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

   Remarque : Dans le cas où $\mathbb{K} = \mathbb{R}$ alors $\|x\|$ représente gémétriquement la distance euclidienne du point x de $\mathbb{R}^n$ à l'origine 0.

2. La "norme 1", notée $\|\cdot\|_1$: pour tout $x \in \mathbb{K}^n$ on pose

$$\|x\|_1 = |x_1| + |x_2| + \cdots |x_n| = \sum_{i=1}^{n} |x_i|$$

3. La "norme ∞", notée $\|\cdot\|_\infty$: pour tout $x \in \mathbb{K}^n$ on pose

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \cdots |x_n|\} = \max\{|x_i|, \quad i \in [[1, N]]\}.$$

4. La "norme α" pour un $\alpha \in [1, +\infty[$ arbitraire, notée $\|\cdot\|_\alpha$: pour tout $x \in \mathbb{K}^n$ on pose

$$\|x\|_\alpha = (|x_1|^\alpha + |x_2|^\alpha + \cdots |x_n|^\alpha)^{1/\alpha} = \left(\sum_{i=1}^{n} |x_i|^\alpha\right)^{1/\alpha}$$

Remarque : En particulier pour $\alpha = 1$ on obtient la "norme 1" et pour $\alpha = 2$ on obtient la norme euclidienne (qu'on peut aussi appeler "norme 2"). On peut aussi montrer que la "norme α" converge vers "norme ∞" si $\alpha \to +\infty$.
Remarque : On va utiliser souvent la notation plus commode $\|x\|$ (sans indice) pour la norme euclidienne $\|x\|_2$, car cette norme sera beaucoup utilisée dans la suite.
On admet sans preuve le résultat suivant :

Proposition 4.1. *Sur $\mathbb{K}^n$ toutes les normes sont équivalentes (plus précisément, deux normes arbitraires sur $\mathbb{K}^n$ sont équivalentes). Plus généralement si E est un espace vectoriel de dimension finie alors toutes les normes sur E sont équivalentes.*

4.1.3 Normes dans l'espace des matrices

On considère ici $E = \mathcal{M}_{m,n}(\mathbb{K})$.
Un manière de définir une norme sur cet espace est la suivante : une matrice arbitraire $A \in \mathcal{M}_{m,n}(\mathbb{K})$ peut être vue comme un vecteur colonne de $\mathbb{K}^{mn}$ (en mettant les colonnes de A les unes après les autres sur une même grande colonne verticale). Alors on peut définir une norme en E sur A comme une norme de A sur l'espace euclidien $\mathbb{K}^{mn}$ (voir la partie précédente 4.1.2). Par exemple en considérant la norme euclidienne sur $\mathbb{K}^{mn}$ on peut définir

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2}.$$

qui s'appelle la **norme de Frobenius** sur E.
Mais dans ce cours on utilisera très souvent une autre manière de définir une norme sur E ; pour simplifier la présentation on va définir cette norme sur l'ensemble des matrices carrées. On commence par le résultat suivant :

Lemme 4.1. *Soit $\|\cdot\|$ une norme sur $\mathbb{K}^n$ et $A \in \mathcal{M}_n(\mathbb{K})$ une matrice. Alors il existe une constante $M \geq 0$ telle que*

$$\|Ax\| \leq M\|x\|, \quad \forall\, x \in \mathbb{K}^n.$$

Démonstration. On va montrer ce résultat en deux étapes.

1. On montre d'abord l'inégalitée demandée si on utilise la "norme 1" au lieu de $\|\cdot\|$. Nous avons pour tout $x \in \mathbb{K}^n$:

$$\|Ax\|_1 = \sum_{i=1}^{n} |(Ax)_i| = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} A_{ij}x_j \Big| \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|\,|x_j|$$

   Soit $M_1 = \max\{|A_{ij}|,\ \ i,j \in [[1,n]]\}$. On obtient alors de l'inégalité précédente :

$$\|Ax\|_1 \leq M_1 \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |x_j| \right) = M_1 \sum_{i=1}^{n} \|x\|_1$$

   ce qui nous donne

$$\|Ax\|_1 \leq M_1 n \|x\|_1, \quad \forall x \in \mathbb{K}^n. \tag{4.1}$$

2. Pour montrer le résultat demandé on utilise l'équivalence entre la norme $\|\cdot\|$ et la norme $\|\cdot\|_1$, qui nous dit : il existe $C_1, C_2 \in \mathbb{R}$ avec $0 < C_1 \leq C_2$ tels que

$$C_1\|x\|_1 \leq \|x\| \leq C_2\|x\|_1, \quad \forall\, x \in \mathbb{K}^n.$$

   En utilisant aussi (4.1) on peut écrire pour tout $x \in \mathbb{K}^n$:

$$\|Ax\| \leq C_2\|Ax\|_1 \leq C_2 M_1 n \|x\|_1 \leq C_2 M_1 n \frac{1}{C_1}\|x\|$$

   ce qui nous donne le résultat en posant $M = \frac{C_2 M_1 n}{C_1}$.

□

Remarque : *On utilisera souvent dans des inégalités concernant des normes sur $\mathbb{K}^n$ l'idée qu'il suffit de montrer l'inégalité pour une norme précise, celle qui nous convient (comme par exemple la "norme 1" dans la preuve précédente). Ensuite on procède comme dans l'étape 2 de la preuve ci-dessus.*

Nous avons la proposition suivante qui donne aussi une définition :

Proposition 4.2. *Soit $\|\cdot\|$ une norme sur $\mathbb{C}^n$ (norme de vecteur).*
Alors l'application $\|\cdot\|_s : \mathcal{M}_n(\mathbb{C}) \to \mathbb{R}$ donnée par

$$\|A\|_s = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|}, \quad \forall\, A \in \mathcal{M}_n(\mathbb{C})$$

est bien définie.
En plus cette application $\|\cdot\|_s$ est une norme sur $\mathcal{M}_n(\mathbb{C})$ appelée **norme matricielle subordonnée** *à la norme vectorielle $\|\cdot\|$ (ou* **norme d'opérateur** *associée à la norme vectorielle $\|\cdot\|$).*

Démonstration. Montrons d'abord que l'application est bien définie. Du Lemme 4.1 on déduit que pour tout $A \in \mathcal{M}_n(\mathbb{C})$ il existe $M \geq 0$ tel que

$$\frac{\|Ax\|}{\|x\|} \leq M, \quad \forall\, x \in \mathbb{C}^n,\ x \neq 0.$$

On déduit que

$$\sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} \leq M < +\infty$$

donc l'application $\|\cdot\|_s$ est bien définie.
Dans la suite on va montrer que $\|\cdot\|_s$ est bien une norme sur $\mathcal{M}_n(\mathbb{C})$). Pour cela on va montrer que les 3 conditions de la définition d'une norme (voir Définition 4.1) sont satisfaites.

1. Il est très facile de voir que

   $$\|A\|_s \geq 0, \quad \forall\, A \in \mathcal{M}_n(\mathbb{C}).$$

   Ensuite comme $0x = 0, \ \forall\, x \in \mathbb{C}^n$ on déduit facilement

   $$\|0\|_s = 0.$$

   Finalement suposons que $A \in \mathcal{M}_n(\mathbb{C})$ est tel que $\|A\|_s = 0$ et il faut montrer que $A = 0$. Nous avons

   $$\frac{\|Ax\|}{\|x\|} \leq 0, \quad \forall\, x \neq 0$$

   Ceci donne

   $$\|Ax\| \leq 0, \quad \forall\, x \neq 0$$

et comme $\|Ax\| \geq 0$ on en déduit (en utilisant aussi $A0 = 0$) :

$$\|Ax\| = 0, \quad \text{donc} \quad Ax = 0 \quad \forall\, x \in \mathbb{C}^n.$$

Nous utilisons maintenant le résultat général : si $A \in \mathcal{M}_{mn}(\mathbb{K})$ alors

$$Ax = 0 \quad \forall\, x \in \mathbb{K}^n \quad \Rightarrow\ A = 0$$

(prendre $x = e_j$ ce qui donne que la j - ème colonne de A est 0 ; comme c'est vrai pour tout $j \in [[1, n]]$ alors $A = 0$).
On déduit donc $A = 0$.

2. Soient $\lambda \in \mathbb{K}$ et $A \in \mathcal{M}_n(\mathbb{C})$. En utilisant $\|\lambda Ax\| = |\lambda|\,\|Ax\|$ on obtient

$$\|\lambda A\|_s = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|\lambda Ax\|}{\|x\|} = \sup_{x \in \mathbb{C}^n, x \neq 0} |\lambda| \frac{\|Ax\|}{\|x\|} = |\lambda| \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = |\lambda|\,\|A\|_s.$$

3. Soient $A, B \in \mathcal{M}_n(\mathbb{C})$. En utilisant l'inégalité triangulaire pour la norme vectorielle nous avons

$$\|(A + B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\|, \quad \forall\, x \in \mathbb{C}^n$$

ce qui nous donne

$$\frac{\|(A + B)x\|}{\|x\|} \leq \frac{\|Ax\|}{\|x\|} + \frac{\|Bx\|}{\|x\|}, \quad \forall\, x \in \mathbb{C}^n, x \neq 0.$$

En passant au "sup" en x pour $x \neq 0$ on obtient

$$\|A + B\|_s \leq \|A\|_s + \|B\|_s$$

ce qui finit la preuve.

□

Exemple 4.1. *En dimension 1 la norme matricielle subordonnée à toute norme n'est autre que la valeur absolue. Plus précisément, soit $A \in \mathbb{K}$ arbitraire qu'on peut voir comme une matrice dans $\mathcal{M}_{11}(\mathbb{K})$. Alors pour toute norme matricielle $\|\cdot\|$ subordonnée en $\mathcal{M}_{11}(\mathbb{K})$ on a*

$$\|A\| = |A|$$

(car si on note toujours par $\|\cdot\|$ la norme sur $\mathbb{K}$ dont est obtenue la norme matricielle subordonnée $\|\cdot\|$ alors on a

$$\|A\| = \sup_{x \in \mathbb{C}, x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{C}, x \neq 0} \frac{|A|\,\|x\|}{\|x\|} = \sup_{x \in \mathbb{C}, x \neq 0} |A| = |A| \ ).$$

La proposition suivante nous donne d'autres expressions pour la norme matricielle subordonnée :

Proposition 4.3. *Soit $\|\cdot\|$ une norme (vectorielle) sur $\mathbb{C}^n$ et $\|\cdot\|_s$ la norme matricielle subordonnée à la norme $\|\cdot\|$. Nous avons alors pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$:*

$$\|A\|_s = \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\| \tag{4.2}$$

et aussi

$$\|A\|_s = \sup_{x\in\mathbb{C}^n, \|x\|\leq 1} \|Ax\|. \tag{4.3}$$

Démonstration. Montrons d'abord (4.2) en montrant la double inégalité.
Pour tout $x \in \mathbb{C}^n, \|x\| = 1$ on a

$$\|Ax\| = \frac{\|Ax\|}{1} = \frac{\|Ax\|}{\|x\|}$$

et on peut écrire

$$\sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\| = \sup_{x\in\mathbb{C}^n, \|x\|=1} \frac{\|Ax\|}{\|x\|} \leq \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} = \|A\|_s.$$

ce qui nous donne l'inégalité "$\geq$" de (4.2).
Pour montrer inégalité inverse, pour tout $x \in \mathbb{C}^n,\ x \neq 0$ on pose $y = \frac{x}{\|x\|}$ et il est facile de voir que $\|y\| = 1$. An a alors

$$\frac{\|Ax\|}{\|x\|} = \left\| \left(\frac{1}{\|x\|}\right) Ax \right\| = \left\| A\left(\frac{x}{\|x\|}\right) \right\| = \|Ay\| \leq \sup_{z\in\mathbb{K}^n, \|z\|=1} \|Az\|.$$

En passant en "sup" en $x \in \mathbb{C}^n$ avec $x \neq 0$ on obtient l'inégalité "$\leq$" de (4.2) ce qui finit la preuve de (4.2).
Montrons maintenant l'égalité (4.3). L'inégalité "$\leq$" est immédiate, en utilisant (4.2). D'autre part de la définition de la norme matricielle subordonnée on a

$$\frac{\|Ax\|}{\|x\|} \leq \|A\|_s, \quad \forall\, x \in \mathbb{C}^n,\ x \neq 0, \ \text{ donc aussi pour tout } \ x \neq 0,\ \|x\| \leq 1.$$

On en déduit

$$\|Ax\| \leq \|A\|_s\|x\| \leq \|A\|_s, \quad \forall\, x \in \mathbb{C}^n,\ x \neq 0,\ \|x\| \leq 1.$$

Comme $\|A0\| = 0$, l'inégalité précédente est valable aussi pour $x = 0$. En passant au "sup" pour $x \in \mathbb{C}^n$ avec $\|x\| \leq 1$ on obtient le résultat. □

Remarque 4.1. *On va utiliser très souvent pour la norme subordonnée la même notation que pour la norme vectorielle à laquelle elle est subordonnée. Par exemple si $A \in \mathcal{M}_n(\mathbb{C})$ et $\alpha \in [1, +\infty]$ alors $\|A\|_\alpha$ désigne la norme matricielle de A subordonnée à la norme vectorielle $\|\cdot\|_\alpha$.*
Une notation qui est parfois utilisée dans la littérature mathémmatique pour la norme matricielle est $|||A|||_\alpha$.

Intérêt principal de la notion de norme matricielle subordonnée est donné par le résultat suivant :

Proposition 4.4. *Soit $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{C})$ subordonnée à une norme vectorielle $\|\cdot\|$ sur $\mathbb{C}^n$. Alors on a*

1.
$$\|I_n\| = 1$$

2.
$$\|Ax\| \leq \|A\| \, \|x\|, \quad \forall\, A \in \mathcal{M}_n(\mathbb{C}), \ \forall\, x \in \mathbb{C}^n$$

3.
$$\|AB\| \leq \|A\| \, \|B\|, \quad \forall\, A, B \in \mathcal{M}_n(\mathbb{C})$$

4.
$$\|A^k\| \leq \|A\|^k, \quad \forall\, A \in \mathcal{M}_n(\mathbb{C}), \quad \forall\, k \in \mathbb{N}^*.$$

Démonstration. 1. On a

$$\|I_n\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} \|I_n x\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} \|x\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} 1 = 1.$$

2. Si $x = 0$ on a
$$\|Ax\| = \|A0\| = 0 = \|A\| \, \|0\|.$$
Si $x \neq 0$ on a
$$\|A\| = \sup_{y \in \mathbb{C}^n, y \neq 0} \frac{\|Ay\|}{\|y\|} \geq \frac{\|Ax\|}{\|x\|}$$
ce qui donne le résultat.

3. On a
$$\|AB\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} \|(AB)x\|.$$
Comme $(AB)x = A(Bx)$ et $\|A(Bx)\| \leq \|A\| \, \|Bx\|$ (utiliser 2.) alors
$$\|AB\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} \|A(Bx)\| \leq \sup_{x \in \mathbb{C}^n, \|x\|=1} \|A\| \, \|Bx\| = \|A\| \sup_{x \in \mathbb{C}^n, \|x\|=1} \|Bx\| = \|A\| \, \|B\|$$
ce qui finit la preuve.

4. La preuve de ce résultat se fait très facilement par récurrence sur k (*Exercice*) en utilisant l'inégalité de la partie 3.

□

Exemples de normes matricielles subordonnées (*exemples fondamentaux*) :
Pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ on a

1.
$$\|A\|_1 = \max_{j \in [[1,n]]} \sum_{i=1}^{n} |A_{ij}|$$

2.
$$\|A\|_\infty = \max_{i\in[[1,n]]} \sum_{j=1}^{n} |A_{ij}|$$

3.
$$\|A\|_2 = \sqrt{\rho(A^*A)}$$
(Rappel : on note par $\rho(B)$ le rayon spectral d'une matrice B).

4. Si A est une matrice hermitienne alors
$$\|A\|_2 = \rho(A).$$

Démonstration. On fera en TD les preuves de 1. , 2. et 3. Faisons la preuve de 4. ; comme $A^* = A$ on a
$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)}.$$
Mais de la Proposition 2.14 on déduit que $\sqrt{\rho(A^2)} = \sqrt{(\rho(A))^2} = \rho(A)$ ce qui finit la preuve. □

Remarque 4.2. *Si $n \geq 2$ alors la norme de Frobénius sur $\mathcal{M}_n(\mathbb{C})$ n'est pas une norme matricielle subordonnée.*

Démonstration. Supposons que le contraire est vrai, donc il existe une norme vectorielle $\|\cdot\|$ telle que pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ on a
$$\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|^2} = \|A\|_s$$
où $\|\cdot\|_s$ désigne la norme matricielle subordonnée à la norme vectorielle $\|\cdot\|$. En prenant dans cette égalité $A = I_n$ et en utilisant la partie 1. de la Proposition 4.4 on obtient
$$\sqrt{n} = 1$$
ce qui est une contradiction. □

Remarque 4.3. *Pour toute norme vectorielle $\|\cdot\|$ sur $\mathbb{R}^n$ on pourrait introduire l'application $\|\cdot\|_s : \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ définie par*
$$\|A\|_s = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|}{\|x\|}, \quad \forall\, A \in \mathcal{M}_n(\mathbb{R})$$
et on peut démontrer que cette application est bien une norme sur $\mathcal{M}_n(\mathbb{R})$. Alors si on considère seulement des matrices réelles, tous les résultats et exemples ce cette section sont encore valables avec cette norme. On va voir plus loin l'intérêt d'avoir défini la norme matricielle subordonnée comme un "sup" sur $\mathbb{C}^n$ au lieu de "sup" sur $\mathbb{R}^n$.

4.2 Conditionnement des matrices (ou des systèmes algébriques linéaires)

On se donne $A \in \mathcal{M}_n(\mathbb{K})$ une matrice et $b \in \mathbb{K}^n$ un vecteur et nous considérons le systèmes algébriques linéaire

$$\text{trouver} \quad x \in \mathbb{K}^n \quad \text{tel que} \quad Ax = b. \tag{4.4}$$

Il y a des situations où une "petite" variation de la donnée b entraine une "grande" variation de la solution x. Cela peut être très génant dans la pratique car souvent les données d'un problème sont le résultats des mesures qui peuvent comporter des erreur, ou des calculs qui peuvent eux aussi comporter des erreurs de précision.
Considérons l'exemple suivant : la matrice $A \in \mathcal{M}_4(\mathbb{R})$ est donné par

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

et le vecteur $b \in \mathbb{R}^4$ est $b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$ On peut vérifier facilement que $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ est une solution du système algébrique linéaire $Ax = b$. D'autre part on peut montrer que $det(A) = 1$ donc la matrice A est inversible et donc le vecteur x ci-dessus est l'unique solution de ce système algébrique linéaire.

Considèrons maintenant un autre vecteur $\tilde{b}$ très proche de b, avec $\tilde{b} = \begin{pmatrix} 32,1 \\ 22,9 \\ 33,1 \\ 30,9 \end{pmatrix}$. En calculant l'unique solution $\tilde{x}$ du système algébrique linéaire $A\tilde{x} = \tilde{b}$ on trouve $x = \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ -1,1 \end{pmatrix}$ qui est assez "loin" de x.
Pour comprendre ce qui se passe, revenons au système général (4.4) avec A une matrice quelconque inversible. Considérons une perturbation $b + \delta b$ de la donnée b et notons par $x + \delta x$ une solution de (4.4) où on remplace b par $b + \delta b$.
Supposons en plus que $b \neq 0$ donc $x \neq 0$.
Nous avons donc

$$A(x + \delta x) = b + \delta b. \tag{4.5}$$

En faisant la différence entre (4.5) et (4.4) on obtient

$$A(\delta x) = \delta b$$

c'est à dire

$$\delta x = A^{-1}\,\delta b.$$

Considérons $\|\cdot\|$ une norme sur $\mathbb{K}^n$ et notons toujours par $\|\cdot\|$ la norme matricielle subordonnée à cette norme vectorielle.
On déduit alors de la dernière égalité :

$$\|\delta x\| \leq \|A^{-1}\|\,\|\delta b\|. \tag{4.6}$$

D'autre part, de l'égalité $Ax = b$ on en déduit

$$\|b\| \leq \|A\|\,\|x\|.$$

Comme $x, b \neq 0$ alors $\|x\| > 0$ et $\|b\| > 0$, donc on peut diviser cette inégalité par $\|x\|\,\|b\|$ et on obtient

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{4.7}$$

En multipliant (4.6) et (4.7) on obtient

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{4.8}$$

Le terme $\frac{\|\delta b\|}{\|b\|}$ s'appelle **l'erreur relative sur la donnée** b et $\frac{\|\delta x\|}{\|x\|}$ s'appelle **l'erreur relative sur la solution** x.
Cette dernière inégalité nous mène à la définition suivante :

Définition 4.3. *Soit $\|\cdot\|$ une norme matricielle subordonnée à une norme vectorielle sur $\mathbb{C}^n$ notée toujours $\|\cdot\|$ et soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice inversible. On appelle* **conditionnement** *de A relatif à la norme $\|\cdot\|$ le nombre noté $Cond(A, \|\cdot\|)$ défini par*

$$Cond(A, \|\cdot\|) = \|A\|\,\|A^{-1}\|.$$

Remarque 4.4.

1. *En général, pour simplifier les notations on va noter ce conditionnement par $Cond(A)$ quand il est clair de quelle norme il s'agit. Dans ce cas on peut dire uniquement "conditionnement de A" au lieu de dire "conditionnement de A relatif à la norme $\|\cdot\|$".*
2. *On utilise l'expression "la matrice A (ou le système algébrique linéaire associé) est* **bien conditionnée***" si le conditionnement de A "n'est pas trop grand". Dans la cas où ce conditionnement est "grand" on dira que la matrice est "***mal conditionnée***".*

Proposition 4.5. *Soit $\|\cdot\|$ une norme matricielle subordonnée sur $\mathcal{M}_n(\mathbb{C})$. Nous avons*

1. $$Cond(\alpha I_n, \|\cdot\|) = 1, \quad \forall\, \alpha \in \mathbb{C} \setminus \{0\}.$$
2. $$Cond(A, \|\cdot\|) \geq 1, \quad \forall\, A \in \mathcal{M}_n(\mathbb{C}) \quad \text{avec } A \text{ inversible.}$$

3. *Supposons en plus que $\|\cdot\|$ est la norme $\|\cdot\|_2$ (rappelons que c'est la norma matricielle subordonnée à la norme vectorielle euclidienne). Alors pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$ hermitienne et inversible on a*

$$Cond(A, \|\cdot\|_2) = \frac{\max\{|\lambda|,\ \lambda \in Sp(A)\}}{\min\{|\lambda|,\ \lambda \in Sp(A)\}}.$$

Démonstration. 1. Comme $(\alpha I_n)^{-1} = \frac{1}{\alpha} I_n$ et $\|I_n\| = 1$ on a

$$Cond(\alpha I_n, \|\cdot\|) = \|\alpha I_n\| \, \|\frac{1}{\alpha} I_n\| = |\alpha| \, \frac{1}{|\alpha|} = 1.$$

2. Comme $I_n = A\, A^{-1}$ alors

$$1 = \|I_n\| = \|A\, A^{-1}\| \leq \|A\| \, \|A^{-1}\| = Cond(A, \|\cdot\|).$$

3. Comme A est hermitienne alors A^{-1} est aussi hermitienne (car $(A^{-1})^* = (A^*)^{-1} = A^{-1}$). Alors les valeurs propres de A sont réelles et non nulles (car A inversible). Si on note $\lambda_1, \lambda_2, \cdots \lambda_k \in \mathbb{R} \setminus \{0\}$ les valeurs propres distinctes de A (avec $k \in \mathbb{N}^*$) alors les valeurs propres distinctes de A^{-1} sont $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots \frac{1}{\lambda_k} \in \mathbb{R} \setminus \{0\}$ (voir la partie **b)** de la Proposition 2.7). Nous avons alors

$$\|A\| = \rho(A) = \max\{|\lambda_1|, |\lambda_2|, \cdots |\lambda_k|\} \tag{4.9}$$

et aussi

$$\|A^{-1}\| = \rho(A^{-1}) = \max\left\{\frac{1}{|\lambda_1|}, \frac{1}{|\lambda_2|}, \cdots, \frac{1}{|\lambda_k|}\right\}$$

ce qui donne

$$\|A^{-1}\| = \frac{1}{\min\{|\lambda_1|, |\lambda_2|, \cdots |\lambda_k|\}} \tag{4.10}$$

□

En multipliant (4.9) et (4.10) on obtient le résultat.

4.3 Méthodes directes pour la résolution des systèmes algébriques linéaires

4.3.1 Généralités

Dans cette section on se donne $A \in \mathcal{M}_n(\mathbb{K})$ une matrice inversible et $b \in \mathbb{K}^n$ un vecteur ; on supposera $n \geq 2$. On veut trouver $x \in \mathbb{K}^n$ l'unique solution du système algébrique linéaire

$$Ax = b \tag{4.11}$$

On peut donner théoriquement la solution x de ce système. Comme A est inversible alors on peut multiplier (4.11) à droite par la matrice A^{-1} et on trouve facilement

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

En général il est difficile de connaitre la matrice A^{-1} surtout pour des problèmes de grand taille (quand n est "grand"). On cherche alors une approximation numérique, la plus précise que possible, de la solution x de (4.11).
Dans cette section on donnera des méthodes **directes** de résolution numérique de (4.11), c'est à dire, des méthodes qui permet d'obtenir une approximation numérique de la solution en un nombre de pas connu à l'avance (à la différence des méthodes qu'on va donner dans la section suivante, où le nombre de pas n'est pas connu à l'avance).

Remarque 4.5. *Si A est une matrice diagonale ou triangulaire supérieure ou triangulaire inférieure alors la résolution numèrique de* (4.11) *se fait manière très simple (remarquons que dans ce cas les éléments diagonaux A_{ii} de la matrice A sont non nuls car $det(A) = A_{11}A_{22}\cdots A_{nn} \neq 0$).*

1. *Le cas le plus simple est le cas où A est diagonale ; alors la solution de* (4.11) *est donnée par*
$$x_i = \frac{b_i}{A_{ii}}, \quad \forall\, i \in [[1, n]].$$

2. *Si A est triangulaire supérieure, alors le système* (4.11) *s'écrit*
$$\begin{cases} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \quad \cdots \quad + A_{1,n-1}x_{n-1} + A_{1n}x_n = b_1 \\ \qquad\quad A_{22}x_2 + A_{23}x_3 + \quad \cdots \quad + A_{2,n-1}x_{n-1} + A_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \cdots \qquad \cdots \qquad\quad \cdots \\ \qquad\qquad\qquad\qquad\qquad A_{n-1,n-1}x_{n-1} + A_{n-1,n}x_n = b_{n-1} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad A_{n,n}x_n = b_n \end{cases}$$
On peut alors résoudre ce système par une **remontée** *: on peut calculer d'abord x_n en utilisant la dernière équation, ensuite on calcule x_{n-1} en utilisant l'avant-dernière équation et le fait que x_n est connu, etc ... jusqu'à x_1. On a alors l'algorithme suivant :*
$$\begin{cases} x_n = \dfrac{b_n}{A_{nn}} \\ Pour \quad i \quad de \quad n-1 \quad à \quad 1 \quad calculer \\ \qquad x_i = \dfrac{1}{A_{ii}}\left[b_i - \displaystyle\sum_{j=i+1}^{n} A_{ij}x_j\right] \\ \quad fin \end{cases}$$

3. Si A est triangulaire inférieure, alors le système (4.11) s'écrit

$$\begin{cases} A_{11}x_1 = b_1 \\ A_{21}x_1 + A_{22}x_2 = b_2 \\ \quad \cdots \qquad \cdots \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n = b_n \end{cases}$$

On peut alors résoudre ce système par une **descente** *: on peut calculer d'abord x_1 en utilisant la première équation, ensuite on calcule x_2 en utilisant la deuxième équation et le fait que x_1 est connu, etc ... jusqu'à x_n. On a alors l'algorithme suivant :*

$$\begin{cases} x_1 = \dfrac{b_1}{A_{11}} \\ Pour \quad i \quad de \quad 2 \quad à \quad n \quad calculer \\ \qquad x_i = \dfrac{1}{A_{ii}} \left[b_i - \displaystyle\sum_{j=1}^{i-1} A_{ij}x_j \right] \\ \quad fin \end{cases}$$

Remarque 4.6. *Si on sait résoudre numériquement tout système algébrique linéaire du type $Ax = b$ alors on peut calculer numériquement A^{-1}. Pour cela on cherche $B \in \mathcal{M}_n(\mathbb{K})$ telle que $AB = I_n$. On peut écrire*

$$B = [B_1,\ B_2, \cdots, B_n]$$

où B_i représente la i - ème colonne de B. Comme

$$AB = [AB_1,\ AB_2, \cdots, AB_n] \quad et \quad I_n = [e_1,\ e_2, \cdots, e_n]$$

alors chaque colonne B_i satisfait le système algébrique linéaire

$$AB_i = e_i, \quad \forall\, i \in [[1, n]].$$

En résolvant numériquement chacun de ces systèmes avec inconnue B_i on obtient une approximation numérique de $B = A^{-1}$.
En particulier si A est une matrice triangulaire supérieure ou triangulaire inférieure alors on peut calculer facilement A^{-1} en résolvant les systèmes ci-dessus par une "remontée" (si A est tringulaire supérieure) ou par une "descente" (si A est tringulaire inférieure).

4.3.2 La méthode de Gauss et la décomposition $A = LU$

L'algorithme de Gauss

Nous devons résoudre numériquement un système algébrique linéaire du type (4.11) qu'on écrit en détail sous la forme

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n = b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n = b_2 \\ \quad\quad\quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \quad\quad\quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n = b_n \end{cases} \tag{4.12}$$

Il est commode d'introduire une matrice $S \in \mathcal{M}_{n,n+1}(\mathbb{R})$ (matrice du sytème) qui s'obtient de A en y ajoutant la colonne b comme $n+1$ - ème colonne ; avec l'écriture par blocks on a

$$S = (A \quad b).$$

L'algorithme de Gauss (ou la méthode de Gauss) consiste à écrire (4.12) sous la forme équivalente d'un système algébrique linéaire avec une matrice triangulaire supérieure ; ce nouveau système va se résoudre facilement par une remontée. L'algorithme de Gauss va comporter $n-1$ étapes : la première étape consiste à "faire 0 sur la première colonne de A à partir du deuxième élément" (c'est à dire écrire le système (4.12) sous une forme équivalente où la matrice du système aura cette propriété). En général à l'étape numéro p on part d'un système (équivalent au (4.12)) dont la matrice est telle que les premières $p-1$ colonnes ont 0 en dessous de la diagonale et on cherche à "faire 0 sur la colonne p en dessous de la diagonale". A la fin de l'étape $n-1$ on aura un système équivalent à celui du départ, mais avec une matrice du système qui est triangulaire supérieure.
Exemple : en classe

La première étape de l'algorithme de Gauss.
Cette étape consiste à éliminer l'inconnue x_1 dans les équations de 2 à n.
On note $A^{(1)} = A$ et $b^{(1)} = b$ et on introduit la matrice $S^{(1)} \in \mathcal{M}_{n,n+1}(\mathbb{R})$ qui s'écrit par blocs

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

L'un au moins des éléments de la première colonne de $A^{(1)}$ est non nul car sinon on aurait $det(A^{(1)}) = 0$ donc la matrice $A^{(1)}$ serait non inversible.
Nous faisons l'hypothèse supplémentaire suivante :

$$A^{(1)})_{11} \neq 0. \tag{4.13}$$

On appellera dans le suite cet élément $(A^{(1)})_{11}$ le **pivot** de l'étape 1 (remarquons qu'on a $(A^{(1)})_{11} = (S^{(1)})_{11}$).

Remarque 4.7. *Dans le cas où cette hypothèse $(A^{(1)})_{11} \neq 0$ n'est pas satisfaite, nous savons qu'il existe un élément $k \in [[2, n]]$ avec $(A^{(1)})_{k1} \neq 0$; alors avant de passer à la suite de l'algorithme, on inverse les lignes 1 et k du système linéaire, donc de la matrice $S^{(1)}$. Ce cas n'est pas considéré dans ce cours.*

On va noter

$$\alpha_i^{(1)} = \frac{(S^{(1)})_{i1}}{(S^{(1)})_{11}}, \quad \forall\, i \in [[2, n]].$$

L'idée est de multiplier la première ligne de la matrice $S^{(1)}$ par $-\alpha_i^{(1)}$ et de l'ajouter à la ligne i, ceci pour tout $i \in [[2, n]]$. Il est facile de voir que cela permet de faire 0 sur la première colonne de notre matrice du système, sous la diagonale. En notant $S^{(2)}$ la matrice du système qu'on obtient après ces opérations on aura

$$(S^{(2)})_{ij} = (S^{(1)})_{ij} - \alpha_i^{(1)}(S^{(1)})_{1j}, \quad \forall\, i \in [[2, n]], \quad \forall\, j \in [[1, n+1]]$$

(remarquer que $(S^{(2)})_{i1} = 0, \quad \forall\, i \in [[2, n]]$).
D'autre part, nous laissons la première ligne inchangée :

$$(S^{(2)})_{1j} = (S^{(1)})_{1j}, \quad \forall\, j \in [[1, n+1]].$$

Tout ceci nous permet d'exprimer les éléments de $S^{(1)}$ en fonction de ceux de $S^{(2)}$:

$$(S^{(1)})_{ij} = (S^{(2)})_{ij} + \alpha_i^{(1)}(S^{(2)})_{1j}, \quad \forall\, i \in [[2, n]], \quad \forall\, j \in [[1, n+1]]$$

Il est facile de voir qu'on a l'écriture matricielle suivante :

$$S^{(1)} = S^{(2)} + F^{(1)} S^{(2)}$$

avec $F^{(1)} \in \mathcal{M}_n(\mathbb{K})$ la matrice donnée par

$$F^{(1)} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \alpha_2^{(1)} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_n^{(1)} & 0 & \cdots & 0 \end{pmatrix}$$

Nous introduisons le vecteur $\alpha^{(1)} = \begin{pmatrix} 0 \\ \alpha_2^{(1)} \\ \alpha_3^{(1)} \\ \cdot \\ \cdot \\ \alpha_n^{(1)} \end{pmatrix}$ et on observe qu'on peut écrire

$$F^{(1)} = \alpha^{(1)} e_1^T.$$

On peut alors écrire

$$S^{(1)} = L^{(1)} S^{(2)}. \tag{4.14}$$

avec $L^{(1)} \in \mathcal{M}_n(\mathbb{K})$ la matrice triangulaire inférieure avec 1 sur la diagonale, définie par

$$L^{(1)} = I_n + F^{(1)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha_2^{(1)} & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_n^{(1)} & 0 & \cdots & 0 & 1 \end{pmatrix} \tag{4.15}$$

Il est facile de voir que

$$det(L^{(1)}) = 1$$

donc la matrice $L^{(1)}$ est inversible. A la fin de cette première étape le système linéaire du départ qui correspond à la matrice $S^{(1)}$ est équivalent au système qui correspond à la matrice $S^{(2)}$ (car la matrice $L^{(1)}$ est inversible). Plus précisément, si on écrit par blocs

$$S^{(2)} = (A^{(2)} \quad b^{(2)})$$

avec $A^{(2)} \in \mathcal{M}_n(\mathbb{K})$ et $b^{(2)} \in \mathbb{K}^n$ alors le système de départ $A^{(1)}x = b^{(1)}$ est équivalent au système

$$A^{(2)}x = b^{(2)}.$$

De (4.14) on déduit :

$$A^{(1)} = L^{(1)} A^{(2)} \quad \text{et} \quad b^{(1)} = L^{(1)} b^{(2)}.$$

Remarque : On peux dire que à la fin de cette première étape on a éliminé l'inconnue x_1 dans les équations de 2 à n.

L'étape numéro p de l'algorithme de Gauss, avec $p \in [[2, n-1]]$.
Cette étape consiste à éliminer l'inconnue x_p dans les équations de $p+1$ à n ; remarquons que cette étape à du sense seulement pour $n \geq 3$.
Nous supposons que dans les étapes précédentes, de 1 à $p-1$, nous avons construit une suite des matrices du système $S^{(1)}, S^{(2)}, \cdots S^{(p)}$ avec

$$S^{(r)} = L^{(r)} S^{(r+1)}, \quad \forall\, r \in [[1, p-1]]$$

où à chaque étape $L^{(r)}$ est une matrice triangulaire inférieure avec 1 sur la diagonale de la forme

$$L^{(r)} = I_n + \alpha^{(r)} e_r^T$$

où $\alpha^{(r)} \in \mathbb{K}^n$ est un vecteur tel que

$$(\alpha^{(r)})_i = 0, \quad \text{si} \quad i \in [[1, r]].$$

Nous avons

$$S^{(r)} = \left(A^{(r)} \;\; b^{(r)}\right), \quad r = 1, \cdots p-1$$

avec les relations de récurrence :

$$A^{(r)} = L^{(r)} A^{(r+1)} \quad \text{et} \quad b^{(r)} = L^{(r)} b^{(r+1)}, \quad \forall\, r \in [[1, p-1]]$$

Comme $det(L^{(r)}) = 1$ nous avons

$$det(A^{(r)}) = det(A^{(r+1)}), \quad \forall\, r \in [[1, p-1]]. \tag{4.16}$$

D'autre part la matrice $S^{(p)} \in \mathcal{M}_{n,n+1}(\mathbb{K})$ est telle que ses premières $p-1$ colonnes ont la valeur 0 sous la diagonale, c'est à dire

$$(S^{(p)})_{ij} = 0, \quad \forall\, i, j \text{ tels que } \; j \in [[1, p-1]], \; j < i \leq n.$$

En écrivant pas blocs

$$S^{(p)} = (A^{(p)} \quad b^{(p)})$$

avec $A^{(p)} \in \mathcal{M}_n(\mathbb{K})$ et $b^{(p)} \in \mathbb{K}^n$, notre système linéaire de départ $A^{(1)}x = b^{(1)}$ est équivalent au système

$$A^{(p)}x = b^{(p)}$$

et remarquons que aussi la matrice $A^{(p)}$ est telle que ses premières $p-1$ colonnes ont la valeur 0 sous la diagonale, c'est à dire

$$(A^{(p)})_{ij} = 0, \quad \forall\, i, j \text{ tels que } \; j \in [[1, p-1]], \; j < i \leq n.$$

De (4.16) on déduit

$$det(A^{(p)}) = det(A^{(1)})$$

donc la matrice $A^{(p)}$ est inversible car $A^{(1)} = A$ l'est.
Dans la suite nous procédons comme dans l'étape 1 pour faire 0 en dessous de la diagonale sur la p - ème colonne de la matrice du système.
En développant le déterminant de $A^{(p)}$ successivement par rapport aux colonnes de 1 à $p-1$ on trouve

$$det(A^{(p)}) = (A^{(p)})_{11}(A^{(p)})_{22} \cdots (A^{(p)})_{p-1,p-1}\, det((A^{(p)})_{i,j \in [[p,n]]})$$

où $(A^{(p)})_{i,j\in[[p,n]]} \in \mathcal{M}_{n-p+1}(\mathbb{K})$ est la matrice obtenue à partir de $A^{(p)}$ en retenant uniquement les lignes et les colonnes de p à n. Comme $det(A^{(p)}) \neq 0$ on déduit que

$$det((A^{(p)})_{i,j\in[[p,n]]}) \neq 0$$

et ceci implique que l'un au moins des éléments $(A^{(p)})_{ip},\ i \in [[p,n]]$ est non nul.
De manière analogue que dans l'étape 1 nous faisons l'hypothèse :

$$(A^{(p)})_{pp} \neq 0 \tag{4.17}$$

et nous appelons cet élément $(A^{(p)})_{pp}$ le **pivot** de l'étape p.

Remarque 4.8. *Dans le cas où cette hypothèse $(A^{(p)})_{pp} \neq 0$ n'est pas satisfaite, nous savons qu'il existe un élément $k \in [[p+1,n]]$ avec $(A^{(p)})_{kp} \neq 0$; alors avant de passer à la suite de l'algorithme, on inverse les lignes p et k du système linéaire, donc de la matrice $S^{(p)}$. Ce cas n'est pas considéré dans ce cours.*

On va poser

$$\alpha_i^{(p)} = \frac{(S^{(p)})_{ip}}{(S^{(p)})_{pp}}, \quad \forall\, i \in [[p+1,n]].$$

L'idée est de multiplier la p - ème ligne de la matrice $S^{(p)}$ par $-\alpha_i^{(p)}$ et l'ajouter à la ligne i, ceci pour tout $i \in [[p+1,n]]$. Il est facile de voir que cela permet de faire 0 sur la p -ème colonne de notre matrice du système, en dessous de la diagonale. En notant $S^{(p+1)}$ la matrice du système qu'on obtient après ces opérations on aura

$$(S^{(p+1)})_{ij} = (S^{(p)})_{ij} - \alpha_i^{(p)}(S^{(p)})_{pj}, \quad \forall\, i \in [[p+1,n+1]], \quad \forall\, j \in [[p,n+1]]$$

et ceci implique que cette dernière égalité est vraie pour tout $j \in [[1,n+1]]$;
remarquer que $(S^{(p+1)})_{ip} = 0, \quad \forall\, i \in [[p+1,n]]$).
Nous laissons les premières p lignes inchangées :

$$(S^{(p+1)})_{ij} = (S^{(p)})_{ij}, \quad \forall\, i \in [[1,p]], \quad \forall j \in [[1,n+1]]$$

ce qui nous permet d'exprimer les éléments de $S^{(p)}$ en fonction de ceux de $S^{(p+1)}$:

$$(S^{(p)})_{ij} = (S^{(p+1)})_{ij} + \alpha_i^{(p)}(S^{(p+1)})_{pj}, \quad \forall\, i \in [[p+1,n]], \quad \forall\, j \in [[1,n+1]]$$

Il est facile de voir qu'on a l'écriture matricielle suivante :

$$S^{(p)} = S^{(p+1)} + F^{(p)}S^{(p+1)}$$

avec $F^{(p)} \in \mathcal{M}_n(\mathbb{K})$ la matrice donnée par

$$F^{(p)} = \alpha^{(p)} e_p^T$$

où nous introduisons le vecteur $\alpha^{(p)}$ donné par

$$(\alpha^{(p)})_i = \begin{cases} 0 & \text{si} \quad i \in [[1,p]] \\ \alpha_i^{(p)} & \text{si} \quad i \in [[p+1,n]]. \end{cases}$$

On peut alors écrire

$$S^{(p)} = L^{(p)} S^{(p+1)}. \tag{4.18}$$

avec $L^{(p)} \in \mathcal{M}_n(\mathbb{K})$ la matrice triangulaire inférieure avec 1 sur la diagonale, définie par

$$L^{(p)} = I_n + F^{(p)}. \tag{4.19}$$

Il est facile de voir que

$$det(L^{(p)}) = 1$$

donc la matrice $L^{(p)}$ est inversible. A la fin de cette p - ème étape le système linéaire du départ qui correspond à la matrice $S^{(p)}$ est équivalent au système qui correspond à la matrice $S^{(p+1)}$. Plus précisément, si on écrit par blocs

$$S^{(p+1)} = (A^{(p+1)} \quad b^{(p+1)})$$

avec $A^{(p+1)} \in \mathcal{M}_n(\mathbb{K})$ et $b^{(p+1)} \in \mathbb{K}^n$ alors le système de départ est équivalent au système

$$A^{(p+1)} x = b^{(p+1)}.$$

De (4.18) on déduit :

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

Remarque 4.9. *Grâce au fait que* $det(L^{(p)}) = 1$ *nous avons*

$$det(A^{(p)}) = det(A^{(p+1)}). \tag{4.20}$$

A la fin de **l'étape** $n-1$ on arrive à une matrice du système $S^{(n)} \in \mathcal{M}_{n,n+1}(\mathbb{K})$ équivalente à la matrice de départ, telle que, en écrivant par blocs

$$S^{(n)} = (A^{(n)} \quad b^{(n)})$$

on a que $A^{(n)} \in \mathcal{M}_n(\mathbb{K})$ est une matrice triangulaire supérieure qui est inversible (c'est à dire, avec tous les éléments diagonaux non nuls). On va noter dans la suite

$$U = A^{(n)} \quad \text{et} \quad c = b^{(n)}$$

donc le système du départ $Ax = b$ est équivalent au système

$$Ux = c.$$

Ce dernier système se résoud très facilement, par une remontée, car U est triangulaire supérieure.

D'autre part l'algorithme de Gauss nous permet de calculer le déterminant de A. Grâce à la formule (4.20) nous avons

$$det(A) = det(U)$$

Comme U est triangulaire supérieure, son déterminant est le produit des éléments diagonaux de U, d'où

$$det(A) = U_{11}U_{22}\cdots U_{nn}.$$

Remarque 4.10. *A chaque étape p de l'algorithme, avec $p \in [[1,n]]$, il peut être préférable de permuter la ligne p de la matrice avec une autre ligne $k \in [[p+1,n]]$ même si on a $(A^{(p)})_{pp} \neq 0$; ceci dans le cas ou $|(A^{(p)})_{pp}|$ est "trop petit" par rapport aux modules d'autres éléments de la colonne p de $A^{(p)}$, car des valeurs de $\alpha_i^{(p)}$ seraient trop "grandes" ce qui peut créer des instabilités de l'algorithme.*
Un choix qui est souvent utilisé est d'utiliser une stratégie appellée **du pivot partiel** *qui consiste à choisir $k \in [[p,n]]$ tel que*

$$|(A^{(p)})_{kp}| = \max_{i\in[[p,n]]} |(A^{(p)})_{ip}|$$

(ensuite il faut inverser les lignes p et k). Dans ce cas les éléments du vecteur $\alpha^{(p)}$ seront de module inférieur ou égal à 1, ce qui est bon pour la stabilité de l'algorithme.
On peut aller encore plus loin et utiliser une stratégie dite **du pivot total** *qui consiste à autoriser des permutations non seulement des lignes (donc des équations) mais aussi des colonnes (donc des inconnues).*

Evaluons maintenant le nombre d'opérations nécessaires dans cet algorithme. A chaque étape p de l'algorithme on effectue $(n-p)$ divisions et $(n+1-p)^2$ multiplications et aditions. On aura donc un nombre de multiplications et additions égal à

$$\sum_{p=1}^{n}(n+1-p)^2 = \sum_{k=1}^{n} k^2 = \frac{1}{3}n(n-\frac{1}{2})(n-1)$$

(on a fait le changement d'indice $k = n+1-p$).
D'autre part on peut facilement montrer que le nombre d'opérations nécessaires pour résoudre le système équivalent triangulaires est d'ordre n^2.
Alors le terme dominant dans le nombre total d'opérations est

$$N_{dom-op} = \frac{1}{3}n^3.$$

La décomposition $A = LU$

Rappelons la relation de récurrence que nous avons obtenu à l'étape p de l'algorithme de Gauss :

$$A^{(p)} = L^{(p)}A^{(p+1)}, \quad \forall\, p \in [[1,n-1]]$$

avec $A^{(1)} = 1$ et $A^{(n)} = U$. Ceci permet d'obtenir facilement

$$A = L^{(1)}L^{(2)} \cdots L^{(n-2)}L^{(n-1)}U. \tag{4.21}$$

Nous avons le résultat suivant, qui permettra de donner une forme plus simple pour le produit des matrices $L^{(1)}L^{(2)} \cdots L^{(n-2)}L^{(n-1)}$.

Lemme 4.2. *Soient $\alpha_1, \alpha_2, \cdots \alpha_{n-1} \in \mathbb{K}^n$ des vecteurs de dimension n tels que pour tout $p \in [[1, n-1]]$ les p premières composantes de α_p soient nulles. Alors on a*

$$(I_n + \alpha_1 e_1^T)(I_n + \alpha_2 e_2^T) \cdots (I_n + \alpha_{n-1} e_{n-1}^T) = I_n + \alpha_1 e_1^T + \alpha_2 e_2^T + \cdots \alpha_{n-1} e_{n-1}^T.$$

Démonstration. Le résultat s'obtient en développant le produit matriciel dans le membre de gauche de l'égalité à demontrer. Il suffit d'observer que $\alpha_i e_i^T \alpha_j e_j^T = 0$ si $i < j$, car $e_i^T \alpha_j = < e_i \, , \, \alpha_j > = 0$. □

Lemme 4.3.

1. *Si $L_1, L_2 \in \mathcal{M}_n(\mathbb{K})$ sont deux matrices triangulaires inférieures alors leur produit $L_1 L_2$ est aussi triangulaire inférieure. Si en plus L_1, L_2 sont triangulaires inférieures avec 1 sur la diagonale alors leur produit $L_1 L_2$ est aussi triangulaire inférieure avec 1 sur la diagonale.*
2. *Si $L \in \mathcal{M}_n(\mathbb{K})$ est une matrice triangulaire inférieure inversible (donc $L_{ii} \neq 0, \quad \forall\, i \in [[1, n]]$) alors l'inverse L^{-1} est aussi triangulaire inférieure. En plus $(L^{-1})_{ii} = \frac{1}{L_{ii}}, \quad \forall\, i \in [[1, n]]$. Cela entraine le fait que si L est une matrice triangulaire inférieure avec 1 sur la diagonale alors L^{-1} est aussi une matrice triangulaire inférieure avec 1 sur la diagonale.*
3. *Si $U_1, U_2 \in \mathcal{M}_n(\mathbb{K})$ sont deux matrices triangulaires supérieures alors leur produit $U_1 U_2$ est aussi triangulaire supérieure. Si en plus U_1, U_2 sont triangulaires supérieures avec 1 sur la diagonale alors leur produit $U_1 U_2$ est aussi triangulaire supérieure avec 1 sur la diagonale.*
4. *Si $U \in \mathcal{M}_n(\mathbb{K})$ est une matrice triangulaire supérieure inversible (donc $U_{ii} \neq 0, \quad \forall\, i \in [[1, n]]$) alors l'inverse U^{-1} est aussi triangulaire supérieure. En plus $(U^{-1})_{ii} = \frac{1}{U_{ii}}, \quad \forall\, i \in [[1, n]]$. Cela entraine le fait que si U est une matrice triangulaire supérieure avec 1 sur la diagonale alors U^{-1} est aussi une matrice triangulaire supérieure avec 1 sur la diagonale.*

Démonstration. Laissée en exercice. □

Définition 4.4. *On dit d'une matrice inversible $A \in \mathcal{M}_n(\mathbb{K})$ qu'elle*
admet une décomposition *LU s'il existe $L, U \in \mathcal{M}_n(\mathbb{K})$ avec L triangulaire inférieure avec 1 sur la diagonale et U triangulaire supérieure et inversible, telles qu'on a*

$$A = LU.$$

Le résultat d'existence de la décomposition $A = LU$ est le suivant :

Théorème 4.1. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice inversible et supposons qu'aucune permutation n'a été faite dans l'algorithme de Gaus apliqué à la matrice A. Alors A admet une décomposition LU. En plus :*
- la matrice U est celle obtenue dans l'algorithme de Gauss
- les éléments de L qui sont en dessous de la diagonale s'obtiennent de la manière suivante : pour tout $p \in [[1, n-1]]$ on a

$$L_{ip} = (\alpha^{(p)})_i, \quad \forall\, i \in [[p+1, n]]$$

où les vecteurs $\alpha^{(p)}$ sont ceux de l'algorithme de Gauss.

Démonstration. La formule (4.21) obtenue à la suite de l'algorithme de Gauss nous donne

$$A = LU$$

avec

$$L = L^{(1)} L^{(2)} \cdots L^{(n-2)} L^{(n-1)}$$

et U matrice triangulaire supérieure.
Le reste de la preuve est une application immédiate du Lemme 4.2 avec
$\alpha_p = \alpha^{(p)},\ p \in [[1, n-1]]$ ainsi que du Lemme 4.3. □

Remarque 4.11. *Il y a des matrices inversibles qui n'admet pas de décomposition LU ; c'est le cas si dans une étape p de l'algorithme de Gauss on a $(A^{(p)})_{pp} = 0$ (par exemple si $A_{11} = 0$).*

Remarque 4.12. *L'utilité des ces décompositions est la suivante : si on doit résoudre plusieurs fois des systèmes linéaire du type $Ax = b$ avec toujours le même matrice A mais avec chaque fois des vecteurs b différents une décomposition de ce type peut beaucoup réduire le temps de calcul. Il n'est pas nécéssaire de refaire tout l'algorithme de Gauss pour chaque vecteur b, il suffit de la faire une seule fois pour trouver la décomposition, ensuite la résolution est plus facile ; si on dispose d'une décomposition LU alors notre système s'écrit*

$$LUx = b.$$

On fait d'abord un changement d'inconnue $Ux = y$ avec $y \in \mathbb{K}^n$ la nouvelle inconnue et le sytème ci-dessus s'écrit

$$Ly = b$$

Comme L est triangulaire inférieure, ce système se résout en y très facilement par une descente. Ensuite, avec y connu on résoud facilement le système

$$Ux = y$$

par une remontée, car U est triangulaire supérieure.

Nous avons le résultat suivant d'unicité de la décomposition LU :

Théorème 4.2. *Si une matrice inversible $A \in \mathcal{M}_n(\mathbb{K})$ admet une décomposition LU, alors cette décomposition est unique.*

Démonstration. Supposons que A admet deux décopositions LU ; nous avons alors

$$A = L_1U_1 = L_2U_2$$

avec L_1, L_2 triangulaires inférieures avec 1 sur la diagonale et U_1, U_2 triangulaires supérieures et inversibles. De l'égalité $L_1U_1 = L_2U_2$ on déduit facilement

$$L_2^{-1}L_1 = U_2U_1^{-1}$$

Du Lemme 4.3 on déduit :
$L_2^{-1}L_1$ est triangulaire inférieure avec 1 sur la diagonale et
$U_2U_1^{-1}$ est triangulaire supérieure et inversible.
Mais une matrice qui est à la fois triangulaire inférieure avec 1 sur la diagonale et triangulaire supérieure ne peut être que la matrice identité I_n. On a donc

$$L_2^{-1}L_1 = U_2U_1^{-1} = I_n$$

ce qui donne $L_1 = L_2$ et $U_1 = U_2$ et finit la preuve du théorème. □

Nous finissons par le résultat d'existence suivant, que nous admetons
(pour la preuve voir [5] Théorème 8 Chapitre 4) :

Théorème 4.3. *Une matrice inversible $A \in \mathcal{M}_n(\mathbb{K})$ admet une décomposition LU si et seulement si toutes les sous-matrices principales A_p de A, avec $p \in [[1, n]]$ (voir la Définition 2.11) sont inversibles. Dans ce cas on a*

$$U_{11} = A_{11}$$

et

$$U_{pp} = \frac{det(A_p)}{det(A_{p-1})}, \quad \forall\, p \in [[2, n]].$$

Remarque 4.13. *Toute matrice définie positive admet une décomposition LU car elle est inversible et en plus toutes les sous-matrice principales de A sont définies positives donc inversibles (voir Proposition 2.15 et Proposition 2.16).*

4.3.3 Décomposition (ou factorisation) de Choleski

Nous admettons le résultat suivant
(pour la preuve voir [5] Théorème 17 Chapitre 4) :

Théorème 4.4. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne et définie positive. Alors il existe une unique matrice $R \in \mathcal{M}_n(\mathbb{K})$ qui est triangulaire supérieure avec $R_{ii} \in \mathbb{R}$ et $R_{ii} > 0, \quad \forall\, i \in [[1, n]]$ telle que*

$$A = R^*\, R. \tag{4.22}$$

La formule (4.22) donne ce qu'on appelle une **décomposition (ou factorisation) de Choleski**.

Remarque 4.14. *1. Il est possible d'obtenir par un algorithme d'identification les éléments de la matrice R à partir des éléments de la matrice A.*

2. L'avantage principal du fait d'avoir une factorisation de Choleski est le même que pour une décomposition LU : cela permet de résoudre très facilement un système algébrique linéaire de la forme $Ax = b$ en faisant d'abord une descente et ensuite une remontée. Un avantage supplémentaire pour la factorisation de Choleski est le fait de pouvoir faire une économie de mémoire : il suffit de stocker les éléments R_{ij} avec $i \leq j$ de la matrice R, qui sont en nombre de $\frac{n(n+1)}{2}$, alors que pour une décomposition LU il faut stocker n^2 éléments.

4.4 Méthodes itératives de résolution des systèmes algébriques linéaires

4.4.1 Généralités sur les méthodes itératives

Dans cette section on cherche encore à résoudre des systèmes algébriques linéaires de la forme

$$Ax = b \tag{4.23}$$

avec $A \in \mathcal{M}_n(\mathbb{K})$ matrice inversible et $b \in \mathbb{K}^n$ vecteur données et vecteur inconnue $x \in \mathbb{K}^n$; on suppose $n \geq 2$. L'inconvenient principal des méthodes directes étudiées à la section précédente est le fait que pour n très grand il peut y avoir une grande accumulation des erreurs d'arrondi d'un pas à l'autre de l'algorithme ; par contre, les méthodes itérative peuvent permettre d'éviter cet inconvenient.
Le principe général des méthodes iteratives est le suivant : on écrit A sous la forme

$$A = M - N \tag{4.24}$$

avec $M, N \in \mathcal{M}_n(\mathbb{K})$ ce qui permet d'écrire (4.23) sous la forme

$$Mx = Nx + b.$$

Il faut faire en sorte que M soit une matrice inversible et "facile à inverser", c'est à dire, la résolution d'un système algébrique linéaire du type : trouver $y \in \mathbb{K}^n$ tel que

$$My = c$$

avec $c \in \mathbb{K}^n$ vecteur donné, soit facile à faire (c'est le cas par exemple si M est diagonale ou triangulaire supérieure ou triangulaire inférieure).
Alors l'idée est de construire par récurrence une suite $\{x^{(k)}\}_{k\in\mathbb{N}} \subset \mathbb{K}^n$ avec $x^{(0)} \in \mathbb{K}^n$ donnée et $x^{(k+1)}$ définit en fonction de $x^{(k)}$ par la formule de récurrence

$$M\, x^{(k+1)} = N\, x^{(k)} + b, \quad \forall\, k \in \mathbb{N} \tag{4.25}$$

ce qui est équivalent à

$$x^{(k+1)} = M^{-1}\,N\,x^{(k)} + M^{-1}\,b, \quad \forall\, k \in \mathbb{N}. \tag{4.26}$$

Remarque 4.15. *La même idée peut être utilisée pour résoudre numériquement un problème nonlinéaire du type : trouver $x \in \mathbb{K}^n$ tel que*

$$F(x) = 0$$

avec $F : \mathbb{K}^n \to \mathbb{K}^n$ une fonction donnée. L'idée sera encore de décomposer F sous la forme

$$F(x) = Mx - G(x)$$

avec M matrice inversible et facile à inverser et écrire notre équation sous la forme équivalente

$$Mx = G(x).$$

Cette équation se résoud numériquement en construisant une suite définie par reccurence par la formule

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

et $x^{(0)}$ donné.

4.4.2 Les méthodes itératives usuelles

Dans la suite pour toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ on utilise la décomposition

$$A = D - E - F \tag{4.27}$$

avec $D \in \mathcal{M}_n(\mathbb{K})$ matrice diagonale donnée par

$$D_{ij} = A_{ij}\delta_{ij}, \quad \forall\, i, j \in [[1, n]]$$

(c'est à dire : $D = diag(A_{11}, A_{22} \cdots A_{nn})$,
$E \in \mathcal{M}_n(\mathbb{K})$ matrice triangulaire inférieure avec 0 sur la diagonale, donnée par

$$E_{ij} = \begin{cases} -A_{ij} & \text{si} \quad i > j \\ 0 & \text{si} \quad i \leq j \end{cases}$$

et $F \in \mathcal{M}_n(A)$ matrice triangulaire supérieure avec 0 sur la diagonale, donnée par

$$F_{ij} = \begin{cases} -A_{ij} & \text{si} \quad i < j \\ 0 & \text{si} \quad i \geq j. \end{cases}$$

Remarque 4.16. *Les matrice D, E et F dependent de A et on devrait donc les noter D_A, E_A et F_A ; mais assez souvent il est clair de quelle matrice A il s'agit et on préfère pour des raisons de simplicité les notations sans indice "A".*

Nous avons les méthodes itératives suivantes :

1. **Méthode de Jacobi**
   Cette méthode consiste à prendre en (4.24) $M = D$ et donc $N = E + F$.
   Pour que cette méthode puisse être utilisée, la matrice D doit être inversible, ce qui est équivalent avec le fait que A doit satisfaire la condition suivante :

$$A_{ii} \neq 0, \quad \forall\, i \in [[1, n]]. \tag{4.28}$$

   Donc la méthode de Jacobi consiste à construire une suite $\{x^{(k)}\}_{k\in\mathbb{N}} \subset \mathbb{K}^n$ avec $x^{(0)} \in \mathbb{K}^n$ donnée et $x^{(k+1)}$ définit en fonction de $x^{(k)}$ par la formule de récurrence

$$D\, x^{(k+1)} = (E + F)\, x^{(k)} + b, \quad \forall\, k \in \mathbb{N} \tag{4.29}$$

   ce qui est équivalent à

$$x^{(k+1)} = D^{-1}\, (E + F)\, x^{(k)} + D^{-1}\, b, \quad \forall\, k \in \mathbb{N}. \tag{4.30}$$

   Comme $D^{-1} = diag(\frac{1}{A_{11}}, \frac{1}{A_{22}} \cdots \frac{1}{A_{nn}})$ alors l'algorithme pour l'obtention des éléments de $x^{(k+1)}$ à partir des éléments de $x^{(k)}$ s'écrit :

$$x_i^{(k+1)} = \frac{1}{A_{ii}} \left[b_i - \sum_{j=1, j\neq i}^{n} A_{ij} x_j^{(k)} \right], \quad \forall\, i \in [[1, n]].$$

   Nous introduisons la notation

$$J = D^{-1}\, (E + F) \in \mathcal{M}_n(A) \tag{4.31}$$

   donc l'égalité (4.30) s'écrit sous la forme

$$x^{(k+1)} = J\, x^{(k)} + D^{-1}\, b. \tag{4.32}$$

   Cette matrice J s'appelle **matrice de Jacobi** relative à la matrice A.

2. **Méthode de Gauss-Seidel**
   Cette méthode consiste à prendre en (4.24) $M = D - E$ et donc $N = F$ (remarquons qu'ici M est une matrice triangulaire inférieure).
   Pour que cette méthode puisse être utilisée, la matrice $D - E$ doit être inversible, ce qui est encore équivalent à (4.28).
   Donc pour la méthode de Gauss-Seidel la formule de récurrence est

$$(D - E)\, x^{(k+1)} = F\, x^{(k)} + b, \quad \forall\, k \in \mathbb{N} \tag{4.33}$$

   ce qui est équivalent à

$$x^{(k+1)} = (D - E)^{-1}\, F\, x^{(k)} + (D - E)^{-1}\, b, \quad \forall\, k \in \mathbb{N}. \tag{4.34}$$

Comme l'élément $(D-E)_{ij}$ est A_{ij} si $j \leq i$ et l'élément F_{ij} est $-A_{ij}$ si $j > i$ alors de la formule (4.33) on déduit l'algorithme suivant, basé sur une **descente**, pour obtenir les éléments de $x^{(k+1)}$ à partir des éléments de $x^{(k)}$:

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{A_{11}} \left[b_1 - \displaystyle\sum_{j=2}^{n} A_{1j} x_j^{(k)} \right] \\ x_i^{(k+1)} = \dfrac{1}{A_{ii}} \left[b_i - \displaystyle\sum_{j>i} A_{ij} x_j^{(k)} - \sum_{j<i} A_{ij} x_j^{(k+1)} \right], \quad \forall\, i \in [[2, n]]. \end{cases} \tag{4.35}$$

Comme pour la méthode de Jacobi, nous introduisons ce qu'on va appeler la **matrice de Gauss-Seidel** relative à A, à savoir la matrice

$$G = (D-E)^{-1} F \in \mathcal{M}_n(A). \tag{4.36}$$

donc (4.34) s'écrit sous la forme

$$x^{(k+1)} = G\, x^{(k)} + (D-E)^{-1} b, \quad \forall\, k \in \mathbb{N}. \tag{4.37}$$

3. **Méthode de relaxation**
Cette méthode consiste à prendre

$$M = \frac{1}{\omega} D - E$$

et donc

$$N = \frac{1-\omega}{\omega} D + F$$

où $\omega > 0$ est un **paramètre de relaxation** qui a le role de "pondérer" les terms diagonaux.
Pour que cette méthode puisse être utilisée, la matrice $\frac{1}{\omega} D - E$ doit être inversible, ce qui est encore équivalent à (4.28) (car ici encore M est une matrice triangulaire inférieure).
Donc pour la méthode de relaxation la formule de récurrence est

$$\left(\frac{1}{\omega} D - E \right) x^{(k+1)} = \left(\frac{1-\omega}{\omega} D + F \right) x^{(k)} + b, \quad \forall\, k \in \mathbb{N} \tag{4.38}$$

Remarque 4.17. *La méthode de Gauss-Seidel est un cas particulier ($\omega = 1$) de la méthode de relaxation.*

De la formule (4.38) on déduit l'algorithme suivant, basé encore sur une **descente**, pour obtenir les éléments de $x^{(k+1)}$ à partir des éléments de $x^{(k)}$:

$$\begin{cases} \dfrac{A_{11}}{\omega} x_1^{(k+1)} = b_1 - \displaystyle\sum_{j=2}^{n} A_{1j} x_j^{(k)} + \frac{1-\omega}{\omega} A_{11} x_1^{(k)} \\ \dfrac{A_{ii}}{\omega} x_i^{(k+1)} = b_i - \displaystyle\sum_{j>i} A_{ij} x_j^{(k)} - \sum_{j<i} A_{ij} x_j^{(k+1)} + \frac{1-\omega}{\omega} A_{ii} x_i^{(k)}, \quad \forall\, i \in [[2, n]]. \end{cases}$$

En multipliant par $\frac{\omega}{A_{11}}$ et respectivement $\frac{\omega}{A_{ii}}$ on peut voir que ceci est équivalent à

$$x_i^{(k+1)} = \omega x_i^{(k+1/2)} + (1-\omega)x_i^{(k)}, \quad \forall\, i \in [[1,n]]$$

où $x_i^{(k+1/2)}$ est une valeur intermédiaire, qui est le "$x_i^{(k+1)}$" obtenu par l'algorithme de Gauss-Seidel, donc le "$x_i^{(k+1)}$" donné par (4.35).

4. **Méthodes par blocs** (description très succincte)
Supposons que A s'écrit par blocs sous la forme

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdot & \cdot & A_{1p} \\ A_{21} & A_{22} & \cdot & \cdot & A_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{p1} & A_{p2} & \cdot & \cdot & A_{pp} \end{pmatrix}$$

où chaque bloc A_{ii} est une matrice carrée.
On peut alors écrire

$$A = D - E - F$$

avec D, E, F comme au début de la partie 4.4.2 mais avec des écritures par blocs.
On introduit alors les méthodes de **Jacobi par blocs**, **Gauss-Seidel par blocs** et **relaxation par blocs** qui sont définies comme les méthodes respectives données en 1., 2. et 3. mais avec des opérations matricielles par blocs (il faut écrire $(A_{ii})^{-1}$ à la place de $\frac{1}{A_{ii}}$; en plus il faut que tous les blocs A_{ii} soient des matrices inversibles).
Ces méthodes sont utiles pour résoudre numériquement des systèmes algébriques linéaires avec des matrices qui s'écrivent facilement par blocs, comme par exemple la matrice du laplacien 2D.

4.4.3 Convergence des méthodes itératives

Dans cette partie nous considérons $B \in \mathcal{M}_n(\mathbb{K})$ et $c \in \mathbb{K}^n$. Nous considérons aussi $x^* \in \mathbb{K}^n$ tel que

$$x^* = B\,x^* + c. \tag{4.39}$$

D'autre part, nous construisons une suite approximante $\{x^{(k)}\}_{k\in\mathbb{N}} \subset \mathbb{K}^n$ définie par la relation de récurrence

$$x^{(k+1)} = B\,x^{(k)} + c, \quad \forall\, k \in \mathbb{N} \tag{4.40}$$

avec $x^{(0)} \in \mathbb{K}^n$ donné.
Nous allons donner des conditions nécessaires et suffisantes pour avoir la convergence de la suite $x^{(k)}$ vers x^* pour $k \to +\infty$; nous allons voir que ces conditions portent uniquement sur la matrice B.

Remarque 4.18. *Remarquons que toutes les méthodes itératives qu'on a considérées pour résoudre numériquement le système algébrique linéaire* (4.23) *consistent à construire une*

suite $\{x^{(k)}\}_{k\in\mathbb{N}} \subset \mathbb{K}^n$ *avec* $x^{(0)} \in \mathbb{K}^n$ *donné et avec toujours une relation de récurrence de la forme* (4.40) *avec* $B = M^{-1}N$ *et* $c = M^{-1}b$ *(voir la décomposition générale* (4.24)*).*
Nous avons $B = J$ *(matrice de Jacobi donnée par* (4.31)*) si on utilise la méthode de Jacobi et* $B = G$ *(matrice de Gauss-Seidel donnée par* (4.36)*) si on utilise la méthode de Gauss-Seidel.*
D'autre part, comme $A = M - N$ *alors* $x^* \in \mathbb{K}^n$ *est solution de* (4.23) *si et seulement si*

$$Mx^* = Nx^* + b$$

c'est à dire (en multipliant à gauche par M^{-1}*)* x^* *est solution de* (4.23) *si et seulement si* x^* *satisfait* (4.39)*, avec encore* $B = M^{-1}N$ *et* $c = M^{-1}b$*.*

Comme nous voulons étudier la convergence de $x^{(k)}$ vers x^*, on va introduire la notation

$$E^{(k)} = x^{(k)} - x^*, \quad \forall\, k \in \mathbb{N}$$

donc $E^{(k)}$ représente **l'erreur d'approximation** de x^* par $x^{(k)}$.
Par définition on va dire que la méthode itérative considérée est **convergente** *si on a* $x^{(k)} \to x^*$*, c'est à dire* $E^{(k)} \to 0$*, donc* $\|E^{(k)}\| \to 0$ *en* $\mathbb{R}$ *pour* $k \to +\infty$*. Ceci doit être vrai pour toute donnée initiale* $x^{(0)} \in \mathbb{K}^n$ *donc pour tout* $E^{(0)} \in \mathbb{K}^n$ *(on peut utiliser ici n'importe quelle norme sur* $\mathbb{K}^n$ *grâce à l'équivalence des normes).*
En faisant la différence entre (4.40) et (4.39) on obtient par linéarité

$$E^{(k+1)} = B\,E^{(k)}, \quad \forall\, k \in \mathbb{N}$$

On a alors pour tout $k \in \mathbb{N}^*$:

$$E^{(k)} = B\,E^{(k-1)} = B\,B\,E^{(k-2)} = \cdots$$

et on peut montrer très facilement par récurrence sur k la formule

$$E^{(k)} = B^k\,E^{(0)}, \quad \forall\, k \in \mathbb{N}. \tag{4.41}$$

Dans la suite on donnera un théorème qui va nous permettre d'étudier la convergence vers 0 de la suite $E^{(k)}$ donné par (4.41). Nous commençons par le lemme suivant :

Lemme 4.4. *Soit* $S \in \mathcal{M}_n(\mathbb{C})$ *une matrice arbitraire. Nous avons*

1. $$\rho(S) \leq \|S\|, \quad \textit{pour toute norme matricielle subordonnée} \quad \|\cdot\|$$
   où $\rho(\cdot)$ *représente le rayon spectral.*
2. *Pour tout* $\epsilon > 0$ *il existe au moins une norme matricielle subordonnée* $\|\cdot\|$ *(qui va dépendre de* S *et de* ϵ*) telle que*
   $$\|S\| \leq \rho(S) + \epsilon.$$

Démonstration. 1. Soit $\lambda \in \mathbb{C}$ tel que $\rho(S) = |\lambda|$ et soit $x \in \mathbb{C}^n,\ x \neq 0$ un vecteur propre de S associé à la valeur propre λ ; on a alors

$$S\,x = \lambda\,x.$$

En appliquant la norme vectorielle $\|\cdot\|$ on déduit

$$|\lambda|\,\|x\| = \|S\,x\|.$$

Comme $\|x\| > 0$ on déduit

$$|\lambda| = \frac{\|S\,x\|}{\|x\|} \leq \|S\|$$

ce qui donne le résultat.

2. Résultat admis.
Remarque : si S est une matrice hermitienne alors ce résultat est vrai avec la norme $\|\cdot\|_2$ car $\|S\|_2 = \rho(S) \leq \rho(S) + \epsilon$.

□

Remarque 4.19. *Ce résultat nous dit qu'on a*

$$\rho(S) = \inf\{\|S\|, \quad \|\cdot\| \textit{ est une norme subordonnée}\},$$

c'est à dire, pour une matrice complexe donnée il y a toujours une norme matricielle subordonnée qui est aussi proche que l'on veut du rayon spectral de la matrice. Remarquons que dans le cas particulier d'une matrice hermitienne il y a une norme matricielle subordonnée (c'est la norme $\|\cdot\|_2$) qui est égale au rayon spectral (l'infimum ci-dessus est atteint).

Grâce au (4.41), pour montrer la convergence de la suite $x^{(k)}$ définie par (4.40) vers x^* donnée par (4.39) on pourra utiliser le théorème général suivant (avec $S = B$) :

Théorème 4.5. *Pour toute matrice fixée $S \in \mathcal{M}_n(\mathbb{C})$, les 3 propositions suivantes sont équivalentes :*

1. *Pour tout vecteur $v \in \mathbb{C}^n$ on a*

$$S^k\,v \to 0 \quad \textit{pour} \quad k \to +\infty$$

2.

$$\rho(S) < 1$$

3. *Il existe une norme matricielle subordonnée $\|\cdot\|$ telle que*

$$\|S\| < 1.$$

Démonstration. **Montrons 1. $\Rightarrow$ 2.**
Soit $\lambda \in \mathbb{C}$ une valeur propre arbitraire de S et $x \in \mathbb{C}^n \setminus \{0\}$ un vecteur propre associé. On a alors

$$Sx = \lambda x.$$

Pour tout $k \in \mathbb{N}^*$ on a

$$S^k x = S^{k-1}\, Sx = S^{k-1}\,(\lambda x) = \lambda\, S^{k-1}\, x = \lambda\, S^{k-2}\,(Sx) = \lambda^2\, S^{k-2}\, x = \cdots$$

et on peut montrer très facilement par récurrence sur k :

$$S^k x = \lambda^k x, \quad \forall\, k \in \mathbb{N}.$$

Comme $\|S^k x\| \to 0$ alors

$$\|\lambda^k x\| \to 0 \quad \text{pour} \quad k \to +\infty. \tag{4.42}$$

D'autre part on a

$$\|\lambda^k x\| = |\lambda^k|\, \|x\| = |\lambda|^k\, \|x\|.$$

Comme $\|x\|$ est un nombre fixé qui est strictement positif (car $x \neq 0$) on déduit de (4.42)

$$|\lambda|^k \to 0 \quad \text{pour} \quad k \to +\infty$$

ce qui nous donne $|\lambda| < 1$. Comme ceci est valable pour toute valeur propre λ de S on obtient $\rho(S) < 1$ donc la partie **2.**
Montrons 2. $\Rightarrow$ 3.
On applique la partie 2. du Lemme 4.4 avec $\epsilon = \frac{1-\rho(S)}{2} > 0$. On déduit l'existence d'une norme matricielle subordonnée $\|\cdot\|$ telle que

$$\|S\| \leq \rho(S) + \frac{1-\rho(S)}{2} = \frac{1+\rho(S)}{2} < 1$$

ce qui finit la preuve.
Montrons 3. $\Rightarrow$ 1.
On notera encore par $\|\cdot\|$ la norme vectorielle à laquelle la norme matricielle subordonnée est associée. Soit $v \in \mathbb{C}^n$ un vecteur arbitraire. En utilisant les parties 2. et 4. de la Proposition 4.4 on a pour tout $k \in \mathbb{N}^*$:

$$\|S^k v\| \leq \|S^k\|\, \|v\| \leq \|S\|^k\, \|v\|$$

ce qui nous donne la partie **1.** car $\|S\| < 1$. □

Remarque 4.20.
1. *C'est dans la preuve de ce théorème qu'on a besoin de la définition de la norme matricielle subordonnée avec "sup" sur $\mathbb{C}^n$ au lieu de "sup" sur $\mathbb{R}^n$ pour une matruice réelle.*
2. *De la preuve 3. $\Rightarrow$ 1. de ce théorème on voit que plus $\|S\|$ est proche de 0 plus la convergence de $S^k v$ vers 0 est rapide. Comme $\rho(S)$ est aussi proche qu'on souhaite d'une norme $\|S\|$ alors on peut dire : "plus $\rho(S)$ est petite plus la convergence de $S^k v$ vers 0 est rapide".*

Dans la suite nous donnons un exemple d'application du Théorème 4.5. Nous avons la définition suivante :

Définition 4.5. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice.*

1. *On dit que A est* **à diagonale dominante** *si*

$$|A_{ii}| \geq \sum_{j=1, j\neq i}^{n} |A_{ij}|, \quad \forall\, i \in [[1, n]].$$

2. *On dit que A est* **à diagonale strictement dominante** *si*

$$|A_{ii}| > \sum_{j=1, j\neq i}^{n} |A_{ij}|, \quad \forall\, i \in [[1, n]].$$

Nous avons le résultat suivant :

Proposition 4.6. *Soit $A \in \mathcal{M}_n(\mathbb{K})$ une matrice à diagonale strictement dominante. On a alors*

1. *A est une matrice inversible.*

2. *La méthode itérative de Jacobi pour la matrice A est bien définie et convergente.*

Démonstration. 1. Vue en TD.

2. Nous considérons la décomposition standard (4.27)

$$A = D - E - F.$$

Du fait que A est à diagonale strictement dominante on déduit que
$|A_{ii}| > 0, \quad \forall\, i \in [[1, n]]$ donc la condition

$$A_{ii} \neq 0, \quad \forall\, i \in [[1, n]]$$

est satisfaite. Ceci nous dit que la matrice diagonale D est inversible, donc la méthode de Jacobi est bien définie.
Considérons la matrice de Jacobi relative à A qui est

$$J = D^{-1}(E + F).$$

Rappelons la norme subordonnée $\|\cdot\|_\infty$ de J :

$$\|J\|_\infty = \max_{i\in[[1,n]]} \sum_{k=1}^{n} |J_{ik}|.$$

Comme $D^{-1} = diag(\frac{1}{A_{11}}, \frac{1}{A_{22}}, \cdots \frac{1}{A_{nn}})$ on a de la définition de J :

$$J_{ik} = \begin{cases} -\frac{A_{ik}}{A_{ii}} & \text{si} \quad k \neq i \\ \\ 0 & \text{si} \quad k = i. \end{cases}$$

(car la ligne i de J s'obtient en multipliant la ligne i de D^{-1}, qui est égale à $\frac{1}{A_{ii}} e_i^T$, par $E + F$ et ceci donne $\frac{1}{A_{ii}}$ multipliée par la ligne i de $E + F$).
Nous avons alors pour tout $i \in [[1, n]]$:

$$\sum_{k=1}^{n} |J_{ik}| = \frac{1}{|A_{ii}|} \sum_{k=1, k\neq i}^{n} |A_{ik}|.$$

En utilisant le fait que A est à diagonale strictement dominante on obtient

$$\sum_{k=1}^{n} |J_{ik}| < 1, \quad \forall\, i \in [[1, n]]$$

ce qui donne

$$\|J\|_\infty < 1.$$

Le Théorème 4.5 nous donnent le résultat de convergence de la méthode de Jacobi. □

Nous avons le résultat général suivant :

Théorème 4.6. *Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice* **hermitienne et définie positive** *(matrice (HDP)). Supposons qu'on a la décomposition*

$$A = M - N$$

avec $M, N \in \mathcal{M}_n(\mathbb{C})$ telles que M est inversible et $M + N^*$ est une matrice (HDP). Alors on a

$$\rho(M^{-1} N) < 1.$$

Démonstration. On donnera seulement l'idée de la preuve ; pour la preuve complète voir [1] Théorème 5.3-1.
On considère l'application $\|\cdot\|_A : \mathbb{C}^n \to \mathbb{R}$ définie par

$$\|x\|_A = \sqrt{< Ax\,,\, x >} = \sqrt{x^* A x}, \quad \forall\, x \in \mathbb{C}^n$$

et on admet que cette application est une norme sur $\mathbb{C}^n$. On notera toujours par $\|\cdot\|_A$ la norme matricielle subordonnée à cette norme vectorielle. On va demontrer qu'on a

$$\|M^{-1} N\|_A < 1 \tag{4.43}$$

ce qui va donner le résultat attendu, en utilisant la partie 1. du Lemme 4.4.
En écrivant $N = M - A$ on a

$$\|M^{-1}\,N\|_A = \sup_{\|x\|_A=1} \|M^{-1}\,(M-A)x\|_A = \sup_{x^*Ax=1} \sqrt{[(I_n - M^{-1}A)x]^* A (I_n - M^{-1}A)x}.$$

On développe l'expression $[(I_n - M^{-1}A)x]^* A(I_n - M^{-1}A)x$ en posant aussi $y = M^{-1}Ax$; avec un peu de calcul on obtient

$$[(I_n - M^{-1}A)x]^* A(I_n - M^{-1}A)x = 1 - y^*(M + N^*)y$$

et on peut montrer que cette dernière expression est $\leq 1-m$ avec $m > 0$ ce qui nous donne (4.43) et finit la preuve. □

Comme exemple d'application de ce résultat on a la proposition suivante :

Proposition 4.7. *Si $A \in \mathcal{M}_n(\mathbb{C})$ est une matrice (HDP) et si $\omega \in]0,2[$ alors la méthode de relaxation pour A est bien définie et converge (donc en particulier la méthode de Gauss-Seidel est bien définie et converge).*

Démonstration. Considérons la décomposition standard (4.27)

$$A = D - E - F.$$

On va appliquer le Théorème 4.6 avec $M = \frac{1}{\omega}D - E$ et $N = (\frac{1}{\omega} - 1)D + F$.
Remarquons d'abord que $A_{ii} \in \mathbb{R}$ et $A_{ii} > 0, \ \forall\, i \in [[1,n]]$ car A est (HDP). On déduit que la matrice diagonale D est une matrice (HDP) donc inversible. On observe que M est une matrice triangulaire inférieure avec $\frac{A_{ii}}{\omega}$, $i = 1, \cdots n$ sur la diagonale, ce qui implique que M est une matrice inversible, donc la méthode de relaxation est bien définie. Nous avons

$$N^* = \left(\frac{1}{\omega} - 1\right) D^* + F^*$$

et remarquons qu'on a aussi $D^* = D$ et $F^* = E$ car A est hermitienne. Ceci nous donne

$$M + N^* = \frac{1}{\omega}D - E + \left(\frac{1}{\omega} - 1\right) D + E = \frac{2-\omega}{\omega} D.$$

Par hypothèse sur ω on a $\frac{2-\omega}{\omega} > 0$ et comme D est (HDP) on déduit que $M + N^*$ est (HDP). On obtient le résultat attendu comme conséquence du Théorème 4.6. □

Chapitre 5

Interpolation polynomiale

5.1 Quelques mots introductifs et rappels sur les polynomes

Il arrive souvent de vouloir remplacer une fonction compliquée f par une fonction plus simple, par exemple un polynome. On cherchera alors à trouver une fonction polynomiale qui "approche" dans un certain sens la fonction donnée f.
Une autre situation qu'on peut rencontrer peut être le fait d'avoir une connaissance incomplète de la fonction f, par exemple on la connait seulement dans un nombre donné de points ; on veux alors trouver un polynome qui va prendre les mêmes valeurs que la fonction f dans ces points.

On suppose encore ici $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$.

Définition 5.1. *1. On appelle* **polynome** *sur* $\mathbb{K}$ *une fonction* $P : \mathbb{K} \mapsto \mathbb{K}$ *telle que* $\exists n \in \mathbb{N}$ *et* $\exists a_0, a_1, \cdots a_n \in \mathbb{K}$ *tels que*

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots a_n x^n, \quad \forall\, x \in \mathbb{K}$$

(on dira que a_j *est le* **coefficient** *d'ordre* j *du polynome).*
Si $a_n \neq 0$ *on dit que* P *est de degré* n *(notation :* $deg(P) = n$*) ; le nombre* a_n *s'appelle aussi* **coefficient dominant** *du polynome* P*. Par convention si* P *est identiquement nul on pose* $deg(P) = -\infty$*.*

2. On note par $\mathcal{P}(\mathbb{K})$ *l'ensemble des polynomes sur* $\mathbb{K}$*.*

3. Pour tout $n \in \mathbb{N}$ *on note par* $\mathcal{P}_n(\mathbb{K})$ *l'ensemble des polynomes sur* $\mathbb{K}$ *de degré inférieur ou égal à* n*.*

4. Soient $P, Q \in \mathcal{P}(\mathbb{K})$ *deux polynomes sur* $\mathbb{K}$*. On dit que* Q *est un* **diviseur** *de* P *(on dit aussi que* P *est* **divisible** *par* Q *ou* Q **divise** P*) s'il existe un autre polynome* R *sur* $\mathbb{K}$ *tel que* $P = QR$
(c'est à dire $P(x) = Q(x)R(x), \quad \forall\, x \in \mathbb{K}$*).*

On rappelle sans preuve le résultat suivant :

Proposition 5.1. *1. L'ensemble $\mathcal{P}(\mathbb{K})$ est un espace vectoriel sur $\mathbb{K}$ avec les opérations habituelles sur les fonctions : "+" (addition) et "$\cdot$" (multiplication par un scalaire).*

2. Pour tout $n \in \mathbb{N}$ l'ensemble $\mathcal{P}_n(\mathbb{K})$ est un espace vectoriel sur $\mathbb{K}$ (vu comme sous-espace vectoriel de $\mathcal{P}(\mathbb{K})$) et il est fini dimensionel et de dimension $n+1$. Une base dans $\mathcal{P}_n(\mathbb{K})$ est $\{1, x, x^2, \cdots x^n\}$ qui s'appelle **la base canonique**.

3. Si les polynomes $A_0, A_1, \cdots A_n \in \mathcal{P}_n(\mathbb{K})$ sont tels que

$$deg(A_j) = j, \quad \forall\, j \in [[0, n]]$$

alors l'ensemble $\{A_0, A_1, \cdots A_n\}$ est une base dans $\mathcal{P}_n(\mathbb{K})$.

4. Si $P, Q \in \mathcal{P}(\mathbb{K})$, $deg(P) = m$ et $deg(Q) = n$ avec $m, n \in \mathbb{N} \cup \{-\infty\}$ alors $deg(PQ) = m + n$.

Définition 5.2. *Soit $P \in \mathcal{P}(\mathbb{K})$ et $x_0 \in \mathbb{K}$.*

1. On dit que x_0 est une **racine** *de P si $P(x_0) = 0$*
(rappel : ceci est équivalent avec le fait que le polynome $(x - x_0)$ est un diviseur de P).

2. On dit que x_0 est une **racine d'ordre** *k de P avec $k \in \mathbb{N}^*$ si $(x - x_0)^k$ est un diviseur de P et $(x - x_0)^{k+1}$ ne l'est pas.*

Rappelons sans preuve le résultat suivant :

Proposition 5.2. *1. Supposons ici $\mathbb{K} = \mathbb{R}$. Alors $a \in \mathbb{R}$ est une racine d'ordre $k \in \mathbb{N}^*$ de P si et seulement si*

$$P(a) = P'(a) = \cdots = P^{(k-1)}(a) = 0 \quad et \quad P^{(k)}(a) \neq 0$$

où $P^{(j)}$ désigne la dérivée à l'ordre $j \in \mathbb{N}$ de P.

2. Soient $l \in \mathbb{N}$ avec $l \geq 2$, $a_1, a_2 \cdots a_l \in \mathbb{K}$ distinctes deux à deux et $k_1, k_2, \cdots k_l \in \mathbb{N}^*$. Alors un polynome $P \in \mathcal{P}(\mathbb{K})$ est divisible par chacun de polynomes $(x - a_1)^{k_1}, (x - a_2)^{k_2}, \cdots, (x - a_l)^{k_l}$ si et seulement si P est divisible par le polynome produit $(x - a_1)^{k_1}(x - a_2)^{k_2} \cdots (x - a_l)^{k_l}$.

Nous rappelons que tout polynome P tel que $deg(P) \geq 1$ admet au moins une racine complexe (c'est le **Théorème de d'Alembert-Gauss**). On en déduit facilement le résultat suivant :

Proposition 5.3. *Soit $P \in \mathcal{P}(\mathbb{K})$ un polynome tel que $deg(P) = n \geq 1$ et soit $a_n \in \mathbb{K}$ avec $a_n \neq 0$ le coefficient de x^n en P. Alors $\exists s \in [[1, n]]$, $\exists \mu_1, \mu_2, \cdots \mu_s \in \mathbb{C}$ distinctes deux à deux et $\exists m_1, m_2, \cdots m_s \in \mathbb{N}^*$ avec $m_1 + m_2 + \cdots + m_s = n$ tels que*

$$P(x) = a_n(x - \mu_1)^{m_1}(x - \mu_2)^{m_2} \cdots (x - \mu_s)^{m_s}, \quad \forall\, x \in \mathbb{K} \tag{5.1}$$

Dans l'égalité (5.1) $\mu_1, \cdots \mu_s$ sont toutes les racines distinctes de P, $m_1, \cdots m_s$ sont leur ordres ou **multiplicités** et s est le nombre des racines distinctes de P.

5.2 Interpolation de Lagrange

5.2.1 La définition du polynome d'interpolation

Soit $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$ une fonction donnée, $n \in \mathbb{N}^*$ et $x_0, x_1, \cdots x_n \in I$ des éléments deux à deux distinctes de l'intervalle I.
On souhaite construire un polynome $P \in \mathcal{P}_n(\mathbb{R})$ tel que

$$P(x_i) = f(x_i), \quad \forall\, i \in [[0, n]]. \tag{5.2}$$

Remarque : *on peut imaginer qu'on ne connait f que dans ces points $x_0, x_1, \cdots x_n$.*

Définition 5.3. *On appelle* **polynomes fondamentaux de Lagrange** *relatifs aux points $x_0, x_1, \cdots x_n$ les polynomes $L_i \in \mathcal{P}_n(\mathbb{R})$, $i \in [[0, n]]$ définis par*

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \quad \forall\, x \in \mathbb{R}, \quad \forall\, i \in [[0, n]]. \tag{5.3}$$

Exemple : en classe

Proposition 5.4. *Nous avons*

1.
$$L_i(x_k) = \delta_{ik}, \quad \forall\, i, k \in [[0, n]]. \tag{5.4}$$

2. *Les polynomes $L_0, L_1, \cdots, L_n$ forment une base dans $\mathcal{P}_n(\mathbb{R})$.*

Démonstration. 1. C'est immédiat !

2. Il suffit de montrer que $L_0, L_1, \cdots, L_n$ forment une famille indépendante (car ils sont en nombre de $n+1$ qui est la dimension de $\mathcal{P}_n(\mathbb{R})$). Alors soient $a_0, a_1, \cdots, a_n \in \mathbb{R}$ tels que
$$\sum_{i=0}^{n} a_i L_i(x) = 0, \quad \forall\, x \in I.$$
En prenant $x = x_k$ et en utilisant 1. on en déduit $a_k = 0$ et ceci pour tout $k \in [[0, n]]$, ce qui finit la preuve.

□

On a le résultat suivant d'existence et unicité du polynome d'interpolation :

Théorème 5.1. *Il existe un unique polynome $P_n \in \mathcal{P}_n(\mathbb{R})$ tel que (5.2) soit vraie. Ce polynome est donné par l'expression*

$$P_n(x) = \sum_{k=0}^{n} f(x_k) L_k(x), \quad \forall\, x \in \mathbb{R} \tag{5.5}$$

et il s'appelle **le polynome d'interpolation de Lagrange** *de f aux points $x_0, x_1, \cdots x_n$.*

Démonstration. Il est clair que le polynome donné par (5.5) appartient à $\mathcal{P}_n(\mathbb{R})$. Pour tout $i \in [[0, n]]$ on a

$$P_n(x_i) = \sum_{k=0}^{n} f(x_k) L_k(x_i) = \sum_{k=0}^{n} f(x_k)\delta_{ik} = f(x_i)$$

donc P_n satisfait (5.2).
Il reste à montrer l'unicité de P_n. Soit $Q_n \in \mathcal{P}_n(\mathbb{R})$ satisfaisant aussi

$$Q_n(x_i) = f(x_i), \quad \forall\, i \in [[0, n]].$$

Ceci nous donne

$$P_n(x_i) = Q_n(x_i), \quad \forall\, i \in [[0, n]].$$

donc

$$P_n(x_i) - Q_n(x_i) = 0, \quad \forall\, i \in [[0, n]]. \tag{5.6}$$

Alors le polynome $P_n - Q_n$ est divisible par $x - x_i$ pour tout $i \in [[0, n]]$ donc il est divisible par le produit de ces polynomes simples. Alors il existe un polynome R tel que

$$P_n(x) - Q_n(x) = R(x) \prod_{i=0}^{n} (x - x_i), \quad \forall\, x \in \mathbb{R}.$$

D'autre part, le fait que $P_n - Q_n \in \mathcal{P}_n(\mathbb{R})$ implique $R = 0$ (car si $R \neq 0$ alors $deg(P_n - Q_n) \geq n + 1$ ce qui est impossible). Ceci nous donne $P_n - Q_n = 0$ d'où l'unicité. □

Remarque 5.1.

1. *Le définition du polynome d'interpolation P_n (initialement défini pour $n \in \mathbb{N}^*$) peut s'étendre au cas $n = 0$. Si on se donne un seul point $x_0 \in I$ il existe un unique polynome $P_0 \in \mathcal{P}_0(\mathbb{R})$ tel que $P_0(x_0) = f(x_0)$; c'est le polynome constant $P_0(x) = f(x_0)$.*
2. *En fait l'égalité (5.5) nous dit que les $f(x_0), f(x_1), \cdots, f(x_n)$ sont les coordonnées de P_n dans la base $L_0, L_1, \cdots, L_n$.*

5.2.2 La formule de Newton

La formule (5.5) est utile de point de vue théorique mais assez peu utilisée dans la pratique. On va donner dans la suite une formule plus pratique basée sur la récurrence.
L'idé est d'écrire le polynome d'interpolation P_n sous la forme

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

avec $a_0, a_1, \cdots a_n \in \mathbb{R}$ des constantes à trouver.

Définition 5.4. *On appelle* **différence divisée** *de f aux points $x_0, x_1, \cdots, x_n$ notée $f[x_0, x_1, \cdots, x_n]$ le coefficient du polynome d'interpolation P_n.*

Théorème 5.2. *1. Nous avons les relations de récurrence suivantes :*

$$f[x_m] = f(x_m), \quad \forall\, m \in [[0, n]] \tag{5.7}$$

and

$$f[x_0, x_1, \cdots, x_{m+1}] = \frac{f[x_1, x_2, \cdots, x_{m+1}] - f[x_0, x_1, \cdots, x_m]}{x_{m+1} - x_0}, \quad \forall\, m \in [[0, n-1]]. \tag{5.8}$$

2. La relation de récurrence suivante est satisfaite : pour tout $m \in [[0, n-1]]$ on a

$$P_{m+1}(x) = P_m(x) + f[x_0, x_1, \cdots, x_{m+1}] \prod_{j=0}^{m} (x - x_j), \quad \forall\, x \in \mathbb{R}. \tag{5.9}$$

3. Nous avons la formule suivante appellée **formule de Newton** *pour le calcul de P_n :*

$$P_n(x) = f(x_0) + \sum_{i=1}^{n} f[x_0, x_1, \cdots, x_i] \prod_{j=0}^{i-1} (x - x_j), \quad \forall\, x \in \mathbb{R}. \tag{5.10}$$

Démonstration. 1. Pour montrer (5.7) il suffit d'observer que $f[x_m] = P(x_m) = f(x_m)$. Pour montrer (5.8) nous introduisons le polynome

$$Q(x) = \frac{x - x_0}{x_{m+1} - x_0} Q_m(x) + \frac{x_{m+1} - x}{x_{m+1} - x_0} P_m(x) \tag{5.11}$$

où $Q_m \in \mathcal{P}_m(\mathbb{R})$ est le polynome d'interpolation de f aux points $x_1, x_2, \cdots x_{m+1}$. Il est facile de voir que $Q \in \mathcal{P}_{m+1}(\mathbb{R})$. Nous avons aussi

$$Q(x_0) = P_m(x_0) = f(x_0)$$

$$Q(x_{m+1}) = Q_m(x_{m+1}) = f(x_{m+1}).$$

D'autre part, dans le cas où $m \geq 1$ (possible uniquement pour $n \geq 2$) comme $Q_m(x_i) = P_m(x_i) = f(x_i), \quad \forall\, i \in [[1, m]]$ on a

$$Q(x_i) = \frac{x_i - x_0}{x_{m+1} - x_0} f(x_i) + \frac{x_{m+1} - x_i}{x_{m+1} - x_0} f(x_i) = f(x_i), \quad \forall\, i \in [[1, m]].$$

Nous déduisons alors par unicité que Q n'est autre que le polynome d'interpolation de f aux points $x_0, x_1, \cdots x_{m+1}$, ce qui nous donne

$$P_{m+1} = Q. \tag{5.12}$$

Par définition le coefficient de x^m dans $P_m(x)$ est $f[x_0, x_1, \cdots x_m]$ et le coefficient de x^m dans $Q_m(x)$ est $f[x_1, x_1, \cdots x_{m+1}]$, ce qui nous dit que le coefficient de x^{m+1} dans $Q(x)$ (donc aussi dans $P_{m+1}(x)$) est $\frac{f[x_1, x_2, \cdots, x_{m+1}] - f[x_0, x_1, \cdots, x_m]}{x_{m+1} - x_0}$. Nous avons donc obtenu l'égalité (5.8).

2. Comme le polynome $P_{m+1} - P_m$ est dans $\mathcal{P}_{m+1}(\mathbb{R})$ et qu'il s'annule aux points $x_0, x_1, \cdots x_m$ alors il existe $a_{m+1} \in \mathbb{R}$ tel que

$$P_{m+1}(x) = P_m(x) + a_{m+1} \prod_{i=0}^{m} (x - x_i), \quad \forall\, x \in \mathbb{R}. \tag{5.13}$$

   En identifiant le coefficient de x^{m+1} dans cette égalité polynomiale on obtient

$$a_{m+1} = f[x_0, x_1, \cdots x_{m+1}]$$

   ce qui avec (5.13) nous donne le résultat attendu.

3. De 2. on déduit :

$$P_n(x) = P_{n-1}(x) + f[x_0, x_1, \cdots, x_n] \prod_{j=0}^{n-1} (x - x_j) =$$

$$= P_{n-2}(x) + f[x_0, x_1, \cdots, x_{n-1}] \prod_{j=0}^{n-2} (x - x_j) + f[x_0, x_1, \cdots, x_n] \prod_{j=0}^{n-1} (x - x_j) = \cdots$$

   ce qui donne "de proche en proche" la formule souhaitée.

□

Pour calculer les coefficients

$$f[x_0],\ f[x_0, x_1],\ f[x_0, x_1, x_2]\ \cdots\ f[x_0, x_1, \cdots x_n]$$

nécessaires au calcul de P_n nous avons l'algorithme suivant :

Initialisation : on pose $f[x_j] = f(x_j), \quad \forall\, j \in [[0, n]]$
pour $i \in [[1, n]]$ *faire*
  pour $k \in [[0, n - i]]$ *faire*

$$f[x_k, \cdots x_{k+i}] = \frac{f[x_{k+1}, \cdots x_{k+i}] - f[x_k, \cdots x_{k+i-1}]}{x_{k+i} - x_k}$$

  fin pour
fin pour

L'avantage principal de la formule de Newton est le suivant : si on a déja calculé P_n correspondant aux points $x_0, \cdots x_n$ et si on ajoute un point x_{n+1} et on veut calculer P_{n+1} on se sert des coeficients obtenus pour le calcul de P_n. Si on utilise la formule (5.5) qui utilise les polynomes fondamentaux de Lagrange, alors pour passer de P_n à P_{n+1} il faut tout recommencer.

5.2.3 Estimation d'erreur

On donne ici une estimation de l'erreur faite quand on approche une fonction f par un polynome d'interpolation de Lagrange. On commence par le résultat suivant :

Lemme 5.1. *Soit $I = [a, b]$ avec $a, b \in \mathbb{R}$, $a < b$, $n \in \mathbb{N}^*$ et $f : I \to \mathbb{R}$ une fonction de classe C^n (donc $f \in C^n(I)$). Soient $x_0, x_1, \cdots x_n \in I$ des points distinctes deux à deux. Alors il existe $\xi \in I$ avec $\min\{x_0, x_1, \cdots x_n\} < \xi < \max\{x_0, x_1, \cdots x_n\}$ tel que*

$$f[x_0, x_1, \cdots, x_n] = \frac{f^{(n)}(\xi)}{n!}. \tag{5.14}$$

Démonstration. On introduit la fonction $E_n : I \to \mathbb{R}$ définie par

$$E_n(x) = f(x) - P_n(x), \quad \forall\, x \in I.$$

où P_n le polynome d'interpolation de Lagrange de f aux points $x_0, x_1, \cdots x_n$. Comme E_n s'annule aux points $x_0, x_1, \cdots x_n$ alors elle a au moins $n+1$ racines distinctes.
Rappelons le Théorèmme de Rolle : pour une fonction de classe C^1, entre deux racines de la function il y a au moins une racine de la dérivée de cette fonction.
En appliquant ce théorème pour E_n on déduit que E_n' a au moins n racines distinctes ; en l'appliquant encore pour E_n' on déduit que E_n'' a au moins $n-1$ racines distinctes .. et ainsi de suite. On en déduit facilement que la fonction $E_n^{(n)}$ a au moins une racine. Alors il existe $\xi \in \mathbb{R}$ tel que

$$E_n^{(n)}(\xi) = 0$$

c'est à dire

$$f^n(\xi) = P_n^{(n)}(\xi). \tag{5.15}$$

D'autre part, comme P_n est un polynome de degrée inférieur ou égal à n on a

$$P_n^{(n)}(x) = a_n\, n!, \quad \forall\, x \in \mathbb{R}$$

où a_n est le coefficient de x^n dans P_n. Or nous savons de la Définition 5.4 que

$$a_n = f[x_0, x_1, \cdots, x_n]$$

donc

$$P_n^{(n)}(x) = n!\, f[x_0, x_1, \cdots, x_n], \quad \forall\, x \in \mathbb{R}.$$

En utilisant (5.15) on déduit facilement (5.14). □

Remarque 5.2. *Dans le cas $n = 1$ le résultat de ce lemme n'est autre que le Théorème d'accroisements finis (TAF). En effet on voit facilement que $f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0}$ ce qui nous donne : il existe $\xi \in I$ avec $\min\{x_0, x_1\} < \xi < \max\{x_0, x_1\}$ tel que*

$$f'(\xi) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

qui est (TAF).

On peut maintenant ennoncer le résultat principal de cette subsection :

Théorème 5.3. *Soit $I = [a, b]$ avec $a, b \in \mathbb{R}$, $a < b$, $n \in \mathbb{N}^*$ et $f : I \to \mathbb{R}$ une fonction de classe C^{n+1}. Soient $x_0, x_1, \cdots x_n \in I$ des points distinctes deux à deux et P_n le polynome d'interpolation de Lagrange de f aux points $x_0, x_1, \cdots x_n$. Alors*

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\theta_n(x)|, \quad \forall\, x \in I \tag{5.16}$$

où par définition

$$\theta_n(x) = \prod_{i=0}^{n} (x - x_i), \quad \forall\, x \in I$$

et

$$M_{n+1} = \max_{a \leq x \leq b} \left| f^{(n+1)}(x) \right|.$$

Démonstration. Si $x \in I$ est l'un des points $x_0, x_1, \cdots x_n$ alors l'inégalité (5.16) est évidente car $|f(x) - P_n(x)| = 0$.
Supposons maintenant que $x \in I$ avec $x \notin \{x_0, x_1, \cdots x_n\}$.
On va noter par Q le polynome d'interpolation de f aux points $x_0, x_1, \cdots x_n, x$, donc $Q \in \mathcal{P}_{n+1}(\mathbb{R})$. Du Théorème 5.2 partie 2. on déduit

$$Q(y) = P_n(y) + f[x_0, x_1, \cdots x_n, x](y - x_0)(y - x_1) \cdots (y - x_n), \quad \forall\, y \in I.$$

En prenant $y = x$ dans cette formule et en utilisant la fait que $Q(x) = f(x)$ on déduit

$$f(x) - P_n(x) = f[x_0, x_1, \cdots x_n, x]\, \theta_n(x). \tag{5.17}$$

En utilisant le Lemme 5.1 avec $n + 1$ à la place de n on déduit l'existence d'un $\xi \in I$ tel que

$$f[x_0, x_1, \cdots x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

ce qui avec (5.17) nous donne immédiatement (5.16). □

5.3 Interpolation d'Hermite

On suppose qu'on a un intervalle $I \subset \mathbb{R}$ et une fonction $f : I \to \mathbb{R}$ avec f de classe C^1. On se donne aussi $x_0, x_1, \cdots x_n \in I$ des points distinctes 2 à 2 et on se propose de trouver un polynome $H_n \in \mathcal{P}_{2n+1}(\mathbb{R})$ tel que

$$\begin{cases} H_n(x_j) = f(x_j), & \forall\, j \in [[0, n]] \\ \\ H_n'(x_j) = f'(x_j), & \forall\, j \in [[0, n]]. \end{cases} \tag{5.18}$$

Nous avons le résultat suivant :

Théorème 5.4. *Il existe un unique polynome $H_n \in \mathcal{P}_{2n+1}(\mathbb{R})$ tel que* (5.18) *soit vraie. Ce polynome s'écrit sous la forme*

$$H_n(x) = \sum_{i=0}^{n} f(x_i)h_i(x) + \sum_{i=0}^{n} f'(x_i)\tilde{h}_i(x), \quad \forall\, x \in I$$

où les polynomes h_i et $\tilde{h}_i$ sont définis par

$$h_i(x) = [1 - 2L_i'(x_i)(x - x_i)]\, L_i^2(x), \quad \forall\, x \in I, \quad \forall\, i \in [[0, n]]$$

et

$$\tilde{h}_i(x) = (x - x_i)\, L_i^2(x), \quad \forall\, x \in I, \quad \forall\, i \in [[0, n]]$$

et où les L_i sont les polynomes fondamentaux de Lagrange donnés par (5.3).

Idée de la preuve :

1. Il est facile de voir que les polynomes h_i et $\tilde{h}_i$ pour $i \in [[0, n]]$ sont dans $\mathcal{P}_{2n+1}(\mathbb{R})$, donc $H_n \in \mathcal{P}_{2n+1}(\mathbb{R})$.
2. On vérifie que pour tous $i, j \in [[0, n]]$ on a

$$\begin{aligned} h_i(x_j) &= \delta_{ij} \\ \tilde{h}_i(x_j) &= 0 \\ h_i'(x_j) &= 0 \\ \tilde{h}_i{}'(x_j) &= \delta_{ij} \end{aligned}$$

   ce qui permet de montrer que H_n satifait (5.18) (*Exercice*) et ceci nous donne la partie existence.
3. Il nous reste à montrer l'unicité d'un tel polynome H_n.
   S'il y a un autre polynome $\bar{H}_n$ satisfaisant (5.18), en notant $H = H_n - \bar{H}_n$ on déduit facilement par linéarité que H satisfait

$$H(x_i) = H'(x_i) = 0, \quad \forall\, i \in [[0, n]].$$

   Alors H divise $(x - x_i)^2$ pour tout $i \in [[0, n]]$ donc il existe un polynome R tel que

$$H(x) = R(x) \prod_{i=0}^{n} (x - x_i)^2, \quad \forall\, x \in I.$$

   Comme $H \in \mathcal{P}_{2n+1}(\mathbb{R})$ alors nécéssairement $R = 0$ (car $R \neq 0$ implique $deg(H) \geq 2n + 2$). Donc $H = 0$, ce qui finit la preuve de l'unicité.

5.4 Interpolation (ou approximation) au sense de moindres carrés discrets

On suppose ici qu'on connait la fonction f en un nombre de points assez grand ; plus précisément, on suppose qu'on connait f aux points $x_0, x_1, \cdots x_n \in \mathbb{R}$ avec $n \in \mathbb{N}^*$ "grand". On souhaite avoir une approximation polynomiale de f qui ne soit pas nécessairement le polynome d'interpolation P_n de f aux points $x_0, x_1, \cdots x_n$, car P_n serait un polynome de degré trop grand ce qui n'est pas toujours convenable (un inconvenient est le fait que P_n peut être très "oscillant" quand n est grand).
L'idée est alors de se "contenter" d'un polynome $Q \in \mathcal{P}_m(\mathbb{R})$ avec $m \in \mathbb{N}, m \leq n$ (par exemple $m = 1$). Il sera alors en général impossible d'avoir $Q = f$ dans tous les points $x_0, x_1, \cdots x_n$ mais on cherchera Q tel que l'erreur qui est faite dans les points $x_0, x_1, \cdots x_n$ en approchant f par Q soit "la plus petite que possible".
On utilise le même concept dans le cas où les valeurs de f en x_i ne sont pas connues de manière très précise (par exemple si elles sont le résultat des mesures expérimentales qui peuvent être entachées d'erreurs). Dans ce cas, calculer le polynome d'interpolation P_n n'a pas de sens car il serait trop "sensible" aux erreurs de précision pour les $f(x_i)$.

La problématique est donc la suivante : on se donne $n \in \mathbb{N}^*$, des éléments $x_0, x_1, \cdots x_n \in \mathbb{R}$ distinctes deux à deux et aussi des éléments $y_0, y_1, \cdots y_n \in \mathbb{R}$ qui sont des valeurs connues ou approchées de la fonction f. On se donne aussi $m \in \mathbb{N}$ avec $m \leq n$ et et on va chercher une approximation Q de f comme un élément de $\mathcal{P}_m(\mathbb{R})$.
En fait on cherche $Q^* \in \mathcal{P}_m(\mathbb{R})$ telle que

$$\sum_{i=0}^{n} |Q^*(x_i) - y_i|^2 \leq \sum_{i=0}^{n} |Q(x_i) - y_i|^2, \quad \forall\, Q \in \mathcal{P}_m(\mathbb{R}).$$

On préfère écrire ceci sous la forme équivalente :

$$\sum_{i=1}^{n+1} |Q^*(x_{i-1}) - y_{i-1}|^2 \leq \sum_{i=1}^{n+1} |Q(x_{i-1}) - y_{i-1}|^2, \quad \forall\, Q \in \mathcal{P}_m(\mathbb{R}).$$

(l'idée est de rendre la plus petite que possible la norme euclidienne du vecteur de $\mathbb{R}^{n+1}$ dont le i - ème élément est l'erreur $Q(x_{i-1}) - y_{i-1},\ i \in [[1, n+1]]$).
Donc Q^* devra minimiser sur $\mathcal{P}_m(\mathbb{R})$ l'application

$$Q \in \mathcal{P}_m(\mathbb{R}) \ \mapsto \ \sum_{i=1}^{n+1} |Q(x_{i-1}) - y_{i-1}|^2 \in \mathbb{R}.$$

D'autre part, un polynome arbitraire $Q \in \mathcal{P}_m(\mathbb{R})$ s'écrit sous la forme

$$Q(x) = u_1 + u_2 x + u_3 x^2 + \cdots u_{m+1} x^m$$

c'est à dire

$$Q(x) = \sum_{j=1}^{m+1} u_j x^{j-1}, \quad \forall\, x \in \mathbb{R}$$

avec $u_1, u_2, \cdots u_{m+1} \in \mathbb{R}$.

En notant $u = \begin{pmatrix} u_1 \\ u_2 \\ . \\ . \\ u_{m+1} \end{pmatrix} \in \mathbb{R}^{m+1}$ et en introduisant la fonction $J : \mathbb{R}^{m+1} \to \mathbb{R}$ telle que

$$J(u) = \sum_{i=1}^{n+1} \left[\sum_{j=1}^{m+1} u_j x_{i-1}^{j-1} - y_{i-1} \right]^2, \quad \forall\, u \in \mathbb{R}^{m+1}$$

tout revient alors à chercher un point de minimum $u^* = \begin{pmatrix} u_1^* \\ u_2^* \\ . \\ . \\ u_{m+1}^* \end{pmatrix} \in \mathbb{R}^{m+1}$ de la fonction J. Autrement dit, on cherche $u^* \in \mathbb{R}^{m+1}$ tel que

$$J(u^*) \leq J(u), \quad \forall\, u \in \mathbb{R}^{m+1}.$$

Il est clair alors que le polynome recherché Q^* va s'obtenir par la formule

$$Q^* = \sum_{j=1}^{m+1} u_j^* x^{j-1}, \quad \forall\, x \in \mathbb{R}.$$

Exemple : droite de regression (en classe)

Remarquons que pour tout $u \in \mathbb{R}^{m+1}$ le réel $J(u)$ s'écrit sous la forme $J(u) = \|w - Y\|^2$ où $w \in \mathbb{R}^{m+1}$ est un vecteur dont la i - ème composante est $w_i = \sum_{j=1}^{m+1} u_j x_{i-1}^{j-1}$ et $Y = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_n \end{pmatrix} \in \mathbb{R}^{n+1}$ (donc $Y_i = y_{i-1},\ i \in [[1, N+1]]$).

On introduit dans la suite la matrice $B \in \mathcal{M}_{n+1,m+1}(\mathbb{R})$ telle que

$$B_{ij} = x_{i-1}^{j-1}, \quad \forall\, i \in [[1, n+1]], \quad \forall\, j \in [[1, m+1]]$$

et on observe que $w_i = (Bu)_i$ donc $w = Bu$. Alors la fonction J s'écrit

$$J(u) = \|Bu - Y\|^2, \quad \forall\, u = \begin{pmatrix} u_1 \\ u_2 \\ . \\ . \\ u_{m+1} \end{pmatrix} \in \mathbb{R}^{m+1}$$

et le problème qu'on cherche à résoudre peut s'écrire sous la forme équivalente :
Trouver $u^* \in \mathbb{R}^{m+1}$ tel que

$$\|Bu^* - Y\| \leq \|Bu - Y\|, \quad \forall\, u \in \mathbb{R}^{m+1}. \tag{5.19}$$

En notant $v = Bu \in Im(B)$ et $v^* = Bu^* \in Im(B)$ ce problème est équivalent au problème

Trouver $v^* \in Im(B)$ tel que

$$\|v^* - Y\| \leq \|v - Y\|, \quad \forall\, v \in Im(B). \tag{5.20}$$

On rappelle (et on admet) le résultat suivant sur la projection dans un espace euclidien général $\mathbb{R}^p$:

Lemme 5.2. *(Projection sur un espace vectoriel) Soit $F \subset \mathbb{R}^p$ où F est un sous-espace vectoriel de $\mathbb{R}^p$ et $a \in \mathbb{R}^p$. Alors il existe un unique $a^* \in F$ qui s'appelle* **la projection** *de a sur F tel que*

$$\|a - a^*\| \leq \|a - x\|, \quad \forall\, x \in F$$

(en fait a^* est l'élément de F qui minimise la distance de a aux éléments de F).
En plus on a

$$< a - a^*\ ,\ x > = 0, \quad \forall\, x \in F$$

(autrement dit, $a - a^*$ est orthogonal à F, ce qui s'écrit aussi $a - a^* \perp F$).

On peut alors montrer le résultat principal de cette section :

Théorème 5.5. *Il existe un unique $u^* \in \mathbb{R}^{m+1}$ solution de (5.19). En plus u^* est l'unique solution du système algébrique linéaire*

$$B^T B u^* = B^T Y. \tag{5.21}$$

Démonstration. On utilise le Lemme 5.2 avec $p = n+1, F = Im(B)$ et $a = Y$. On déduit qu'il existe un unique $v^* \in Im(B)$ tel qu'on a (5.20) ; on a en plus

$$< Y - v^* \ , \ v > = 0, \quad \forall \, v \in Im(B).$$

On déduit qu'il existe $u^* \in \mathbb{R}^{m+1}$ tel que

$$< Y - Bu^* \ , \ Bu > = 0, \quad \forall \, u \in \mathbb{R}^{m+1}.$$

ce qui donne

$$< B^T Y - B^T B u^* \ , \ u > = 0, \quad \forall \, u \in \mathbb{R}^{m+1}.$$

Ceci nous dit que u^* satisfait (5.21).
Il nous reste à démontrer que la matrice $B^T B \in \mathcal{M}_{m+1}(\mathbb{R})$ est inversible ; pour cela il suffit de montrer que $Ker(B^T B) = \{0\}$. Considérons alors $z \in \mathbb{R}^{m+1}$ tel que

$$B^T B z = 0.$$

En faisant le produit scalaire avec z on déduit

$$< B^T B z \ , \ z > = 0.$$

Comme $< B^T B z \ , \ z > = < Bz \ , \ Bz > = \ \|Bz\|^2$ alors z satisfait $\|Bz\| = 0$ donc

$$Bz = 0. \tag{5.22}$$

D'autre part on peut démontrer que $rang(B) = m+1$. En effet si on considère la sous-matrice carrée B_1 de B de taille $m+1$ donnée par

$$(B_1)_{ij} = x_{i-1}^{j-1}, \quad \forall \, i, j \in [[1, m+1]]$$

(ceci est possible car $m \leq n$) alors $det(B_1) \neq 0$, car les points $x_0, x_1, \cdots x_n$ sont distinctes 2 à 2 (c'est un déterminant de Vandermonde).
On déduit alors que $Ker(B) = \{0\}$ (la matrice B est injective) ce qui avec (5.22) nous donne $z = 0$ ce qui finit la preuve. □

Remarque 5.3. *Nous pouvons donner une généralisation du résultat qu'on vient de voir. Supposons que nous avons une fonction $f : \mathbb{R}^p \to \mathbb{R}^q$ (avec $p, q \in \mathbb{N}^*$) et nous voulons trouver une approximation de f de la forme $g(x, \alpha)$ avec $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_s)^T \in \mathbb{R}^s$ avec $s \in \mathbb{N}^*$ en utilisant le fait qu'on connait des mesures $y^{(0)}, y^{(1)}, \cdots y^{(n)} \in \mathbb{R}^q$ de f aux points*

respectifs $x^{(0)}, x^{(1)}, \cdots, x^{(n)} \in \mathbb{R}^p$. *La méthode de moindres carrées ici consiste à chercher* $\alpha \in \mathbb{R}^s$ *qui minimise la fonction* $J : \mathbb{R}^s \to \mathbb{R}$ *définie par*

$$\alpha \ \mapsto \ J(\alpha) = \sum_{i=0}^{n} \|g(x^{(i)}, \alpha) - y^{(i)}\|^2.$$

On dit aussi qu'on cherche α *qui permet "d'ajuster" le modèle* $f = g(x, \alpha)$ *aux données* $\{(x^{(i)},\ y^{(i)})\}_{i\in[[0,n]]}$.
La théorie de l'interpolation au sens de moindres carrées discrètes vue dans ce chapitre est un cas particulier de cette situation (avec $p = q = 1, \alpha = u, s = m + 1$ *et* $g(x, \alpha) = g(x, u) = \sum_{j=1}^{m+1} u_j x^{j-1}$*).*

Un autre exemple d'un tel modèle ce sont les **réseaux des neurones** *qui sont des fonctions de la forme* $g : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ *données par*

$$g(x, \alpha) = h(x \cdot \alpha - a)$$

avec $h : \mathbb{R} \to \mathbb{R}$ *une fonction donnée et* $a \in \mathbb{R}$ *donné (c'est l'exemple le plus simple de réseau de neurones appellé " à simple couche" ou "perceptron").*

Chapitre 6

Intégration numérique

6.1 Introduction

On se donne un intervalle $[a, b] \subset \mathbb{R}$ avec $a, b \in \mathbb{R}$, $a < b$ et $f : [a, b] \to \mathbb{R}$ une fonction continue. On va noter

$$I = I(f) = \int_a^b f(x)\,dx. \tag{6.1}$$

En général il est difficile de calculer exactement $I(f)$. Le but de ce chapitre est d'apprendre à en calculer une approximation numérique.
Nous connaissons une approximation numérique de $I(f)$ à l'aides de sommes de Rieman. On donnera ici des approximations plus performantes et on étudiera des erreurs d'approximations.

6.2 Formules simples

6.2.1 Introduction aux formules simples

Pour donner une formule approximative de l'intégrale I, on va utiliser des points $x_0, x_1, \cdots x_p \in [a, b]$ distinctes deux à deux appellés **noeuds** et des nombres
$\mu_0, \mu_1, \cdots \mu_p \in \mathbb{R}$ appellés **poids**, où $p \in \mathbb{N}$.

Définition 6.1. *On appelle* **formule d'intégration numérique** *(on l'appelle aussi* **formule de quadrature***) pour le calcul de $I(f)$ une formule du type*

$$I_p(f) = \sum_{j=0}^{p} \mu_j f(x_j). \tag{6.2}$$

Remarque 6.1. *Pour désigner le fait que $\sum_{j=0}^{p} \mu_j f(x_j)$ approche $\int_a^b f(x)\,dx$ on utilise souvent la notation suivante : $I \sim I_p$ c'est à dire*

$$\int_a^b f(x)\,dx \sim \sum_{j=0}^{p} \mu_j f(x_j).$$

On va noter dans la suite

$$e_p(f) = I(f) - I_p(f)$$

qui est **l'erreur d'approximation** de $I(f)$ par $I_p(f)$.
Dans la suite on va supposer que les noeuds $x_0, x_1, \cdots x_p$ sont fixés et on choisira les poids $\mu_0, \mu_1, \cdots \mu_p$ tels que l'erreur soit la plus petite que possible.

Définition 6.2. *On dit que la formule de quadrature* (6.2) *pour approcher* (6.1) *(c'est à dire pour approcher $I(f)$ donnée par* (6.1)*) est* **exacte** *pour une fonction donnée f si*

$$I_p(f) = I(f) \qquad \textit{c'est à dire} \qquad e_p(f) = 0.$$

Définition 6.3. *On dit que la formule de quadrature* (6.2) *pour approcher* (6.1) *est*

1. **d'ordre au moins** *m avec $m \in \mathbb{N}$ (on peut dire aussi* **de degré de précision** *au lieu de* **ordre***) si elle est exacte pour tout polynome $g \in \mathcal{P}_m(\mathbb{R})$ (on peut dire : si elle est exacte sur $\mathcal{P}_m(\mathbb{R})$), c'est à dire si*

$$I_p(g) = I(g), \quad \forall\, g \in \mathcal{P}_m(\mathbb{R}). \tag{6.3}$$

2. **d'ordre** *m avec $m \in \mathbb{N}$ si elle est exacte sur $\mathcal{P}_m(\mathbb{R})$ et en plus*

$$I_p(x^{m+1}) \neq I(x^{m+1})$$

(c'est à dire, elle n'est pas exacte sur $\mathcal{P}_{m+1}(\mathbb{R})$).

Remarque 6.2. *Grâce à la linéarité on a que* (6.3) *est équivalent au système de $m+1$ égalités*

$$I_p(x^k) = I(x^k), \quad \forall\, k \in [[0, m]].$$

Nous avons le résultat suivant

Proposition 6.1. *On suppose que $p \in \mathbb{N}$; Alors il existe des poids uniques $\mu_0, \mu_1, \cdots \mu_p \in \mathbb{R}$ tels que la formule* (6.2) *pour approcher* (6.1) *soit d'ordre au moins p. En plus ces poids sont données par les formules*

$$\mu_0 = b - a, \quad si \quad p = 0 \tag{6.4}$$

$$\mu_k = \int_a^b L_k(x)\, dx, \quad \forall\, k \in [[0, p]], \qquad si \quad p \in \mathbb{N}^* \tag{6.5}$$

où $L_0, L_1, \cdots L_p$ sont les polynomes fondamentaux de Lagrange aux points $x_0, x_1, \cdots x_p$.

Démonstration. De la Remarque 6.2 il faut trouver $\mu_0, \mu_1, \cdots \mu_p \in \mathbb{R}$ tels que

$$I_p(x^k) = I(x^k), \quad \forall\, k \in [[0, p]]$$

ce qui nous donne

$$\sum_{j=0}^{p} \mu_j x_j^k = b_k, \quad \forall\, k \in [[0, p]]$$

où on a noté

$$b_k = \int_a^b x^k \, dx, \quad \forall\, k \in [[0, p]].$$

Il est facile de voir que ceci est équivalent au système algébrique linéaire :

$$\textit{Trouver } \mu = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \cdot \\ \cdot \\ \mu_p \end{pmatrix} \in \mathbb{R}^{p+1} \textit{ tel que}$$

$$A\mu = b$$

où $A \in \mathcal{M}_{p+1}(\mathbb{R})$ est la matrice carrée donnée par

$$A_{kj} = x_j^k, \quad \forall\, k, j \in [[0, p]]$$

et $b = \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ b_p \end{pmatrix} \in \mathbb{R}^{p+1}$ est un vecteur donné.

En rappelant que $det(A)$ est un déterminant de Van der Monde et qu'il est non nul on déduit que A est inversible, ce qui donne l'existence et l'unicité du vecteur μ donc des poids $\mu_0, \mu_1, \cdots \mu_p$.

Pour le calcul des poids μ_j nous avons :
Cas 1. Si $p = 0$ alors μ_0 est tel que

$$\mu_0 1 = \int_a^b 1 \, dx = b - a$$

ce qui donne bien (6.4).
Cas 2. Si $p \in \mathbb{N}^*$ alors comme $L_0, L_1, \cdots L_p$ sont dans $\mathcal{P}_p(\mathbb{R})$ la formule (6.2) pour approcher (6.1) est exacte pour tous les $L_0, L_1, \cdots L_p$. On a alors

$$I_p(L_k) = I(L_k), \quad \forall\, k \in [[0, p]]$$

donc

$$\sum_{j=0}^{p} \mu_j L_k(x_j) = I(L_k) = \int_a^b L_k(x) \, dx, \quad \forall\, k \in [[0, p]].$$

Comme $L_k(x_j) = \delta_{kj}$ on obtient (6.5), ce qui finit la preuve. □

Exemple 6.1. *On va prendre ici $p = 1$ et les noeuds $x_0 = a$ et $x_1 = b$. On cherche les poids $\mu_0, \mu_1 \in \mathbb{R}$ telle que la formule d'approximation*

$$I_1(f) = \mu_0 f(x_0) + \mu_1 f(x_1)$$

soit d'ordre au moins 1 pour approcher (6.1).
Les polynomes fondamentaux de Lagrange aux points $x_0 = a$ *et* $x_1 = b$ *sont*

$$L_0(x) = \frac{x-b}{a-b}, \quad \forall\, x \in \mathbb{R}$$

et respectivement

$$L_1(x) = \frac{x-a}{b-a}, \quad \forall\, x \in \mathbb{R}.$$

En utilisant la formule (6.5) *on a*

$$\mu_0 = \int_a^b L_0(x)\,dx = \frac{1}{a-b}\int_a^b (x-b)\,dx = \frac{1}{a-b}\left[\frac{-(b-a)^2}{2}\right]$$

ce qui donne

$$\mu_0 = \frac{b-a}{2}.$$

De la même manière, en utilisant la formule

$$\mu_1 = \int_a^b L_1(x)\,dx$$

on trouve

$$\mu_1 = \frac{b-a}{2}.$$

On a alors la formule d'approximation

$$I_1(f) = \frac{b-a}{2}[f(a) + f(b)] \tag{6.6}$$

qui est une formule d'ordre au moins 1 pour approcher (6.1).
La formule (6.6) *s'appelle* **la formule du trapèze** *pour approcher* (6.1).
Interprétation graphique : *en classe*

Exemple 6.2. *On considère ici* $p = 2, x_0 = a, x_1 = \frac{a+b}{2}$ *et* $x_2 = b$. *On cherche les poids* $\mu_0, \mu_1, \mu_2 \in \mathbb{R}$ *telle que la formule d'approximation*

$$I_2(f) = \mu_0 f(x_0) + \mu_1 f(x_1) + \mu_2 f(x_2)$$

soit d'ordre au moins 2 pour approcher (6.1).
Les polynomes fondamentaux de Lagrange aux points $x_0 = a$, $x_1 = (a+b)/2$ *et* $x_2 = b$ *sont*

$$L_0(x) = \frac{[x-(a+b)/2](x-b)}{[a-(a+b)/2](a-b)} = \frac{2}{(a-b)^2}\left[x^2 - \frac{a+3b}{2}x + \frac{1}{2}b(a+b)\right], \quad \forall\, x \in \mathbb{R}$$

$$L_1(x) = \frac{(x-a)(x-b)}{[(a+b)/2-a]\,[(a+b)/2-b]} = -\frac{4}{(b-a)^2}\left[x^2 - (a+b)x + ab\right], \quad \forall\, x \in \mathbb{R}$$

et

$$L_2(x) = \frac{(x-a)[x-(a+b)/2]}{(b-a)[b-(a+b)/2]} = \frac{2}{(b-a)^2}\left[x^2 - \frac{3a+b}{2}x + \frac{1}{2}a(a+b)\right], \quad \forall\, x \in \mathbb{R}.$$

(Pour simplifier les notations on utilise les mêmes notations pour les polynomes fondamentaux de Lagrange qu'à l'exemple précédent, même si ces polynomes sont différents).
En utilisant la formule (6.5) *on a*

$$\mu_0 = \int_a^b L_0(x)\,dx = \frac{2}{(a-b)^2}\left[\int_a^b x^2\,dx - \frac{a+3b}{2}\int_a^b x\,dx + \frac{1}{2}b(a+b)\int_a^b 1\,dx\right]$$

$$= \frac{2}{(a-b)^2}\left[\frac{b^3-a^3}{3} - \left(\frac{a+3b}{2}\right)\left(\frac{b^2-a^2}{2}\right) + \frac{1}{2}b(a+b)(b-a)\right].$$

En utilisant les formules de factorisation :

$$b^3 - a^3 = (b-a)(a^2+ab+b^2) \quad \textit{et} \quad b^2 - a^2 = (b-a)(b+a)$$

on factorise la grande paranthèse par $(b-a)$ *ce qui permet de simplifier l'expression de* μ_0 *par* $(b-a)$. *On obtient*

$$\mu_0 = \frac{2}{b-a}\left[\frac{a^2+ab+b^2}{3} - \frac{a^2+4ab+3b^2}{4} + \frac{ab+b^2}{2}\right] = \frac{2}{12(b-a)}(a^2-2ab+b^2)$$

ce qui donne

$$\mu_0 = \frac{b-a}{6}.$$

Le poid μ_1 *s'obtient par la formule*

$$\mu_1 = \int_a^b L_1(x)\,dx$$

ce qui donne par le même type de calculs qu'avant

$$\mu_1 = \frac{2}{3}(b-a).$$

Finalement on a

$$\mu_2 = \int_a^b L_2(x)\,dx$$

ce qui donne comme ci-dessus

$$\mu_2 = \frac{b-a}{6}.$$

On a alors la formule d'approximation

$$I_2(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{6.7}$$

qui est une formule d'ordre au moins 2 pour approcher (6.1).
La formule (6.7) *s'appelle* **la formule de Simpson** *pour approcher* (6.1).

6.2.2 Formules de Newton-Cotes

Nous considérons ici le cas particulier où les noeuds sont équidistants sur l'intervalle $]a, b[$. On se donne $p \in \mathbb{N}^*$ et on pose

$$h = \frac{b-a}{p}.$$

Dans la suite on va considérer deux variantes.
Variante 1.
On pose

$$x_k = a + kh, \quad k \in [[0, p]]$$

donc $x_0, x_1, \cdots x_p$ sont des noeuds équidistantes sur l'intervalle $[a, b]$. On considère alors la formule de quadrature pour approcher (6.1) :

$$I_p(f) = \sum_{k=0}^{p} \mu_k f(x_k) \tag{6.8}$$

où les poids μ_k sont données par la formule (6.5), donc

$$\mu_k = \int_a^b L_k(x)\, dx, \quad \forall\, k \in [[0, p]]$$

où $L_0, L_1, \cdots L_p$ sont les polynomes fondamentaux de Lagrange aux points $x_0, x_1, \cdots x_p$.
La formule (6.8) pour approcher (6.1) s'appelle **formule de Newton-Cotes fermée** à $(p+1)$ points. On lui dit "fermée" parce que les deux extrémitées a et b de l'intervalle font partie des noeuds.
La Proposition 6.1 nous dit que la formule (6.8) est d'ordre au moins p.
Variante 2.
On suppose ici $p \geq 2$. On pose

$$x_k = a + kh, \quad k \in [[1, p-1]]$$

donc $x_1, x_2, \cdots x_{p-1}$ sont des noeuds équidistantes sur l'intervalle $[a+h,\, b-h]$. On considère alors la formule de quadrature pour approcher (6.1) :

$$I_{p-2}(f) = \sum_{k=1}^{p-1} \mu_k f(x_k) \tag{6.9}$$

où les poids μ_k sont donnés par les formules

$$\mu_0 = b - a, \quad \text{si} \quad p = 2 \tag{6.10}$$

$$\mu_k = \int_a^b L_k(x)\, dx, \quad \forall\, k \in [[1, p-1]], \qquad \text{si} \quad p \geq 3 \tag{6.11}$$

et $L_1, L_2, \cdots L_{p-1}$ sont les polynomes fondamentaux de Lagrange aux points $x_1, x_2, \cdots x_{p-1}$ (ces polynomes sont différents de ceux de la Variante 1).
La formules (6.9) pour approcher (6.1) s'appelle **formule de Newton-Cotes ouverte** à $(p-1)$ points. On lui dit "ouverte" parce que les deux extrémitées a et b de l'intervalle ne font pas partie des noeuds.
La Proposition 6.1 nous dit que la formule (6.9) est d'ordre au moins $p-2$.
Exemples :

1. La formule de trapèzes (6.6) est une formule du type Newton-Cotes fermée à deux points.
2. La formule de Simpson (6.7) est une formule du type Newton-Cotes fermée à 3 points.
3. La formule dite **du rectangle** définie par

$$I_0(f) = (b-a) f\left(\frac{a+b}{2}\right) \tag{6.12}$$

   est une formule de Newton-Cotes ouverte à un seul point.

La proposition suivante, que nous donnons sans preuve, dit que dans certains cas l'ordre de ces formules peut être amélioré par rapport à ce que donne la Proposition 6.1.

Proposition 6.2. *Pour tout $p \in \mathbb{N}^*$ la formule de Newton-Cotes fermée à p points ainsi que la formule de Newton-Cotes ouverte à p points sont d'ordre :*
$p-1$ si p est un nombre pair
p si p est un nombre impair.
(on gagne un ordre si p est impair).

Exemple 6.3. *la formule de Simpson est d'ordre 3 alors que les formules du trapèze et du rectangle sont d'ordre 1.*

Remarque : En choisissant de manière "optimale" les points $t_0, t_1, \cdots t_p$ on peut beaucoup améliorer l'ordre de la formule (les points de Gauss, voir TD ?)

6.2.3 Estimation d'erreur

On considère de nouveau la formule de quadrature (6.2) pour approcher (6.1). Nous notons dans la suite

$$e_p(f) = I(f) - I_p(f)$$

qui est **l'erreur d'approximation** de (6.1) par (6.2).
On cherchera dans la suite à obtenir des estimations de l'erreur $e_p(f)$.
Remarque notation : nous utilisons dans la suite la convention $y^0 = 1, \quad \forall\, y \geq 0$ *; donc on pose aussi* $0^0 = 1$ *ce qui est assez inhabituel.*

Définition 6.4. *Pour tout* $m \in \mathbb{N}$ *on appelle* **noyau de Peano** *à l'ordre* m *relatif à la formule de quadrature* (6.2) *la fonction* $G_m : [a,b] \to \mathbb{R}$ *donnée par*

$$G_m(y) = \int_a^b \left[(x-y)^+\right]^m \, dx - \sum_{k=0}^{p} \mu_k \left[(x_k - y)^+\right]^m$$

où $z^+ = \max\{x, 0\}$ *est la partie positive de tout nombre réel* z.

Lemme 6.1. *1. La fonction* $z \in \mathbb{R} \;\mapsto\; z^+ \in \mathbb{R}$ *(fonction partie positive) est une fonction continue sur* $\mathbb{R}$.

2. La fonction G_m *est bien définie et continue.*

Démonstration. 1. *Exercice facile*

2. Par composition de fonctions continues la fonction $y \in [a,b] \;\mapsto\; [(x_k - y)^+]^m$ est continue donc c'est pareil pour la fonction $y \in [a,b] \;\mapsto\; \sum_{k=0}^{p} \mu_k [(x_k - y)^+]^m$.
D'autre part, pour tout $y \in [a,b]$ la fonction $x \in [a,b] \mapsto [(x-y)^+]^m$ est continue et bornée. On déduit alors que l'intégrale $\int_a^b [(x-y)^+]^m \, dx$ est bien définie.
Finalement pour tout $x \in [a,b]$ la fonction $y \in [a,b] \mapsto [(x-y)^+]^m$ est continue, ce qui nous donne la continnuité de la fonction $y \mapsto \int_a^b [(x-y)^+]^m \, dx$ (résultat admis !) Ceci finit la preuve.

□

Remarque 6.3. *On peut écrire*

$$G_m(y) = e_p(g_{m,y})$$

où $g_{m,y}(x) = [(x-y)^+]^m, \quad \forall\, x \in [a,b]$.

Nous avons le résultat suivant :

Lemme 6.2. *Supposons que la formule de quadrature* (6.2) *pour approcher* (6.1) *est d'ordre au moins* $m \in \mathbb{N}$ *et suppososns en plus que* f *est de classe* C^{m+1} *sur l'intervalle* $[a,b]$. *Alors on a*

$$I(f) - I_p(f) = \frac{1}{m!} \int_a^b f^{(m+1)}(y) G_m(y) \, dy$$

où G_m *est le noyau de Peano donné en Définition 6.4.*

Démonstration. On utilise le développement de Taylor de f avec reste intégral à l'ordre m autour du point a ; nous avons

$$f(x) = Q_m(x) + R_m(x), \quad \forall\, x \in [a,b]$$

avec

$$Q_m(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots + \frac{1}{m!}f^{(m)}(a)(x-a)^m$$

et

$$R_m(x) = \int_a^x \frac{f^{(m+1)}(y)}{m!}(x-y)^m\, dy.$$

Nous observons qu'on peut écrire

$$R_m(x) = \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m\, dy.$$

Par linéarité on a

$$e_p(f) = e_p(Q_m) + e_p(R_m).$$

D'autre part on a $e_p(Q_m) = 0$ car $Q_m \in \mathcal{P}_m(\mathbb{R})$ et la formule de quadrature est exacte pour tout polynome de $\mathcal{P}_m(\mathbb{R})$. Nous avons alors

$$e_p(f) = e_p(R_m) = \int_a^b \left[\int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m\, dy\right] dx - \sum_{k=0}^{p} \mu_k \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x_k-y)^+\right]^m\, dy \tag{6.13}$$

En utilisant le théorème de Fubini on peut intervertir les intégrales, ce qui donne

$$\int_a^b \left[\int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m\, dy\right] dx = \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[\int_a^b \left[(x-y)^+\right]^m\, dx\right] dy.$$

Nous avons aussi

$$\sum_{k=0}^{p} \mu_k \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x_k-y)^+\right]^m\, dy = \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[\sum_{k=0}^{p} \mu_k \left[(x_k-y)^+\right]^m\right] dy$$

Ceci nous donne de (6.13) :

$$e_p(f) = \int_a^b \frac{f^{(m+1)}(y)}{m!} G_m(y)\, dy$$

et finit la preuve. □

6.3 Formules composées

Pour avoir une très bonne approximation de $\int_a^b f(x)\,dx$ par la formule de quadrature (6.2) la solution n'est pas d'augmenter le nombre p de noeuds ou poids (le Lemme 6.2 ne permet pas de déduire en général que $|I_p(f) - I(f)| \to 0$ si $p \to +\infty$).
La solution qui est utilisée est de diviser l'intervalle $[a, b]$ en un nombre "grand" de sous-intervalles et d'utiliser une formule de quadrature d'ordre pas trop élevé sur chaqun des sous-intervalles.
Nous considérons alors des nombres $a_0, a_1 \cdots, a_n$ avec $n \in \mathbb{N}^*$ et

$$a = a_0 < a_1 < a_2 < \cdots < a_{n-1} < a_n = b.$$

On a alors

$$I(f) = \int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{a_{i-1}}^{a_i} f(x)\,dx.$$

Pour tout $i \in [[1, n]]$ on va faire le changement des variables $x = a_{i-1} + t(a_i - a_{i-1})$. Nous considérons pour tout $i \in [[1, n]]$ la fonction $\theta_i : \mathbb{R} \to \mathbb{R}$ définie par

$$\theta_i(t) = a_{i-1} + t(a_i - a_{i-1}), \quad \forall\, t \in \mathbb{R}$$

et nous remarquons que θ_i est une bijection de classe C^1 de $[0, 1]$ dans $[a_{i-1},\, a_i]$. Nous notons aussi $g_i = f \circ \theta_i$. En faisant le changement des variables $x = \theta_i(t)$ on arrive à

$$\int_{a_{i-1}}^{a_i} f(x)\,dx = \int_0^1 f(\theta_i(t))\,\theta_i'(t)\,dt = \int_0^1 g_i(t)(a_i - a_{i-1})\,dt$$

ce qui nous donne

$$I(f) = \sum_{i=1}^{n} (a_i - a_{i-1}) \int_0^1 g_i(t)\,dt. \tag{6.14}$$

On utilisera dans la suite une formule simple de quadrature sur l'intervalle $[0, 1]$; nous considérons $p + 1$ noeuds distinctes deux à deux $t_0, t_1, \cdots, t_p \in [0, 1]$ (avec $p \in \mathbb{N}$) et les poids $\mu_0, \mu_1, \cdots, \mu_p \in \mathbb{R}$. Pour toute fonction continue $g : [0, 1] \to \mathbb{R}$ on va approcher $\int_0^1 g(t)\,dt$ par la formule de quadrature

$$I_p(g) = \sum_{k=0}^{p} \mu_k g(t_k). \tag{6.15}$$

On revient à la formule (6.14) et on approche $\int_0^1 g_i(t)\,dt$ en utilisant la formule simple de quadrature (6.15) avec g_i à la place de g. On obtient une approximation I_{np} de $I(f)$ donnée par la **formule composée** suivante :

$$I_{np}(f) = \sum_{i=1}^{n} (a_i - a_{i-1}) \sum_{j=0}^{p} \mu_j f(a_{i-1} + t_j(a_i - a_{i-1})). \tag{6.16}$$

Remarque 6.4. *On pourrait aussi dire qu'on approche* $\int_{a_{i-1}}^{a_i} f(x)\,dx$ *par une formule simple de quadrature du type* $\sum_{k=0}^{p} \nu_k f(x_k^{(i)})$ *avec*

$$\nu_k = (a_i - a_{i-1})\mu_k, \quad x_k^{(i)} = a_{i-1} + t_k(a_i - a_{i-1}).$$

Remarque 6.5. *Très souvent on considère des points* $a_0, a_1, \cdots a_n$ *équidistantes, c'est à dire, on pose*

$$h = \frac{b-a}{n} \quad et \quad a_i = a + ih, \quad i \in [[0, n]].$$

Dans ce cas on utilise très souvent les formules composées suivantes :

1. **La formule du trapèze composée** *qui s'obtient en utilisant la formule du trapèze simple sur* $[0,1]$ *(prendre* $p = 1, t_0 = 0, t_1 = 1, \mu_0 = \mu_1 = \frac{1}{2}$*), ce qui donne*

$$\int_a^b f(x)\,dx \sim \sum_{i=1}^{n} h\frac{1}{2}[f(a_{i-1}) + f(a_i)]$$

c'est à dire

$$\int_a^b f(x)\,dx \sim h\left[\frac{f(a)+f(b)}{2} + \sum_{i=1}^{n-1} f(a+ih)\right] \tag{6.17}$$

2. **La formule du Simpson composée** *qui s'obtient en utilisant la formule du Simpson simple sur* $[0,1]$ *(prendre* $p = 2, t_0 = 0, t_1 = \frac{1}{2}, t_2 = 1, \mu_0 = \mu_2 = \frac{1}{6}, \mu_1 = \frac{2}{3}$*), ce qui donne*

$$\int_a^b f(x)\,dx \sim \frac{h}{6}\sum_{i=1}^{n}\left[f(a_{i-1}) + 4f\left(\frac{a_{i-1}+a_i}{2}\right) + f(a_i)\right]. \tag{6.18}$$

On posera dans la suite

$$e_{np}(f) = I(f) - I_{np}(f)$$

qui donne **l'erreur d'approximation** que nous faisons en approchant $I(f)$ par $I_{np}(f)$. On pose aussi

$$h = \max_{i\in[[1,n]]}(a_i - a_{i-1})$$

et on va prendre en général h "petit". On a le résultat suivant concernant l'erreur d'approximation

Théorème 6.1. *Supposons que la formule simple de quadrature* (6.15) *sur* $[0,1]$ *est d'ordre au moins* $m \in \mathbb{N}$*. Alors pour toute fonction* $f \in C^{m+1}([a,b])$ *on a la majoration suivante pour l'erreur d'approximation :*

$$|e_{np}(f)| \leq h^{m+1}\frac{(b-a)S_{m+1}}{m!}\int_0^1 |G_m(t)|\,dt$$

où G_m *est le noyau de Peano de la formule* (6.15) *et*

$$S_{m+1} = \max_{y\in[a,b]} |f^{(m+1)}(y)|$$

Démonstration. En utilisant (6.14) et (6.16) on déduit

$$e_{np}(f) = \sum_{i=1}^{n} (a_i - a_{i-1}) e_i(f) \tag{6.19}$$

avec

$$e_i(f) = \int_0^1 g_i(t)\, dt - \sum_{j=0}^{p} \mu_j g_i(t_j), \quad \forall\, i \in [[1, n]].$$

En utilisant le Lemma 6.2 on déduit

$$e_i(f) = \frac{1}{m!} \int_0^1 g_i^{(m+1)}(t)\, G_m(t)\, dt$$

et comme

$$g_i^{(m+1)}(t) = (a_i - a_{i-1})^{m+1} f^{(m+1)}(a_{i-1} + t(a_i - a_{i-1}))$$

on obtient pour tout $i \in [[1, n]]$:

$$|e_i(f)| \leq \frac{(a_i - a_{i-1})^{m+1}}{m!} \int_0^1 |f^{(m+1)}(a_{i-1} + t(a_i - a_{i-1}))|\, |G_m(t)|\, dt$$

où G_m est le noyau de Peano à l'ordre m relatif à la formule de quadrature (6.15).
En utilisant la définition de h on déduit

$$|e_i(f)| \leq \frac{h^{m+1}}{m!} S_{m+1} \int_0^1 |G_m(t)|\, dt, \quad \forall\, i \in [[1, n]].$$

Avec (6.19) cela donne

$$|e_{np}(f)| \leq \frac{h^{m+1}}{m!} S_{m+1} \int_0^1 |G_m(t)|\, dt \sum_{i=1}^{n} (a_i - a_{i-1})$$

ce qui donne le résultat. □

Remarque 6.6.

1. *Le noyau de Peano G_m dépend uniquement de $t_0, t_1, \cdots t_p$ et $\mu_0, \mu_1, \cdots \mu_p$ donc il ne dépend pas de n ou h. On peut donc voir $\int_0^1 |G_m(t)|\, dt$ comme une constante.*
2. *On déduit que sous les hypothèses du Théorème 6.1 il existe une constante $C \geq 0$ telle que*

$$|e_{np}(f)| \leq C h^{m+1}$$

En général on cherche à avoir h petit (donc n grand) et m assez grand pour avoir une bonne erreur d'approximation.

3. *Dans le cas des points $a_0, a_1, \cdots a_n$ équidistantes (cas traité dans la Remarque 6.5) nous avons :*
pour la formule du trapèze composée :

$$|e_{np}(f)| \leq Ch^2$$

car l'ordre de la formule du trapèze simple est $m = 1$,
pour la formule de Simpson composée :

$$|e_{np}(f)| \leq Ch^4$$

car l'ordre de la formule du Simpson simple est $m = 3$.

Chapitre 7

Résolution numérique des systèmes algébriques non-linéaires

7.1 Introduction

On se donne un ensemble $D \subset \mathbb{R}^n$ et une fonction $f : D \to \mathbb{R}^n$ (f est un champ de vecteurs de dimension n).
On va chercher des **racines** r de f, c'est à dire, on cherche $r \in D$ tel que

$$f(r) = 0. \tag{7.1}$$

Remarque 7.1.

1. *Il n'est pas toujours possible de résoudre* (7.1) *"à la main" en donnant une solution "exacte"; on va chercher alors à résoudre ce problème numériquement, par des algorithmes spécifiques.*
2. *Si* $n = 1$ *alors* (7.1) *est une équation algébrique* **scalaire**, *c'est à dire, une équation avec une inconnue.*
3. *Si* $n \geq 2$ *alors* (7.1) *est un système algébrique comportant* n *équations et* n *inconnues.*
4. *En particulier* f *peut être une* **fonction affine** *de la forme* $f(x) = Ax - b$, $\forall\, x \in D$ *où* $A \in \mathcal{M}_n(\mathbb{R})$ *est une matrice carrée de taille* n *et* $b \in \mathbb{R}^n$. *Alors* (7.1) *est équivalent au système algébrique linéaire* $Ax = b$ *déjà étudié dans des chapitres précédents.*
5. *Un exemple de situation où on est amenés à résoudre un système du type* (7.1) *apparait dans des problèmes* **d'optimisation sans contraintes**. *Si* $x^* \in \mathbb{R}^n$ *est un point de minimum d'une fonction* $J : \mathbb{R}^n \to \mathbb{R}$ *de classe* C^1 *alors* x^* *satisfait l'équation* **d'Euler** *qui s'écrit*

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

C'est donc un problème du type (7.1) *avec* $f = \nabla J$.

7.2 La méthode de dichotomie (ou de bissection)

On supposera dans cette section que $n = 1$, que D est un intervalle dans $\mathbb{R}$ et que $f : D \to \mathbb{R}$ est une fonction **continue**.
On suppose aussi qu'on a trouvé $a, b \in D$ avec $a < b$ tels que $f(a)f(b) < 0$ (donc $f(a)$ et $f(b)$ ont des signes opposés).
Par le **théorème des valeurs intermédiaires** on sait qu'il existe au moins un élément $r \in]a, b[$ (donc $r \in D$ car D est un intervalle), tel que $f(r) = 0$. La méthode de **dichotomie** (ou bissection) est une méthode pour approcher numériquement une telle racine r de f. Cette méthode consiste à encadrer r par des intervalles de longueur de plus en plus petites. Initialement on sait que $r \in]a, b[$ donc cette intervalle $]a, b[$ est un premier encadrement pour r.
Ensuite on pose $m = \frac{a+b}{2}$ et on calcule $f(m)$.
· Si $f(m) = 0$ alors $r = m$ et on arrête l'algorithme.
· Si $f(m) \neq 0$ alors comme f est continue, la racine r va se trouver
  · dans $]a, m[$ si $f(a)\, f(m) < 0$
  · dans $]m, b[$ si $f(m)\, f(b) < 0$
(il est impossible d'avoir $f(a)f(m) > 0$ et $f(m)f(b) > 0$ car en multipliant les deux inégalitées on trouve $f(a)f(b) > 0$ ce qui est contradictoire avec l'hypothèse faite). Alors un nouveau encadrement de r sera soit l'intervalle $]a, m[$ soit $]m, b[$ qui est chacun un intervalle de longueur $\frac{b-a}{2}$ (la moitié de la longueur de l'encadrement initial). Et ainsi de suite ...
Ceci nous donne l'algorithme suivant :
Algorithme
On considère $a, b \in D$ avec $a < b$ et $\epsilon > 0$ (ϵ est le seuil de précision).
On suppose $f(a)f(b) < 0$; on pose $\eta = b - a$.
Tant que $\eta > \epsilon$
  On pose $m = \frac{a+b}{2}$
    Test : si $f(m) = 0$ alors $r = m$ STOP
      sinon si $f(a)f(m) < 0$ alors $b = m$
      sinon $a = m$
    Fin test
  On pose $\eta = \frac{\eta}{2}$.
Fin "tant que"
On pose $r = a$ (on pourrait aussi poser $r = b$).

Cet algorithme construit deux suites $a^{(k)}$ et $b^{(k)}$ (avec $a^{(0)} = a$ et $b^{(0)} = b$) tels que

$$a^{(k)} \leq r \leq b^{(k)}$$

et $b^{(k)} - a^{(k)} = \frac{b^{(0)}-a^{(0)}}{2^k}$. Alors la distance entre la vraie solution qui est r et l'approximation $a^{(k)}$ de r est inférieure ou égale à $\frac{b^{(0)}-a^{(0)}}{2^k}$, quantité qui tend vers 0 si $k \to +\infty$.

7.3 Méthodes de point fixe

7.3.1 Généralités

Nous écrivons l'équation (7.1) sous une forme équivalente : trouver $r \in D$ tel que

$$g(r) = r \tag{7.2}$$

avec $g : D \to \mathbb{R}^n$ une nouvelle fonction (qui dépend de f). En fait il faut introduire g tel que r est solution de (7.1) si et seulement si r est solution de (7.2).

Exemple 7.1. *1. $f(r) = 0 \iff f(r) + r = r \iff g(r) = r$ où g est définie par $g(x) = f(x) + x, \ \forall\, x \in D$.*

2. On considère $h : D \to \mathbb{R}$ avec $h(x) \neq 0, \ \forall\, x \in D$ (en particulier h peut être une constante non nulle). Alors

$$f(r) = 0 \iff h(r)f(r) = 0 \iff h(r)f(r) + r = r.$$

Donc l'équation de départ (7.1) est équivalente à l'équation $g(r) = r$ avec $g(x) = h(x)f(x) + x, \ \forall\, x \in D$.

Définition 7.1. *Soit $\Omega \subset \mathbb{R}^p$ et $\varphi : \Omega \to \mathbb{R}^p$.*
a) *On dit qu'un élément $y \in \Omega$ est un* **point fixe** *de φ si*

$$\varphi(y) = y.$$

b) *On dit qu'un sous ensemble S de Ω ($S \subset \Omega$) est un ensemble* **stable** *pour la fonction φ si $\varphi(S) \subset S$.*

Alors avec les notations précédentes, chercher une racine r de f (voir (7.1)) revient à chercher un point fixe r de g (voir (7.2)).

Exemple 7.2. *Supposons qu'on veut résoudre l'équation suivante*

$$x^2 - 2x - 3 = 0. \tag{7.3}$$

Il est facile de voir que cette équation admet deux racines réelles qui sont $r_1 = 3$ et $r_2 = -1$.
Il y a plusieurs possibilités d'écrire (7.3) comme un problème de point fixe :
possibilité 1. *On ajoute x ce qui donne : (7.3) $\iff x^2 - 2x - 3 + x = x$. Alors x est solution de (7.3) si et seulement si x est point fixe de g_1 avec $g_1 : \mathbb{R} \to \mathbb{R}$ définie par*

$$g_1(x) = x^2 - x - 3, \quad \forall\, x \in \mathbb{R}.$$

possibilité 2. *Nous avons*

$$(7.3) \iff 2x = x^2 - 3 \iff x = \frac{x^2 - 3}{2}$$

donc x est solution de (7.3) *si et seulement si x est point fixe de g_2 avec $g_2 : \mathbb{R} \to \mathbb{R}$ définie par*

$$g_2(x) = \frac{x^2 - 3}{2}, \quad \forall\, x \in \mathbb{R}.$$

possibilité 3. *Nous avons*

$$(7.3) \iff x^2 = 2x + 3.$$

Cette dernière égalité est équivalente à l'équation $x = \sqrt{2x+3}$ si on se "contente" des racine positives uniquement. On dira alors : x est racine **positive** *de* (7.3) *si et seulement si x est un point fixe de g_3 avec $g_3 : [0, +\infty[\to \mathbb{R}$ définie par*

$$g_3(x) = \sqrt{2x+3}, \quad \forall\, x \in [0, +\infty[.$$

Une manière naturelle d'essayer d'approcher un point fixe d'une fonction g est la **méthode d'approximations successives** (appellée aussi **méthode de point fixe**). Cette méthode consiste à construire par récurrence une suite $\{x^{(k)}\}_{k\in\mathbb{N}} \subset D$ définie par

$$\left\{ \begin{array}{rcl} x^{(k+1)} & = & g(x^{(k)}), \quad \forall\, k \in \mathbb{N} \\ x^{(0)} & \in & D \quad \text{donné.} \end{array} \right. \tag{7.4}$$

Une telle suite va s'appeller **suite des approximations successives** relatif à la fonction g.

Remarque 7.2.

1. *Pour qu'une telle suite soit définie on doit avoir $x^{(k)} \in D \ \ \forall\, k \in \mathbb{N}$. Une condition suffisante qui nous assure qu'une telle suite $x^{(k)}$ est bien définie pour tout $x^{(0)} \in D$ est*

$$g(D) \subset D.$$

   (donc D est un ensemble stable pour D).

2. *Si g est une fonction* **continue** *et si on sait que la suite $x^{(k)}$ est* **convergente**, *avec $x^{(k)} \to y$ pour $k \to +\infty$ avec $y \in D$, alors y est un point fixe de g (c'est évident en passant à la limite $k \to +\infty$ dans la première égalité de* (7.4)*).*

7.3.2 Convergence de la méthode des approximations successives

On commence par définir la notion de fonction strictement contractante. Une telle fonction a la propriété "qu'elle rapproche les images". La définition exacte est la suivante :

Définition 7.2. *Soit $E \subset \mathbb{R}^p$ et $\phi : E \to \mathbb{R}^p$ une fonction. On dit que ϕ est* **strictement contractante** *sur E (on dit aussi : une* **contraction stricte** *sur E) s'il existe une constante α avec $0 \leq \alpha < 1$ et une norme $\|\cdot\|$ sur $\mathbb{R}^p$ telle que*

$$\|\phi(x) - \phi(y)\| \leq \alpha \|x - y\|, \quad \forall\, x, y \in E. \tag{7.5}$$

(ou équivalent, ϕ est une fonction lipschitzienne avec une constante de Lipschitz strictement inférieure à 1).

Remarque 7.3. *Il est évident que toute fonction qui est strictement contractante est aussi* **continue** *car elle est lipschitzienne.*

Exemple 7.3. *Soit* $\phi : \mathbb{R} \to \mathbb{R}$ *avec* ϕ *de classe* C^1 *et telle que*

$$|\phi'(x)| \leq \alpha, \quad \forall\, x \in \mathbb{R}$$

avec $\alpha \in [0, 1[$. *Alors* ϕ *est une contraction stricte sur* $\mathbb{R}$
(car $\forall\, x, y \in \mathbb{R}$ *en utilisant le Théorème d'accroissement finis (TAF) on a*

$$\phi(x) - \phi(y) = \phi'(z)(x - y), \quad \textit{avec} \quad z \in \mathbb{R}$$

ce qui donne $|\phi(x) - \phi(y)| = |\phi'(z)|\,|(x - y)| \leq \alpha |x - y|$ *et ceci donne le résultat).*
Par exemple la fonction $x \mapsto \frac{1}{2}\cos x$ *est telle que le module de sa dérivée est toujours inférieur ou égal à* $\frac{1}{2}$, *donc c'est une contraction stricte sur* $\mathbb{R}$.

Nous avons le résultat suivant :

Théorème 7.1. *(Théorème du point fixe de Banach-Picard)*
Soit $E \subset \mathbb{R}^p$ *un ensemble* **fermé** *et* $\phi : E \to \mathbb{R}^p$ *satisfaisant les deux hypothèses suivantes :*

$$\phi(E) \subset E$$

(donc E *est un ensemble stable pour* Φ*)* *et*

$$\phi \quad \textit{est une contraction stricte sur} \quad E.$$

Soit alors $\alpha \in [0, 1[$ *tel que* (7.5) *soit satisfaite. Alors*
a) *Il existe un unique point fixe* $y \in E$ *de* ϕ.
b) *Pour tout* $x^{(0)} \in E$ *la suite* $\{x^{(k)}\}_{k \in \mathbb{N}} \subset E$ *définie par récurrence par la relation*

$$x^{(k+1)} = \phi(x^{(k)}), \quad \forall\, k \in \mathbb{N} \tag{7.6}$$

converge vers la point fixe y *de* ϕ. *En plus nous avons*

$$\|x^{(k)} - y\| \leq \alpha^k \|x^{(0)} - y\|, \quad \forall\, k \in \mathbb{N} \tag{7.7}$$

(donc on a une convergence au moins **géométrique** *de la suite des approximations successives* $x^{(k)}$ *vers le point fixe* y*).*

Démonstration. Grâce à la propriété de stabilité $\Phi(E) \subset E$ de Φ la suite $x^{(k)}$ est bien définie. Pour tout $k \in \mathbb{N}^*$ nous avons

$$\|x^{(k+1)} - x^{(k)}\| = \|\phi(x^{(k)}) - \phi(x^{(k-1)})\| \leq \alpha \|x^{(k)} - x^{(k-1)}\|$$

ce qui nous donne par récurrence

$$\|x^{(k+1)} - x^{(k)}\| \leq \alpha \|x^{(k)} - x^{(k-1)}\| \leq \alpha^2 \|x^{(k-1)} - x^{(k-2)}\| \leq \cdots \leq \alpha^k \|x^{(1)} - x^{(0)}\|.$$

Nous avons donc

$$\|x^{(k+1)} - x^{(k)}\| \leq \alpha^k \|x^{(1)} - x^{(0)}\|, \quad \forall\, k \in \mathbb{N}.$$

Alors pour tous $k, m \in \mathbb{N}$ avec $m > k$ on a en utilisant l'inégalité triangulaire :

$$\|x^{(m)} - x^{(k)}\| = \Big\| \sum_{j=k}^{m-1} [x^{(j+1)} - x^{(j)}] \Big\| \leq \sum_{j=k}^{m-1} \|x^{(j+1)} - x^{(j)}\|$$

ce qui avec l'inégalité précédente nous donne

$$\|x^{(m)} - x^{(k)}\| \leq \|x^{(1)} - x^{(0)}\| \left(\alpha^k + \alpha^{k+1} + \cdots + \alpha^{m-1}\right).$$

D'autre part nous avons

$$\alpha^k + \alpha^{k+1} + \cdots + \alpha^{m-1} = \alpha^k(1 + \alpha + \cdots \alpha^{m-k-1}) \leq \alpha^k \frac{1}{1-\alpha}$$

car $0 \leq \alpha < 1$. Nous avons donc

$$\|x^{(m)} - x^{(k)}\| \ \leq \ \frac{\|x^{(1)} - x^{(0)}\|}{1-\alpha}\, \alpha^k, \quad \forall\, k, m \in \mathbb{N} \quad \text{avec} \quad k < m. \tag{7.8}$$

Cette dernière inégalité nous dit que la suite $x^{(k)}$ est une suite de Cauchy. Comme le suite est dans $\mathbb{R}^p$ alors elle est convergente donc il existe $y \in \mathbb{R}^p$ tel que

$$x^{(k)} \to y \quad \text{pour} \quad k \to +\infty.$$

Comme l'ensemble E est fermé et la suite $x^{(k)}$ est dans E alors $y \in E$.
En passant à la limite $k \to +\infty$ dans l'égalité (7.6) et en utilisant la continuité de ϕ nous déduisons que y est un point fixe de ϕ.
L'inégalité (7.7) s'obtient en observant que pour tout $k \in \mathbb{N}^*$ on a

$$\|x^{(k)} - y\| = \|\phi(x^{(k-1)}) - \phi(y)\| \leq \alpha \|x^{(k-1)} - y\|$$

donc

$$\|x^{(k)} - y\| \leq \alpha \|x^{(k-1)} - y\| \leq \alpha^2 \|x^{(k-2)} - y\| \leq \cdots \leq \alpha^k \|x^{(0)} - y\|$$

ce qui nous donne (7.7).
Il nous reste à montrer l'unicité de y. Supposons que $z \in E$ est un autre point fixe de ϕ. Nous avons

$$\|y - z\| = \|\phi(y) - \phi(z)\| \leq \alpha \|y - z\|$$

ce qui donne $\|y - z\|(1 - \alpha) \leq 0$ donc $\|y - z\| = 0$ (car $\alpha < 1$) ce qui donne $y = z$; ceci finit la preuve. □

On considère ici $D \in \mathbb{R}^n$ avec D un ensemble **ouvert** et $g : D \to \mathbb{R}^n$ une fonction de classe C^1. On suppose que $r \in D$ est un point fixe de g donc r satisfait l'égalité (7.2). Considérons la suite $x^{(k)}$ définie par (7.4) (en supposant que cette suite est bien définie).

Remarque 7.4. *Si $x^{(0)} = r$ alors la suite $x^{(k)}$ est bien définie et elle est constante égale à r. De même, s'il existe $j \in \mathbb{N}$ tel que $x^{(j)} = r$ alors la suite est bien définie et est constante égalle à r pour tout rang supérieur ou égal à j.*

Nous considérons un élément $x^{(0)}$ arbitraire dans D.
On va noter dans la suite

$$e^{(k)} = x^{(k)} - r, \quad \text{pour tout } k \in \mathbb{N} \text{ tel que } x^{(k)} \text{ est définie.}$$

(c'est l'erreur d'approximation de r par $x^{(k)}$). En faisant la différence entre la première égalité de (7.4) et (7.2) on trouve

$$e^{(k+1)} = g(x^{(k)}) - g(r). \tag{7.9}$$

Commençons par le cas particulier de la dimension 1 ($n = 1$). En faisant un développement de Taylor de g autour du point fixe r nous avons

$$g(x^{(k)}) = g(r) + g'(r)(x^{(k)} - r) + o(e^{(k)}).$$

Nous avons alors de (7.9) en "négligeant" le terme $o(e^{(k)})$:

$$e^{(k+1)} \sim g'(r)e^{(k)}$$

(où "$\sim$" signifie "proche de"). On en déduit

$$\left|e^{(k+1)}\right| \sim |g'(r)|\left|e^{(k)}\right|.$$

Alors l'erreur d'approximation peut diminuer d'un pas à l'autre (en valeur absolue) si on a

$$|g'(r)| \ < \ 1.$$

(c'est uniquement une "intuition" qu'il faudra montrer rigoureusement).
Supposons maintenant qu'on est en dimension quelconque $n \in \mathbb{N}^*$; nous avons un raisonnement qui est très analogue au cas $n = 1$. Dans ce cas aussi on a un développement de Taylor de g autour de r, qui s'écrit

$$g(x^{(k)}) = g(r) + J_g(r)(x^{(k)} - r) + o(\|e^{(k\|)})$$

où $J_g(r)$ est la matrice Jacobienne de g en r. Nous avons encore

$$e^{(k+1)} \sim J_g(r)e^{(k)}$$

et on peut voir que la condition pour avoir la convergence vers r est

$$\|J_g(r)\| \ < \ 1$$

où $\|\cdot\|$ doit être une norme matricielle subordonnée à la norme vectorielle notée encore $\|\cdot\|$.

Avant d'énoncer un résultat rigoureux, donnons la définition suivante :

Définition 7.3. *On appelle* **bassin d'attraction** *pour la méthode itérative* (7.4) *et pour le point fixe r de g l'ensemble des $x^{(0)} \in D$ tels que la suite $x^{(k)}$ donnée par* (7.4) *est bien définie et en plus*

$$x^{(k)} \to r \quad pour \quad k \to +\infty.$$

Remarque 7.5.

1. *Le bassin d'attraction de* (7.4) *pour le point fixe r contient toujours le point r (car si $x^{(0)} = r$ alors $x^{(k)} = r, \ \forall\, k \in \mathbb{N}$).*
2. *En général on souhaite savoir si le bassin d'attraction contient autre chose que le point r ; par exemple on veut savoir s'il contient un voisinage de r en D.*
3. *Si le bassin d'attraction est égal à l'ensemble D, qui est le domaine de définition de g, alors on dira que la méthode itérative* (7.4) **converge globalement** *vers le point fixe r.*
   Si le bassin d'attraction contient un voisinage de r dans D alors on dira que la méthode itérative (7.4) **converge localement** *vers le point fixe r.*

Remarquons que si une méthode itérative converge globalement alors elle converge localement.

Nous avons le résultat suivant :

Théorème 7.2. *Soit $D \subset \mathbb{R}^n$ avec D ouvert, $g : D \to \mathbb{R}^n$ une fonction de classe C^1 et $r \in D$ un point fixe de g. Considérons $\|\cdot\|$ une norme en $\mathbb{R}^n$ et notons toujours par $\|\cdot\|$ la norme matricielle subordonnée associée. Soit $x^{(k)}$ la suite définie par* (7.4) *avec $x^{(0)} \in D$ donnée.*

a) *Supposons qu'on a*

$$\|J_g(r)\| < 1. \tag{7.10}$$

Alors il existe un rayon $\beta > 0$ avec $\overline{B(r,\beta)} \subset D$ tel que si $x^{(0)} \in \overline{B(r,\beta)}$ alors la suite $x^{(k)}$ est bien définie avec $x^{(k)} \in \overline{B(r,\beta)} \ \ \forall\, k \in \mathbb{N}$ et

$$x^{(k)} \ \to \ r \quad pour \quad k \to +\infty$$

(Ici $B(r,\beta)$ est définie en utilisant la norme vectorielle $\|\cdot\|$ de l'hypothèse) .
En plus la convergence est au moins géométrique, c'est à dire il existe $\alpha \in [0,1[$ et une constante $C \geq 0$ tels que

$$\|x^{(k)} - r\| \leq C\alpha^k, \quad \forall\, k \in \mathbb{N}.$$

On dira alors, quand l'hypothèse (7.10) *est satisfaite, que r est un point fixe* **attractif** *de g.*

b) *Supposons ici que $n = 1$, D est un intervalle ouvert dans $\mathbb{R}$ et*

$$|g'(r)| > 1. \tag{7.11}$$

Alors la suite $x^{(k)}$ si elle est bien définie, ne converge pas vers r (sauf dans le cas où il existe $j \in \mathbb{N}$ tel que $x^{(j)} = r$). On dira alors, quand l'hypothèse (7.11) *est satisfaite, que r est un point fixe* **repulsif** *de g.*

Démonstration. **a)** De l'hypothèse (7.10), grâce à la continuité de l'application $x \mapsto \|J_g(x)\|$ (car g est de classe C^1), on déduit qu'il existe $\alpha \in [0, 1[$ et $\beta > 0$ tels que

$$\|J_g(x)\| \leq \alpha, \quad \forall\, x \in \overline{B(r, \beta)}$$

(prendre $\alpha \in]\|J_g(r)\|\,, 1[$).
Le théorème d'accroissements finis (TAF) nous donne

$$\|g(x) - g(y)\| \leq \sup_{z \in \overline{B(r,\beta)}} \|J_g(z)\|\, \|x - y\|, \quad \forall\, x, y \in \overline{B(r, \beta)}.$$

On obtient donc

$$\|g(x) - g(y)\| \leq \alpha \|x - y\|, \quad \forall\, x, y \in \overline{B(r, \beta)}. \tag{7.12}$$

De (7.12) on déduit (prendre $y = r$)

$$\|g(x) - g(r)\| \leq \alpha \|x - r\| \leq \|x - r\|, \quad \forall\, x \in \overline{B(r, \beta)}.$$

Comme $g(r) = r$ cette dernière inégalité nous donne : $\|x - r\| \leq \beta \Rightarrow \|g(x) - r\| \leq \beta$, ce qui veut dire :

$$g(\overline{B(r, \beta)}) \subset \overline{B(r, \beta)}. \tag{7.13}$$

On déduit de (7.12) et (7.13) que la restriction de g sur $\overline{B(r, \beta)}$ satisfait les hypothèses du Théorème de point fixe de Banach-Picard (Théorème 7.1) avec $\phi = g$ (en fait $\phi =$ la restriction de g sur $\overline{B(r, \beta)}$) et $E = \overline{B(r, \beta)}$. En appliquant ce théorème, on obtient immédiatement le résultat attendu.
b) Supposons par absurd que $x^{(k)} \neq r$, $\forall\, k \in \mathbb{N}$ et que $x^{(k)} \to r$ pour $k \to +\infty$. Par continuité de la fonction $x \mapsto |g'(x)|$ on déduit de l'hypothèse (7.11) qu'il existe $\epsilon > 0$ tel que

$$|g'(x)| \geq 1, \quad \forall\, x \in]r - \epsilon, r + \epsilon[. \tag{7.14}$$

D'autre part, comme $x^{(k)} \to r$ il existe un rang $k_0 \in \mathbb{N}$ tel que

$$r - \epsilon < x^{(k)} < r + \epsilon, \quad \forall\, k \geq k_0.$$

Comme $g(r) = r$, en utilsant (TAF), nous avons pour tout $k \geq k_0$:

$$x^{(k+1)} - r = g(x^{(k)}) - g(r) = g'(y_k)(x^{(k)} - r), \quad \text{avec} \quad r - \epsilon < y_k < r + \epsilon$$

ce qui nous donne, grâce à (7.14)

$$|x^{(k+1)} - r| \geq |x^{(k)} - r|, \quad \forall\, k \geq k_0.$$

Comme $|x^{(k)} - r| > 0$ pour tout $k \in \mathbb{N}$, alors la suite réelle $|x^{(k)} - r|$ est strictement positive, croissante à partir du rang k_0 et elle tend vers 0, ce qui est impossible ; ceci finit la preuve.

□

Remarque 7.6. *1. En dimension 1 de l'espace, nous savons que la norme matricielle subordonnée est la valeur absolue (voir Exemple 4.1). Alors dans le cas $n = 1$ l'hypothèse* (7.10) *s'écrit*

$$|g'(r)| < 1.$$

2. La partie **a)** *du Théorème 7.2 nous dit que si le point fixe r de g est attractif alors on a la convergence locale de la méthode de point fixe respective.*

Exemples : (*en classe*).

7.3.3 Ordre de convergence pour la méthode des approximations successives

Nous donnons d'abord la notion générale d'ordre de convergence .

Définition 7.4. *Soit $\{y_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^p$ une suite convergente, donc il existe $y \in \mathbb{R}^p$ tel que*

$$y_k \to y \quad pour \quad k \to +\infty.$$

On dit que y_k converge vers y **à l'ordre au moins** *m avec $m \in [1, +\infty[$ s'il existe $k_0 \in \mathbb{N}$ et $C \geq 0$ constante tels que*

$$\|y_{k+1} - y\| \leq C\|y_k - y\|^m, \quad \forall\, k \geq k_0. \tag{7.15}$$

Dans la suite nous justifions l'affirmation suivante : *plus l'ordre de convergence est grand, plus la convergence de y_k vers y est "rapide".*
En effet, supposons que $m > 1$ et soit $C_1 \geq 0$ tel que la constante C de (7.15) s'écrit $C = C_1^{m-1}$ (donc $C_1 = C^{1/(m-1)}$). En multipliant (7.15) par C_1 on arrive à

$$C_1\|y_{k+1} - y\| \leq (C_1\|y_k - y\|)^m, \quad \forall\, k \geq k_0.$$

Notons $z_k = C_1\|y_k - y\|$ pour tout $k \in \mathbb{N}$. Alors z_k est une suite réelle positive avec $z_k \to 0$ pour $k \to +\infty$. L'inégalité précédente s'écrit

$$z_{k+1} \leq z_k^m, \quad \forall\, k \geq k_0.$$

Nous avons alors pour tout $k \geq k_0$ assez grand :

$$z_k \leq z_{k-1}^m \leq (z_{k-2}^m)^m = z_{k-2}^{m^2} \leq (z_{k-3}^m)^{m^2} = z_{k-3}^{m^3} \leq \cdots$$

et nous obtenons facilement par récurrence

$$z_k \leq z_l^{m^{k-l}}, \quad \forall\, k, l \in \mathbb{N} \quad \text{avec} \quad k > l \geq k_0.$$

Comme la suite z_k converge vers 0 alors on peut fixer l assez grand tel que z_l soit aussi petit que l'on souhaite ; on peut prendre par exemple $z_l \leq \frac{1}{2}$. On déduit alors de l'inégalité précédente

$$z_k \leq \left(\frac{1}{2}\right)^{m^{k-l}}, \quad \forall\, k \geq l. \tag{7.16}$$

Comme $m > 1$ cela donne une convergence très rapide de z_k vers 0, donc de la suite y_k vers la limite y en $\mathbb{R}^p$; en plus on observe que "plus m est grande, plus la convergence sera rapide".
Par exemple si $m = 2$ alors (7.16) nous donne la majoration

$$\|y_k - y\| \leq \frac{1}{C_1}\frac{1}{2^{2^{k-l}}}, \quad \forall\, k \geq l.$$

Nous avons le résultat suivant concernant la méthode des approximations successives :

Proposition 7.1. *Soit $D \subset \mathbb{R}$ un intervalle ouvert, $g : D \to \mathbb{R}$ une fonction de classe C^2 et $r \in D$ un point fixe de g. Supposons en plus qu'on a*

$$g'(r) = 0. \tag{7.17}$$

Alors il existe $\beta > 0$ avec $[r - \beta,\, r + \beta] \subset D$ tel que pour tout $x^{(0)} \in [r - \beta,\, r + \beta]$ la suite $x^{(k)}$ donnée par (7.4) est bien définie et converge vers le point fixe r. En plus le convergence est au moins d'ordre 2 (on dit aussi que la convergence est au moins **quadratique***).*

Démonstration. Les hypothèses de la partie **a)** du Théorème 7.2 sont satisfaites (car $|g'(r)| = 0 < 1$). On déduit de ce théorème qu'il existe $\beta > 0$ avec $[r - \beta,\, r + \beta] \subset D$ tel que pour tout $x^{(0)} \in [r - \beta,\, r + \beta]$ la suite $x^{(k)}$ donnée par (7.4) est bien définie, reste dans l'intervalle $[r - \beta,\, r + \beta]$ et converge vers le point fixe r.
Il nous reste à montrer que la convergence est d'ordre au moins 2. Pour cela nous faisons le développement de Taylor à l'ordre 2 de g autour de r, ce qui nous donne

$$g(x^{(k)}) = g(r) + g'(r)\, e^{(k)} + \frac{1}{2} g''(r + \theta_k e^{(k)})(e^{(k)})^2$$

avec $\theta_k \in]0, 1[$ (rapppel : $e^{(k)} = x^{(k)} - r$). Comme $g(r) = r$ et $g'(r) = 0$ ceci nous donne

$$x^{(k+1)} - r = \frac{1}{2} g''(r + \theta_k e^{(k)})(e^{(k)})^2. \tag{7.18}$$

On va noter

$$M_2 = \sup_{x\in[r-\beta,r+\beta]} |g''(x)|$$

et on a $M_2 < +\infty$ car $g \in C^2(D)$. Remarquons aussi que
$r + \theta_k e^{(k)} \in [r-\beta, r+\beta]$ pour tout $k \in \mathbb{N}$. On déduit alors de (7.18) :

$$|e^{(k+1)}| \ \leq \ M_2\, |e^{(k)}|^2, \quad \forall\, k \in \mathbb{N}$$

ce qui nous donne le résultat. □

Exemple : (en classe)

7.4 La méthode de Newton

Dans cette section on suppose $n = 1$. Soit $D \subset \mathbb{R}$ un intervalle et $f : D \to \mathbb{R}$ une fonction de classe C^1. Soit $r \in D$ une racine de f, donc r satisfait (7.1). On considère ici une manière particulière d'obtenir une suite approximante de r. Pour un $x^{(0)} \in D$ donné on considère la droite tangente en $(x^{(0)},\, f(x^{(0)}))$ au graph de la fonction f ; l'équation de cette droite est

$$y = f(x^{(0)}) + (x - x^{(0)}) f'(x^{(0)}).$$

On va considérer la point d'abscisse x où cette droite intersecte l'axe horizontale, c'est à dire, on pose $y = 0$ dans l'égalité précédente ; cela donne

$$x = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}$$

et cette abscisse x sera l'élément $x^{(1)}$ de la suite approximante. Il faut pour cela avoir $f'(x^{(0)}) \neq 0$. De la même manière on obtient $x^{(2)}$ à partir de $x^{(1)}$, etc ..
Alors la suite que nous construisons se defini par la relation de réccurence suivante :

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad \forall\, k \in \mathbb{N} \\ x^{(0)} &\in D \quad \text{donné.} \end{aligned} \tag{7.19}$$

Cette méthode s'appelle **méthode** (ou **algorithme**) **de Newton**.

Remarque 7.7. *Pour que la suite $x^{(k)}$ construite par cette méthode soit bien définie il faut*

$$x^{(k)} \in D \quad \textit{et} \quad f'(x^{(k)}) \neq 0, \quad \forall\, k \in \mathbb{N}.$$

Dans la suite nous posons

$$D_1 = \{x \in D, \quad f'(x) \neq 0\} \subset D$$

et nous introduisons la fonction $\varphi : D_1 \to \mathbb{R}$ définie par

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \quad \forall\, x \in D_1. \tag{7.20}$$

Nous avons pour tout $x \in D_1$:

$$x = \varphi(x) \Longleftrightarrow x = x - \frac{f(x)}{f'(x)} \Longleftrightarrow f(x) = 0$$

c'est à dire $x \in D_1$ est point fixe de φ si et seulement si x est racine de f.

Remarque 7.8. *Supposons que la racine r de f est telle que $f'(r) \neq 0$, c'est à dire $r \in D_1$. Alors r est un point fixe de φ. En plus la relation de réccurence (7.19) n'est autre que la méthode des approximations successives relative à la fonction φ, car (7.19) s'écrit $x^{(k+1)} = \varphi(x^{(k)})$.*

Nous finissons ce chapitre par le résultat suivant de convergence pour la méthode de Newton :

Théorème 7.3. *Soit $D \subset \mathbb{R}$ un intervalle ouvert, $f : D \to \mathbb{R}$ de classe C^2 et $r \in D$ telle que*

$$f(r) = 0.$$

Supposons en plus que

$$f'(r) \neq 0.$$

Alors il existe $\beta > 0$ avec $[r - \beta, r + \beta] \subset D$ tel que pour tout $x^{(0)} \in [r - \beta, r + \beta]$ la suite $x^{(k)}$ donnée par la méthode itérative (7.19) est bien définie et on a $x^{(k)} \to r$ pour $k \to +\infty$. En plus la convergence de $x^{(k)}$ vers r est au moins d'ordre 2.

Démonstration. On fera la preuve sous l'hypothèse plus forte $f \in C^3(D)$ et on va admettre le résultat pour $f \in C^2(D)$.
Considérons la fonction φ définie par (7.20). Comme $f'(r) \neq 0$, par continuité de f' il existe un intervalle ouverte V qui contient r avec $V \subset D$ tel que

$$f'(x) \neq 0, \quad \forall\, x \in V$$

(donc $V \subset D_1$). Alors il est clair que φ est bien définie sur V et $\varphi \in C^2(V)$ (car $f \in C^3(D)$).
Rappelons que r est un point fixe de φ (voir Remarque 7.8).
D'autre part nous avons pour tout $x \in V$:

$$\varphi'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}$$

Ceci nous donne

$$\varphi'(r) = 0.$$

On peut alors appliquer la Proposition 7.1 avec V à la place de D et φ à la place de g et on obtient le résultat. □

Chapitre 8

Aspects théoriques et numérique sur les systèmes d'équations différentielles ordinaires (EDO)

8.1 Cadre général théorique

8.1.1 Introduction

On se donne $I \subset \mathbb{R}$ un intervalle ouvert, un nombre $d \in \mathbb{N}^*$, un ensemble ouvert $U \subset \mathbb{R}^d$ et une fonction **continue**

$f : I \times U \mapsto \mathbb{R}^d$ avec $f = \begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ f_d \end{pmatrix}$ où $f_k = I \times U \mapsto \mathbb{R}$ pour tout $k = 1, 2, \cdots d$.

On va noter les variables de f par $t \in I$ et $x = (x_1, x_2, \cdots x_d) \in U$, donc on a $f(t, x)$ ou $f(t, x_1, x_2, \cdots x_n)$.

On cherche un intervalle ouvert $J \subset I$ et une fonction $y : J \mapsto U$ où

$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_d \end{pmatrix}$ et $y_k : J \mapsto \mathbb{R} \ \ \forall\, k = 1, 2, \cdots d$ avec $y \in C^1(J)$ et tels que

$$y'(t) = f(t, y(t)), \qquad \forall t \in J \tag{8.1}$$

c'est à dire, pour tout $t \in J$:

$$\left\{ \begin{array}{rcl} y_1'(t) & = & f_1(t, y_1(t), y_2(t), \cdots y_d(t)) \\ y_2'(t) & = & f_2(t, y_1(t), y_2(t), \cdots y_d(t)) \\ \cdot & & \\ \cdot & & \\ y_d'(t) & = & f_n(t, y_1(t), y_2(t), \cdots y_d(t)). \end{array} \right. \tag{8.2}$$

On dira que (8.1) ou (8.2) est un **système d'équations différentielles ordinaire (EDO)** d'ordre 1 ou encore une **équation différentielle ordinaire vectorielle** d'ordre 1.

Remarque 8.1.
1. *En général on cherche J le plus "grand" que possible ; quand J peut être pris égal à I alors on dit que y est une solution* **globale**.
2. *Dans le cas $d = 1$ le système se réduit à une seule équation ; on dit alors qu'on a une* **équation différentielle scalaire d'ordre 1.**
3. *On dit que l'EDO* (8.1) *est d'ordre 1 car il fait intervenir la dérivée de l'inconnue y à l'ordre 1 seulement.*

En général on peut avoir une infinité des solutions de (8.1).
Le plus souvent on cherche à résoudre le système (8.1) avec ce qu'on appelle une "condition initiale", c'est à dire, on se donne $t_0 \in I$ et $y^0 \in U$, $y^0 = (y_1^0, y_2^0, \cdots y_d^0)^T \in U$ et on cherche un intervalle ouvert $J \subset I$ tel que $t_0 \in J$ et une fonction $y : J \mapsto U$ satisfaisant (8.1) ainsi que la condition initiale

$$y(t_0) = y^0. \tag{8.3}$$

On appelle alors **problème de Cauchy** le système (8.1) et (8.3).
On verra dans cette section un résultat d'existence et unicité d'une solution pour les problèmes de Cauchy, sous des hypothèses appropriées.

Remarque 8.2. *Si f est indépendante de t (donc $f : U \to \mathbb{R}^d$) alors le système* (8.1) *s'écrit $y' = f(y)$ et s'appelle système d'équations différentielles* **autonome**.

Nous avons à faire très souvent à des équations différentielles scalaires d'ordre supérieur à 1. Ce sont des équations différentielles du type suivant : on se donne une fonction continue $g : I \times \mathbb{R}^d \mapsto \mathbb{R}$, $g(t, x_1, x_2, \cdots x_d)$ et on cherche un intervalle ouvert $J \subset I$ et une fonction $z : J \to \mathbb{R}$ de classe C^d tels que

$$z^{(d)}(t) = g(t, z(t), z'(t), z''(t), \cdots z^{(d-1)}(t)) \quad \forall\, t \in J. \tag{8.4}$$

On dira alors que (8.4) est une **l'équation différentielle scalaire d'ordre** d.

Remarque 8.3. *On peut ramener la résolution de* (8.4) *à la résolution d'un système de type* (8.1) *ou* (8.2) *de la manière suivante : on introduit des nouvelles inconnues*

$$\begin{aligned} z_1(t) &= z(t) \\ z_2(t) &= z'(t) \\ &\cdot \\ &\cdot \\ z_d(t) &= z^{(d-1)}(t) \end{aligned}$$

Si z satisfait (8.4) *alors $(z_1, z_2, \cdots z_d)$ satisfait le système différentiel d'ordre 1 suivant :*

$$\begin{aligned} z_1' &= z_2 \\ z_2' &= z_3 \\ &\cdot \\ &\cdot \\ z_{d-1}' &= z_d \\ z_d' &= g(t, z_1, z_2, \cdots z_d) \end{aligned}$$

et réciproquement, si $z_1, z_2, \cdots z_d$ satisfont ce dernier système alors en posant $z = z_1$ alors z satisfait (8.4).
En posant $Z = (z_1, z_2, \cdots z_d)^T$, le système ci-dessus s'écrit sous la forme vectorielle

$$Z'(t) = F(t, Z)$$

où la fonction $F : I \times \mathbb{R}^d \to \mathbb{R}^d$ est définie par

$$(t, x_1, x_2, \cdots x_d)^T \mapsto F(t, x_1, x_2, \cdots x_d) = \begin{pmatrix} x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_d \\ g(t, x_1, x_2, \cdots x_d) \end{pmatrix}.$$

Exemple en mécanique :
Nous considérons le mouvement rectiligne d'un corp sous l'action d'une force donnée, qui dépend uniquement de la position et de la vitesse du corp.
Nous notons $t \in \mathbb{R}$ le temps, $z(t)$ la position et $z'(t)$ la vitesse du corp. En supposant que la masse du corp est égale à 1 et en appliquant la deuxième loi de Newton on voit que z satisfait l'EDO d'ordre 2 suivante :

$$z''(t) = g(z(t), z'(t))$$

où $g : \mathbb{R}^2 \to \mathbb{R}$ est une fonction donnée ; $g(x_1, x_2)$ représente la force qui s'exerce sur un corp se trouvant à la position x_1 et ayant la vitesse x_2.

En notant $z_1 = z$ et $z_2 = z'$ cette EDO d'ordre 2 est équivalente au système EDO d'ordre 1 suivant :

$$\begin{cases} z_1'(t) &= z_2(t) \\ z_2'(t) &= g(z_1(t),\, z_2(t)) \end{cases}$$

En notant $Z = (z_1, z_2)^T$, ce système peut s'écrire sous la forme vectorielle

$$Z'(t) = f(t, Z(t))$$

avec $f : \mathbb{R}^3 \mapsto \mathbb{R}^2$ la fonction donnée par

$$f(t, x_1, x_2) = \begin{pmatrix} x_2 \\ g(x_1, x_2) \end{pmatrix} \qquad \forall\, (t, x_1, x_2) \in \mathbb{R}^3.$$

Remarquer que f est indépendante du temps t et donc ce système EDO est autonome.

8.1.2 Quelques cas particuliers de résolution "à la main" des EDO

Le cas linéaire à coefficients constants

C'est le cas où dans (8.1) on a $U = \mathbb{R}^d$, $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ avec

$$f(t, x) = Ax + b(t), \quad \forall\, (t, x) \in I \times \mathbb{R}^d.$$

Ici A et b sont données, $A \in \mathcal{M}_d(\mathbb{R})$ matrice constante et $b : I \to \mathbb{R}^d$ est une fonction continue. Le système EDO (8.1) s'écrit alors

$$y' = Ay + b(t) \tag{8.5}$$

Rappelons que dans ce cas il existe une solution unique du problème de Cauchy (8.5), (8.3) qui est définie sur tout l'intervalle I et qui est donnée par la **formule de Duhamel** :

$$y(t) = e^{(t-t_0)A}\, y^0 + \int_{t_0}^{t} e^{(t-s)A}\, b(s)\, ds, \quad \forall\, t \in I \tag{8.6}$$

(**Rappel :** par définition $\forall\, p \in \mathbb{N}^*$, $\forall\, B \in \mathcal{M}_p(\mathbb{R})$ on a $e^B = \sum_{j=0}^{\infty} \frac{1}{j!} B^j \in \mathcal{M}_p(\mathbb{R})$).

Remarquons que si dans la formule ci-dessus on prend $p = 1$ et $B \in \mathbb{R}$ alors e^B n'est autre que l'exponentielle habituelle d'un nombre réel.
Rappelons que pour déduire (8.6) on montre d'abord que la solution de (8.5), (8.3) s'écrit comme la somme de la solution générale de l'équation homogène correspondante (celle obtenue en prenant $b = 0$ en (8.5)) et d'une solution particulière de (8.5) qui peut s'obtenir en utilisant la méthode de **variation de la constante**.

Le cas linéaire en dimension 1

C'est le cas où dans (8.1) on prend $d = 1, U = \mathbb{R}$ et $f : I \times \mathbb{R} \to \mathbb{R}$ est donnée par

$$f(t, x) = a(t)x + b(t), \quad \forall\, (t, x) \in I \times \mathbb{R}$$

avec $a, b : I \to \mathbb{R}$ des fonctions continues données.
L'EDO va donc s'écrire

$$y' = a(t)y + b(t) \tag{8.7}$$

Rappelons que dans ce cas aussi il existe une solution unique du problème de Cauchy (8.7), (8.3) qui est définie sur tout l'intervalle I et qui est donnée par la version suivante de la **formule de Duhamel** :

$$y(t) = \exp\left(\int_{t_0}^{t} a(s)ds\right) y^0 + \int_{t_0}^{t} \exp\left(\int_{s}^{t} a(\tau)d\tau\right) b(s)\, ds, \quad \forall\, t \in I. \tag{8.8}$$

Remarque 8.4. *Si a est une constante $a \in \mathbb{R}$ alors $\int_s^t a(\tau)\, d\tau = a(t-s)$ et (8.8) devient*

$$y(t) = e^{(t-t_0)a}\, y^0 + \int_{t_0}^{t} e^{(t-s)a}\, b(s)\, ds, \quad \forall\, t \in I$$

qui est aussi le cas particulier $d = 1$ et $A = a$ de (8.6).

Remarque : Il n'existe pas de méthode standard pour résoudre un système EDO à coefficients variables de la forme

$$y' = A(t)y + b(t)$$

si $d \geq 2$ avec la matrice A dépendante de t.

Le cas des variables séparées

Dans cette partie on a $d = 1$ et $f : I \times U \to \mathbb{R}$ de la forme : on considère I et U deux intervalles ouvertes dans $\mathbb{R}$

$$f(t, x) = \alpha(t)\, \beta(x), \quad \forall\, (t, x) \in I \times U$$

avec $\alpha : I \to \mathbb{R}$ et $\beta : U \to \mathbb{R}$ des fonctions données continues.
Alors (8.1) devient l'EDO scalaire : trouver J intervalle avec $J \subset I$ et $y : J \to \mathbb{R}$ de classe C^1 tels que

$$y'(t) = \alpha(t)\, \beta(y(t)), \quad \forall\, t \in J. \tag{8.9}$$

On appelle une telle équation **équation différentielle à variables séparées**.
Dans la suite nous montrons comment résoudre en général une équation du type (8.9).
Nous supposons que β n'est pas la fonction constante 0.
Nous notons par $M \subset U$ l'ensemble de toutes les racines de β, donc

$$M = \{x \in U, \quad \beta(x) = 0\}.$$

($M \subset U$ et $M \neq U$ et peut être l'ensemble vide $\emptyset$). L'ensemble M est fermé (car $M = \beta^{-1}(\{0\})$ et le singleton $\{0\}$ est un ensemble fermé) donc l'ensemble $U \setminus M$ est ouvert.

Il est évident que si M n'est pas l'ensemble vide alors pour tout $r \in M$ la fonction constante $y(t) = r \quad \forall\, t \in I$, est une solution de l'équation différentielle (8.9) (l'intervalle ouvert J où y est définie sera alors $J = I$).

Pour trouver d'autres solutions que les fonctions constantes, nous considérons B une primitive de la fonction $\frac{1}{\beta}$ sur $U \setminus M$, donc B est une fonction de classe C^1 sur l'ensemble ouvert $U \setminus M$ tel que

$$B'(x) = \frac{1}{\beta(x)}, \quad \forall\, x \in U \setminus M.$$

(une telle primitive existe car $\frac{1}{\beta}$ est une fonction continue sur $J \setminus M$).
D'autre part, soit A une primitive de α sur I, c'est à dire :

$$A'(t) = \alpha(t), \quad \forall\, t \in I.$$

Nous avons

Théorème 8.1. *Soit $J \subset I$ un intervalle ouvert et $y : J \to U$ une fonction de classe C^1 telle que*

$$\beta(y(t)) \neq 0, \quad \forall\, t \in J$$

(autrement dit : $y(t) \in U \setminus M, \quad \forall\, t \in J$).
Alors y est une solution de (8.9) si et seulement si il existe une constante $C \in \mathbb{R}$ tel que

$$B(y(t)) = A(t) + C, \quad \forall\, t \in J. \tag{8.10}$$

Démonstration. " $\Longrightarrow$ " Si y est solution de (8.9) alors en divisant cette équation par $\beta(y(t))$ (car non nulle) on a

$$\frac{y'(t)}{\beta(y(t))} = \alpha(t), \quad \forall\, t \in J$$

ce qui est équivalent à

$$\frac{d}{dt}B(y(t)) = \frac{d}{dt}A(t), \quad \forall\, t \in I_0.$$

Il est clair alors qu'il existe $C \in \mathbb{R}$ tel qu'on a (8.10).
" $\Longleftarrow$ " Si y est solution de (8.10) alors en dérivant cette égalité en t on trouve immédiatement (8.9). □

Exemple 8.1. *On se propose de résoudre l'équation différentielle*

$$y' = t\, y^2. \tag{8.11}$$

C'est une équation différentielle du type (8.9) (à variables séparées) avec

$$I = \mathbb{R}$$

$$U = \mathbb{R}$$
$$\alpha(t) = t, \quad \forall\, t \in \mathbb{R}$$
$$\beta(x) = x^2, \quad \forall\, x \in \mathbb{R}$$

donc $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *donnée par*

$$f(t, x) = tx^2, \quad \forall\, (t, x) \in \mathbb{R}^2.$$

Cherchons d'abord les racines de β *: nous avons* $\beta(r) = 0 \iff r^2 = 0 \iff r = 0$*, donc* $M = \{0\}$*. Nous avons alors immédiatement une solution de* (8.11) *:*
$J = \mathbb{R}$ *et* $y(t) = 0 \quad \forall\, t \in \mathbb{R}$*.*

Pour trouver d'éventuelles solutions non nulles, nous divisons l'équation (8.11) *par* $y^2(t)$ *(en supposant* $y(t) \neq 0 \quad \forall\, t$*). On a alors*

$$\frac{y'(t)}{y^2(t)} = t$$

qui est équivalent à

$$\frac{d}{dt}\left(-\frac{1}{y(t)}\right) = \frac{d}{dt}\left(\frac{t^2}{2}\right).$$

Ceci est équivalent à : $\exists C \in \mathbb{R}$ *tel que*

$$-\frac{1}{y(t)} = \frac{t^2}{2} + C \tag{8.12}$$

Remarque : *On aurait pu arriver à* (8.12)*, en appliquant directement le Théorème 8.1 ; pour ceci on introduit une primitive de* $\frac{1}{x^2}$*, c'est à dire* $B : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ *avec*

$$B(x) = -\frac{1}{x}, \quad \forall\, x \neq 0$$

et une primitive de t*, c'est à dire* $A : \mathbb{R} \to \mathbb{R}$ *avec*

$$A(t) = \frac{t^2}{2}, \quad \forall\, t \in \mathbb{R}.$$

Alors (8.12) *est une conséquence directe de l'égalité* (8.10) *du Théorème 8.1.*

Pour trouver les solutions non nulles de (8.11) *il suffira alors d'exprimer* $y(t)$ *en fonction de* t *à partir de* (8.12)*. Nous avons*

$$y(t) = -\frac{1}{t^2/2 + C} = -\frac{2}{t^2 + 2C}$$

Donc pour toute constante $C \in \mathbb{R}$ *nous avons une solution*

$$y(t) = -\frac{2}{t^2 + 2C}. \tag{8.13}$$

Il reste à trouver l'intervalle ouvert J de définition de la solution y. Nous avons les cas suivants :
Cas 1) : $C > 0$.
Dans ce cas comme la solution (8.13) *est définie pour tout* $t \in \mathbb{R}$ *on a* $J = \mathbb{R}$.
Cas 2) : $C = 0$.
Dans ce cas comme la solution (8.13) *est définie pour tout* $t \neq 0$ *on a* $J =]-\infty,\ 0[$ *ou* $J =]0,\ +\infty[$.
Cas 3) : $C < 0$.
Comme dans ce cas la solution (8.13) *est définie pour tout* $t \neq \pm\sqrt{-2C}$ *on a* $J =]-\infty,\ -\sqrt{-2C}[$ *ou* $J =]-\sqrt{-2C},\ \sqrt{-2C}[$ *ou* $J =]\sqrt{-2C},\ +\infty[$.

D'autre part, considérons l'équation différentielle (8.11) *avec condition initiale*

$$y(1) = 2 \tag{8.14}$$

(on a donc un problème de Cauchy; ici $t_0 = 1$ *et* $y^0 = 2$*). Nous considérons d'abord la solution constante* $y(t) = 0 \quad \forall t \in \mathbb{R}$ *et nous observons que cette solution ne peut pas satisfaire* (8.14).
Regardons alors les solutions données par (8.13). *On est amenés à résoudre l'équation avec inconnue* C *:*

$$-\frac{2}{1+2C} = 2$$

ce qui donne comme seule solution $C = -1$. *Comme* $C < 0$ *on est dans le* **Cas 3)** *et il y a 3 choix possibles pour l'intervalle* J *; on prendra celui qui contient* $t_0 = 1$, *donc* $J =]-\sqrt{2},\ \sqrt{2}[$. *Donc en conclusion la solution du problème de Cauchy* (8.11) *et* (8.14) *est donnée par* $y :]-\sqrt{2},\ \sqrt{2}[\to \mathbb{R}$ *avec*

$$y(t) = \frac{2}{2-t^2}, \quad \forall\, t \in]-\sqrt{2},\ \sqrt{2}[.$$

8.1.3 Résultat téoriques d'existence et unicité pour le problème de Cauchy

Nous admettons le résultat suivant d'existence et unicité pour le problème de Cauchy (8.1) et (8.3) :

Théorème 8.2. *(Théorème de Cauchy-Lipschitz local). Supposons que* f *est continue et qu'elle est "localement lipschitzienne en* x*", c'est à dire, il existe un voisinage* I_1 *de* t_0 *avec* $I_1 \subset I$, *un voisinage* U_1 *de* y^0 *avec* $U_1 \subset U$ *et aussi une constante* $L \geq 0$ *tels que*

$$\|f(t,u) - f(t,v)\| \leq L\,\|u - v\|, \quad \forall\, t \in I_1,\ \ \forall\, u, v \in U_1$$

(n'importe quelle norme $\|\cdot\|$ *sur* $\mathbb{R}^d$ *peut être utilisée).*
Alors il existe $r > 0$ *avec* $]t_0 - r,\ t_0 + r[\subset I_1$ *et* $y :]t_0 - r,\ t_0 + r[\to\ U$ *avec* $y \in C^1(]t_0 - r,\ t_0 + r[)$, *solution de* (8.1) (8.3) *; en plus cette solution* y *est unique sur l'intervalle* $]t_0 - r,\ t_0 + r[$. *En plus pour tout* $m \in \mathbb{N}^*$ *si* f *est de classe* C^m *alors* y *est de classe* C^{m+1}.

Remarque 8.5. *1. Ce théorème nous donne seulement une solution "locale", c'est à dire, définie sur un voisinage* $]t_0-r,\, t_0+r[$ *de* t_0 *; ce résultat ne permet pas de savoir si cette solution peut se "prolonger" en dehors de ce voisinage, éventuellement sur tout le ouvert* I *pour obtenir une solution globale.*

2. Les hypothèses de ce théorème sont satisfaites dans le cas (très fréquent) où $f \in C^1$ *; c'est une conséquence du (TAF).*

Exemple 8.2. *Considérons le problème de Cauchy vu dans l'Exemple 8.1 ; ici* $d = 1$, $I = U = \mathbb{R}$, $f(t,x) = t\,x^2 \;\; \forall\, (t,x) \in \mathbb{R}^2$, $t_0 = 1$ *et* $y^0 = 2$. *Comme* $f \in C^1$ *le théorème 8.2 s'applique ce qui nous donne l'existence d'un* $r > 0$ *et d'une solution unique* y *définie au moins sur l'intervalle* $]1-r,\, 1+r[$ *(ceci est confirmée par le calcul de la solution "à la main" fait dans l'Exemple 8.1. Mais dans ce cas cette solution ne se prolonges pas à une solution globale (c'est à dire, définie sur* $\mathbb{R}$*).*

Une question importante qui se pose pour le problème de Cauchy (8.1) et (8.3) donné est : a-t-on une solution globale ?
Il y a deux types de résultats connus qui permettent de répondre positivement à cette question :

- Un résultat appellé **théorème des bouts** qui permet de prolonger la solution locale donnée par le Théorème 8.2 à une solution globale. C'est basé sur le fait de pouvoir obtenir des estimations appropriées pour la solution locale ; ce résultat dépasse le cadre de ce cours.
- On a un résultat qui donne directement une solution globale ; c'est le théorème suivant que nous admettons sans preuve.

Théorème 8.3. *(Théorème de Cauchy-Lipschitz global). Supposons qu'on a :*
$U = \mathbb{R}^d$, f *est une fonction continue et en plus* f *est une fonction globalement lipschitzienne en* x*, c'est à dire : il existe* $L \geq 0$ *telle que*

$$\|f(t,u) - f(t,v)\| \leq L\,\|u - v\|, \quad \forall\, t \in I, \;\; \forall\, u, v \in \mathbb{R}^d. \tag{8.15}$$

Alors il existe $y : I \to \mathbb{R}^d$ *avec* $y \in C^1(I)$ *solution de* (8.1) *et* (8.3) *et cette solution est unique sur* I*. En plus pour tout* $m \in \mathbb{N}^*$ *si* f *est de classe* C^m *alors* y *est de classe* C^{m+1}*.*

Remarque 8.6. *L'hypothèse "*f *est globalement lipschitzienne en* x*" est beaucoup plus restrictive que l'hypothèse "*f *est localement lipschitzienne en* x*" du Théorème 8.2. Le fait d'avoir* $f \in C^1$ *ne suffit pas pour déduire que* f *est globalement lipschitzienne en* x*. Par exemple la fonction* f *de l'Exemple 8.1 est de classe* C^1 *mais elle n'est pas globalement lipschitzienne en* x *(pour voir cela observer que*

$$\frac{|f(1,n+1) - f(1,n)|}{n+1-n} = (n+1)^2 - n^2 = 2n+1 \to +\infty, \quad si \quad n \to +\infty.)$$

Donc le Théorème 8.3 ne s'applique pas pour cet exemple et cela est cohérent avec le fait qu'il n'y a pas de solution globale pour le problème de Cauchy (8.11) *et* (8.14)*.*

Exemple 8.3. *Considérons le système EDO linéaire avec coefficients constants qui est le problème de Cauchy* (8.5) *et* (8.3) *avec* $U = \mathbb{R}^d$ *et* $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ *donnée par*

$$f(t,x) = Ax + b(t), \quad \forall\, (t,x) \in I \times \mathbb{R}^d$$

avec $A \in \mathcal{M}^d(\mathbb{R})$ *matrice constante et* $b = I \to \mathbb{R}^d$ *fonction continue.*
Il est clair que f *est une fonction continue. D'autre part, pour tous* $t \in I$ *et* $u, v \in \mathbb{R}^d$ *on a*

$$\|f(t,u) - f(t,v)\| = \|Au + b(t) - Av - b(t)\| = \|A(u-v)\| \leq \|A\|\,\|u - v\|$$

où la norme utilisée pour écrire $\|A\|$ *est la norme matricielle subordonnée à la norme vectorielle arbitraire* $\|\cdot\|$. *Ceci nous dit que l'hypothè* (8.15) *est satisfaite avec* $L = \|A\|$. *Alors le Théorème 8.3 s'applique et on a l'existence et l'unicité globale d'une solution pour ce problème de Cauchy.*

8.2 Résolution numérique des systèmes EDO

8.2.1 Généralités

Dans la plupart des cas il n'est pas possible d'obtenir des solution "à la main" (ou des solutions explicites) pour les EDO ; on cherche alors à donner une approximation numérique pour la solution du problème qui nous intéresse.
Nous considérons dans cette section le problème de Cauchy (8.1) et (8.3) et nous supposons que ce problème admet une solution $y : J \to U$, $y \in C^1(J)$ avec $J \subset I$ intervalle ouvert. Soit $T > 0$ et supposons que la solution $y(t)$ est définie sur l'intervalle $[t_0, t_0 + T]$ (donc on a forcément l'inclusion $[t_0, t_0 + T] \subset J$). Pour obtenir une approximation numérique de y nous faisons une discrétisation de l'intervalle $[t_0, t_0 + T]$; nous prennons $N \in \mathbb{N}^*$ et $t_0, t_1, \cdots t_N$ avec

$$t_0 < t_1 < \ \cdots \ < t_N = t_0 + T$$

et tels que si nous notons $h_n = t_{n+1} - t_n$ pour tout $n \in [[0, N-1]]$ nous avons que $\max_{n \in [[0,N-1]]} h_n$ est "petite" (donc forcément N est "grand").
Le plus souvent on prend h_n constant, c'est à dire on pose $h = \frac{T}{N}$ et on pose

$$t_n = t_0 + n\,h, \quad \forall\, n \in [[0, N]]$$

donc $h_n = h$ pour tout n.
On construira dans chaque point t_n de la discrétisation de l'intervalle $[t_0, t_0 + T]$ une approximation de $y(t_n)$ que nous notons par $y^{(n)} \in U$. On va écrire $y^{(n)} \sim y(t_n)$. Comme $y(t_0) = y^0$ connue alors il est naturel de poser

$$y^{(0)} = y^0.$$

On va procéder par récurrence sur n : on suppose $y^{(n)}$ connue et on va construire $y^{(n+1)}$, ceci pour tout $n \in [[0, N-1]]$. Ce type de méthodes s'appellent **méthodes à un pas** ; il y a aussi des méthodes à deux ou plusieurs pas qui ne seront pas abordées dans ce cours.

8.2.2 Les méthodes explicites usuelles

La méthodes d'Euler explicite

Nous décrivons comment construire $y^{(1)}$ à partir de $y^{(0)} = y^0$ connue. Comme t_1 est "proche" de t_0 alors pour tout $i \in [[1, d]]$ $y_i(t_1)$ sera "proche" de $z_i(t_1)$ où $z_i(t)$ est la tangente au point $(t_0, y_i(t_0))$ au graph de la fonction $y_i(t)$. On a

$$z_i(t_1) = y_i(t_0) + y_i'(t_0)(t_1 - t_0)$$

et on va prendre $y_i^{(1)} = z_i(t_1)$. Comme $y_i'(t_0) = f_i(t_0, y^{(0)})$ on a alors

$$y_i^{(1)} = y_i^{(0)} + h_0 f_i(t_0, y^{(0)})$$

et comme cette égalité est vraie pour tout $i \in [[1, d]]$ on obtient l'égalité vectorielle

$$y^{(1)} = y^{(0)} + h_0 f(t_0, y^{(0)})$$

On va procéder de la même manière pour construire $y^{(2)}$ à partir de $y^{(1)}$, ce qui nous donne

$$y^{(2)} = y^{(1)} + h_1 f(t_1, y^{(1)})$$

et ainsi de suite ..
Nous avons en général la relation suivante de récurrence qui donne $y^{(n+1)}$ en fonction de $y^{(n)}$:

$$\begin{cases} y^{(n+1)} = y^{(n)} + h_n f(t_n, y^{(n)}), \quad n \in [[0, N-1]] \\ y^{(0)} = y^0 \quad \text{donnée.} \end{cases}$$

C'est la **méthode d'Euler explicite**.

La méthode de Taylor

C'est une méthode qui utilise le développement de Taylor à l'ordre 2. Nous supposons ici $f \in C^2$ donc la solution exacte y est dans C^3. Pour tout $n \in [[0,\ N-1]]$ la solution y satisfait

$$y(t_{n+1}) = y(t_n) + h_n y'(t_n) + \frac{1}{2} h_n^2 y''(t_n) + O(h_n^3). \tag{8.16}$$

Nous avons pour tout $t \in I$:

$$y'(t) = f(t, y(t))$$

(donc $y'(t_n) = f(t_n, y(t_n))$) et

$$y''(t) = \frac{d}{dt} y'(t) = \frac{d}{dt} f(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + J_{f,x}(t, y(t))\, y'(t)$$

où $J_{f,x}$ est la matrice Jacobienne de f par rapport à x, ce qui donne pour tout $n \in [[0, N-1]]$:

$$y''(t_n) = \frac{\partial f}{\partial t}(t_n, y(t_n)) + J_{f,x}(t_n, y(t_n))\, f(t_n, y(t_n))$$

En utilisant (8.16) nous obtenons :

$$y(t_{n+1}) = y(t_n) + h_n f(t_n, y(t_n)) + \frac{h_n^2}{2}\left[\frac{\partial f}{\partial t}(t_n, y(t_n)) + J_{f,x}(t_n, y(t_n))\, f(t_n, y(t_n))\right] + O(h_n^3).$$

En approchant $y(t_n)$ par $y^{(n)}$ nous obtenons la **méthode de Taylor** qui s'écrit

$$\begin{cases} y^{(n+1)} = y^{(n)} + h_n f(t_n, y^{(n)}) + \dfrac{h_n^2}{2}\left[\dfrac{\partial f}{\partial t}(t_n, y^{(n)}) + J_{f,x}(t_n, y^{(n)})\, f(t_n, y^{(n)})\right], \quad n \in [[0, N-1]] \\ \quad y^{(0)} = y^0 \quad \text{donnée} \end{cases}$$

Exemple 8.4. *Si* $d = 1$, $I = U = \mathbb{R}$ *et* $f(t,x) = tx \quad \forall\, (t,x) \in \mathbb{R}^2$ *alors comme* $\frac{\partial f}{\partial t}(t,x) = x$ *et* $J_{f,x}(t,x) = \frac{\partial f}{\partial x}(t,x) = t$, *la méthode de Taylor s'écrit*

$$y^{(n+1)} = y^{(n)} + h_n t_n y^{(n)} + \frac{h_n^2}{2}\left[y^{(n)} + t_n^2 y^{(n)}\right], \quad n \in [[0, N-1]].$$

Les méthodes de Runge-Kutta explicites

Nous présentons maintenant une autre manière de construire un schéma numérique. En intégrant l'EDO (8.1) pour t de t_n à t_{n+1} on obtient

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt, \quad \forall\, n \in [[0, N-1]]. \tag{8.17}$$

L'idée est d'approcher l'intégrale qui apparait dans cette égalité en utilisant des méthodes d'approximations des intégrales, méthodes vues au Chapitre 6. En remplaçant ensuite $y(t_k)$ par $y^{(k)}$ on obtient un schéma numérique.
Par exemple on peut utiliser une formule de rectange qui utilise le point de gauche de l'intervalle, ce qui donne

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \sim (t_{n+1} - t_n) f(t_n,\, y(t_n))$$

et nous retrouvons le schéma d'Euler explicite

$$y^{(n+1)} = y^{(n)} + (t_{n+1} - t_n) f(t_n,\, y^{(n)}).$$

Nous utilisons maintenant une méthode de rectangle qui utilise le point de milieu de l'intervalle, ce qui donne

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \sim (t_{n+1} - t_n) f(t_{n+1/2},\, y(t_{n+1/2}))$$

où nous notons $t_{n+1/2} = t_n + \frac{h_n}{2}$, le milieu de l'intervalle $[t_n,\, t_{n+1}]$.
Ensuite $y(t_{n+1/2})$ sera approché en utilisant un schéma d'Euler explicite sur l'intervalle $[t_n,\, t_{n+1/2}]$:

$$y(t_{n+1/2}) \sim y(t_n) + \frac{h_n}{2} f(t_n; y(t_n)).$$

Nous avons alors l'approximation

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \sim h_n f\left(t_n + \frac{h_n}{2},\, y(t_n) + \frac{h_n}{2} f(t_n,\, y(t_n))\right)$$

qui nous donne le schéma suivant :

$$y^{(n+1)} = y^{(n)} + h_n f\left(t_n + \frac{h_n}{2},\, y^{(n)} + \frac{h_n}{2} f(t_n,\, y^{(n)})\right), \quad \forall\, n \in [[0, N-1]] \tag{8.18}$$

avec toujours $y^{(0)} = y^0$ donnée.
Il est plus pratique d'écrire ce schéma en deux étapes (ou "étages") :

$$\begin{cases} k_1 = f(t_n,\, y^{(n)}) \\ k_2 = f\left(t_n + \frac{h_n}{2},\, y^{(n)} + \frac{h_n}{2} k_1\right) \\ y^{(n+1)} = y^{(n)} + h_n k_2 \end{cases} \tag{8.19}$$

Le schéma (8.18) ou (8.19) s'appelle schéma de **Runge-Kutta explicit d'ordre 2 (ou à 2 étages)**.

Dans la suite nous allons généraliser ce schéma. Considérons $n \in [[0, N-1]]$ fixé. Nous introduisons une sous-division de l'intervalle $[t_n,\, t_{n+1}]$; pour cela nous considérons $s \in \mathbb{N}$ avec $s \geq 2$ et des nombres réels $c_1, c_2, \cdots c_s$ avec

$$0 = c_1 \leq c_2 \leq c_3 \leq \cdots \leq c_s \leq 1.$$

Nous posons

$$t_{n,i} = t_n + c_i h_n, \quad \forall\, i \in [[1, s]]$$

donc nous avons $t_n = t_{n,1} \leq t_{n,2} \leq t_{n,3} \leq t_{n,s} \leq t_{n+1}$.
En intégrant l'EDO (8.1) pour t de t_n à $t_{n,i}$ on obtient

$$y(t_{n,i}) = y(t_n) + \int_{t_n}^{t_{n,i}} f(t, y(t))\, dt, \quad \forall\, i \in [[2, s]]. \tag{8.20}$$

Nous allons approcher l'intégrale ci-dessus en utilisant comme noeuds les points $t_{n,1},\, t_{n,2}, \cdots t_{n,i-1}$ et comme poids des nombres de la forme $h_n a_{i,1},\, h_n a_{i,2}, \cdots h_n a_{i,i-1}$ où $(a_{i,j})_{j \in [[1,\, i-1]]}$ sont des nombres à fixer. Nous avons alors

$$y(t_{n,i}) \sim y(t_n) + h_n \sum_{j=1}^{i-1} a_{i,j} f(t_{n,j},\, y(t_{n,j})) \quad \forall\, i \in [[2, s]].$$

Nous notons par $y^{(n,i)}$ une approximation de $y(t_{n,i})$ et l'égalité précédente nous suggère la définition suivante (récursive en i) pour $y^{(n,i)}$:

$$y^{(n,1)} = y^{(n)}$$

et

$$y^{(n,i)} = y^{(n)} + h_n \sum_{j=1}^{i-1} a_{i,j} f(t_{n,j},\, y^{(n,j)}), \quad \forall\, i \in [[2, s]]. \tag{8.21}$$

Notons maintenant par k_i une approximation de $f(t_{n,i},\, y(t_{n,i}))$ donnée par

$$k_i = f(t_{n,i},\, y^{(n,i)}), \quad \forall\, i \in [[1, s]].$$

Nous avons

$$k_1 = f(t_n, y^{(n)})$$

et grâce à (8.21) nous obtenons la relation de récurrence en i :

$$k_i = f(t_{n,i},\, y^{(n)} + h_n \sum_{j=1}^{i-1} a_{i,j} k_j), \quad \forall\, i \in [[2, s]].$$

Finalement, nous utilisons (8.17) et nous approchons l'intégrale qui apparait dans cette égalité en utilisant comme noeuds les points $t_{n,1},\, t_{n,2}, \cdots t_{n,s}$ et comme poids des nombres de la forme $h_n b_1,\, h_n b_2, \cdots h_n b_s$ avec $b_1, \cdots b_s$ à fixer. Ceci nous donne le schéma suivant appellé **schéma de Runge-Kutta explicite à s étages :**

$$\begin{cases} k_1 = f(t_n,\, y^{(n)}) \\ k_2 = f\left(t_n + c_2 h_n,\, y^{(n)} + h_n a_{2,1} k_1\right) \\ k_3 = f\left(t_n + c_3 h_n,\, y^{(n)} + h_n (a_{3,1} k_1 + a_{3,2} k_2)\right) \\ \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ k_s = f\left(t_n + c_s h_n,\, y^{(n)} + h_n (a_{s,1} k_1 + a_{s,2} k_2 + \cdots + a_{s,s-1} k_{s-1})\right) \\ y^{(n+1)} = y^{(n)} + h_n (b_1 k_1 + b_2 k_2 + \cdots + b_s k_s) \end{cases} \tag{8.22}$$

avec les données : $s \in \mathbb{N},\ s \geq 2$ et

$$(c_i)_{i \in [[2,s]]}, \quad (b_i)_{i \in [[1,s]]}, \quad (a_{i,j})_{1 \leq j < i \leq s}$$

des nombres réels tels que $0 \leq c_2 \leq \cdots \leq c_s \leq 1$. On identifie un tel schéma de Runge-Kutta par la donnée de tous ces nombres qui sont les paramètres du schéma. On peut mettre ces donnés (donc le schéma de Runge-Kutta respectif) dans un tableau de la forme

Par exemple le schéma de Runge-Kutta explicit d'ordre 2 (8.19) est un schéma de Runge-Kutta explicit à 2 étages dont le tableau est le suivant :

Ici on a $s = 2, c_2 = \frac{1}{2}, b_1 = 0, b_2 = 1$ et $a_{21} = \frac{1}{2}$.

Un autre exemple beaucoup utilisé dans la pratique est le schéma dont le tableau est le suivant :

On a donc

$$s = 4, \quad c_2 = c_3 = \frac{1}{2},\ c_4 = 1, \quad b_1 = b_4 = \frac{1}{6},\ b_2 = b_3 = \frac{1}{3}$$

et

$$a_{2,1} = a_{3,2} = \frac{1}{2},\ \ a_{3,1} = a_{4,1} = a_{4,2} = 0,\ \ a_{4,3} = 1.$$

Le schéma s'écrit alors

$$\begin{cases} k_1 = f(t_n,\, y^{(n)}) \\ k_2 = f\left(t_n + \frac{1}{2}h_n,\, y^{(n)} + \frac{1}{2}h_n k_1\right) \\ k_3 = f\left(t_n + \frac{1}{2}h_n,\, y^{(n)} + \frac{1}{2}h_n k_2\right) \\ k_4 = f\left(t_n + h_n,\, y^{(n)} + h_n k_3\right) \\ y^{(n+1)} = y^{(n)} + h_n\left(\frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4\right) \end{cases}$$

et il est connu sous le nom de **schéma de Runge-Kutta explicite d'ordre 4**.

8.2.3 Méthodes implicites

Remarquons d'abord qu'il y a une autre manière d'obtenir le schéma d'Euler explicite. Considérons de nouveau l'égalité (8.17) obtenue en intégrant (8.1) en t de t_n à t_{n+1}. On appproche ensuite $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ avec une méthode de rectangle qui utilise le point de gauche

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \sim (t_{n+1} - t_n) f(t_n, y(t_n)).$$

Le schéma d'Euler explicite s'obtient alors en remplaçant $y(t_k)$ par $y^{(k)}$, ce qui donne $y^{(n+1)} = y^{(n)} + h_n f(t_n,\, y^n)$.

Si maintenant on appproche $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ avec une méthode de rectangle qui utilise le point de droite :

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \sim (t_{n+1} - t_n) f(t_{n+1}, y(t_{n+1}))$$

et on remplace encore $y(t_k)$ par $y^{(k)}$ on obtient ce qu'on appelle le **schéma d'Euler implicit** qui s'écrit

$$y^{(n+1)} = y^{(n)} + h_n f(t_{n+1},\, y^{(n+1)}), \quad \forall\, n \in [[0, N-1]]. \tag{8.23}$$

avec toujours $y^{(0)} = y^0$ donnée.
Ici nous supposons encore que $y^{(n)}$ est connue et nous voulons trouver $y^{(n+1)}$. Le calcul de $y^{(n+1)}$ à partir de l'égalité (8.23) n'est pas immédiat, car il faut résoudre une équation algébrique avec inconnue $y^{(n+1)}$, en fait un système algébrique non-linéaire avec d équations et d inconnues.
On peut utiliser pour cela des méthodes numériques apprises au Chapitre 7, par exemple la méthode des approximations successives.

Il existe encore des méthodes de Runge-Kutta implicites, mais elles dépassent le cadre de ce cours.

8.2.4 Estimations d'erreur pour les schémas explicites

On va supposer dans cette subsection que $U = \mathbb{R}^d$ et que f est une fonction **continue**. On va considérer ici des schémas numériques générales de la forme

$$y^{(n+1)} = y^{(n)} + h_n \Phi(t_n, y^{(n)}, h_n), \quad n \in [[0, N-1]] \tag{8.24}$$

avec toujours

$$y^{(0)} = y^0 \quad \text{donnée.} \tag{8.25}$$

Ici la fonction Φ est telle que $\Phi : [t_0,\, t_0 + T] \times \mathbb{R}^d \times [0, h^*] \to \mathbb{R}^d$ avec $0 < h^* < T$ et h^* "assez petit".

Remarque 8.7. *Tous les schémas explicites vu dans la partie 8.2.2 sont des cas particuliers de* (8.24). *Par exemple :*

- *Le schéma d'Euler explicite correspond au* $\Phi(t, x, h) = f(t, x)$
- *Pour le schéma de Runge-Kutta explicite d'ordre 2 on a* $\Phi(t, x, h) = f(t + \frac{h}{2},\, x + \frac{h}{2} f(t, x))$.

Définition 8.1. *On appelle* **erreur de consistance** *du schéma* (8.24) *pour approcher* (8.1) *(on peut préciser : au temps t pour une solution y et pour le pas de temps h) l'expression*

$$R(t, y, h) = \frac{1}{h}[y(t + h) - y(t)] - \Phi(t, y(t), h)$$

avec $0 < h \leq h^*$, $t_0 \leq t \leq t_0 + T - h$ *et* y *solution de* (8.1).

Remarque 8.8. *On a pour tout* $n \in [[0, N-1]]$

$$R(t_n, y, h_n) = \frac{1}{h_n}\{y(t_n + h_n) - [y(t_n) + h_n\Phi(t_n, y(t_n), h_n)]\} = \frac{1}{h_n}[y(t_{n+1}) - z^{(n+1)}]$$

où $z^{(n+1)}$ *serait l'approximation en* t_{n+1} *si on utilisait la solution exacte en* t_n *Donc pour obtenir l'erreur de consistance on divise par* h_n *l'erreur qu'on ferait au temps* t_{n+1} *si on avait la solution exacte en* t_n*.*

Définition 8.2. *On dit que le schéma* (8.24) *pour approcher* (8.1) *est*

— **consistant** *si pour toute solution* y *de* (8.1) *on a*

$$\sup_{t\in[t_0,t_0+T[} \|R(t, y, h)\| \to 0 \quad \textit{pour} \quad h \to 0.$$

— **consistant à l'ordre au moins** p *avec* $p \in \mathbb{N}^*$*, si pour toute solution* y *de* (8.1) *il existe une constante* $C \geq 0$ *telle que pour tous* t, h *avec* $0 < h \leq h^*$*,* $t_0 \leq t \leq t_0 + T - h$ *on a*

$$\|R(t, y, h)\| \leq Ch^p$$

(ici la constante C *est indépendante de* t *et* h*).*

Par exemple, pour le schéma d'Euler explicite nous avons

$$R(t, y, h) = \frac{1}{h}[y(t + h) - y(t)] - \Phi(t, y(t), h) = \frac{1}{h}[y(t + h) - y(t)] - f(t, y(t)).$$

Comme y est une solution de (8.1) on a

$$R(t, y, h) = \frac{1}{h}[y(t + h) - y(t)] - y'(t)$$

et en utilisant (TAF) on obtient

$$R(t, y, h) = y'(t + \theta(t)h) - y'(t) \tag{8.26}$$

avec $\theta(t) \in]0, 1[$.
Comme y est de classe C^1 sur le compact $[t_0, t_0 + T]$ alors y' est uniformément continue ; ceci nous donne

$$\sup_{t\in[t_0,t_0+T[} \|R(t, y, h)\| \to 0 \quad \text{pour} \quad h \to 0$$

donc le schéma est consistent.
D'autre part, si on suppose que f est de classe C^1 alors y est de classe C^2 sur $[t_0, t_0 + T]$ et on déduit de (8.26) :

$$\|R(t, y, h)\| \leq \sup_{t\in[t_0,t_0+T]} \|y''(t)\|\, h.$$

En conclusion, si f est de classe C^1 alors le schéma d'Euler explicte pour approcher (8.1) est consistant à l'ordre au moins 1.

Nous avons le résultat suivant :

Proposition 8.1. *(condition suffisante de consistance)*
Supposons que Φ *est une fonction* **continue** *et en plus*

$$\Phi(t,x,0) = f(t,x), \quad \forall\, t \in [t_0, t_0+T], \quad \forall\, x \in \mathbb{R}^d. \tag{8.27}$$

Alors le schéma (8.24) *pour approcher* (8.1) *est consistant.*

Démonstration. Soit y une solution de (8.1) et $t \in [t_0, t_0+T[$. Alors pour $h > 0$ assez petit et pour tout $i \in [[1,d]]$ on a en appliquant (TAF) :

$$R_i(t,y,h) = y_i'(t+\theta_i h) - \Phi_i(t, y(t), h)$$

avec $\theta_i \in]0,1[$. Ceci nous donne

$$R_i(t,y,h) = R_{i1}(t,y,h) + R_{i2}(t,y,h) \tag{8.28}$$

avec

$$R_{i1}(t,y,h) = y_i'(t+\theta_i h) - y_i'(t)$$

et

$$R_{i2}(t,y,h) = y_i'(t) - \Phi_i(t, y(t), h).$$

Nous avons

$$\sup_{t\in[t_0,t_0+T[} |R_{i1}(t,y,h)| \to 0 \quad \text{pour} \quad h \to 0 \tag{8.29}$$

grâce au fait que y_i' est une fonction uniformément continue sur $[t_0, t_0+T]$.
D'autre part

$$R_{i2}(t,y,h) = y_i'(t) - \Phi_i(t,y(t),h) = f_i(t,y(t)) - \Phi_i(t,y(t),h) = \Phi_i(t,y(t),0) - \Phi_i(t,y(t),h).$$

Nous avons

$$\sup_{t\in[t_0,t_0+T[} |R_{i2}(t,y,h)| \to 0 \quad \text{pour} \quad h \to 0 \tag{8.30}$$

grâce au fait que la fonction $(t,h) \in [t_0, t_0+T] \times [0, h^*] \;\mapsto\; \Phi_i(t, y(t), h)$ est une fonction uniformément continue.
En utilisant (8.28), (8.29) et (8.30) on obtient le résultat. □

Exemple 8.5. - *Pour le schéma d'Euler explicite on a*

$$\Phi(t,x,0) = f(t,x), \quad \forall\, (t,x) \in [t_0,\, t_0+T] \,\times \mathbb{R}^d$$

ce qui nous donne la consistance.
- Pour le schéma de Runge-Kutta explicite d'ordre 2 on a

$$\Phi(t,x,0) = f(t+0, x+0) = f(t,x), \quad \forall\, (t,x) \in [t_0,\, t_0+T] \,\times \mathbb{R}^d$$

ce qui nous donne la consistance de ce schéma aussi.

Définition 8.3. *On dit que le schéma* (8.24) *est* **stable** *si "deux schémas voisins donnent des résultats voisins", c'est à dire : si* $y^{(0)},\ y^{(1)}, \cdots y^{(N)}$ *sont données par* (8.24) *et* $z^{(0)},\ z^{(1)}, \cdots z^{(N)}$ *sont données par la relation de récurrence :*

$$z^{(n+1)} = z^{(n)} + h_n \Phi(t_n, z^{(n)}, h_n) + \epsilon_n, \quad n \in [[0, N-1]] \tag{8.31}$$

avec

$$z^{(0)} = z^0 \in \mathbb{R}^d \quad \textit{donnée}$$

alors il existe des constantes $C \geq 0$ *et* h_* *(indépendantes de* N*) avec* $0 < h_* \leq h^*$ *, tels que si* $0 \leq h_n \leq h_*\ \forall\, n \in [[0, N-1]]$ *alors on a*

$$\|y^{(n)} - z^{(n)}\| \ \leq\ C \left[\|y^{(0)} - z^{(0)}\| + \sum_{n=0}^{N-1} \|\epsilon_n\| \right], \quad \forall\, n \in [[0, N]].$$

Supposons dans la suite que les points t_n sont équidistantes, c'est à dire

$$h_n = h = \frac{T}{N}, \quad \forall\, n \in [[0, N-1]].$$

Nous avons

Proposition 8.2. *(condition suffisante de stabilité) Supposons qu'il existe des constantes* $L \geq 0$ *et* h_* *avec* $0 < h_* \leq h^*$ *tels que*

$$\|\Phi(t, u, h) - \Phi(t, v, h)\| \leq L\, \|u - v\|, \quad \forall\, t \in [t_0, t_0 + T],\ h \in [0, h_*],\ u, v \in \mathbb{R}^d. \tag{8.32}$$

Alors le schéma (8.24) *est stable.*

Démonstration. En faisant le différence entre (8.24) et (8.31) et en utilisant l'hypothèse (8.32) on obtient

$$\|y^{(n+1)} - z^{(n+1)}\| \leq \|y^{(n)} - z^{(n)}\| + hL\|y^{(n)} - z^{(n)}\| + \|\epsilon_n\|.$$

En notant $e^{(n)} = y^{(n)} - z^{(n)}$ on obtient

$$\|e^{(n+1)}\| \leq (1 + hL)\|e^{(n)}\| + \|\epsilon_n\| \tag{8.33}$$

et ceci pour tout $n \in [[0, N-1]]$. Nous avons donc pour un $n \in [[1, N]]$ fixé

$$\|e^{(n)}\| \leq (1 + hL)\|e^{(n-1)}\| + \|\epsilon_{n-1}\| \tag{8.34}$$

ensuite

$$\|e^{(n-1)}\| \leq (1 + hL)\|e^{(n-2)}\| + \|\epsilon_{n-2}\| \tag{8.35}$$

etc ... jusqu'à

$$\|e^{(1)}\| \leq (1 + hL)\|e^{(0)}\| + \|\epsilon_0\| \tag{8.36}$$

On fait la somme de ces inégalitées après avoir multiplié (8.35) par $1+hL$, l'inégalité suivante par $(1+hL)^2$ etc ... et finalement (8.36) par $(1+hL)^{n-1}$. On obtient après plusieurs simplifications :

$$\|e^{(n)}\| \leq (1+hL)^n \|e^{(0)}\| + \sum_{k=0}^{n-1} (1+hL)^{n-1-k} \|\epsilon_k\|. \tag{8.37}$$

D'autre part on a pour tout $j \in [[0, N]]$

$$(1+Lh)^j \leq \left(1 + L\frac{T}{N}\right)^N \leq e^{LT}.$$

(on utilise le fait que $1 + L\frac{T}{N} \leq \exp(L\frac{T}{N})$ donc $\left(1+L\frac{T}{N}\right)^N \leq (\exp(L\frac{T}{N}))^N = e^{LT}$).
On obtient donc de (8.37)

$$\|e^{(n)}\| \leq e^{LT}\left(\|e^{(0)}\| + \sum_{k=0}^{n-1}\|\epsilon_k\|\right) \leq e^{LT}\left(\|e^{(0)}\| + \sum_{k=0}^{N-1}\|\epsilon_k\|.\right)$$

Ceci est vrai pour tout $n \in [[1, N]]$ et de manière évidente pour $n = 0$ aussi, ce qui donne le résultat. □

Exemple 8.6. *Si f satisfait les hypothèses du Théorème 8.3 (Théorème de Cauchy-Lipschitz global) alors le schéma de Euler explicite est stable, car*

$$\|\Phi(t,u,h) - \Phi(t,v,h)\| = \|f(t,u) - f(t,v)\| \leq L\,\|u-v\|, \quad \forall\, t \in [t_0, t_0+T],\ u, v \in \mathbb{R}^d.$$

Dans la suite nous supposons l'existence et l'unicité d'une solution y du problème de Cauchy (8.1), (8.3), définie sur l'intervalle $[t_0, t_0+T]$.

Définition 8.4. *1. On dit que le schéma général* (8.24), (8.25) *est* **convergent** *pour le problème de Cauchy* (8.1), (8.3) *si*

$$\max_{n\in[[0,N]]} \|y(t_n) - y^{(n)}\| \to 0 \quad \textit{pour} \quad N \to +\infty \quad (\textit{ou équivalent pour } h \to 0).$$

2. On dit que le schéma général (8.24), (8.25) *est* **convergent à l'ordre au moins** *p, pour le problème de Cauchy* (8.1), (8.3), *avec $p \in \mathbb{N}^*$, si*

$$\max_{n\in[[0,N]]} \|y(t_n) - y^{(n)}\| \leq C\,h^p.$$

où $C \geq 0$ est une constante indépendante de N (ou de h).

Le résultat principal de cette partie est le suivant :

Théorème 8.4. *Si le schéma numérique* (8.24), (8.25) *pour le problème de Cauchy* (8.1), (8.3) *est consistant et stable, alors il est convergent.*
Si en plus il est consistant à l'ordre au moins p avec $p \in \mathbb{N}^*$ alors il est convergent à l'ordre au moins p.

Démonstration. (rappel : ici $h_n = h = \text{constant}$). De la définition de l'erreur de consistance R nous avons

$$\frac{y(t_{n+1}) - y(t_n)}{h} = R(t_n, y, h) + \Phi(t_n, x(t_n), h)$$

ce qui donne

$$y(t_{n+1}) = y(t_n) + h\,\Phi(t_n, x(t_n), h) + h\,R(t_n, y, h).$$

Nous utilisons la Définition 8.3 de la stabilité avec $z^{(n)} = y(t_n)$ et $\epsilon_n = hR(t_n, y, h)$. On déduit qu'il existe une constante $C \geq 0$ indépendante de N telle que

$$\|y(t_n) - y^{(n)}\| \leq C \left[\|y^0 - y^{(0)}\| + \sum_{n=0}^{N-1} h \|R(t_n, y, h)\| \right]$$

Comme $y^0 = y^{(0)}$ et $\sum_{n=0}^{N-1} h = Nh = T$ nous obtenons

$$\|y(t_n) - y^{(n)}\| \leq C\,T \max_{n \in [[0, N-1]]} \|R(t_n, y, h)\| \leq C\,T \sup_{t \in [t_0, t_0+T[} \|R(t, y, h)\|. \tag{8.38}$$

Comme le schéma est consistant (respectivement consistant à l'ordre p), nous obtenons les deux résultats souhaités en utilisant la Définition 8.2. □

Bibliographie

[1] P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson 1988

[2] J.P. Demailly, *Analyse Numérique et Equations différentielles*, Collection Grenoble Sciences 1996

[3] F. Filbet, *Analyse Numérique, modélisation, algorithme et étude mathématique*, Dunod 2009

[4] A. Fortin, *Analyse Numérique pour Ingénieurs*, Presses Internationales Polytechnique, 2001

[5] P. Lascaux, R. Théodor, *Analyse Numérique matricielle appliquée à l'art de l'ingénieur, 1. Méthodes directes*, Dunod 2000

[6] P. Lascaux, R. Théodor, *Analyse Numérique matricielle appliquée à l'art de l'ingénieur, 2. Méthodes itératives*, Dunod 2000